# Concise *Encyclopedia* of Biostatistics for Medical Professionals

**Abhaya Indrayan**
**Martin P. Holt**

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

# Concise *Encyclopedia* of Biostatistics for Medical Professionals

# Concise *Encyclopedia* of Biostatistics for Medical Professionals

**Abhaya Indrayan**

Delhi University College of Medical Sciences, Delhi, India

**Martin P. Holt**

Halcyon, Leicester, UK

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742



Printed on acid-free paper
Version Date: 20160418

International Standard Book Number-13: 978-1-4822-4387-1 (Hardback)



**Library of Congress Cataloging-in-Publication Data**

Names: Indrayan, Abhaya, 1945- author. | Holt, Martin Patrick, author.
Title: Concise encyclopedia of biostatistics for medical professionals /
Abhaya Indrayan and Martin Patrick Holt.
Description: Boca Raton : Taylor & Francis, 2016. | Includes bibliographical
references.
Identifiers: LCCN 2016015564 | ISBN 9781482243871 (alk. paper)
Subjects: | MESH: Biostatistics | Encyclopedias
Classification: LCC R853.S7 | NLM WA 13 | DDC 610.72/7--dc23
LC record available at https://lccn.loc.gov/2016015564

**Visit the Taylor & Francis Web site at**
**http://www.taylorandfrancis.com**

**and the CRC Press Web site at**
**http://www.crcpress.com**

*To Asha, for her unstinted support all through my life.*

**Abhaya**

*For Audrey, without whom this would not have been possible.*

*Sincerely yours,*

**Martin**

# Contents

# Additional Terms

Appear in italics in the text: These terms are mentioned or briefly explained in the text without details.

# Preface

The science of biostatistics has proliferated enormously during the past few decades, and its medical applications seem to have increased exponentially. Our interaction with a large number of medical practitioners and researchers reveals that they are increasingly perplexed over the biostatistical methodology and terminology used in medical literature. Understanding biostatistical concepts is sometimes a challenge for medical professionals because they often lack formal training in this area. Encyclopedias are commonly used as a reference to quickly learn the essentials of a specific topic. We believe that our concise-encyclopedia (ency) will be a useful resource for anyone needing to quickly understand the basics of a particular biostatistical concept with a minimal level of mathematical knowledge. The topics, including alternative phrases, are arranged alphabetically, and the reader can easily locate the topic of interest and understand the meaning of the terms involved, their applicability to health sciences, and their medical interpretation. We have tried to orient the discussion of the topics firmly to the needs and backgrounds of medical and health professionals so that it resonates with them. Like any encyclopedia, this one, too, is not a replacement for a textbook, as the matter is not arranged in a logical sequence. For a regular book in a similar mode, see the fourth edition of *Medical Biostatistics* by Abhaya Indrayan (CRC Press, 2017), which takes you through most of these topics in a logical flow from beginning to end using medical uncertainties as the common theme for navigation.

Our target audience is those medical professionals who want to adhere to an evidence-based approach to medical practice. This requires a good understanding of biostatistical methods for proper interpretation of results reported in scientific articles. These professionals include physicians, surgeons, dentists, pharmacists, nursing professionals, nutritionists, and epidemiologists. Academia, especially graduate students in medical and health sciences, may also find this ency useful as a reference.

We have conceived this encyclopedia as relatively short, hence the term *concise*. It focuses on conceptual knowledge and practical advice in an easy-to-read format rather than mathematical details. This approach hopefully enhances the book's usefulness as a reference for medical and health professionals. Limiting the book to a single volume will also make it more affordable and attractive to individual professionals and not just libraries. Because of the limited size, the explanations provided for various terms in this volume cannot be comprehensive, and some readers may feel that the types of details they need are missing. For some terms, such as *regression* and *analysis of variance*, full chapters in many books and, in fact, full books are available. Our explanation is limited to the essential features, applications, and interpretations we perceive as relevant to medical sciences. However, in this process, we may not have been able to meet the exacting standards of some statisticians.

It is not possible to cover everything in one volume, yet this ency may be quite comprehensive for the target audience. It defines and explains more than 1000 commonly and not-so-commonly used key biostatistical terms and methods, ranging from simple terms such as *mean* and *median* to advanced terms such as *multilevel models* and *generalized estimating equations*. Those health topics that have significant biostatical components are also presented. These include terms relating to community health, and social and mental health.

We have listed alternative terms to make searching easier and also to help our medical colleagues by removing some of the confusion surrounding multiple terms for the same method or topic. We have also tried to explain how they are related, if at all. For example, an independent variable in a regression setup is also called a regressor, an explanatory variable, a predictor, a covariate, a confounder, a concomitant variable, etc., depending upon the context. Similarly, a dependent variable is also called a response, a target, or an outcome variable. We discuss all of them together under one unifying topic, "dependent and independent variables," so that their subtle distinctions can be explained. All the related terms for each topic are listed for completeness but refer the reader to the unifying topic. For these few terms, the length of the entry is considerably increased. On the other hand, a topic that needs to be extensively described with varied applications, such as chi-square, is split into several relatively smaller topics to retain the focus and to keep each topic within a manageable limit. The purpose is to remain short and crisp for the convenience of our medical readers. For some topics, such a split may have hindered providing a correct perspective. Also, we have tried to describe each topic in a self-contained manner for independent reading so that the reader does not have to continually flip through the book, although some cross-referencing for the involved terms seemed unavoidable. Thus, there is some duplication, which was inevitable in this kind of work, although we have tried to minimize this, sometimes by combining two related topics.

Given that the target audience is medical professionals, formulas and mathematical details are limited to high school algebra. Thus, formulas for only those terms that do not require complicated mathematics have been provided. For others, a heuristic approach is adopted to explain and describe the rationale and applications. We are aware that heuristics can generate imprecision and sometimes even errors—thus, imperative care has been taken. Many topics are illustrated with the help of medical examples and figures, including references to contemporary medical literature. The list of references for topics is not comprehensive, though, and is restricted to the most relevant, so that the reader does not feel burdened. Web links have been provided for references wherever we could locate them so that the reader can directly access the referenced material. This could not be done for the referenced books. The date of last access by us is written only for those

websites that are likely to change. For ostensibly permanent websites such as those of journals, the date of last access is not provided. These links provide an indispensable platform for further study and research.

We realized the difficulty in communicating complex mathematical issues in a simple nonmathematical language and accepted this challenge. Efforts all throughout this ency are in simple language that can be appreciated by the target audience, who have less grounding in mathematics. We have not used formal language; instead, the tone is conversational. In the process, though, we may not have been very accurate in explaining a method or a concept. This admittedly is a limitation of this book. But our focus is firmly on medical applications and concerns. For intricate methods, where we have explained only the concept and applications but not the analysis, a reference is provided wherein one can find the required details. Historical briefs, biographical snippets, and photographs of prominent statisticians are added to provide a broader context for the topics covered.

Within each topic, those terms that are explained elsewhere in this volume are in bold when they appear first, particularly when details of that term are likely to help the reader to understand the current topic. There might be some minor variation in the topic title as required for the current text. Other statistical terms appear in italics and are separately listed so that these can also be located. Phrases that we want to emphasize also are in italics. To those of our medical colleagues who are not familiar with statistical notations, we advise that they see the topic "notations" in this ency so that they feel comfortable in reading and understanding this ency. We have tried to follow a uniform system of notations across the topics.

Almost all other encyclopedias are written by several contributing authors and edited by experts. Perhaps ours is the first attempt to prepare an encyclopedia by two authors. This has advantages and disadvantages. The first advantage is uniformity of the level of discussion—when based on contributions by different authors, some topics are discussed at too high a level and others at too low a level, depending on the concerned authors' understanding of the topic. The second advantage is our consistent style. A disadvantage is that this book may not delve as deeply into some topics as individual experts would prefer.

Despite our best care, errors may have occurred. We are sure that the book will be critically examined by the reviewers. Their feedback will help in improving the next edition.

# Acknowledgments

Our sincere thanks to John Whittington for reviewing and revising several sections of this book, which substantially added value to this work.

We are grateful to Ms. Laxmi Raghuvanshi for helping us with many figures in Excel, in checking the calculations, and providing other help from time to time.

**Abhaya Indrayan**
**Martin P. Holt**

# A

## abortion rate/ratio

Abortion *rate* is the number of abortions per 1000 women of reproductive age group, whereas abortion *ratio* is the number of abortions per 1000 live births:

$$\text{Abortion rate in a year x} = \frac{\text{abortions in year x}}{\text{females of reproductive age-group in that year}} * 1000$$

$$\text{Abortion ratio in year x} = \frac{\text{abortions in year x}}{\text{live births in that year}} * 1000$$

The reproductive age group for females varies from population to population, but for comparability, this is considered to be 15–49 years. In the United States, this is calculated per 1000 women aged 15–44 years [1].

Abortion is the termination of a pregnancy before the fetus becomes viable, generally before completing 24 weeks of pregnancy. It could be deliberate to avoid an unplanned birth or due to a medical complication (called *induced*), or spontaneous (called *miscarriage*) mostly due to poor health of the mother or an accident. In developed countries, induced abortions are dominant, whereas in underdeveloped countries, spontaneous losses also could be substantial. Both types of abortions are mostly preventable. Induced abortions in case of life-threatening situations sometime raise a value judgment question regarding whose life is more important to save—mother or child. The abortion rate as well as the abortion ratio can be calculated separately for induced and spontaneous abortions or for the combined number.

According to the Centers for Disease Control (CDC), induced abortions were nearly 14 per 1000 women in the United States in 2011, and the abortion ratio was 239 per 1000 live births [1]. Most of these abortions took place in women in their early 20s. These numbers signify the distinction between the two indicators of **incidence** of abortions besides telling us the magnitude of the "problem."

When the count of stillbirths is also available so that you know total pregnancies, this rate can also be calculated as the percentage of pregnancies. For example, in India, in the year 2012, an estimated 25% of pregnancies resulted in abortions [2]. Some of these are ascribed to avoiding the birth of a female child because of the stigma attached to girls in some sections of Indian society. This may have occurred in China as well, where the one-child policy was enforced and many preferred a boy.


1. CDC. *Abortion surveillance 2011.* http://www.cdc.gov/reproductive health/data_stats/#Abortion, last accessed May 15, 2015.
2. Johnston WR. *India abortion percentages by state and territory, 1971–2013.* http://www.johnstonsarchive.net/policy/abortion/india/ab-indias2.html, last accessed May 16, 2015.


## absolute risk reduction, *see* attributable risk (AR)

## accelerated failure time model, *see also* proportional hazards

The accelerated failure time (AFT) model is used in **survival analysis** for studying the effect of a covariate that either accelerates or decelerates survival or failure. This is an alternative to the popular **proportional hazards** model, which works when the covariate affects the incidence of failure by a fixed multiplicative constant irrespective of time. Note the difference—one is for the covariate, such as an intervention, which affects the quickness with which the outcome occurs, whereas the second is for the covariate that affects the number of outcomes. Hazard ratio is difficult to interpret for a covariate that affects the speed of occurrence. The results of the AFT model may be more relevant as they can be directly translated into expected reduction or prolongation of the time to event. This can provide a useful framework for research, particularly in relation to aging.

The AFT model is a parametric model as it depends on the distribution of the failure time, whereas the proportional hazards model is semiparametric as it does not require distribution of the failure time; the proportionality of hazards alone is enough. AFT utilizes the entire survival curve, whereas the proportional hazards model compares only the mortality rates. For details of the differences between the two, see Patel et al. [1].

In a study on factors associated with time to next attack in neuromyelitis optica in South Korea, an AFT model revealed that the interattack interval naturally (without intervention) increased by 1.31 times as the cumulative number of attacks increased with time but also increased independently 4.26 times in those on rituximab treatment [2]. Note in this case how the number of attacks and the treatment decelerated the time to next attack (time interval increased). In the case of antiretroviral therapy in HIV cases in Cameroon, the AFT model has been used since the survival time rapidly increased after start of the therapy [3].

Under the AFT model, the effect of intervention is to shift the entire survival curve almost uniformly. This can be mathematically expressed as $S_1(ct) = S_0(t)$, where $S_0(t)$ is the survival at time $t$ in the control group and $S_1(ct)$ is the survival at time $ct$ in the intervention group. If the intervention increases the duration of survival, the value of $c$ will exceed 1, and if the intervention decreases the duration of survival, the value of $c$ will be less than 1. This could be understood as the acceleration or deceleration factor. The AFT model helps us in estimating the value of $c$. However, that is valid only when the effect of intervention on the duration of survival is consistent all through one's life. The model can be used only when the survival duration follows a specified distribution. This can be **exponential, Weibull**, or any such distribution. For further details, see Swindell [4].

A

1. Patel K, Kay R, Rowell L. Comparing proportional hazards and accelerated failure time models: An application in influenza. *Pharm Stat* 2006 Jul–Sep;5(3):213–24. http://www.ncbi.nlm.nih.gov/pubmed/17080754
2. Kim SM, Park J, Kim SH, Park SY, Kim JY, Sung JJ, Park KS, Lee KW. Factors associated with the time to next attack in neuromyelitis optica: Accelerated failure time models with random effects. *PLoS One* 2013 Dec 16;8(12):e82325. http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0082325
3. Poka-Mayap V, Pefura-Yone EW, Kengne AP, Kuaban C. Mortality and its determinants among patients infected with HIV-1 on antiretroviral therapy in a referral centre in Yaounde, Cameroon: A retrospective cohort study. *BMJ Open* 2013 Jul 13;3(7). pii: e003210. http://bmjopen.bmj.com/content/3/7/e003210.full.pdf+html
4. Swindell WR. Accelerated failure time models provide a useful statistical framework for aging research. *Exp Gerontol* Mar 2009;44(3):190–200. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2718836/

## acceptable risk

As in common parlance, acceptable risk in health and medicine too is highly contextual and greatly varies from situation to situation and person to person. Nonetheless, for some kind of uniformity and to remove subjectivity, the *acceptable risk* of any intervention can be defined as gains from the intervention outweighing its cost. Each of these can be computed from different perspectives. For example, in public health, a risk is acceptable if the gain to society as a whole outweighs the adverse health impacts and resources required from society to reduce that risk. Assessment of these may require systematic studies and expert opinions. For environment regulations in the United States, a lifetime risk of less than one in a million is considered an acceptable risk of developing cancer from a risk factor, whereas in the United Kingdom, this is the acceptable limit for 1-year risk. This illustrates that acceptable risk varies according to individual and collective perceptions. Generally, it depends on the current norms prevalent in a setup, and these can quickly change. For details of the concept of acceptable risk, see Hunter and Fewtrell [1].

Sometimes, a higher risk than is really acceptable is tolerated by individuals and societies, such as with smoking. Thus, there is a distinction between accepted risk and acceptable risk. Also, determination of acceptable risk could be highly subjective, influenced by local perceptions and culture, as discussed by Hsu [2] for surgical masks as opposed to N95 respirators for prevention of nosocomial Middle East respiratory syndrome coronavirus (MERS-CoV), which has high **case fatality**. Similarly, Park et al. [3] have mentioned acceptable risk in the context of revisional bariatric surgery performed via a laparoscopic approach without really quantifying it.

For statistical decisions, though, the acceptable risks are fairly well defined in terms of probability of **Type I error** and of **Type II error**. These errors delineate the acceptable risk of a false-positive result and a false-negative result when based on appropriate statistical methods of data analysis. However, these, too, depend on the availability of correct data and the use of correct methods.

1. Hunter PR, Fewtrell L. Acceptable risk, in: *Water Quality: Guidelines, Standards and Health*. World Health Organization. (Eds. Fewtrell L, Bartram J). IWA Publishing, 2001. http://www.who.int/water_sanitation_health/dwq/iwachap10.pdf, last accessed May 25, 2015.
2. Hsu LY. Respiratory precautions for MERS-CoV: Acceptable risk–benefit determination. *Singapore Med J* 2014;55(6):293. http://sma.org.sg/UploadedImg/files/SMJ/5506/5506ed1.pdf
3. Park JY, Song D, Kim YJ. Causes and outcomes of revisional bariatric surgery: Initial experience at a single center. *Ann Surg Treat Res* 2014 Jun;86(6):295–301. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4062454/

## acceptance and rejection regions

*Acceptance and rejection regions* are typical statistical terms and are used for the set of values of the **test criterion** in a **test of hypothesis** setup that decides whether to reject or not to reject a **null hypothesis**. The possible values of the test criterion are divided into two mutually exclusive sets: one with very little chance of being consistent with the null hypothesis and the other with a bigger chance. The set of values that have very little chance is termed the *rejection region*, and the other, which is the complement, is called the *acceptance region*.

For example, for the ***z*-test**, a value of $|z| \geq 1.96$ is the rejection region for a **two-sided alternative** at a 5% **level of significance**. Half of this region is in left tail, and the other half is in the right tail of the distribution (Figure A.1a). For **chi-square** at 5 degrees of freedom, the rejection region at a 5% level of significance is $\chi^2 \geq 11.07$ (Figure A.1b). This is a **one-tailed** critical region. Our *z*-test and chi-square test examples illustrate that different tests have different procedures to obtain the acceptance and the rejection regions. You can see that these regions are determined by the level of significance, the sampling distribution of the test criterion, and whether the test is one-tailed or two-tailed.

Technically, a null hypothesis is never accepted (it is just not rejected)—thus, the term *acceptance region* is inappropriate. But this term was commonly used till some time ago. The values in this region are not sufficiently against the null hypothesis. The rejection region is also called the **critical region**, and the value that separates the two regions is called the **critical value**. The critical values in our examples are 1.96 and 11.07, respectively. However, these terms are seldom used in the modern literature since almost all statistical software directly give the exact ***P*-value** corresponding to the value of the test criterion for a given sample—thus, the region or the critical value is not needed. This *P*-value is compared with the prefixed level of significance to come to the conclusion to either reject or not reject the null hypothesis.

## acceptance sampling, *see* lot quality assurance scheme

## accuracy

In the context of a *single* value, accuracy is the nearness to its true value. "Age 26 years, 2 months, and 21 days" is more accurate than just "26 years." In a conventional sphygmomanometer, calibrations are at intervals of 2 mmHg; thus, most blood pressure (BP) measurements

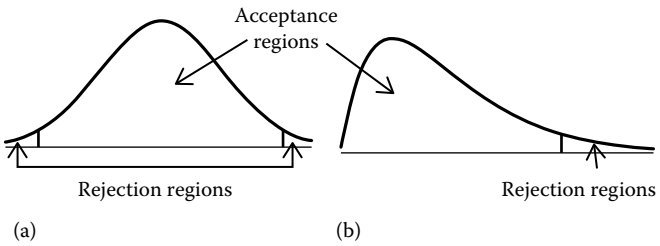



**FIGURE A.1** Acceptance and rejection regions: (a) *z*-test two-tailed; (b) $\chi^2$-test one-tailed.

also have only that kind of accuracy. Therefore, this instrument does not provide very accurate readings. Perhaps better accuracy is not needed for BP values. Only as much accuracy is needed in medical measurements as is required for clinical assessment.

Statistically, more accurate measurement will have less error component. But realize that results cannot be more accurate than the measurements themselves. Although it is customary to state summaries like mean and standard deviation to one more decimal place than are the data, they must be finally interpreted as if they have the same accuracy as the measurements. Intermediary calculations are done to several more decimals so that the final result retains the accuracy of the measurements. Remember that no such calculation can ever add to the accuracy. For ***P*-values** also, the old convention was to say "$P < 0.05$." This is less accurate than something like "$P = 0.045$," which is now commonly done. The more decimal places, the better the accuracy. But in this case, too, having more than three decimals does not help in a statistical assessment of the result. The other statistical example is the use of a **continuity correction**, as for discrete values to continuous values (see **Gaussian probability** as an example). This correction helps in getting more accurate results.

For many medical measurements, accuracy beyond a certain degree does not add to the prediction of occurrence or nonoccurrence of outcomes, but there are instances when a minor variation can cause a substantial ripple effect and a major difference in the outcome. For details, see the topic **butterfly effect**, where an example of cancer is also provided, although in that case, it is not the inaccuracy but the effect of a slight change in the input value.

The term *accuracy* is also used to describe the quality of a method. The accuracy of a method is its ability to reach the true value almost each time. This is evaluated by comparing the values obtained by a method with the known standard values. A more accurate method will provide readings with less error most of the time. For example, Lock and Ronald [1] found the FreeStyle Freedom Life system to be the most accurate method out of the four they compared in meeting the International Organization for Standardization (ISO) standards for blood glucose measurements.

You may find an overlap of usage among the terms *accuracy*, *reliability*, and validity. In scientific literature, this happens because of poor understanding or because of carelessness. For example, Rebel et al. [2] have mixed up the use of the terms *accuracy* and *reliability* while evaluating point-of-care monitoring of blood glucose level. In fact, accuracy is very different from reliability (reliability is the same as precision) and slightly different from validity. Statistical validity of the sample values refers to the average being able to hit the target, whereas accuracy is generally used for single measurements. For further details, see **reliability** and **validity**.

1. Lock JP, Ronald NG. Accuracy of four blood glucose monitoring systems. *Abott Diabetes Care* 2010. https://abbottdiabetescare.co.uk/images/uploads/documents/CP074.pdf, last accessed May 20, 2015.
2. Rebel A, Rice MA, Fahy BG. The accuracy of point-of-care glucose measurements. *J Diabetes Sci Technol* 2012;6(2):396–411. http://anest.ufl.edu/files/2011/09/Rebel-A-The-accuracy-of-point-of-care-glucose-measurements.pdf

## actuarial method, *see* expectation of life and the life table

## adaptive designs for clinical trials

Adaptive designs allow flexibility to redesign a clinical trial midstream, guided by the interim experience. This can make the trial more efficient by saving time, money, and patients. Suppose that after due consideration of available knowledge, you plan a clinical trial on 1000 subjects in each **arm** with 1-year follow-up. Two months into the trial, you find from **interim analysis** that your anticipations at the planning stage were incorrect and that the trial will give you confirmatory results on efficacy (or a lack of it) in just 6 months (or you need to extend it to 16 months to get an adequate number with the desired end point); or only 700 would suffice (or 1500 would be needed); or that the doses you are trying are too high (or too low) and you need to have a middling dose; or that a concomitant treatment is needed; or that a particular subgroup, such as males of age 70+ years, needs to be excluded. Sometimes, even the baseline information on the enrolled subjects may indicate that modifications in the design are needed. For example, this may tell you that the expected kind of subjects are not being enrolled and that the eligibility criteria need to be changed. In such situations, you would not want to waste resources by sticking to the original plan in the hope of wonders; instead, you would want to adapt the trial to the realities revealed by actual experience.

The need for adaptation arises from the fact that sometimes, all the information needed to plan a near-perfect trial is not available in advance. Besides medical issues such as side effects, biostatistical issues that can affect planning are the variance of the estimate of the **effect size** you wish to consider, expected compliance, and a conjecture of the anticipated effect size itself. When these are not available with a reasonable degree of assurance, you may want to fall back on an adaptive design that allows flexibility. Adaptive designs incorporate practical considerations of possibly not getting the things right first time in the preparation of the design.

Adaptation is planned in advance by anticipating different scenarios that could unfold in the planned trial. Thus, it can handle only limited issues and not those that were unforeseen or ignored. The issues must arise from interim analysis and not from extraneous sources. Changing the features of the trial due to other considerations is not adaptation in the sense we are discussing here. Adaptation is not a substitute for a sloppily planned trial. On the other hand, implicit in all this is that the trial has an immaculate design that has considered all the available information, including what adaptations are to be made in case specific issues are found.

Despite the clear advantages of adaptive trials, they have been rarely used so far. The first difficulty is that the adaptive methodology is still evolving. Secondly, this could involve many features of the design and can be done possibly at several stages of an ongoing trial. It is difficult to visualize in advance where and what adaptations would be required. Thirdly, there is confusion on what all should be considered for adaptation, how to keep **Type I** and **Type II errors** under control despite periodic evaluations, and how to handle logistic problems that arise from adaptation of an ongoing trial. The timing of interim analysis will be stated in the trial's protocol: it will occur at a predefined time, and care is taken that such appraisal does not undermine the integrity and validity of the trial.

When adaptations are done, the analysis becomes computationally complex, for which specially tailored software packages are being developed. Nonetheless, the adaptive strategy is being explored with enthusiasm by drug companies, regulators, and researchers. As experience accumulates, adaptive trials can be better designed and will have wider acceptability. An adaptive trial can quickly move from phase II to phase III since lessons learnt are already incorporated. This can expedite the process of product development. For sure, this strategy can address frustration arising in conventional structured designs when the trial gives negative results and you wish that if something could be done differently, the results would not be so disappointing. Adaptive trials may soon become industry standard as the problems in their implementation are sorted out.

A

Among various adaptations, the biostatistically most relevant is to **reestimate the sample size** on the basis of the actual effect size found at interim stages. You can imagine that the reestimated sample size should only increase and not decrease, although a decision to stop the trial either for efficacy or for futility can be made. That, however, does not mean that the trial is deliberately designed to have a small size. Reestimation requires intricate statistical inputs as this is done to preserve the **level of significance** and the statistical **power**. Such statistical adaptation does not cause many ethical problems; rather, it seems to enhance ethics by keeping a provision to stop the trial early in case convincing evidence of efficacy or of futility (see **stopping rules**).

Interim analysis has the potential to unblind an otherwise blinded trial. Un**blinding** is necessary to assess whether the treatment arm is giving evidence of sufficient efficacy relative to the control arm. Steps are specified at the time of the designing of the trial to determine who will be unblinded (such as the **Data Safety and Monitoring Board**), and how the investigators and the subjects will continue to remain blinded for the trial to go unhindered. The person/team unblinded for interim analysis must be independent of the team of investigators and comprise people who have nothing to gain or lose from the interim analysis. This team should make only indirect statements, such as "the trial will continue for 8 more months" or "will enroll 130 more subjects," so that the integrity of the trial remains intact.

Englert and Kieser [1] have discussed how adaptive designs in phase II cancer clinical trials help in reducing the required number of subjects for coming to a decision. For more details of designs with interim appraisals, see Chow and Chang [2].


1. Englert S, Kieser M. Optimal adaptive two-stage designs for phase II cancer clinical trials. *Biom J* 2013 Nov;55(6):955–68. http://www.ncbi.nlm.nih.gov/pubmed/23868324
2. Chow S-C, Chang M. *Adaptive Design Methods in Clinical Trials*. Chapman & Hall/CRC, 2006.


## added variable plot

The added variable plot is the plot of **residuals** from a **regression** model that excludes a particular **independent variable** of interest versus the residuals when this particular independent variable is regressed on the remaining variables. Thus, both regressions have the same set of regressors, but the dependent in one case is the actual outcome, and the dependent in the other case is the regressor of interest itself. Consider, for example, regression of $y$ on $x_1, x_2, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_K$ and you are interested in investigating the role of a particular variable $x_k$ as an added variable. The procedure would be to find residuals of regression of $y$ on $x_1, x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_K$ (without $x_k$) and the residuals of regression of $x_k$ on the same set of regressors, namely, $x_1, x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_K$ (without $x_k$). Plot the first residuals against the second residuals to get an added variable plot. For further details, see Dayal [1].

Thorough study of an added variable plot for patterns might provide useful clues first about any utility of this added variable and second about the functional form of the relationship of the added variable with the outcome, and whether or not there are few influential values. You can decide whether to keep these influential values or to exclude them. Their exclusion may help in achieving better parsimony, and inclusion can give more realistic results.

The information provided by the added variable plot has been found to be more valid than the information provided by the residual plot against this new regressor. This is considered a valuable tool to evaluate all kinds of regressions: ordinary least squares, logistic, and Cox. Statistical software packages generally have a provision to generate added variable plots for all these regressions.

Possibly due to the expertise needed to read a variable added plot, this is rarely used in medical research. We could not find a single medical article in the PubMed database that has used the term "variable added plot." We are still including it here because of its statistical relevance for medical research. An example on this is provided by Chen and Wang [2] in the context of the Cox proportional hazards model.


1. Dayal M. Why added variable plots? *Applied Statistics and Computing Lab*. Indian School of Business, Hyderabad. https://www.academia.edu/2519488/WHY_ADDED_VARIABLE_PLOTS_, last accessed May 15, 2015.
2. Chen C-H, Wang PC. Diagnostic plots in Cox's regression model. *Biometrics* 1991;47(3):841–50. http://www.jstor.org/discover/10.2307/2532643?uid=2134&uid=2474376657&uid=2&uid=70&uid=3&uid=2474376647&uid=60&sid=21103282253817


## additive effect/factors

Suppose you find for your subjects that an increase in **body mass index (BMI)** in men from 32 to 33 kg/m$^2$ is associated with an increase in systolic level of blood pressure (BP) by 3 mmHg on average and that an increase in triglyceride level (TGL) from 150 to 160 mg/dL is associated with 4 mmHg increase in this BP on average. Does this mean that a group of subjects with BMI = 33 kg/m$^2$ and TGL = 160 mg/dL will have an average systolic BP 3 + 4 = 7 mmHg more than the average of group with BMI = 32 kg/m$^2$ and TGL = 150 mg/dL? If yes, the effects of BMI and TGL on systolic BP are called additive. Quite often, these effect sizes do not add up, and the average difference would be either substantially higher or substantially lower. In that case, the factors are said to have an **interaction**. In other words, the absence of interaction indicates that the factors and their effects are additive.

Statistical methods such as **two-way analysis of variance (ANOVA)** and **regression** are relatively simple when the factors are additive. If not, an interaction term also needs to be considered. When you are considering two or more factors together, it is desirable to explore their interaction. Even when factors are additive, results from considering them one at a time can give different results compared to considering them together. For example, results of two one-way ANOVAs are not the same as one two-way ANOVA even when the factors are additive.

The concept of additive effect can be readily extended to the administration of two (or more) drugs. For example, according to Govorov et al. [1], tadalafil and alpha-blocker cotherapy in Russia for lower urinary tract symptoms due to benign prostatic hyperplasia (LUTS/BPH) suggests an additive effect, though such cotherapy is not recommended yet. On the other hand, Iverson et al. [2] found in a small study in the United States that the effect of ingestion of two amino acids (leucine and glycine) on the decrease in glucose response was not additive.


1. Govorov A, Kasyan G, Priymak D, Pushkar D, Sorsaburu S. Tadalafil in the management of lower urinary tract symptoms: A review of the literature and current practices in Russia. *Cent European J Urol* 2014;67(2):167–77. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4132596/

2. Iverson JF, Gannon MC, Nuttall FQ. Interaction of ingested leucine with glycine on insulin and glucose concentrations. *J Amino Acids* 2014:521941. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4121211/

## adjusted effects, *see also* adjusted mean, adjusted odds ratio, adjusted $R^2$

When the effect of an **antecedent** X on **outcome** Y is also affected by another factor Z, the effect obtained after removing the effect of Z is called the Z-adjusted effect of X on Y.

Whenever two or more factors are considered together for their effect on an outcome, their independent effects most likely will be different from their joint effect. The effect of most medical factors depends on what other factors are present. The presence of other factors can be synergistic (enhance the effect) or antagonistic (reduce the effect). For example, the risk of death of a person of age 70 years with kidney disease and diabetes is more than the sum of the risk of death with kidney disease alone and diabetes alone. When both diseases are present, the risk increases steeply. In this case, you can find the effect of diabetes alone after adjusting for the effect of kidney diseases using statistical methods such as **logistic regression**. This will be called the effect of diabetes on the risk of death adjusted for kidney disease. Similarly, you can find the adjusted effect of kidney diseases. Almost any measure of **effect size** can be adjusted for the presence of other factors by suitable statistical methods when the relevant data are available. For example, you can have **adjusted mean**, **adjusted odds ratio**, **adjusted $R^2$**, etc. But be careful in their interpretation. Sometimes, such adjusted effects are interpreted as a net effect or pure effect. This may not be so, since adjustment is done only for those factors that have been included in the analysis. There might be other factors, not included in the analysis, that also might be making a negative or positive contribution. That part is ignored in this kind of statistical adjustment.

## adjusted mean

This is the mean obtained after removing the effect of differential composition of two or more groups so that the means become comparable. Suppose that in a survey of 300 adults, the mean serum folate level (nmol/L) in current, former, and never smokers of different age groups is as given in Table A.1. The age composition of groups is different, and this might affect the mean folate level in different smoking groups. Thus, the mean folate level in current, former, and never smokers is not comparable. They can be made more comparable by proper adjustment for age variation.

The means in column D are calculated with due consideration to the different numbers of subjects in different age groups as is needed for grouped data. This is called the **weighted mean**. For example, the mean serum folate level for current smokers is

$$\frac{6.5\times 80 + 7.0\times 40 + 8.5\times 10}{80+40+10} = 6.8 \text{ nmol/L}.$$

Of the 70 former smokers, 40 are in the age group of 60+ years. This number is only 10 out of 130 in the current smokers. This differential can affect the mean serum folate level because the level depends on age. This mean in column D is unadjusted in the sense that the difference in age distribution is overlooked. The difference in these unadjusted means across smoking status categories is not necessarily real but could be partially due to the difference in their age structure. The effect of age disparity can be removed by calculating the age-adjusted mean. For this, a standard age distribution is required. The age distribution of total subjects (last row) can serve as the standard in this example. When this standard is used on the unadjusted means, the following is obtained:

**TABLE A.1**
**Mean Serum Folate Level (nmol/L) by Age and Smoking Status**

| | Age (Years) | | | Unadjusted Mean for All Adults | Age-Adjusted Mean |
|---|---|---|---|---|---|
| | 20–39 | 40–59 | 60+ | | |
| **Smoking Status** | **A** | **B** | **C** | **D** | **E** |
| Current smokers | 6.5 | 7.0 | 8.5 | 6.8 | 7.1 |
| | (80) | (40) | (10) | (130) | |
| Former smokers | 7.5 | 8.5 | 9.0 | 8.6 | 8.1 |
| | (10) | (20) | (40) | (70) | |
| Never smokers | 7.0 | 8.0 | 9.0 | 7.6 | 7.7 |
| | (50) | (40) | (10) | (100) | |
| **Total** | **6.8** | **7.7** | **8.9** | **7.5** | |
| | **(140)** | **(100)** | **(60)** | **(300)** | |

*Note:* The number of subjects is in parentheses.

The age-adjusted mean serum folate level for current smokers is

$$\frac{6.5\times 140 + 7.0\times 100 + 8.5\times 60}{140+100+60} = 7.1 \text{ nmol/L}.$$

This is obtained by applying the age-specific rates in the group to the standard age structure. This procedure is similar to the procedure used to calculate the directly standardized death rate. Similarly, age-adjusted mean can be calculated for the former smokers and the never smokers. These are given in the last column of Table A.1.

Note that the large difference between the unadjusted means of folate level in current and former smokers decreased considerably in this example after the age adjustment. This implies that much of this difference was due to the differential in the age structure of the subjects in these two categories. This adjustment brought the groups to a common base with respect to age and made them comparable.

The difference between age-adjusted and age-standardized effects is that age adjustment is done for "locally" available age distribution whereas age standardization is done with respect to the standard age structure of the population, which may be extraneous or even imaginary. For example, *age-standardized death rates* across countries can be obtained with respect to an imaginary world standard population, such as the one devised by the World Health Organization [1].

1. Omar B, Ahmad OB, Boschi-Pinto C, Lopez AD, Murray CJL, Lozano R, Inoue M. *Age Standardization of Rates: A New WHO Standard*. GPE Discussion Paper Series: No.31. EIP/GPE/EBD World Health Organization 2001. http://www.who.int/healthinfo/paper31.pdf

## adjusted odds ratio (adjusted OR), *see also* odds ratio

As the name implies, the odds ratio (OR) is the ratio of the **odds** of the presence of an antecedent in those with a positive outcome to

A

the odds in those with a negative outcome. This ratio needs to be adjusted when the outcome is suspected to be affected by other factors. For example, the presence of oral cancer in any case is affected by whether or not the person consumed smokeless tobacco for a long duration. Let this be the antecedent of interest. But oral cancer can also occur or accelerate by insufficient intake of fruits and vegetables and the presence of leukoplakia. When such concomitant variables, called *covariates*, are ignored altogether, what is obtained is called *unadjusted OR*. But when the analysis is geared to remove the influence of these covariates, adjusted OR is obtained. For example, you may have an unadjusted OR = 12.7 for smokeless tobacco in cases of oral cancer, and after adjustment of the covariates, this may reduce to 8.9. The balance of 3.8 can be attributed to the covariates in the model. Adjusted OR will decrease if the covariates tend to increase the incidence of disease and will increase if the covariates decrease the incidence.

The most common method for finding the adjusted OR is **logistic regression**. In this method, the **logistic coefficients** are the logarithms of respective ORs. This regression can be run with several covariates in the logistic model. Each coefficient provides ln(OR) for that factor, and this is automatically adjusted for the other covariates in the model. In our example, the coefficient of leukoplakia will provide an OR adjusted for the use of smokeless tobacco and insufficient intake of fruits and vegetables, and the coefficient of the use of smokeless tobacco will provide an OR adjusted for insufficient intake of fruits and vegetables and the presence of leukoplakia. The same regression model can provide several adjusted ORs.

Adjusted ORs are often misused. It is incorrect to interpret the adjusted OR as the *net effect* of the antecedent. Sometimes, this is termed the *independent contribution* of the antecedent. This term is also suspect. The adjustment is only for those covariates that are included in your model. There might be other factors that are not present in the model, including those that are unknown but may influence the outcome. Since no adjustment is done for these covariates, adjusted OR does not measure the net effect, or even the independent effect. The best way to understand this is the effect adjusted for the other covariates in the model. In our example, 8.9 is the OR adjusted for intake of fruits and vegetables and the presence of leukoplakia. This is not adjusted for other factors such as family history and exposure to sun, and this is not the net effect either.

Another fallacy commonly occurs in interpreting the adjusted OR when only the **linear effect** of the covariates is considered. Almost invariably, logistic regression is run with linear terms of the covariates. Thus, the adjustment also is for linear effect. If the effect of a covariate has a quadratic, logarithmic, exponential, or any such nonlinear form, that will not be adjusted unless such forms are included in the model.

## adjusted $R^2$, *see also* multiple correlation

In order to understand adjusted $R^2$, first recall $R^2$ itself. This is the square of the **multiple correlation** coefficient, which is defined as the Pearsonian **correlation coefficient** between a variable and the best linear combination of other relevant variables. This is generally obtained between a **dependent variable** $y$ and a combination of **independent variables** $x_1, x_2, \ldots, x_K$ in a multiple linear regression setting. This linear combination is chosen in such a manner that $R^2$ is maximum. The value of $R^2$ is obtained as 1 − SSE/SST, where SSE is the **sum of squares** due to error and SST is the total sum of squares (see **multiple linear regression** to understand these sums of squares). The value of $R^2$ is used to assess how good that linear regression is in explaining or predicting $y$.

One property of $R^2$ is that it improves as more variables are added to the regression. However, adding new variables also results in the loss of **degrees of freedom** (*df's*), and therefore, **statistical significance** does not necessarily improve by adding new regressors—it can even decline. Thus, an adjustment is required for the df's before the value of $R^2$ can be realistically assessed. These df's depend on the number of **regressors** in the model and the sample size. The adjustment is done by dividing the sums of squares by the respective df's. Thus,

$$\text{adjusted } R^2 = 1 - \frac{\text{SSE}/\text{df}_e}{\text{SST}/\text{df}_t},$$

where $\text{df}_e$ is the df for SSE and $\text{df}_t$ is for SST. In terms of the usual $R^2$, this becomes

$$\text{adjusted } R^2 = 1 - \frac{(1-R^2)(n-1)}{n-K-1},$$

where $n$ is the sample size and $K$ is the number of regressors. This adjustment would be substantial if $n$ is small. Since this is adjusted for the number of independent variables, adjusted $R^2$ can also be used to compare two or more regressions with a different number of regressors. For example, in a study on echocardiographic predictors of exercise and mortality in chronic obstructive pulmonary disease in Denmark, lung functions alone (in addition to the baseline variables) provided adjusted $R^2 = 0.475$ for prediction of distance performance in a 6-minute walking test, whereas this became adjusted $R^2 = 0.511$ when echocardiographic variables were also added [1]. The number of regressors in the two models is different—thus, adjusted $R^2$ is the right method to compare the performance of the two models.

A word of caution. Just as with $R^2$, adjusted $R^2$ is also used only for *linear* regressions. If the regression is nonlinear, the corresponding parameter is called the **coefficient of determination** and is denoted by $\eta^2$, as separately explained.

1. Schoos MM, Dalsgaard M, Kjærgaard J, Moesby D, Jensen SG, Steffensen I, Iversen KK. Echocardiographic predictors of exercise capacity and mortality in chronic obstructive pulmonary disease. *BMC Cardiovasc Disord* 2013 Oct 12;13:84. http://www.biomedcentral.com/content/pdf/1471-2261-13-84.pdf

## adjusting for baseline values, *see also* analysis of covariance (ANCOVA)

Baseline adjustment is the adjustment in the results due to a differential baseline in two or more groups. Baseline adjustment is a worthwhile exercise—sometimes an essential requirement—for evaluating any change. This is because change can substantially depend on the baseline values. An effective hematinic regimen may be able to improve hemoglobin level from a baseline of 8–12 mg/dL but would not show similar improvement when the baseline level is already 13 mg/dL. Such a differential response due to a differential baseline is generally anticipated before conducting the study and not afterward. In any case, a rational judgment should be made to decide that the outcome variable for analysis would be the raw values as observed or as the change from baseline. This would mostly depend on the interpretability. In either case, adjustment for baseline should be considered.

Three basic methods are available for baseline adjustment. First is the **stratification** of the subjects by baseline values. This may work when stratification can be done on some objective basis and the sample size is large enough to yield a reasonable size for each stratum. Thus, this should be done before conducting the study. Homogeneity of baseline values within each stratum helps to rule out the effect of baseline. The results should be presented stratum wise in this case. Second is **analysis of covariance (ANCOVA)** as discussed under this topic. You can also include the stratification variable as a covariate if stratification has been done. For ANCOVA, the covariate must be quantitative and should have been suspected to be a variable that has a fairly strong relation with the outcome. Third is a simple analysis of change scores. Change can also be measured in terms of the percentage of the baseline values if that is more appropriate and can be analyzed either by parametric methods that are primarily based on Gaussian distribution such as *t*-test or by nonparametric methods such as the Wilcoxon test when Gaussianity is violated.

ANCOVA generally considers a linear effect of the baseline values, and the tests of hypotheses and confidence intervals require that the distribution of the residuals is Gaussian. When these conditions are met, ANCOVA continues to be the method of choice for baseline adjustment. But this may fail if the change occurs before baseline assessment or if the outcome variable is unreliable or unstable [1]. A good strategy is to analyze data with and without baseline adjustment, and see how adjustment has or has not been helpful.

1. Glymour MM, Weuve J, Berkman LF, Kawachi I, Robins JM. When is baseline adjustment useful in analyses of change? An example with education and cognitive change. *Am J Epidemiol* 2005;162:267–78. http://aje.oxfordjournals.org/content/162/3/267.full.pdf+html

## adolescent health (epidemiological indicators of)

Epidemiologically, adolescent health is generally measured in terms of the adequacy of physical growth and sexual maturation. Vital changes occur during adolescence that set the basics of adulthood. A spurt in gain in height and weight occurs, and genitals take shape. Pubic hair grows. Menarche occurs, and breasts develop in girls. Variation, as always, remains an integral part of all these developments. Relatively few physical sicknesses occur in this phase of life. Although adolescent health can also be assessed in terms of the parameters that apply to adults, such as smoking and other health behaviors, we restrict our discussion here to the more commonly used epidemiological indicators.

### *Growth in Height and Weight in Adolescence*

Most countries have not developed standards for growth in height and weight during adolescence. The National Center for Health Statistics (NCHS) data, obtained for US children and adolescents [1], are often used for comparison. Maximum growth in median height in 1 year occurs during the 13th year in US boys (7.5 cm) and during the 11th year in girls (7.5 cm). This maximum in British children occurs around the 14th year and the 12th year, respectively [2]. Many children show slower or delayed growth mostly due to genetic and nutritional factors. A child's measurement can be compared with the NCHS **growth chart** or the chart of the child's native country (where available) to assess the progress of growth. The assessment is in terms of the **percentile** achieved.

Preece and Baines [3] have developed models that can be used to evaluate height parameters such as height for age, velocity at start of growth spurt, age at peak height velocity, and peak height velocity. These parameters have special relevance to the adolescent group.

Short stature in adolescent girls that persists into adulthood is associated with increased risk of adverse reproductive outcomes. The risk of low-birth-weight babies, cephalopelvic disproportion, dystocia, and cesarean section increases in short mothers. No specific health risk is known for short-statured boys.

### *Sexual Maturity Rating*

Breast development in females, the appearance of pubic hair in both boys and girls, and the development of male genitalia are graded into stages from 1 to 5, where the first is the preadolescent stage and the last is the fully matured adult stage. One can think of these stages as scores and use them as measures of the extent of pubertal development. This can be related to age to find whether or not development is on course. Clear guidelines on this are not yet available and need to be developed separately for each population depending on the rate of sexual maturation generally seen in healthy adolescents in that population.

1. CDC. *United States: Growth Charts, 2000*. http://www.cdc.gov/growthcharts, last accessed June 1, 2015.
2. Cole TJ, Freeman JV, Preece MA. British 1990 growth reference centiles for weight, height, body mass index and head circumference fitted by maximum penalized likelihood. *Stat Med* 1998;17:407–29. http://www.ncbi.nlm.nih.gov/pubmed/9496720
3. Preece MA, Baines MJ. A new family of mathematical models describing the human growth curve. *Ann Hum Biol* 1978;51:1–24. http://informahealthcare.com/doi/abs/10.1080/03014467800002601?journalCode=ahb

## adult health (epidemiological indicators of)

The health of an adult can be measured in several dimensions. The discussion in this biostatistics book is restricted to some epidemiological indicators of general health and excludes particular diseases. Nonetheless, physiological functions such as of the kidney, liver, lung, and heart can be measured to assess health at an individual level.

Among epidemiological indicators of adult health, the most prominent is **obesity**. As described under that topic, this can be measured by body mass index, waist–hip ratio, skinfold thickness, and several other indicators. Second is smoking. For comprehensively measuring this, see the **Indrayan smoking index**, which includes not just cigarette-years but also age at the start of smoking; intermittent smoking; smoking of other products such as cigars, pipes, and *bidi*; years since cessation by ex-smokers; etc. Different aspects of **quality of life** are measured in social, mental, and physical health domains. Several other such measures can be devised to meet specific objectives.

## adult literacy rate, *see* education indicators

## adverse effects and adverse patient outcomes

Adverse effects of medical interventions are those effects of an intervention that are detrimental to the health of the person who gets the intervention. These effects could be transient or long term.

A

These may also be called **side effects**. The intervention could be advice, a drug, surgery, a device, or any such maneuver. No intervention is completely free of side effects—thus, the convention is to refer to relatively serious side effects as adverse effects. This makes it subjective, and no widely acceptable definition is available for severity. However, there is some consensus regarding when a side effect can be called rare and when it is common. This is as follows:

Very common: ≥1/10
Common: ≥1/100 and <1/10
Uncommon: ≥1/1000 and <1/100
Rare: ≥1/10,000 and <1/1000
Very rare: <1/10,000

Adverse effects are gradually gaining prominence, perhaps even more than efficacy, as awareness is increasing. Patients do not want to end up with some other ailment while undergoing treatment of their disease. They also want to be fully warned of the adverse effects up front so that an informed decision can be made to go ahead with the treatment or not, or to look for an alternative strategy. For this reason, the pharmaceutical and medical device companies are now more sensitive to such effects of their product. Regulatory agencies also seek data on adverse effects from the companies who want license to produce and market such products.

Although phase III trials (see **phases of clinical trials**) are designed to gather the data on adverse effects, the follow-up cannot be very long in this phase. Some adverse effects show up after sustained use of the drug for chronic diseases such as hypertension and diabetes that need almost lifelong treatment. Some adverse effects show up long after the treatment is discontinued. Thus, **postmarketing surveillance** (phase IV trial) is considered the most dependable methodology to discover such effects. There are many instances when a drug or a treatment strategy is retracted when such adverse effects are reported. After years of intake of aspirin as a preventive medication for minimizing the chance of heart attacks, the Food and Drug Administration (FDA) of the United States now concludes that the data do not support the use of aspirin as a preventive medication by people who have not had a heart attack, stroke, or cardiovascular problems [1]. In such people, the benefit has not been established, but risks—such as dangerous bleeding into the brain or stomach—are still present. You can see how adverse effects can remain unnoticed for years when millions have adopted the drug in good faith.

Another related term is *adverse patient outcomes*. Here, the focus is on the outcome instead of the intermediary events. Instead of being improved, if the health of a patient deteriorates or experiences adverse effects after the intervention, this is called adverse patient outcome. This can occur either because of inadequate care in a health facility or due to inappropriate intervention. Lapses such as late or missed medication can also cause adverse patient outcomes. Nosocomial and urinary tract infections in inpatients and complaints from patients are commonly used criteria for classifying them as adverse patient outcomes. The magnitude of such outcomes is a potent tool for measuring the quality of services and of interventions. Statistically, this is used as an index for **quality control** in a health setup. Sometimes, adverse patient outcomes form a basis for insurance claims or any such compensation to the patient.

Beside categorization into care- and intervention-related categories, adverse effects may also be classified as perceived or reported; preventable or nonpreventable; temporary or permanent; and also minor, major, or grave. For quality control purposes, preventable adverse outcomes are important. Two major sources attributed to such adverse outcomes are poor nursing services and toxic effects of drugs. For nurses, this could be due to either heavy workload [2] or inadequate training [3]. This source obviously applies to a hospital setup only, whereas toxic effects of drugs will mostly occur in a domiciliary setup.

1. FDA. *Can an Aspirin a Day Help Prevent a Heart Attack?* http://www.fda.gov/ForConsumers/ConsumerUpdates/ucm390539.htm, last accessed May 24, 2015.
2. Hinno S, Partanen P, Vehviläinen-Julkunen K. Nursing activities, nurse staffing and adverse patient outcomes as perceived by hospital nurses. *J Clinic Nurs* 2012;21:1584–93. http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2702.2011.03956.x/abstract
3. Youngberg BJ. *Patient Safety Handbook*. Jones & Bartlett, 2012.

## affinity, *see also* measures of dissimilarity and similarity

In statistics, affinity is the closeness of one set of values with another set. The term is generally used for groups rather than individuals. Similarity seems a more appropriate term for individual values. However, affinity between groups depends on similarity of individual values belonging to different groups.

Conventional similarity measures are applicable to quantitative values. The value 12 is closer to the value 14 than the value 8. This is easily seen in a univariate case such as this but not as obvious in a **multivariate** situation. Consider aspartate aminotransferase (AST) units per liter, alanine aminotransferase (ALT) units per liter, and total bilirubin (mg/dL) in persons suspected to have liver disease. Let these values in three persons be (22, 35, 3.2), (20, 38, 2.8), and (25, 27, 3.2). In such a multivariate setup, which two subjects have more similarity with regard to these values? Or, which of the other two persons has more similarity with the first person with respect to these parameters? See **measures of dissimilarity and similarity** in this volume, which can be used to assess similarity between individuals in different groups. Some of these can be used for assessing affinity in groups in a multivariate setup.

For affinity in groups with qualitative values, separate measures are available for binary categories and for polytomous categories. This is measured by associations as described under the topics **association between dichotomous characteristics** and **association between polytomous characteristics**.

Affinity measures are needed for **cluster analysis** and used in some other statistical setups also. But the term is frequently used in the medical context as well with varied meaning. Affinity of chromatography columns is commonly assessed for, say, binding of drugs. Galli et al. [1] discuss the affinity of endocrine active substances and their metabolites toward the ligand binding domain of the androgen receptor in three distantly related species, namely, human, rat, and zebrafish. This needs quite complex mathematics. Chaudhury et al. [2] talk about B-cell affinity maturation, which explains enhanced antibody cross-reactivity induced by a malaria vaccine. These illustrate the varied uses of the term in medical literature.

1. Galli CL, Sensi C, Fumagalli A, Parravicini C, Marinovich M, Eberini I. A computational approach to evaluate the androgenic affinity of iprodione, procymidone, vinclozolin and their metabolites. *PLoS One* 2014 Aug 1;9(8):e104822. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4128724/

2. Chaudhury S, Reifman J, Wallqvist A. Simulation of B cell affinity maturation explains enhanced antibody cross-reactivity induced by the polyvalent malaria vaccine AMA1. *J Immunol* 2014 Sep 1;193(5):2073–86. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4135178/

## age–period–cohort analysis

Age–period–cohort analysis is used to segregate the effect of age, period, and cohort on an outcome in a **time series** setup where observations at different ages are available for people born over a long period of time. Age, the calendar date of the study (called *period*), and the year of birth (called *cohort*) all can have an independent effect on health parameters. For example, people born in 1930 may have had different health problems at the age of 65 years than those born in 1960. This can happen because the nutrition pattern may have changed in these 40 years; people may have become more conscious about exercise, life may have become more stressful, etc. In other words, **birth cohort** itself may have an independent effect. The age effect is already well known—health problems at the age of 65 years are different than, say, at the age of 30 years. *Period* is the term used for the calendar year of the end point. For those born in 1930, age 65 years is reached in 1995, and for those born in 1960, this period is the year 2025. Results may differ just because the period differs. This can happen due to factors such as changes in the inclusion criteria, different **sampling techniques**, and migration in surveys at different points in time. Those aged 65 years in 2025 may be exposed to much better connectivity, and this could have revolutionized their attitude and care compared to those aged 65 in 1995.

In this setup, the change in values at the end point relative to the starting point contains the effect of age, period, and cohort. The statistical problem is to find the independent effect of each of these factors. The challenge arises due to the fact that period = cohort + age. Given any two, the third can be exactly obtained. Because of this rather perfect **collinearity**, the usual **regression** techniques fail, and all three cannot be simultaneously estimated within the usual regression setup. A method called **partial least squares** is generally used in this setup. This method extracts components of independent variables that have the highest covariance with the dependent variable. For example, Jiang et al. [1] studied trends in body mass index (BMI) in males and females of different ages in Ireland by measuring them in the years 1998, 2002, and 2007. They found the first extracted component $w_1$ = 0.029 * age + 0.039 * period − 0.026 * cohort for males. This shows that men born later had lower BMI than those born earlier (cohort effect = −0.026 units per year as per the coefficient in this component), older men had higher BMI than younger men at the time of examination (+0.029 units per year), and the period effect at +0.039 units per year in these subjects was the highest among these three [1]. The last shows that, possibly, the survey methods at different points in time were different, and that also contributed to the decrease in BMI over this period.

Another example of the cohort effect is as follows. Recently, a controversy erupted over a cardiovascular risk model in the US population of age 40–79 years. It was found to overestimate the risk by almost twice as much as it should in some cases. It appears that the problem arose because the cohorts underlying the model are over a generation old, with enrollment years beginning in the late 1980s and early 1990s. Major population changes have occurred in the past two decades. People now smoke less, and awareness about the harms of high cholesterol has increased. Thus, the risk one generation ago is unlikely to represent the risk of contemporary 40- to 79-year olds [2]. This is in addition to the fact that those who were 40 years old at that time are 70 years old now (although they continue to be in the age bracket of 40–79 years) and have different risk due to aging and longer exposure to the risk factors.

In place of partial least squares, the **mixed effects models** can also be used for age–period–cohort analysis [3]. The method is too intricate for this book for medical professionals.

1. Jiang T, Gilthorpe MS, Shiely F, Harrington JM, Perry IJ, Kelleher CC, Tu YK. Age–period–cohort analysis for trends in body mass index in Ireland. *BMC Public Health* 2013;13:889. http://www.biomedcentral.com/1471-2458/13/889
2. Kovalchik S. *New Cardiovascular Risk Calculator Is Too Risky*. Statslife, Royal Statistical Society. http://www.statslife.org.uk/significance/1075-new-cardiovascular-disease-risk-calculator-is-too-risky
3. Jaacks LM, Gordon-Larsen P, Mayer-Davis EJ, Adair LS, Popkin B. Age, period and cohort effects on adult body mass index and overweight from 1991 to 2009 in China: The China Health and Nutrition Survey. *Int J Epidemiol* 2013;42:828–37. http://ije.oxfordjournals.org/content/42/3/828.full.pdf+htm

## age-specific death rates, *see* death rates

## agglomerative methods, *see* hierarchical clustering

## agreement assessment (overall), *see also* Bland–Altman method of agreement

When a measurement is taken by two methods, at two times, at two sites, by two observers, etc., on the same set of subjects, the assessment of agreement between these measurements can be useful in many situations. For example, new instruments are invented and new methods are discovered that measure anatomical and physiological parameters with less inconvenience and at a lower cost. We do want to assess how much they agree with the existing method. Acceptance of any new method depends on a convincing demonstration that it is nearly as good as the established method, although the established method may also be in error.

Irrespective of what is being measured, it is highly unlikely that the new method would give exactly the same reading in each case as the old method even if they are equivalent. Some differences would necessarily arise. How do you decide that the new method is interchangeable with the old? The problem is described as one of agreement. This is different from evaluating which method is better. The assessment of "better" is done with reference to a gold standard. Assessment of agreement does not require any such standard.

The term *agreement* is used in several different contexts. The following discussion is mostly restricted to a setup where a pair of observations ($x$, $y$) is obtained by measuring the same characteristic on the same subject by two different methods, by two different observers, by two different laboratories, at two anatomical sites, etc. Later, we briefly describe a method for assessing agreement when the measurements are different, such as body mass index (BMI) and skinfold thickness for assessing obesity.

The measurement could be qualitative or quantitative. The method of assessing agreement in these two setups is different.

A

## Agreement in Quantitative Measurements

The problem of agreement in quantitative measurement can arise in at least five different types of situations: (i) comparison of self-reported values with the instrument-measured values, for example, urine frequency and bladder capacity by patient questionnaire and frequency–volume chart; (ii) comparison of measurements at two different sites, for example, paracetamol concentration in saliva and that in serum; (iii) comparison of two methods, for example, bolus and infusion methods of estimating hepatic blood flow in patients with liver disease; (iv) comparison of two observers, for example, duration of electroconvulsive fits reported by two psychiatrists on the same group of patients, or of two laboratories when, for example, aliquots of the same sample are sent to two different laboratories for analysis; (v) **intraobserver consistency** in evaluating reliability of the method or of the observer, for example, measurement of anterior chamber depth of an eye segment two times by the same observer using the same method. In the first four cases, the objective is to find whether a simple, safe, less expensive procedure can replace an existing procedure. In the last case, it is an evaluation of the reliability of the method.

The statistical problem in all these cases is to check whether or not a $y = x$ type of relationship exists in *individual* subjects. This looks like a regression setup $y = a + bx$ with $a = 0$ and $b = 1$, but that really is not so. The difference is that, in regression, the relationship is between $x$ and the average of $y$. In an agreement setup, the concern is with individual values and not with averages. Nor should agreement be confused with high correlation. Correlation would be nearly 1 even if there is a systematic bias and nearly the same difference occurs in every subject. The following example illustrates the distinction between $y = x$ regression and agreement.

The following are hemoglobin (Hb) values reported by two laboratories for the same blood samples:

| | |
|---|---|
| Lab 1 ($x$) | 11.3 12.0 13.9 12.8 11.3 12.0 13.9 12.8 |
| Lab 2 ($y$) | 11.5 12.4 14.2 13.2 11.1 11.6 13.6 12.4 |

$$\bar{x} = 12.5 \quad \bar{y} = 12.5 \quad r = 0.945$$

$$\hat{y} = x \text{ , i.e., } b = 1 \quad \text{and} \quad a = 0$$

The two laboratories have the same mean for these eight samples and a very high correlation (0.945). The **intercept** $a$ is 0, and **slope** $b$ is 1.00. Thus, the regression is $y = x$. Yet there is no agreement in any of the subjects. The difference or error ranges from 0.2 to 0.4 g/dL. This is substantial for Hb level in the context of the present-day technology. Thus, equality of means, a high degree of correlation, and regression $y = x$ are not enough to conclude agreement. Special methods are required.

If you notice carefully, the first four values of $x$ in this example are the same as the last four values. The first four values of $y$ are higher and the last four values are lower by same margin. Thus, for each $x$, $\bar{y} = x$, giving rise to the regression $\hat{y} = x$. In this particular case, the correlation coefficient also is nearly 1.

## Approaches for Measuring Quantitative Agreement

The conventional method for measuring agreement in quantitative measurements is the popular **Bland–Altman method** [1], also called the limits of disagreement approach. The other approach is **intraclass correlation** [2]. Both of these are described in this volume under the respective terms. Indrayan [3] has compared these two approaches and listed their merits and demerits. Alternative methods are as follows.

The limits of disagreement approach of Bland and Altman is based on the average difference and has the same limitation as applicable to all averages. For example, this approach does not work if the bias or error is proportional. Fasting blood glucose level varies from 60 to 300 mg/dL or more. Five percent of 60 is 3 and of 300 is 15. The limits of disagreement approach considers them different and ignores that both are 5% and proportionately the same. Also, if one difference is 10 and the other is 2, not necessarily proportional, the limits of disagreement approach generally considers the average. Individual differences tend to be overlooked as outliers. This ignores the fact that a few unusually large differences distort the average and are not properly accounted for except by disproportional inflation of the standard deviation (SD).

To account for small and big individual differences as well as proportional bias, it may be prudent to set up a clinical limit that can be tolerated for individual differences without affecting the management of the condition. Such limits are required anyway for the limits of agreement approach also, albeit for the average. These clinical limits of indifference can be absolute or in terms of percentage. If not more than a prespecified, say, 4% of individual differences are beyond these limits in a large sample, you can be safe in assuming adequate agreement. This does not require any calculation of mean and SD. You may want to add a condition, such as that none of the differences should be more than two times the limit of indifference. Any big difference, howsoever isolated, raises alarm. A plot of $y$ versus $x$ can track that the differences are systematic or random.

As an example, consider the following data on fasting blood sugar level in 10 blood samples.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method 1 ($x$) | 86 | 172 | 75 | 244 | 97 | 218 | 132 | 168 | 118 | 130 |
| Method 2 ($y$) | 90 | 180 | 73 | 256 | 97 | 228 | 138 | 172 | 116 | 132 |
| $d = x - y$ | –4 | –8 | +2 | –12 | 0 | –10 | –6 | –4 | +2 | –2 |
| 5% of $x$ | 4.30 | 8.60 | 3.75 | 12.20 | 4.85 | 10.90 | 6.60 | 8.40 | 5.90 | 6.50 |

Suppose method 1 is the current standard, although this can also be in error. Method 2 is desperately cheap and gives instant results. For example, method 1 could be the usual venous blood–based estimation of glucose level, and method 2, based on capillary blood. Suppose also that clinicians are willing to accept 5% error in view of distinct advantages of method 2. Note that this indifference is in percentage and not an absolute value.

None of the differences exceed the clinical limit of indifference in this sample. Thus, method 2 can be considered in agreement with method 1, although a larger sample is required to be confident. However, most differences are negative, indicating that method 1 generally provides lower values. The average difference is 4.2 mg/dL in absolute terms and nearly 3% of $y$ in relative terms. This suggests the correction factor for bias. If you decide to subtract 3% of the level obtained by method 2, you can get very close to the value obtained by method 1 in most cases.

Now forget about 5% tolerance, and note that some differences are small and some are quite large in this example. Since the SD of the differences $s_d = 2.80$ in this case, the Bland and Altman limits of disagreement are

$$-4.2 \pm 2 \times 2.80, \quad \text{or} \quad -9.8 \quad \text{to} \quad +1.4.$$

These limits may look too wide and beyond clinical tolerance, particularly on the negative side. These limits do not allow larger error for larger values that proportionate considerations would allow. Also, these are based on the average and do not adequately consider individual differences. If 1 out of 10 values shows a big difference, this can distort the mean, inflate the SD and provide unrealistic limits of disagreement. The alternative approach just suggested can be geared to allow not more than 5% individual differences beyond the tolerance limit, and you can impose an additional condition that none should exceed by, say, 10% of the base value. Since it is based on individual differences and not the average, this alternative approach may be more appealing too.

Another modification of the Bland–Altman method is in comparison of two entirely different measurements. For example, you might be interested in finding out whether assessing obesity in individuals by BMI and skinfold thickness gives the same result or not. If they give the same result, BMI is much simpler. These two measurements are on entirely different scales, and the usual Bland–Altman method is not applicable. However, both these can be converted to ***z*-score** or ***T*-score** for each person. Once this is done, both are on the same scale and can be compared. Now you can use the Bland–Altman method on these converted scores and assess agreement as usual.

### *Agreement in Qualitative Measurements*

Assessment of optical disk characteristics by two or more observers, results of Lyme disease serological testing by two or more laboratories, and comparison of x-ray images with Doppler images are examples of the problem of qualitative agreement. The objective is to find the extent of agreement between two or more methods, observers, laboratories, etc. In some cases, for example, in a comparison of two laboratories, agreement has the same interpretation as **reproducibility**. In the case of comparison of observers, it is termed **interrater reliability**. In all these cases, only one group of subjects is assessed twice. Thus, this is a matched pairs setup.

For simplicity, for the time being, consider only the presence or absence of a characteristic assessed by two observers in the same group of subjects. An example is the presence or absence of a lesion in x-rays read by two observers. Suppose the observations are as given in Table A.2.

The two observers agree on a total of (29 + 11 =) 40 cases in this example. This is the sum of the frequencies in the leading diagonal cells. In the other 20 cases, the observers do not agree. Apparently, the agreement = 40/60 = 66.7%. But part of this agreement is due to chance, which might happen if both are dumb observers and randomly allocate subjects to present and absent categories. This chance agreement is measured by the cell frequencies expected in the diagonal when the observers' ratings are independent of one another. These expected frequencies are obtained by multiplying the respective marginal totals and dividing by the grand total, as obtained for calculating the **chi-square**.

**TABLE A.2**
**Presence or Absence of a Lesion Assessed by Two Radiologists in X-Rays of 60 Suspected Cases**

| | Observer 1 | | |
|---|---|---|---|
| **Observer 2** | **Present** | **Absent** | **Total** |
| Present | 29 | 7 | **36** |
| Absent | 13 | 11 | **24** |
| **Total** | **42** | **18** | **60** |

For the data in Table A.2, the chance-expected frequencies are $36 \times 42/60 = 25.2$ and $24 \times 18/60 = 7.2$ in the two diagonal cells. The total of these two is 32.4. Agreement on so many cases is expected by chance alone. Thus, agreement in excess of chance is in only 40 – 32.4 = 7.6 cases. The maximum possible excess is 60 – 32.4 = 27.6. A popular measure of agreement is the ratio of the observed excess to the maximum possible excess, in this case, 7.6/27.6 = 0.275 or 27.5%. Thus, the two observers in this case do not really agree much on rating of x-rays for the presence or absence of lesion. Most of their agreement is due to chance.

The method just described is for 2 × 2 tables. A general method for assessing quantitative agreement is **Cohen kappa**. This can be used when each measurement is categorized into more than two categories. In the case of ordinal categories, it also raises the possibility of near agreement, such as one method rating an object as excellent and the other rating it as good. They do not exactly match but are close. For this, one can use weighted kappa, as discussed under Cohen kappa or the **Bangdiwala *B***-statistic [4]. There are several other measures of **association between dichotomy characteristics**, as separately discussed. These include the Jaccard coefficient and Yule *Q*. These actually are for association in dichotomy categories but can be construed to measure agreement as well.


1. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;i:307–10. http://www-users.york.ac.uk/~mb55/meas/ba.pdf, last accessed November 21, 2013.
2. Lee J, Koh D, Ong CN. Statistical evaluation of agreement between two methods for measuring a quantitative variable. *ComputBiol Med* 1989;19:61–70. http://www.ncbi.nlm.nih.gov/pubmed/2917462
3. Indrayan A. *Medical Biostatistics*, Third Edition, CRC Press, Boca Raton, FL, 2012.
4. Munoz SR, Bangdiwala SI. Interpretation of Kappa and B statistics measures of agreement. *J Appl Stat* 1997;24(1):105–12. http://www.tandfonline.com/doi/abs/10.1080/02664769723918#.UvD4ydKSxIA


## agreement charts

These charts graphically depict the extent of **agreement** between two **ordinal** measures. This applies when the same attribute is assessed by two methods, two observers, two sites, etc. on *n* different subjects by using an ordinal scale, and the objective is to depict the extent of concordance. The corresponding numerical measure is **Cohen kappa**. This chart serves as a good complement to the value of kappa but cannot replace it.

Table A.3 contains data on diagnoses of the same 69 suspected cases of multiple sclerosis by two independent neurologists. The diagnosis is in four ordinal categories from "certain" to "no." The agreement between the neurologists is clear for the cases in the diagonal, i.e., 5, 11, 3, and 14. These are shown by dark rectangles in Figure A.2. This would be called an agreement chart even when the shades of gray are absent. Such a chart without shades of gray can be drawn for nominal categories also. Shades of gray can be explained as follows for ordinal categories.

For ordinal categories as in Table A.3, we can go a step further, realizing that there is some degree of agreement between contiguous

A

**TABLE A.3**
**Assessment of Multiple Sclerosis in 69 Cases by Two Neurologists**

| Multiple Sclerosis Diagnosis | | Winnipeg Neurologist | | | | Total |
|---|---|---|---|---|---|---|
| | | Certain | Probable | Possible | No | |
| New Orleans neurologist | Certain | 5 | 3 | 0 | 0 | **8** |
| | Probable | 3 | 11 | 4 | 0 | **18** |
| | Possible | 2 | 13 | 3 | 4 | **22** |
| | No | 1 | 2 | 4 | 14 | **21** |
| **Total** | | **11** | **29** | **11** | **18** | **69** |

*Source:* Bangdiwala SI, Shankar V. *BMC Med Res Methodol* 2013 Jul 29;13:97. http://www.biomedcentral.com/content/pdf/1471-2288-13-97.pdf.

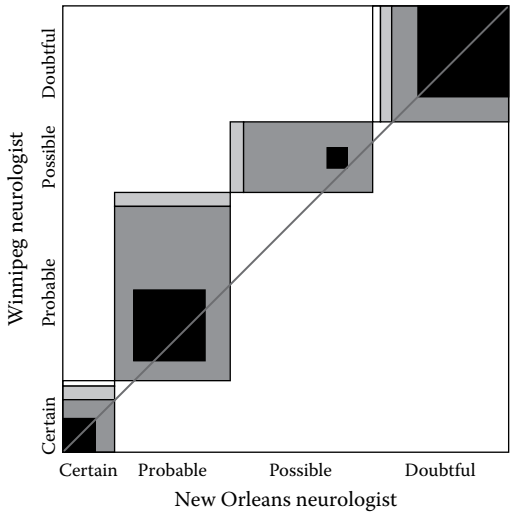



**FIGURE A.2** Agreement chart for comparing multiple sclerosis diagnosis by two neurologists. (From Bangdiwala SI, Shankar V. *BMC Med Res Methodol* 2013 Jul 29;13:97. http://www.biomedcentral.com/content/pdf/1471-2288-13-97.pdf.)

categories and less agreement between distant categories. This means that if one observer calls it "probable" and the other "possible," there is some agreement compared with if one says "probable" and the other says "no." This can be termed *partial agreement* of different degrees and can be shown by progressively declining shades of gray as in Figure A.2. The areas are proportional to the numbers in different cells of Table A.3.

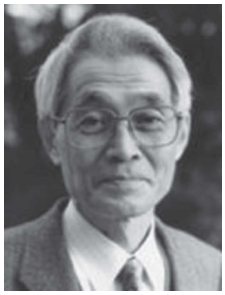
The actual procedure for an agreement chart is to draw $K$ rectangles of dimensions proportional to the respective row and column totals at the diagonal position from the lower left to the upper right inside an $n \times n$ square, where $K$ is the number of categories and $n$ is the total number of subjects. Then these rectangles are filled up with black in the proportion of the complete agreement and shades of gray for partial agreement of different degrees. These charts can be directly obtained through some statistical software packages, such as SAS.

1. Bangdiwala SI, Shankar V. The agreement chart. *BMC Med Res Methodol* 2013 Jul 29;13:97. http://www.biomedcentral.com/content/pdf/1471-2288-13-97.pdf

## AIC, *see* Akaike information criterion (AIC) and general AIC (GAIC)

## Akaike information criterion (AIC) and general AIC (GAIC)

The Akaike information criterion (AIC) is used to select one of many potential **models** such that the chosen model is relatively simple yet is a good fit to the data. A paper on this criterion was first published by Hirotugu Akaike in 1974 [1], but it became popular after its English generalization appeared in 2002 [2]. One of its early applications is in evaluating pharmacokinetic equations [3].


Hirotugu Akaike

AIC is based on the method of **maximum likelihood**. This method provides those estimates of the parameters of a model that make the observed sample values most likely. For computational convenience, log-likelihood is maximized in place of the likelihood itself. Logarithms tend to linearize the likelihood function in most cases. Since likelihood is a probability between 0 and 1, log-likelihood will always be negative. Thus, instead of log-likelihood, $-2\ln L$ is used, where $L$ is the notation for likelihood. This not only makes it positive but also is easy to handle as the distribution of $-2\ln L$ is known (chi-square) for many situations. Because of the minus sign in this, those estimates are chosen in a model that minimizes $-2\ln L$. This is the same as maximizing $L$.

In the case of multiparameter models such as **regression**, it is known that the larger the number of independent variables, the better the fit. In this case, the regression coefficient of each independent variable is a parameter. However, an increased number of parameters also makes the model more complex, less parsimonious, and more difficult to adopt in practice. Thus, the aim is to choose a model that has a minimum number of parameters, yet the likelihood

remains relatively high. To balance the number of parameters and the likelihood, the number of parameters is used as a *penalty* on the likelihood. This gives rise to the AIC, defined as

$$\text{AIC} = -2\ln L + 2K,$$

where $K$ is the number of free parameters. This $K$ is similar to the *degrees of freedom* in other setups. Note that AIC penalizes likelihood for a higher number of parameters. The parameter estimates obtained by this method are called **penalized likelihood** estimates. This method obtains those estimates that minimize AIC. This is an approach in between goodness of fit and simplicity where parsimony is balanced with good predictivity. The following comments may be helpful in application of AIC to real-life situations.

- To answer the question "From which model have our observed values most likely come?", calculate AIC for all potential models and select the one with minimum AIC. Statistical packages are available that do this calculation for you.
- AIC is asymptotic, i.e., it works well only for large $n$.
- AIC is mostly used for regressions or models depicting trend. This is considered an improvement over **stepwise methods** since some aspects of stepwise methods are subjective.

The penalty in AIC is $2K$. In some situations, a higher number of parameters severely compromises parsimony. In those situations, you can incorporate a higher penalty for the number of parameters. When this is done, this is called generalized AIC (GAIC). In general, you can have

$$\text{GAIC}(p) = -2\ln L + pK.$$

For example, in the case of the **Box–Cox power exponential (BCPE) method** of finding centile curves for the growth of children, $p = 3$ is preferred. In the usual AIC, $p = 2$.


1. Hirotugu A. A new look at the statistical model identification. *IEEE Trans Automat Control* 1974;19 (6):716–23. http://link.springer.com/chapter/10.1007%2F978-1-4612-1694-0_16#page-2
2. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Second Edition. Springer-Verlag, 2002.
3. Yamaoka K, Nakagawa T, Uno T. Application of Akaike's information criterion (AIC) in the evaluation of linear pharmacokinetic equations. *J Pharmacokinet Biopharm*1978;6:165–75. http://link.springer.com/article/10.1007/BF01117450#page-1


## aleatory uncertainties, *see also* epistemic uncertainties

Aleatoricism is the incorporation of chance into the process of creation. The word derives from Latin word *alea*, the rolling of dice. Aleatory uncertainties can be understood as those arising from factors internal to the system. They are inherent, unpredictable, and stochastic in nature. For further details, see Indrayan [1]. The other type of uncertainty is epistemic, which arises from limitation of knowledge. The details of **epistemic uncertainties** are separately given. We can divide them into the following categories for effective control.

- Biological—nonmodifiable (age, gender, heredity or genetic makeup, birth order, height, etc.)
- Biological—modifiable (anthropological, physiological, biochemical parameters)
- Socioeconomic factors (income, education, and occupation) that can affect personal hygiene, nutrition, and self-esteem
- Cultural, behavioral, and psychological (mental status, family system, faith in prayers, sexual practices, addictions, personality traits, tension–anxiety–stress, etc.)
- Observers, instruments, and laboratories, i.e., avoidable variation in measurements
- Environmental (climate, dust, mosquitoes, flies, pollution, sanitation, water supply, infection load, quality and quantity of health facilities, family and societal support, communication, traffic, laws and their enforcement, etc.)
- Multifactorial (lifestyle, hygiene, nutrition, knowledge–attitude–practices, susceptibility, utilization of health services, etc.; importantly, **sampling errors/fluctuations**)

One important feature of aleatory uncertainties is that they are **empirical** and can be evaluated by **probability**. The second aspect of aleatory uncertainty, particularly in medical research, is its control. This is achieved by developing a design that can provide evidence largely free of aleatory variation encumbrances. Since the sources of aleatory uncertainties are known, an appropriate design can indeed be developed. **Uncertainty analysis** is one of the tools that help to delineate aleatory uncertainties by providing intervals within which the results are likely to lie under varying conditions.


1. Indrayan A. *Fundamentals of Medical Research: For Emerging Researchers.* Lap Lambert Academic Publishing, 2012.


## algorithm (statistical)

An algorithm is a set of systematic rules for accomplishing a particular objective. The diagnostic algorithm, for example, is to elicit signs and symptoms, order and assess laboratory and radiological investigations, place all the information in the context of the patient concerned, and evaluate the probabilities of various competing diagnoses. In biostatistics, the term is generally used for the specifics of a step-by-step procedure to be followed in a statistical method. Many times, they refer to the computer coding for various calculations. If this is your interest, you may want to see Weihs et al. [1] for statistical algorithms in freely available R packages.

In the case of **stepwise regression**, for example, the algorithm could either be forward selection, backward elimination, stepwise, or best subset. Each of these has steps of what to do. In **classification and regression trees**, the tree algorithm, as the name suggests, devises split nodes that generate branches by using if-then rules. **Path analysis** provides an algorithm for decomposing effect into direct and indirect within the postulated model. For **cluster analysis**, two broad categories of algorithms are hierarchical agglomerative and hierarchical divisive. For **reestimation of sample size** in case of **adaptive trials**, many algorithms are available, such as those devised by **O'Brien–Fleming** and **Lan–deMets**. A recent development is statistical algorithms for monitoring gene expression (see, for example, Ref. [2]). There are many other statistical algorithms that can be cited.

A

1. Weihs C, Mersmann O, Ligges U. *Foundations of Statistical Algorithms: With References to R Packages.* Chapman & Hall/CRC Press, 2013.
2. Affymetrix. *New Statistical Algorithms for Monitoring Gene Expression on GeneChip® Probe Arrays.* Technical Note. http://sgdlite.princeton.edu/statistical_algorithms_technote.pdf, last accessed May 24, 2015.

## allocation of subjects, *see also* random allocation

The term *allocation of subjects* is generally used in the context of **clinical trials** where the trialist has to decide which person or patient will receive which "treatment." These treatments could be a **placebo**, an existing standard, or a new regimen under trial and its variations such as different dosages or different routes of administration. Subjects receiving a particular treatment form a group. Thus, allocation of subjects is done to various groups with some specific objectives in mind.

The method followed for allocation of subjects to different groups has significant bearing on the results. If mostly severe cases or mostly old-age cases happen to receive the new treatment, the trial may unnecessarily provide evidence against the **efficacy** of the treatment. Thus, a preset and well-thought-out method is followed for allocation of the subjects. This method is explicitly stated in the **protocol** itself as part of the design of the trial much before the trial begins.

The most popular method and perhaps the most appropriate is **random allocation**. This gives an equal chance to the eligible subjects of being allocated to any of the groups under trial. Thus, many unknown or unaccounted-for factors that can influence the results tend to equalize among the groups, and this helps in getting unbiased results. Subjective allocation, such as alternating between the groups, is not considered random even if the investigator justifies it by saying that he/she allocated without any bias. It is necessary that a scientifically accepted method of random allocation be adopted.

Random allocation can be done in various ways, such as simple, **block, cluster, and stratified randomization**. Sometimes due to exigencies, **quasi-random allocation** is done. In addition, the allocation is, many times, concealed from the participants, resulting in **blinding**, so that unbiased responses can be elicited. The term *participants* in blinding refers to not only the subjects but also the assessors and possibly the data analysts, who can also be biased.

## alternative hypothesis, *see* null and alternative hypotheses

## alpha error, *see* level of significance

## alpha-spending function, *see also* Lan–deMets procedure

In some medical studies, particularly in a **clinical trial** setup, the data are sometimes periodically examined in an ongoing trial as per the predetermined scheme to assess if it is worth continuing or it is better to stop. Stopping can be either because sufficient evidence of the efficacy of the regimen under test has already emerged or because the interim data convincingly indicate that the regimen lacks the required efficacy—thus, it is futile to continue. When such **interim analyses** are done, each analysis commits its own **Type I error** of wrongly rejecting a **null hypothesis**. Thus, the total Type I error, also called alpha error, can exceed its stipulated limit. For keeping it under control, each comparison at interim stages is done at a much smaller level. In other words, a small portion of the available alpha error is spent at each stage of analysis. The rule that sets the limit of alpha error at each interim analysis is called the alpha-spending function. This function tells us how much alpha error can be allowed at the first interim analysis, how much at the second interim analysis, etc.

Among popular alpha-spending functions are the **O'Brien–Fleming procedure** and **Lan–deMets procedure**. These are described as separate topics. The alpha-spending function is flexible, and its details and those of the interim analysis proposed to be used should be fully specified with complete justification in the **protocol** itself so that no subjectivity creeps in at a later stage.

## analysis of covariance (ANCOVA)

This is a method of adjusting the effect of a covariate on a quantitative outcome while preserving the effect of the real factors under study. You may be aware that in both **regression** and **analysis of variance (ANOVA)**, the dependent variable is quantitative. However, in the case of regression, the independents too are quantitative, whereas in the case of ANOVA, they are qualitative. What should one do if the independent set contains both quantitative and qualitative factors and the dependent set continues to be quantitative as before? The answer is analysis of covariance (ANCOVA). You can see that ANCOVA sits between regression and ANOVA. The primary purpose of ANCOVA is to adjust the results for covariates that can affect the results—thus providing unbiased results for the effect of the main variables under study. This method is commonly used also for **adjusting for baseline** imbalances. ANCOVA is a generalized procedure and is mathematically intricate even when only the linear effect is considered. As for other mathematically intricate topics, we are explaining only the underlying principles and their application, and not the formulas.

Consider determinants of body mass index (BMI) of adults of aged 20–49 years. This is the dependent variable $y$ in this setup and is calculated exactly as weight/height.[2] Among several determinants of BMI that can be considered, the important ones are gender (male or female); physical activity (none, mild, moderate, or heavy); and fat and carbohydrate intake (g/day). Note that these variables affect each other, and any conclusion based on any one variable in isolation may be misleading. Thus, all the variables should be considered together. If the primary interest is in knowing the relationship between BMI and physical activity and gender, it needs to be adjusted for the accompanying differentials in dietary intake. In this situation, dietary intake of fat and carbohydrate are the covariates. This adjustment is done by ANCOVA. It is similar to ANOVA of BMI on gender and physical activity on **residuals** after removing the effect of dietary differentials. Adjusted analysis can give very different results from unadjusted analysis. The result also depends on what and how many covariates are included in the analysis.

For running ANCOVA, the quantitative covariates remain as they are, but the qualitative covariates are assigned values (0, 1) through the **indicator variable**. In the BMI example, suppose $x_1$ is the fat intake and $x_2$ the carbohydrate intake. For sex, the indicator variable could be $x_3$, with $x_3 = 0$ for males and $x_3 = 1$ for females. For physical activity, since there are four categories, the indicator variables can be defined as follows:

For no physical activity: $x_4 = 0$, $x_5 = 0$, $x_6 = 0$
For mild physical activity: $x_4 = 1$, $x_5 = 0$, $x_6 = 0$
For moderate physical activity: $x_4 = 0$, $x_5 = 1$, $x_6 = 0$
For heavy physical activity: $x_4 = 0$, $x_5 = 0$, $x_6 = 1$

The number of indicator variables is one less than the number of categories. But this takes care of all the categories, and no category is left out in the cold. Now a regression is run with $x_1, x_2$, $x_3$, $x_4$, $x_5$, and $x_6$ as the regressors. The corresponding **regression coefficients** measure the contribution of each variable and each category, and this contribution is now adjusted for the other variables included in the model. Note the following:

- All covariates must be prespecified in the trial **protocol**. If an unanticipated covariate appears at a later stage, the statistical and clinical reasons to include this in ANCOVA should be explicitly documented. Previous experience may also suggest the functional form of the role of these covariates in explaining or predicting the outcome. As much as possible, this form should also be specified in the protocol. Also think of interactions. The number of covariates should be small for clarity and should be chosen with care. A large number of covariates tends to complicate the inference besides, of course, requiring a large sample. This is more so when interactions are also included. Similarly, a large number of cross-classifications in the case of categorical covariates can cause problems even when the covariates are not many. In this case, examine whether some categories can be collapsed without sacrificing the utility of results. Computationally, a large number of categories does not cause problems when an adequate sample size is available.
- The method becomes complex when, for example, the interest is in the effect of $x_1$ and $x_3$ on $y$ keeping $x_2$, $x_4$, $x_5$, and $x_6$ fixed. This can also be stated as the effect of $x_1$ and $x_3$ on $y$ after eliminating the effect of $x_2$, $x_4$, $x_5$, and $x_6$—in our example, the effect of fat intake and sex on BMI in people with the same carbohydrate intake and same the physical activity. Typically, covariates are quantitative, but in this case, physical activity is qualitative. The method of **general linear models** allows this also, and statistical packages do all the required calculations once the design command is properly specified. For example, Murphy et al. [1] studied the association between domestic physical activity and leanness controlled for age, gender, socioeconomic status, and smoking status using ANCOVA.
- Just as in regression and ANOVA, ANCOVA also is generally restricted to a linear effect and requires Gaussian distribution of residuals for the confidence interval and test of significance of the estimates of the parameters. Linearity is not in the restricted sense but in the wider sense where the square, cube, logarithm, etc., of independent variables are permissible but not of parameters.
- You can see that ANCOVA, as just explained, is another statistical method to reduce the specter of uncertainty. It adjusts the comparison across groups for imbalances in the covariates and enhances the precision by accounting for these imbalances. To check the validity of ANCOVA, homogeneity of slopes and absence of **interaction** between covariates and qualitative variables are tested. Thus, do ANCOVA with abundant caution. ANCOVA can be extended to include interaction between covariates and qualitative variables (that is, varying slopes) although this requires much more statistical expertise both for running the statistical package as well as for interpreting the output.

1. Murphy MH, Donnelly P, Breslin G, Shibli S, Nevill AM. Does doing housework keep you healthy? The contribution of domestic physical activity to meeting current recommendations for health. *BMC Public Health* 2013 Oct 18;13(1):966. http://www.biomedcentral.com/content/pdf/1471-2458-13-966.pdf

## analysis of variance (ANOVA), *see also* one-way ANOVA, two-way ANOVA, repeated measures ANOVA

Analysis of variance (ANOVA) is a method for testing the significance of the difference in means in three or more groups, analogous to the **Student *t*-test** for two independent groups. The method grew out of Ronald Fisher's investigations at the Rothamsted Research Centre of the effect of fertilizers on crop yield in the 1920s. This has now been extended to a large number of other complex problems such as multifactor setup, repeated measures, fixed and random effects, and regression setup. All these are discussed in this volume under the respective topics.

The method remains conceptually simple but becomes mathematically complex. As always in this book, we avoid complex mathematical expressions and concentrate on explanations that may help in understanding the basic concepts, in being more judicious in choosing an appropriate method for a particular set of data, in realizing the limitations of the methods, and in interpreting the results properly.

The Student *t*-test is valid for almost any underlying distribution if $n$ is large but requires a **Gaussian** pattern if $n$ is small. A similar condition also applies to the ANOVA method of comparison of means in three or more groups. The test criterion now used is called $F$. As in the case of the Student $t$, this test also requires that the outcome variable be quantitative and the variance in different groups be nearly the same. This property is called **homoscedasticity**. The third, and most important, prerequisite for validity of $F$ is independence of observations. Serial measurements, taken over a period of time on the same unit, generally lack independence because a measurement depends on its value on the previous occasions. A different set of methods, called hierarchical or **repeated measures** and akin to a paired $t$, is generally applied for analyzing such data. But the dependence of measurements at different anatomical sites, such as electrical activity at different sites of, brain is more difficult to handle and may require **multivariate methods**.

Some kind of **random sampling** is required, as always for statistical methods, for validity of conclusions from the ***F*-test**. The usual ANOVA considers groups on a **nominal scale** with no order or quantitative implications. This implies that if your groups are like cases with mild disease, moderate disease, a serious form of disease, and critical cases, usual ANOVA would ignore this gradient in the groups. Further analysis or an alternative method such as **scoring** would be needed if you want to study the effect of this gradient on the outcome.

The name *analysis of variance* comes from the fact that the total variance in the subjects in all the groups combined is partitioned into components arising from different sources such as within-groups variance and between-groups variance. Between-groups variance is the systematic variation occurring due to group differentials. For example, ANOVA may reveal that 60% variation in P3 amplitude in

A

**TABLE A.4**
**ANOVA Table for a Two-Way ANOVA for Effect of BMI and WHR Categories on Triglyceride Level in 100 Subjects**

| Source | Sum of Squares | df | Mean Square | *F* | *P*-Value |
|---|---|---|---|---|---|
| Between BMI categories | 2369.334 | 3 | 789.778 | 4.079 | 0.009 |
| Between WHR categories | 3417.571 | 2 | 1708.786 | 8.826 | 0.000 |
| Interaction BMI * WHR | 872.474 | 6 | 145.412 | 0.751 | 0.610 |
| Error | 17037.355 | 88 | 193.606 | | |
| **Corrected Total** | **30412.960** | **99** | | | |

healthy adults is due to genetic differentials, 10% due to age differentials, and the remaining 30% due to other factors. Such a residual left after the extraction of the factor effects of interest is considered a random component arising due to intrinsic biologic variability between individuals and the unaccounted-for factors. This is the within-groups variance. If genuine group differentials are present, then the between-groups variance should be substantially large relative to the within-groups variance. Thus, the ratio of these two components of variance can be used as a criterion to find whether the group means are different. This is what the *F*-test does. The numerator of the variance is the sum of squares. The within-groups sum of squares has other names also, such as **error sum of squares** and *residual sum of squares*.

The results of ANOVA are presented in what is called an *ANOVA table* (for an example, see Table A.4), which contains various sums of squares, their respective degrees of freedom (df's), mean squares, the values of *F*, and their statistical significance in terms of *P*-value. Sum of squares has various types such as Type I, Type II, and Type III, as explained under **sum of squares**. Table A.4 shows results of a two-way ANOVA for triglyceride level as affected by body mass index (BMI) and waist–hip ratio (WHR) in a random sample of 100 subjects. Such a table can be easily obtained by using almost any statistical software package. Mean square is obtained by dividing the sum of squares by the respective df, and under the conventional setup, *F* is obtained by dividing each mean square by mean square due to error (MSE). The *P*-value comes from the *F* distribution under the null hypothesis of equality of means.

The ANOVA method in actual applications depends on the design. For example, a **one-way ANOVA** is used for comparing means in three or more groups without considering any other factor. A **two-way ANOVA** is used when the subjects are simultaneously divided by two factors, such as patients divided by sex and severity of disease. In Table A.4, these are divided by BMI and WHR categories. Repeated measures ANOVA is used when the same subjects are assessed on the same medical parameters on two or more occasions, such as pain score before the anesthesia for a surgery and at 30 s, 1 min, and 2 min after the anesthesia. Finding the sum of squares due to each regressor and the residual in a **regression** setup is also called ANOVA. There are several other applications of the ANOVA method. For further details of the ANOVA method, see Doncaster and Davey [1].

1. Doncaster CP, Davey AJH. *Analysis of Variance and Covariance: How to Choose and Construct Models for the Life Sciences*. Cambridge University Press, 2007.

## analysis (statistical)

Many people wonder what this so-called statistical analysis is all about. Analysis literally means breaking down a complex entity into smaller components so that it can be properly understood, comprehended, and examined. The term *statistical analysis* is applied mostly to data whose micro and macro features are studied through statistical methods. For example, you may be interested in sieving signals from noise in the data and arrive at a result that has credibility. The term *data analysis* is used every day for statistical analysis, but there is a subtle distinction. Statistical analysis necessarily has an inference component that requires that some conclusion be made regarding the population as a whole from which the data have been extracted. Data analysis may or may not have this component. Data analysis may help draw a conclusion regarding the data on hand without any extrapolation to the "population."

All statistical methods help in the statistical analysis of data. Different methods are used for different purposes depending on the objective of the study. It could be exploratory, descriptive, or inferential. The objective of **exploratory analysis** is to examine the data with regard to their adequacy for inferential purposes. This includes, among others, studying the **validity** of the data, assessment of missing values, and identifying the outliers. **Descriptive analysis** is mostly in terms of tabulation and graphs. These help in summarizing the data, finding the pattern in the data, and identifying the right methods of further analysis. The core statistical analysis is inferential, which includes finding the strength and form of relationships or trends, confidence intervals, and the tests of hypotheses, for the parameters of interest. Steps taken before the availability of data, such as designs of the study, framing of the forms for recording, and collection of data, are also statistical but not part of statistical analysis.

To give you some examples of core statistical analyses, regression *analysis* is done to find the best-suited relationship between a dependent variable and a set of regressors, cluster *analysis* is done to discover affinity structure in the data, survival *analysis* is done to find survival patterns and contributory factors, *analysis* of covariance is done to find the effect of factors and covariates, discriminant *analysis* is done to find the best function that separates groups from one another, etc.

Statistical analysis is recognized as an important ingredient in any empirical science that depends on observed values and not theorems. Medicine and health are particularly vulnerable because of high interindividual and intraindividual variations. This is aggravated by the need to be exact in conclusions regarding such vital aspects of life.

## analytical studies

Two broad types of data-based studies are descriptive and analytical. The objective of **descriptive studies** is to find the health status of a group of subjects without worrying about the reasons or contributory factors. These are enumerative in nature since the objective mostly is to give counts and percentages in different segments of the subjects. They are context specific as they relate intimately to the locale and the characteristics of the population. Census and sample surveys are examples of descriptive studies. On the other hand, there are relatively small-scale studies with the objective to discover antecedent–outcome types of relationships. These are called analytical studies as they have cause–effect overtones and the effect of each contributory factor is sought to be delineated. They have wider implications for similar subjects in other populations. They have at least one factor that can be considered antecedent and another factor that can be considered outcome. One precedes the other, although the sequence may not be clear to begin with. Analytical studies try

to answer the *why* and *how* of a phenomenon. Investigations on the association between two or more factors also come under this category, but the effort generally is to estimate the exact contribution of each antecedent to the outcome. Analytical studies are statistically much more challenging than descriptive studies.

Analytical studies almost invariably have a comparison group, although this can arise from within. Sometimes, the values before an intervention serve as the comparison group.

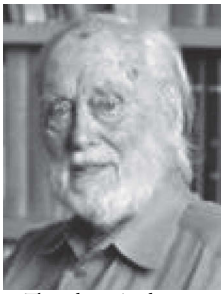
As described for **designs of medical studies**, analytical studies are basically of two types—observational and experimental. Observational studies are on the natural course of events without human intervention, and the data are examined regarding what might have caused or contributed to a particular outcome. These could be retrospective (e.g., case–control), prospective (e.g., cohort), or cross-sectional. See the topic **observational studies** for details. On the other hand, a basic feature of experimental studies is intentional human intervention to see its effect on the outcome. As per the details given for **experimental studies**, these could be laboratory experiments, clinical trials, or field trials. **Clinical trials** are the most glaring example of medical experiments and most challenging among the analytical studies in health and medicine.

## ANCOVA, *see* analysis of covariance (ANCOVA)

## Anderberg dichotomy coefficient, *see* association between dichotomous characteristics (degree of)

## Anderson–Darling test

The Anderson–Darling (A-D) test can be used to check whether sample values fit well in to a specified statistical **distribution**. They proposed this test in 1952 [1]. Although the test can be used for any distribution, its most common application is to check whether the observed values could have come from a **Gaussian distribution**. For correct implementation of many statistical procedures such as ANOVA and regression, fulfillment of the requirement of Gaussianity must be checked, particularly for obtaining valid confidence intervals and for correct testing of hypothesis results. For checking Gaussianity, the A-D test can be used for $n$ as low as 8. If the test gives a statistically significant ***P*-value**, you cannot assume that the values have come from a Gaussian distribution.



Theodore Anderson

The A-D test is a modification of the popular nonparametric **Kolmogorov–Smirnov (K-S) test** and uses the weighted sum of the squares of the differences between the **cumulative frequencies** based on the hypothesized and the observed cumulative frequencies. The K-S test, in its usual form, requires a large sample, whereas the A-D test does not require a large sample. However, the A-D test is parametric as its $P$-value depends on the distribution under test. If the distribution under test is Gaussian, then the $P$-value is different compared to if the distribution under test is, for example, a Weibull distribution. Because of mathematical intricacies, we are not giving the details of the calculations required for an A-D test. This test is widely available in statistical packages, and you would not have much problem in using this test for your data when needed.

If you are interested in a non-Gaussian application of the A-D test, see the work of MacKenzie [2], who used it for checking the Poisson distribution of intersuicide duration in Cornell and the Massachusetts Institute of Technology (MIT).

1. Anderson TW, Darling DA. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *Ann Math Stat* 1952;23:193–212. http://projecteuclid.org/euclid.aoms/1177729437
2. MacKenzie DW. Applying the Anderson–Darling test to suicide clusters: Evidence of contagion at U.S. universities? *Crisis* 2013 Jan 1;34(6):434–7. http://www.ncbi.nlm.nih.gov/pubmed/23502060

## ANOVA, *see* analysis of variance (ANOVA)

## antagonism, *see* interaction

## antecedent factors

An antecedent is a factor that precedes an outcome and is suspected to have contributed in shaping the outcome. This could be an exposure, such as sexual intercourse with a subject suffering from a sexually transmitted disease, or could be a risk factor, such as short stature for age at puberty. The other names for antecedent are *cause*, *predisposing factor*, and *determinant*. In a regression setup, they are called regressors, predictors, or independent variables. Much of **empirical** research in medical sciences is directed toward finding the cause of a particular outcome. While cause attribution is difficult in an empirical setup, antecedent–outcome relationships are relatively easy to study. For **analytical studies**, this relationship is a defining feature.

In a gestational age–birth weight setup, gestational age is clearly an antecedent. A woman's nutrition level, body mass index, smoking, age, parity, previous history, and heredity may also be examined as antecedents for birth weight. But what about unconventional factors such as blood group and the month of conception? If a researcher has reasons to suspect that these also may have some role in determining birth weight, these also can be examined as antecedents. Thus, an antecedent is not necessarily a cause but can be any factor that may have been directly or indirectly instrumental or believed to be affecting an outcome. In this case, the study investigates **association** in place of cause.

In some situations, the distinction between an outcome and an antecedent is blurred. Smoking can lead to depression, and depression can lead to smoking. Gender and blood group are genetically determined simultaneously, and neither is antecedent to the other. Anxiety can cause disease, and vice versa. Also, the same factor can be an antecedent in one setup and an outcome in another setup. Excessive alcohol intake is an antecedent for liver diseases but is an outcome of peer pressure or depression. What is clear to you as an antecedent may not be clear to others. Thus, the **protocol** of the study must specify all the antecedents along with the method for their measurement. In any case, this specification is needed to provide exactitude to the research.

A large number of antecedents can be identified for an outcome, but ethics, feasibility, and parsimony require that only a few relevant ones be included in the study. The general advice is to prefer those that are directly influencing the outcome or those that are suspected

A

to make a prominent contribution. However, in the present era when such kind of obvious factors have already been fully investigated, those making an indirect contribution and those that are obscure can also be investigated. Sometimes, unsuspected antecedents are investigated for fun in the hope of an unexpected result. Note that only the known or suspected factors, even speculative, can be investigated, but those in the **epistemic** domain can never be investigated—and those may be more important. Thus, all medical studies have limited implications.

It is not necessary that the value of antecedent factors is already known in a study and that of outcome elicited. In **retrospective studies** such as **case–control**, the antecedent is elicited, and outcome is already known. Thus, the antecedent is a response in this setup. In **cross-sectional studies**, both antecedent and outcome are elicited simultaneously. However, in prospective studies, the antecedent is known, and the outcome appears during the follow-up.

## APACHE score

An acronym for *acute physiology and chronic health evaluation*, the APACHE score is used to assess the criticality of patients admitted to the intensive care unit (ICU). In a medical care setup, caregivers, patients, and their families all are keen to know how severe the condition is. This helps in taking appropriate steps and to forward a reasonable prognosis. Among many severity scoring systems for critical conditions, APACHE scores are the most commonly used and deserve some understanding. The basic premise in these scores is that the worst physiological derangement noted during the first 24 h after admission in an ICU more or less determines the chance of hospital survival. These derangements define organ insufficiency. This implies, though, that treatment and care are not of much consequence as they are nearly the same in hospitals across the United States, where this scoring system evolved. In other setups, care may be a strong determinant, and this is not part of the APACHE score.

William Knaus

APACHE I was proposed in 1981 by Knaus et al. [1] and was found to be surprisingly accurate in predicting mortality in patients in a variety of ICUs. An exception noted later was patients requiring a coronary bypass graft, where the physiological derangement was high but mortality was low. APACHE I considered 34 physiological measurements that are routinely collected in most US hospitals. Thus, no extra efforts were required. Each of these measurements was assigned points according to the severity of derangement. For example, a serum pH value of either <7.15 or ≥7.70 has +4 points as both are considered equally grave, whereas a normal value between 7.33 and 7.49 has 0 points as this is no derangement [2]. A slightly higher pH value between 7.50 and 7.59 was assigned +1 point, but on the lower side, a value between 7.15 and 7.32 was assigned +2 points as this is considered relatively more harmful. The sum of such points over the 34 measurements is the APACHE score. The higher the score, the greater the chance of death. However, this version was found too complex for adoption.

APACHE II [2] is the simplified version of APACHE I and includes only 12 physiological measurements. But it added points for age (ranging from 0 for age <45 years to 6 for age ≥75 years) and previous history (5 points for nonoperative or postoperative emergency patients and 2 points for elective postoperative patients in the past). The maximum possible score is 71, although in practice, none exceeds 55. A score of 40 or more has been seen to be strongly associated with hospital death. These scores were woven into a **logistic regression** with mortality as the outcome using data from a large number of ICUs across the United States. The equation derived is

$$\ln\left(\frac{R}{1-R}\right) = -3.517 + (0.146 * \text{APACHE-II score}) + (0.603, \text{only if postoperative surgery}) + (\text{diagnostic category weight}),$$

where $R$ is the predicted risk of death and diagnostic category weight was separately derived for each of 50 disease groups. For example, this weight for asthma/allergies is −2.108 and for cardiogenic shock is +0.393. A negative weight implies that the risk of mortality is less, and positive weight implies that the risk is greater. These weights have been described by Knaus et al. [2].

Subsequently, APACHE III appeared between 1991 and 2002 in several different versions. This included 17 physiological variables, adjusted for location and length of stay before ICU admission, and used splines for statistical modeling. The last version of APACHE III covered 96 disease groups. APACHE IV appeared in 2006 and included 116 disease groups. This revised the prediction equation, used five new predictors, extended splines, and made prior length of stay continuous in terms of minutes and not just in integer days. The details of APACHE IV have been described by Zimmerman and Kramer [3].

In short, APACHE III and IV are more complex and only marginally increase the predictive accuracy over APACHE II. Thus, many still prefer APACHE II. The percentage of ICU patients correctly classified into survive/death by APACHE II was observed as 85.5% in US hospitals, and the **area under the ROC curve** was 0.863 [2].

This scoring system is applicable to critical cases wherein survival is known not to exceed 80%. If somebody naive uses this percentage as predictivity for survival without using any scoring system, he/she would be right in about 80% of cases. Since the predictivity of APACHE II is about 85%, this adds just about 5% to the accuracy of prediction. ICU admission itself requires screening that reasonably predicts the chance of survival without using the APACHE score. Note that correct prediction is not as high as the hype around this scoring system. Moreover, not many studies are available that can give guidance regarding the use of APACHE in developing countries, where ICU care and mortality could be very different. Not many realize its limitations.

APACHE scoring is useful in many other setups. You can legitimately compare severity of cases admitted in the ICU of one hospital with that in another hospital, or in two or more groups, such as severity in people with different occupations. Similarly, if a regimen is effective in 72% of cases with an APACHE II score between 20 and 24, and another is effective in 78% with the same score, you are confident of a 6% difference in efficacy in such cases. Also, if the average APACHE II score in critical cases admitted

in a hospital during 2005–2009 is 17 and the average rises to 21 in cases admitted during 2010–2014, it would be legitimate to say that the cases admitted later are more severe. The actual utility of APACHE scores lies in this kind of comparison rather than in predicting survival.


1. Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE. APACHE—Acute physiology and chronic health evaluation: A physiologically based classification system. *Crit Care Med* 1981;9(8):591–7. http://www.ncbi.nlm.nih.gov/pubmed/7261642
2. Knaus WA, Draper EA, Wagnwer DP, Zimmerman JE. APACHE II: A severity of disease classification system. *Crit Care Med* 1985;13:818–29. http://www.ncbi.nlm.nih.gov/pubmed/3928249
3. Zimmerman JE, Kramer AA. Outcome prediction in critical care: The Acute Physiology and Chronic Health Evaluation models. *Curr Op in Crit Care* 2008;14:491–7. http://www.ncbi.nlm.nih.gov/pubmed/18787439


## Apgar score

The Apgar score is just about the first test a newborn undergoes immediately after birth. A newborn is evaluated on appearance, pulse, grimace, activity, and respiration. The first letters of these make up the term *APGAR*, but it was first proposed by Virginia Apgar in 1952 [1].

Virginia Apgar

Appearance is evaluated by skin color, pulse by heart rate, grimace by reflexes, activity by muscle tone, and respiration by breathing effort. Each of these is assigned a score of 0, 1, or 2 at 1 min after birth and again 5 min after birth to see if there is any improvement. The sum of these scores is called the Apgar score at 1 and 5 min, respectively. The maximum is 10 for an absolutely healthy child, although that would be rare, and 0 for a dead child. A healthy child will have a score of 7 or more. A low score (say, <4) is regarded as an indication of poor prognosis and a need for immediate special care of the child, such as resuscitation.

Li et al. [2] investigated the relevance of Apgar score in the United States after 50 years of use in the context of steep advancements in care and technology, and report that it has continuing value for predicting neonatal and postneonatal adverse outcomes in term as well as preterm infants. They also found that this is applicable to twins and in various race/ethnic groups as well. The Apgar score is not so appropriate for long-term prediction of the health of a child [3], although Naeser et al. [4] found in a co-twin control study in Denmark that a high Apgar score was a risk factor for atopic dermatitis. The authors called this a novel finding.

**Scoring systems** come under the domain of biostatistics. For example, the **APACHE score** is based on **logistic regression**. But Apgar is an example of a simple scoring system as a sum of five individual scores and serves the purpose well.


1. Apgar V. A proposal for a new method of evaluation of the newborn infant. *Curr Res Anesth Analg* 1953;32(4):260–7. http://apgar.net/virginia/Apgar_Paper.html, last accessed May 13, 2015.
2. Li F, Wu T, Lei X, Zhang H, Mao M, Zhang J. The Apgar score and infant mortality. *PLoS One* 2013 Jul 29;8(7):e69072. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3726736/
3. Ehrenstein V. Association of Apgar scores with death and neurologic disability. *Clin Epidemiol* 2009 Aug 9;1:45–53. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2943160/
4. Naeser V, Kahr N, Stensballe LG et al. Apgar score is related to development of atopic dermatitis: Cotwin control study. *J Allergy (Cairo)* 2013;2013:712090. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3809604/


## area diagram

This diagram represents the data in terms of areas. One such illustration is Figure A.3a. It shows the distribution by cholesterol level of subjects of different body mass index (BMI) values in a hypertension clinic. But the BMI levels are stacked over one another. This is an area diagram as the sequential addition of the percentages gives the cumulative distribution. Not merely stacking but also the sequence of stacking must have a physical meaning in such an area diagram. In this example, the area below the second polygon depicts the cholesterol level distribution of those who have BMI < 30 kg/m$^2$, and the area below the top line is the cholesterol level distribution of all the subjects with any BMI. The figure also shows that more obese subjects commonly have

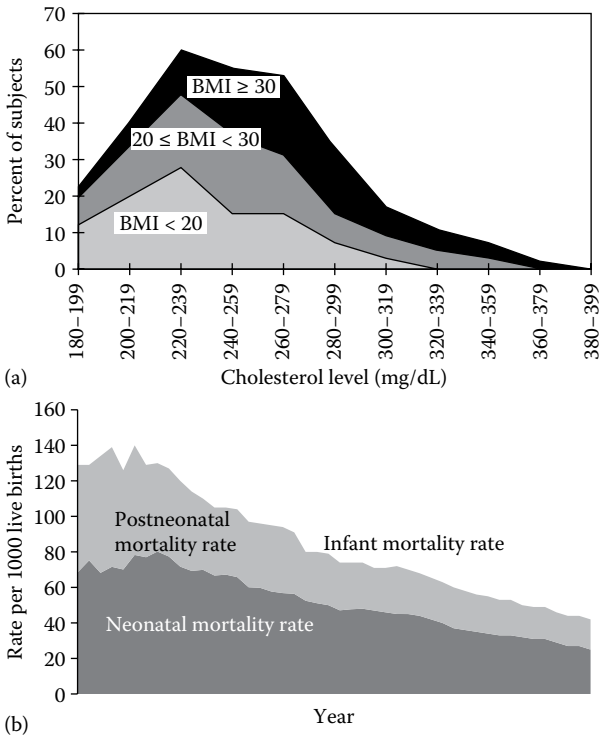



**FIGURE A.3** (a) An area diagram for cholesterol levels in different body mass index (BMI) categories. (b) An area diagram showing trend of neonatal, postneonatal, and infant mortality rate in a developing country.

A

higher cholesterol levels, and the distribution changes shape from one BMI category to another.

Figure A.3b is a different type of area diagram. This shows the trend of **neonatal mortality rate** (NMR) and postneonatal mortality rates (PNMR) in a developing country over a period of 40 years. But the top line also has a physical meaning since that depicts the trend of **infant mortality rate** (IMR). This is possible in this case since NMR + PNMR = IMR, and all are computed per 1000 live births.

## area sampling

As the name implies, in area sampling, geographical areas are sampled, and all eligible persons or units of interest in the selected areas are included in the sample.

Sometimes, particularly in large-scale field studies, it is convenient to sample geographical areas instead of individuals or families. This happens when maps with appropriate divisions are available but the **sampling frame** of the actual **units** of investigation is not. In some areas, households migrate from one place to another, new households spring up quickly, and no updated records are available. Even when the full listing is available, area sampling is easy to execute when a map is available (Figure A.4).

In area sampling, the **sampling unit** is specified in advance depending on what spatial divisions are available or applicable to the study. For example, you can have census blocks or villages as your sampling unit if the map has a demarcation for these areas. If a district has 2500 census blocks, select as many as decided on, and everybody in those selected census blocks will be in your sample. If the unit of inquiry is an old person of age 65 or above, visit every house in the selected areas and locate persons of this age.

Area sampling is a variation of **cluster sampling** where, also, people contiguous in some sense are selected. As in the case of cluster sampling, area sampling is most efficient when units within selected areas are diverse. This helps in achieving representation of a cross-section of the population. But it is likely that units within an area are similar since they belong to the same area. In that case, a large number of areas are selected so that some areas with different kind of people are also included. The sample size may steeply increase, but the cost may still be low since traveling cost and travelling time are drastically reduced.

In most situations, the areas are selected randomly. For this, the areas are numbered, and selection can be done with the help of random numbers. However, as in all random sampling, particularly for small sample size, not all types of areas may be represented. In Figure A.3, northeast areas are not represented. If it is expected, for example, that people living in one type of area are different from others, use **stratified random sampling** with different types of areas as strata. This will ensure that all kinds of people are included in the sample.

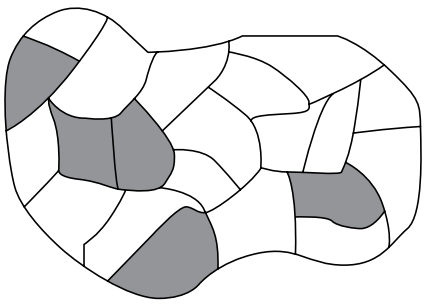


**FIGURE A.4** Selection of five areas from a town.

## area under the concentration curve (AUC curve)

In pharmacokinetic studies, drug concentration in the system is measured at different points in time, and a curve is obtained (Figure A.5a). This is called the *concentration curve*. The objective is to find the pattern and extent of **bioavailability** of the drug over a period of time. This is the total amount of the drug reaching the body. Whereas the pattern is obtained by the shape of the curve, the extent of bioavailability is obtained by what is called the area under the concentration curve (AUC curve). This has special significance when the concentration is obtained at varying time intervals. In the case of regular time intervals, AUC is conceptually similar to the mean or sum of all the values. In Figure A.5a, the concentration rapidly increases till 180 min and then gradually starts to decline. Figure A.5b compares AUC for two regimens.

AUC is also used in animal experiments and clinical trials when the response is studied at different points in time. The response could be intensity of pain after analgesia, physical stress during exercise, efficacy of a regimen over time, or any other such measure. Many times, concentration is measured in logarithmic units because concentration is mostly multiplicative.

AUC can be obtained for each person under study, but the convention is to draw it for averages over the subjects. This is valid so long as the values across persons are homogeneous. If one person shows a widely different trend from another, the average conceals this variation. Such an individual-specific response may have clinical implications that are lost in averages.

The method generally followed for calculating the area under the curve is called the *trapezoid rule*. The total area is divided into small trapezoids (Figure A.5a). The area of each trapezoid can be calculated easily, e.g., area of trapezoid ABCD = (AB + CD)/2 * BC. Sum the areas of all the trapezoids and get the area under the curve. The last point is joined with concentration 0 for this purpose if the period of observation does not extend to the

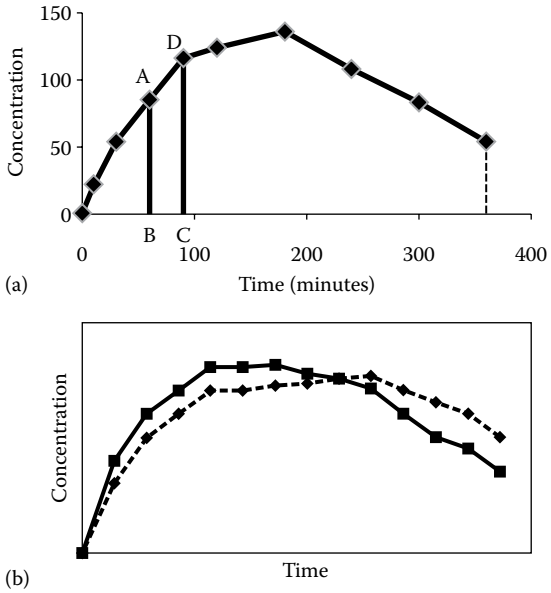



**FIGURE A.5** (a) Area under the concentration curve. (b) Markedly different curves with the same area.

time when concentration of the drug falls to zero (undetectable) levels. Thus, this will be the area truncated at the last time point. For example, Gaudreault et al. [1] evaluated truncated area under the curve as a measure of the relative extent of bioavailability. Many statistical packages plot this kind of curve and calculate the area under the curve once the concentrations at different points in time are entered.

A difficulty with AUC is that it has limited physical meaning. When used in isolation, it may fail to provide an adequate assessment of the trend, even on average, because it is neither specific nor sensitive to the changes in the patterns. In the case of comparison, curves with markedly different trends can give the same area (Figure A.5b). For this reason, AUC is not a valid measure of **bioequivalence**. Also, as the time passes, some patients drop out or are cured—thus, the average response at different points in time is based on different values of $n$. This is forgotten while studying AUC.

The AUC curve can give valid results when the average response in one group is better than that in the other at almost every time point. If they crisscross, then the conclusion can be very wrong. In this case, the investigator may have to be more specific on what exactly he/she is looking for. Merely stating a trend or pattern does not help. Depending on what is most relevant for the objective of the study, particularly in **pharmacokinetic** studies, it could be the time taken to reach the peak ($T_{max}$), the time taken to return to the initial level, the peak level attained ($C_{max}$), the response level after a specific time gap, **half-life**, etc. In many situations, the AUC needs to be considered in conjunction with other parameters such as $T_{max}$ and $C_{max}$ for a valid conclusion. In a study on treatment efficacy, the interest may be in the value at the end relative to the initial value. If the study is on effectiveness of an analgesic, the interest may be in $T_{max}$, $C_{max}$, and possibly time to reach to some specified critical level. No matter what parameter is used, it should always be decided beforehand on the basis of clinical utility and not after inspecting the data. AUC is considered a comprehensive measure of performance, but further insight is obtained by $T_{max}$, $C_{max}$, etc.

The AUC curve is different from the area under the **receiver operating characteristic (ROC) curve**. The term *area under the curve* is also used in the context of probabilities in a distribution. For one such use, see **Gaussian distribution**.

1. Gaudreault J, Potvin D, Lavigne J, Lalonde RL. Truncated area under the curve as a measure of relative extent of bioavailability: Evaluation using experimental data and Monte Carlo simulations. *Pharm Res* 1998;15:1621–9. http://link.springer.com/article/10.1023/A%3A1011971620661#page-1

## area under the ROC curve, *see* *C*-statistic

## arithmetic mean, *see* mean (arithmetic, geometric, and harmonic)

## ARMA and ARIMA models, *see* autoregressive moving average (ARMA) models

## Armitage–Doll model

The Armitage–Doll model seeks to statistically explain the onset of epithelial carcinoma as a multistage process in humans, triggered by exposure to carcinogenic substances and undergoing several successive qualitatively different cellular changes (such as mutations) over time. This model was first proposed by Armitage and Doll in 1954 [1].

The Armitage–Doll model explains the rapid rise in incidence and hence mortality due to some cancers over age and also the long latent period before a tumor develops after the exposure. The model states that

$$\text{hazard of cancer at time } t\text{: } h(t) = \alpha t^{K-1},$$

where $K$ is the number of stages of cellular changes required for cancer to appear and $\alpha$ is a constant. $K$ and $\alpha$ will change from cancer to cancer. The model is easier to understand if log transformation is used, since then,

$$\log h(t) = \log\alpha + (K-1)\log t.$$

This is a linear relationship between the log of hazard and the log of time that is an outcome of **Weibull distribution** of the time for appearance of cancer. The value of $K$ for many cancers is between 5 and 7, suggesting that cancer development requires 5–7 cellular changes [2]. The model mimics a two-stage model for, say, breast cancer, with exponential proliferation of aberrant cells. But it fails to account for the steep fall in cancer incidence after the age of 80 years [3]. The model has been recently applied to pancreatic cancer data that found that this cancer possibly takes about 17 years and requires a series of five mutations for cells to become malignant [4].

1. Armitage, P. Doll, R. The age distribution of cancer and a multistage theory of carcinogenesis. *Br J Cancer* 1954;8(1):1–12. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2007940/pdf/brjcancer00386-0010.pdf
2. Armitage P. Multistage models of carcinogenesis. *Env Health Persp* 185;63:195–201. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1568502/pdf/envhper00446-0191.pdf
3. Ritter G, Wilson R, Pompeia F, Burmistrov D, The multistage model of cancer development: Some implications. *Toxicol Industrial Health* 2003;19:125–45. http://users.physics.harvard.edu/~wilson/publications/pp876.pdf
4. Mdzinarishvili T, Sherman S. Basic equations and computing procedures for frailty modeling of carcinogenesis: Application to pancreatic cancer data. *Cancer Inform* 2013;12:67–81. http://www.la-press.com/basic-equations-and-computing-procedures-for-frailty-modeling-of-carci-article-a3548-abstract

## arms of a trial

Various groups receiving different regimens in a **clinical trial** are called arms of a trial. A clinical trial generally will have two arms—a test arm and a control arm—but a test also can have more than one arm if two or more regimens are under investigation. In this case, this becomes a **multiarm trial**. A test arm can also be called a treatment arm if the regimen is for treatment of a disease or an experimental arm if it is an experiment.

A clinical trial necessarily contains a group on whom the regimen is tried to find its **efficacy** and side effects. By itself, this group may provide an estimate of the efficacy but would not provide any comparison. Many researchers use the preintervention status of the subjects for comparison. This could provide an estimate of the change brought about by the intervention if nothing else has changed between the initial and final measurements. Part of this effect could be due to the **placebo** and **Hawthorne effects**. You cannot say how much of it is due to pure psychological effect and how much is due to the regimen. Thus, a **control group** is also investigated, who receive a placebo. A group receiving the existing regimen can also serve as

A

a control. The purpose of the control arm is to serve as a comparator so that the effect can be meaningfully interpreted. These different groups in a trial form arms of the trial.

## association and its degree

Association has the same statistical meaning as in day-to-day language. Two characteristics are said to be associated when change in one is accompanied by some change in the other in the long run. However, this is not necessarily causal—that is, the effect in every case is not necessary. Association between two characteristics can occur because of the presence of other conducive factors. If glaucoma occurs more commonly in females in some population, sex and occurrence of glaucoma are associated in that population. This can be due to anemia in relatively more females than males or for some other reason. Being female is not the cause of glaucoma. This is just *association*, which is relatively a much milder term than cause.

When no association exists, you can say that the characteristics are statistically **independent**. This, in effect, means that the chance of occurrence of one does not affect the chance of occurrence of the other. The occurrence of glaucoma in one woman does not alter the chance of occurrence in another woman unless, possibly, they belong to the same family. Statistically, the term *association* is used primarily for **qualitative** characteristics rather than **quantitative** characteristics. The corresponding term for quantities is **correlation**.

The **chi-square test** is the method of choice to find whether two characteristics are associated or not. This is a large-sample test and can be used both for **dichotomous** and **polytomous categories**. A **statistically significant** chi-square only says that an association is present but does not say how much. Some associations are strong, and some are weak. For assessing the degree, methods for **association between dichotomous characteristics** are different from the methods for **association between polytomous characteristics**. See these topics for details. For example, you can find an association between occupation and site of cancer in cancer patients. Since these categories are on a **nominal scale**, the sign of association, negative or positive, depends on the order in which they are stated in a **contingency table**. However, an **association between ordinal characteristics** can be factually negative—as one becomes higher, the other becomes lower. The association between age groups (young, middle, and old) and physical health (poor, fair, good, and excellent) among adults is generally negative in the sense that as age increases, physical health decreases. When one characteristic is qualitative and the other quantitative, you can explore whether a **coefficient of determination** can be used to measure the degree of association.

Whereas **experimental studies** and **clinical trials** are aimed at cause–effect relationships, **observational studies**, such as prospective, retrospective, and cross-sectional, are expected to provide evidence only of association and not of cause–effect. Cause–effect is concluded when several other stringent criteria are met, as mentioned under **cause–effect relationship**.

## association between dichotomous characteristics (degree of)

When one characteristic of the subjects tends to change with another characteristic, this is called an association between these two characteristics. The most common situation for investigation of an association in the case of **qualitative variables** is that of dichotomy. The bivariate binary observations in this case are generally summarized in the form of a 2 × 2 table of the type given in Table A.5.

**TABLE A.5**
**Cross-Classification of Subjects by Two Binary Variables**

| | Variable 1 | |
|---|---|---|
| **Variable 2** | **Present** | **Absent** |
| Present | $a$ | $b$ |
| Absent | $c$ | $d$ |

It is customary in the case of qualitative variables to use the term *association* in preference to *correlation*. The most widely used measure of the strength of association in such a setup is either **relative risk (RR)** or **odds ratio (OR)**, depending on whether the design is **prospective** or **retrospective/cross-sectional**. If variable 1 is the **outcome** and variable 2 is the **antecedent** in Table A.5, these are calculated as follows.

$$\text{RR} = \frac{a/(a+b)}{c/(c+d)} \quad \text{and} \quad \text{OR} = \frac{ad}{bc}.$$

In some situations, ln(RR) and ln(OR) are preferred because of the linearizing property of the logarithm in this case. This transformation also assigns a negative or positive sign to the relationship as appropriate, making the interpretation easy.

Consider the cross-sectional data in Table A.6 on gender and blindness in patients coming to a cataract clinic where the definition of blindness is visual acuity (VA) < 1/60 in at least one eye. What is the degree of association between gender and blindness in these data?

The answer lies in the OR. Since females have a higher rate of blindness in this case, it is better to compute OR in females relative to males. This will somewhat reverse the notations and is given by

$$\text{OR} = \frac{110 \times 419}{101 \times 370} = 1.23.$$

As far as these data are concerned, females with blindness in at least one eye are 1.23 times as likely to have a cataract as males, or the odds are nearly 5:4. This, in a way, measures the degree of association as well. A higher value of OR indicates stronger association.

The conventional measures of association for 2 × 2 tables are the following dichotomy coefficients:

$$\text{Positive matching dichotomy coefficient: } S_1 = \frac{a}{a+b+c+d}.$$

This is the proportion of pairs with both values present.

**TABLE A.6**
**Gender and Blindness in Patients Coming to a Cataract Clinic**

| Gender | Blind | Not Blind | Total |
|---|---|---|---|
| Male | 101 | 419 | **520** |
| Female | 110 | 370 | **480** |
| **Person** | **211** | **789** | **1000** |

$$\text{Jaccard dichotomy coefficient: } S_2 = \frac{a}{a+b+c}.$$

This is the proportion of pairs with both values present given that at least one occurs. This is not symmetric to the presence and absence of disease.

$$\text{Simple matching dichotomy coefficient: } S_3 = \frac{a+d}{a+b+c+d}.$$

This is the proportion of pairs where the values of both variables agree.

$$\text{Anderberg dichotomy coefficient: } S_4 = \frac{a}{a+2(b+c)}.$$

This is basically the same as $S_2$ but is standardized for all possible patterns of agreement and disagreement.

$$\text{Tanimoto dichotomy coefficient: } S_5 = \frac{a+b}{a+2(b+c)+d}.$$

This is $S_3$ standardized for all patterns of agreement and disagreement.

It is not easy to choose one among these. They will most likely give different results. You may have to carefully examine the data and the objective of the relationship that you want to measure. For example, $S_2$ and $S_4$ are appropriate when the absence of an attribute in both variables ($d$ in Table A.5) does not convey any useful information.

Another popular measure of degree of association for dichotomous categories is

$$\text{Yule } Q: = \frac{ad-bc}{ad+bc}.$$

In terms of OR, this is $Q$ = (OR − 1)/(OR + 1). This lies between −1 and +1 and can be interpreted like a **correlation coefficient** for assessing the strength of relationship.

**Association between polytomous characteristics** and **association between ordinal characteristics** are discussed separately.

## association between ordinal characteristics (degree of)

Association between two characteristics is the tendency of change in one characteristic of subjects to be accompanied by change in the other characteristic. Most common is the **association between dichotomous characteristics**, which is presented separately. Now consider two characteristics on an **ordinal scale**, such as severity of disease and obesity. One may have $R = 4$ categories, and the other, $C = 3$ categories. The interest is in measuring the strength of association between these two ordinal characteristics. The data can be arranged in an $R \times C$ table.

The association is high if the higher category of one is more frequently seen with the higher category of the other. The association between severity of disease and obesity is high if more severe cases are obese. This is called **concordance**. If less severe cases are mostly obese, this is discordance. However, in this case, you need to consider all possible pairs of pairs. If one pair is $(x_1, y_1)$ and the other pair is $(x_2, y_2)$, they are concordant if $x_1 < x_2$ and $y_1 < y_2$ or if $x_1 > x_2$ and $y_1 > y_2$, and disconcordant if $x_1 < x_2$ but $y_1 > y_2$ or if $x_1 > x_2$ but $y_1 < y_2$. Also, pairs are tied for $x$ if $x_1 = x_2$ irrespective of $y$, and tied for $y$ if $y_1 = y_2$ irrespective of $x$. In ordinal data, ties are quite common. For $n$ subjects, there are a total of $n(n-1)/2$ pairs since the pair $[(x_1, y_1), (x_2, y_2)]$ is considered same as $[(x_2, y_2), (x_1, y_1)]$. For a $3 \times 2$ table (Table A.7), this can be explained as follows.

Total number of persons $a + b + c + d + e + f = n$
Total number of pairs of pair = $n(n-1)/2 = T$
Concordant pairs of subjects = $a(e + f) + bf = P$
Discordant pairs of subjects = $c(d + e) + bd = Q$
Pairs tied on characteristic 1 ($x$) alone = $ad + be + cf = X_0$
Pairs tied on characteristic 2 ($y$) alone = $a(b + c) + bc + d(e + f) + ef = Y_0$
Pairs tied on both $x$ and $y$ = $a(a-1)/2 + b(b-1)/2 + c(c-1)/2 + d(d-1)/2 + e(e-1)/2 + f(f-1)/2 = (XY)_0$

$(XY)_0$ does not contribute to the measure of association, and these are ignored except for calculating the total number of pairs.

Now, various measures of ordinal association can be defined as follows. You may find varying definitions in the literature.

$$\text{Kendall Tau-a: } \tau_a = \frac{P-Q}{n(n-1)/2}.$$

This is the surplus of concordant pairs over discordant pairs as a proportion of the total pairs. If the agreement in pairs is perfect, $Q = 0$, and tau-a = 1, assuming no ties. If all are discordant pairs, $P = 0$ and tau-a = −1. Thus, this ranges from −1 to +1. If ties are present, this modifies to

$$\text{Kendall Tau-b: } \tau_b = \frac{P-Q}{\sqrt{(P+Q+X_0)(P+Q+Y_0)}}.$$

The denominator is now partially adjusted for ties, and $P$ and $Q$ will also be automatically adjusted by definition. Tau-b works well for square tables where the number of categories for one characteristic is the same as for the other characteristics (i.e., $R = C$). Tau-b = +1 if the table is diagonal and Tau-b = −1 if all diagonal elements are zero. If the table is not square, the corresponding adjustment for the size is

$$\text{Kendall Tau-c: } \tau_c = \frac{2(P-Q)R}{n(n-1)(R-1)},$$

where $R$ is the smaller of number of rows and number of columns. This is also called *Stuart Tau-c*.

**TABLE A.7**
**Cross-Classification of Persons with Two Ordinal Characteristics**

| | Characteristic 1 ($x$) | | |
|---|---|---|---|
| **Characteristic 2 ($y$)** | **Low** | **Medium** | **High** |
| Low | $a$ | $b$ | $c$ |
| High | $d$ | $e$ | $f$ |

A

Beside these variations of tau, there are two other popular measures of ordinal association.

$$\text{Goodman–Kruskal gamma: } \gamma = \frac{P-Q}{P+Q}.$$

This completely excludes ties from the numerator as well as from the denominator. This also ranges from −1 to +1. If the number of discordant pairs is the same as concordant pairs, $\gamma = 0$.

All these measures are symmetric in the sense that it does not matter which characteristic is in the rows and which is in the columns (or, in other words, there is no distinction between antecedent and outcome). Both are treated the same way. For a directional hypothesis such that *x* predicts *y* (or *x* is the antecedent and *y* is the outcome), use

$$\text{Somer } d = \frac{P-Q}{P+Q+Y_0},$$

where $Y_0$ is the number of pairs tied for *y*. Note that only the pairs tied for *x* are excluded from the denominator.

Suppose Table A.8 is obtained regarding smoking and drinking in cancer cases.

For this table,

$n = 51$ and $T = 51 \times 50/2 = 1275$
Concordant pairs $P = 35(2 + 3) + 4 \times 3 = 187$
Discordant pairs $Q = 1(6 + 2) + 4 \times 6 = 32$
Tied on *x* (smoking) $X_0 = 35 \times 6 + 4 \times 2 + 1 \times 3 = 221$
Tied on *y* (drinking) $Y_0 = 35(4 + 1) + 4 \times 1 + 6(2 + 3) + 2 \times 3 = 215$
Tied on both *x* and *y* $(XY)_0 = ½ (35 \times 34 + 4 \times 3 + 1 \times 0 + 6 \times 5 + 2 \times 1 + 3 \times 2) = 620$

These numbers give

$$\text{Tau-a} = \frac{187-32}{1275} = 0.12$$

$$\text{Tau-b} = \frac{187-32}{\sqrt{(187+32+221)(187+32+215)}} = 0.35$$

$$\text{Tau-c} = \frac{2\times(187-32)\times 2}{51\times 50\times 1} = 0.24$$

**TABLE A.8**
**Smoking and Drinking in Cancer Cases**

| | | Smoking | | |
|---|---|---|---|---|
| **Drinking** | **Nonsmokers** | **<3.5 Pack/ Week** | **≥3.5 Pack/ Week** | **Total** |
| Nondrinkers | 35 | 4 | 1 | **40** |
| Drinkers | 6 | 2 | 3 | **11** |
| **Total** | **41** | **6** | **4** | **51** |

$$\text{Goodman–Kruskal } \gamma = \frac{187-32}{187+32} = 0.71$$

$$\text{Somer } d = \frac{187-32}{187+32+215} = 0.36$$

Note how widely different values are obtained by different measures for the same data. For this reason, many workers do not rely much on these measures.

## association between polytomous characteristics (degree of)

Association is the tendency of two characteristics of subjects moving together at least in some subjects. The change in one relative to the other may be small or large. **Association between dichotomous categories** of the two characteristics and **association between ordinal characteristics** are discussed separately. This section is on association between two **polytomous** characteristics.

Consider classification of disabled persons by type of disability and the degree of disability (Table A.9). The type of disability is in five categories, and the degree of disability, in three categories. If **odds ratio (OR)** is used as a measure of association in this example, you need to compute OR for each pair of degree-of-disability categories—i.e., OR of category 1 versus category 2, of category 1 versus category 3, and of category 2 versus category 3—for each type of disability. Such multiple ORs may be useful in some cases but can be confusing in many cases. Moreover, a measure of overall association will not be available.

An alternative is to compute the usual **chi-square** $\chi^2 = \Sigma(O - E)^2/E$ and use this as a measure of association. It is true that a higher degree of association will yield a higher value of chi-square, but the value of chi-square also depends heavily on the sample size *n*. It increases without bound as *n* increases. To counter this, the following measure is proposed:

$$\text{Phi coefficient}: \phi = \sqrt{\chi^2/n},$$

where the square root is taken to offset the effect of squares in chi-square. Note that the concept of negative association is not relevant in the case of polytomous **nominal categories** except for the order in which they are stated—and this order has no meaning for nominal

**TABLE A.9**
**Type of Disability and Its Degree in 1000 Disabled Persons**

| | Degree of Disability | | | |
|---|---|---|---|---|
| **Type of Disability** | **Mild** | **Moderate** | **Severe** | **Total** |
| Visual | 19 | 69 | 22 | **110** |
| Locomotor | 39 | 142 | 29 | **210** |
| Hearing | 46 | 325 | 89 | **460** |
| Mental | 21 | 98 | 51 | **170** |
| Other | 7 | 3 | 20 | **50** |
| **Total** | **132** | **657** | **211** | **1000** |

categories. The phi coefficient can exceed unity. Also, $\phi$ depends on the size of the table. A modification of this measure is

$$\text{Tschuprow coefficient } T = \sqrt{\frac{\phi^2}{\sqrt{(R-1)(C-1)}}},$$

where $R$ and $C$ are the number of rows and number of columns, respectively, in the contingency table. This takes care of the size of the table. Another modification is

$$\text{Contingency coefficient: } C = \frac{\sqrt{\chi^2}}{\sqrt{\chi^2 + n}}.$$

This cannot exceed unity but could also never be one, even when the association is perfect. The value of $\chi^2$, and hence of $\phi$ and of $C$, can be severely affected by the cutoff points of categories when they are for a variable on a **metric scale**. Different cutoff points may give different values.

The most popular of such chi-square-based measures is the Cramer *V*.

$$\text{Cramer } V = \sqrt{\frac{\chi^2}{n * \min(R-1, C-1)}}.$$

This provides a good measure for tables of any size, and ranges between 0 and 1. For 2 × 2 tables, $V = \phi$. Cramer *V* is especially suitable for a square contingency table (i.e., when number of rows $R$ = number of columns $C$).

The value of $\chi^2$ for the data in Table A.9 is 37.14. If the **cell frequencies** are proportionately decreased to one-fifth, rounded off to the nearest integer, so as to have a total of 200, you get $\chi^2 = 7.39$. The large difference between this value of $\chi^2$ for $n = 200$ and the previous value for $n = 1000$ illustrates that $\chi^2$ is heavily dependent on $n$. A proportionate decrease (or increase) in cell frequencies does not affect the degree of association but affects the value of $\chi^2$. For these data,

(a) For $n = 1000$, as in Table A.9, $\chi^2 = 37.14$, $\phi = 0.19$, $C = 0.19$, and $V = 0.14$.
(b) For $n = 200$ (proportionate cell frequencies), $\chi^2 = 7.39$, $\phi = 0.19$, $C = 0.19$, and $V = 0.14$.

Note that $\phi$, $C$, and $V$ for the two $n$'s are the same, while the value of $\chi^2$ is very different. In this example, the values of $\phi$ and $C$ seem equal for each of the two $n$'s because of the rounding off, but this generally will not be the case.

A major objection to the measures $\phi$, $T$, $C$, and $V$ is that they lack underlying substantive meaning. A measure of the degree of relationship in polytomous categories with a useful interpretation is **proportional reduction in error (PRE)**, as explained separately. The primary purpose of PRE is to measure the utility of one characteristic in predicting the other.

## association in prospective, retrospective, and cross-sectional studies, *see also* chi-square test for 2 × 2 tables

Antecedents and outcomes are defining features of all analytical studies, including prospective, retrospective, and cross-sectional studies. This section is restricted to two characteristics: one **antecedent** and the other **outcome**, both being **dichotomous**. The results in this setup are easily described by a 2 × 2 contingency table, as shown in Table A.10. The notation in this table is as follows: $O_{rc}$ is the observed frequency in the $(r, c)$th cell ($r = 1, 2$; $c = 1, 2$), and inside parentheses in each cell in the table are the corresponding probabilities. The dot in the subscript is for the corresponding total. This is shown as though the relationship between antecedent and outcome is under study, and we presume this in this section, but these can be relaxed to include any two dichotomous categories. This is also known as a *fourfold table*. **Prospective, retrospective, and cross-sectional studies** provide three different situations for such a table, as described next.

**TABLE A.10**
**General Structure of a 2 × 2 Contingency Table**

| Variable 2 (Outcome) | Variable 1 (Antecedent) | | |
|---|---|---|---|
| | **Present** | **Absent** | **Total** |
| Present | $O_{11}$ ($\pi_{11}$) | $O_{12}$ ($\pi_{12}$) | $\boldsymbol{O_{1\bullet}}$ **(**$\boldsymbol{\pi_{1\bullet}}$**)** |
| Absent | $O_{21}$ ($\pi_{21}$) | $O_{22}$ ($\pi_{22}$) | $\boldsymbol{O_{2\bullet}}$ **(**$\boldsymbol{\pi_{2\bullet}}$**)** |
| **Total** | $\boldsymbol{O_{\bullet 1}}$ **(**$\boldsymbol{\pi_{\bullet 1}}$**)** | $\boldsymbol{O_{\bullet 2}}$ **(**$\boldsymbol{\pi_{\bullet 2}}$**)** | $\boldsymbol{n}$ |

### *Structure in a Prospective Study*

Because the investigation is from antecedent to outcome in a prospective study, the column totals $O_{\bullet 1}$ and $O_{\bullet 2}$ are fixed in advance. They can also be denoted by $n_1$ and $n_2$, respectively. These are the numbers of exposed and nonexposed subjects followed up for appearance of outcome. The row totals $O_{1\bullet}$ and $O_{2\bullet}$ become known only after the investigation is over. In terms of notations in Table A.10, the relevant **null hypothesis** in this case is $H_0$: $\pi_{11} = \pi_{12}$. This states that the incidence rate in the two groups is the same. In this case, $\pi_{11} + \pi_{21} = 1$ and $\pi_{12} + \pi_{22} = 1$. Thus, $H_0$ is equivalent to $\pi_{21} = \pi_{22}$.

### *Structure in a Retrospective Study*

The direction of the investigation in a retrospective study is from outcome to antecedent. Thus, the row totals $O_{1\bullet}$ and $O_{2\bullet}$, say with and without disease, are fixed in advance, and the column totals $O_{\bullet 1}$ and $O_{\bullet 2}$ are obtained through the study. The fixed row totals can also be denoted by $n_1$ and $n_2$. The null hypothesis now is that the rate of presence of antecedent in those with a positive outcome is the same as in those with a negative outcome, i.e., $H_0$: $\pi_{11} = \pi_{21}$. In this case, $\pi_{11} + \pi_{12} = 1$ and $\pi_{21} + \pi_{22} = 1$. The $H_0$ implies $\pi_{12} = \pi_{22}$ also.

### *Structure in a Cross-Sectional Study*

In this case, $n$ subjects are simultaneously cross-classified by the antecedent and outcome. Neither the column totals nor the row totals are fixed in advance, and both become known only after study of the subjects is over. In this case, $\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22} = 1$. According to the multiplicative **law of probabilities**, the antecedent and outcome are independent if and only if $H_0$: $\pi_{rc} = \pi^*_{r\bullet}\, \pi_{\bullet c}$ ($r, c = 1, 2$) holds, where $\pi_{r\bullet} = \pi_{r1} + \pi_{r2}$ and $\pi_{\bullet c} = \pi_{1c} + \pi_{2c}$.

The $H_0$ in the prospective and retrospective setup is called *hypothesis of homogeneity* (column homogeneity and row homogeneity, respectively), and the $H_0$ in the cross-sectional setup is called *hypothesis of independence*. All these situations can be viewed as subjects divided by two qualitative characteristics with the objective to investigate if one characteristic has any association with the

A

other—whether one is occurring more commonly with the other than expected by chance. These hypotheses are tested by **chi-square test for 2 × 2 tables**. See that topic for details.

## assumptions (statistical)

Assumptions actually are *requirements* for a method to be valid. All statistical methods are valid under some conditions. For example, the **Student *t*-test** requires that the observations follow a **Gaussian distribution**, and the ANOVA ***F*-test** requires **homoscedasticity** across groups plus some other conditions. These are examples of requirements for the validity of a method but are statistically called assumptions. The term *assumptions* has come to stay, although the term *requirements* is a better description.

Different statistical methods have different requirements, but perhaps the most common is that individuals are randomly selected and are **independent**, and the values follow a Gaussian pattern. Among others are homoscedasticity, linearity, lack of **multicollinearity**, etc., depending upon which statistical method is being used.

All statistical methods require **random sampling** in case a sample is taken. This is the reason that random sampling is so pervasive in statistical texts. This helps in working with probabilities—the backbone of statistical science. Independence is violated when two or more observations belong to the same person or to an identifiable affinity group. Many physiological parameters and pathologies are shared among members of the same family, and children in the same class share a teaching environment. Variables concerned with these will not be independent. Generally, however, my blood pressure (BP) level is hardly affected by what the BP of my neighbor is if we do not have the same lifestyle. Thus, these are practically independent. Gaussianity of observations can be easily checked (see **Gaussianity, how to check**) for quantitative data, but realize that symmetry and single mode are no guarantee that the observations follow a Gaussian distribution. There might be deviation in **kurtosis** as examined under the **Box–Cox power exponential (BCPE) method**. In many cases with large samples, Gaussianity is not a strict requirement, because of the **central limit theorem**. This applies to sample means and several other measures. Those statistical methods that work reasonably well in situations where assumptions are mildly violated are called **robust** methods.

The user should be familiar with the validity conditions of the method he/she is using and is expected to satisfy himself/herself that these conditions are nearly fulfilled, if not fully fulfilled. Results based on invalid methods would also be invalid. In case the conditions are not satisfied, alternative approaches such as **nonparametric** methods are adopted. These are valid under milder conditions but generally have less statistical **power**.

## asymptotic relative efficiency, *see* relative efficiency of estimators

## asymptotic regression, *see* regression (types of)

## attack rate

The extent of morbidity in a group of people can be measured in several ways. In the case of acute conditions, particularly infections, it is easier to talk in terms of attacks than **incidence**. The same person can have two or three attacks of diarrhea or of cold in 1 year. Some noncommunicable conditions such as angina and asthma also have the same feature. Thus, the emphasis here is on disease spells rather than on affected persons.

$$\text{Attack rate} = \frac{\text{new spells during a specified time interval}}{\text{total population at risk during the same interval}} * 100.$$

This is commonly used when the exposure is for a limited period, such as during an epidemic, but can be used otherwise too. This can also be calculated per *person-year*.

You can also calculate secondary attack rate (SAR) based on new spells among those who were exposed.

$$\text{Secondary attack (SAR)} = \frac{\text{new spells within the range of incubation period among those exposed to primary cases}}{\text{subjects exposed to the primary cases that can spread the disease}} * 100.$$

For this rate, the denominator and the numerator are restricted to the susceptible contacts. This rate is generally used for diseases such as measles and chickenpox that are infective for only a short period. SAR measures the intensity of the spread of infection or risk among the susceptible contacts after exposure to an infective case. When the primary case, also called the **proband**, is infective for a long period, as in tuberculosis, the duration of exposure becomes important. SAR then is computed per 100 person-weeks, person-months, or person-years of exposure.

SAR is a useful measure not only for infectious diseases but also for diseases of unknown etiology, such as Hodgkin disease, to find out whether it is communicable. SAR is also useful in evaluating the effectiveness of control measures such as isolation and immunization.

SAR can be used in other contexts as well. As an example, consider the risk of diabetes in nondiabetic siblings of children diagnosed with type 1 diabetes. Steck et al. [1] analyzed the family history of 1586 patients in Colorado with type 1 diabetes diagnosed before 16 years of age and interviewed during 1999–2002. Probands are those who were initially affected, and secondary cases are those who appeared later in the family of probands. SAR by age 20 years in siblings was 4.4%, but it was significantly higher in siblings of probands diagnosed under age 7 years than in those diagnosed later. In the parents, too, the SAR by age 40 years was higher when the proband was diagnosed at less than 7 years of age.

This example illustrates the use of the concept of SAR in a very different context. It is not related to the repeated episodes yet delineates the communicability. Such usage is not uncommon, and it is nice to be familiar with different contexts in which a term can be used.

1. Steck AK, Barriaga KJ, Emery LM, Fiallo-Scharer RU, Gottlieb PA, Rewers MJ. Secondary attack rate of type 1 diabetes in Colorado families. *Diabetes Care* 2005;28:296–300. http://care.diabetesjournals.org/content/28/2/296.long

## attenuated correlation

Attenuated correlation is the underestimation of the correlation coefficient due to random measurement errors. This applies to those measurements that cannot be made with sufficient **accuracy**. You should be able to anticipate this error in your investigation. Behavioral variables are glaring examples of variables that can hardly be measured accurately; medical measurements such as blood pressure and pulse

rate are also known to contain errors when measured, particularly when measured by human beings. In the laboratory, the quality of reagents and chemicals can affect accuracy. A lower value of a correlation than expected is an appropriate situation where you can suspect attenuated correlation. The Pearson **correlation coefficient** can be disattenuated as follows:

$$\text{Disattenuated correlation between } x \text{ and } y = \frac{\text{observed } r_{xy}}{\sqrt{\text{Rel}(x)*\text{Rel}(y)}},$$

where $r_{xy}$ is the correlation coefficient, and Rel($x$) and Rel($y$) are **reliabilities** of the observed values of $x$ and $y$. Reliabilities depend on measurement error. If the observed correlation is 0.7 and the reliability of $x$ and $y$ are 0.8 each, the actual correlation is $0.7/\sqrt{(0.8*0.8)} = 0.875$. This correction may not work well when reliability drops below 0.70. Reliability of observed values should be measured using a valid procedure. Mostly, **Cronbach alpha** is used for this purpose. If you have an opportunity to find the errors ($e_x$ and $e_y$) by comparing with standard values or otherwise, you can find reliability as follows:

$$\text{Rel}(x) = \frac{\text{var}(x) - \text{var}(e_x)}{\text{var}(x)},$$

where var is the sample variance. And similarly for $y$.

Similar adjustment can also be suggested for partial correlation, regression coefficient, and multiple correlation [1]. However, note the following for a disattenuated correlation:

- This can exceed 1.0. In that case, the value is considered equal to 1.0.
- This is not suitable for finding the confidence interval and hypothesis testing, since its **sampling distribution** is not known.
- This is not a substitute for accurate measurements. It should be used only when accurate measurements are not possible or are exceedingly difficult or costly.
- The difference between disattenuated and attenuated correlation provides a good assessment of the magnitude of measurement errors.

An application of a disattenuated correlation is seen in a microarray cross-platform study by Archer et al. [2]. The authors considered microarray gene expression in two platforms as $x$ and $y$. Fluorescent intensities for scanned microarray images were used as surrogates for true gene expression values. These surrogates obviously have errors. They noted that correlation attenuation for C3B arrays across platforms was 0.386, which happened to be more than half of the variability attributed to measurement error, and the disattenuated correlations are much higher. This suggested that the cross-platform correlations reported in previous studies without this correction were underestimates. The authors concluded that it is essential to evaluate intraplatform reproducibility while estimating a cross-platform correlation.

1. Jason WO. Effect sizes and the disattenuation of correlation and regression coefficients: Lessons from educational psychology. *Practical Assessment, Research & Evaluation*, 2003;8(11). http://pareonline.net/getvn.asp?v=8&n=11

2. Archer KJ, Dumur CI, Taylor GS et al. Application of a correlation correction factor in a microarray cross-platform reproducibility study. *BMC Bioinformatics* 2007 Nov 15;8:447. http://www.biomedcentral.com/content/pdf/1471-2105-8-447.pdf

## attributable fraction, *see* attributable risk (AR) fraction

## attributable risk (AR), *see also* attributable risk (AR) fraction, population attributable risk, and population attributable fraction

Attributable risk (AR), also known as *absolute risk difference*, is the difference in the **risk** among the exposed and the nonexposed subjects. If the risk of lung cancer among those smoking for 10 years is 7.6% and in nonsmokers of the same age-gender is 1.2%, the risk attributable to smoking is the difference, i.e., 6.4%. This can also be understood as the *risk difference* and is also called **absolute risk reduction**. For AR to be valid, it is necessary that the groups are similar with respect to all factors except the risk factor under review, i.e., no other factor is present to alter the risk. For studying any risk, information on outcome is obtained for any given risk. This implies that this should be a **prospective study**. Risk can be any exposure or any other **antecedent**.

The method of calculating AR is different for independent samples than for matched pairs.

### *AR in Independent Samples*

For independent samples, the method can be explained with the help of the notations in Table A.11. In this table, *a, b, c, d*, and $O_1$., $O_2$. are the observed cell frequencies, and $\pi$'s are the corresponding probabilities. Sample size is $n_1$ for those with the antecedent and $n_2$ for those without the antecedent.

In terms of probabilities in Table A.11, for independent samples,

$$\text{AR} = \pi_{11} - \pi_{12},$$

provided $\pi_{11} + \pi_{21} = 1$ and $\pi_{12} + \pi_{22} = 1$. This is estimated as

$$\text{AR} = \frac{a}{n_1} - \frac{b}{n_2} = p_1 - p_2.$$

Since AR is the difference between risks in two groups, the confidence interval (CI) for large $n$ can be obtained by the usual method for

**TABLE A.11**
**Structure of a Study for Attributable Risk: Independent Samples**

| | Antecedent | | |
|---|---|---|---|
| **Outcome** | **Present** | **Absent** | **Total** |
| Present | $a$ ($\pi_{11}$) | $b$ ($\pi_{12}$) | $\boldsymbol{O_1.}$ ($\boldsymbol{\pi_1.}$) |
| Absent | $c$ ($\pi_{21}$) | $d$ ($\pi_{22}$) | $\boldsymbol{O_2.}$ ($\boldsymbol{\pi_2.}$) |
| **Total** | $\boldsymbol{n_1}$**(1)** | $\boldsymbol{n_2}$**(1)** | |

A

**TABLE A.12**
**Comparison of Relative Risk (RR) and Attributable Risk (AR) of Death Due to Lung Cancer and Cardiovascular Diseases in Heavy Smokers among British Male Physicians**

| | Annual Death Rate/1000 | | | |
|---|---|---|---|---|
| **Cause of Death** | **Nonsmokers** | **Heavy Smokers** | **RR** | **AR** |
| Lung cancer | 0.07 | 2.27 | 32.43 | 2.20 |
| Cardiovascular disease | 7.32 | 9.93 | 1.36 | 2.61 |

CI for difference in proportions. The test of hypothesis $H_0$: AR = 0 is equivalent to $H_0$: $\pi_{11} = \pi_{12}$ as well as to $H_0$: RR = 1, where RR stands for relative risk. These can be tested by chi-square test or the *z*-test. These procedures are equivalent for large *n*. The hypothesis that AR is a specified quantity $\pi_0$ can also be tested by the ***z*-test**, as explained for proportions.

RR and AR can sometimes lead to very different conclusions. Consider the data in Table A.12, which are from the famous Doll and Hill study of British doctors. It compared the mortality from lung cancer and cardiovascular disease in nonsmokers and heavy smokers (>25 cigarettes per day) from 1951 to 1961.

The RR for lung cancer was high, i.e., 32.43. This indicates a very strong **association** of lung cancer deaths with heavy smoking and underscores the importance of smoking in the etiology of lung cancer. The association of cardiovascular disease death with heavy smoking was mild, only 1.36. But the AR in the two cases was nearly the same. During 1951–1961, elimination of heavy smoking among British male doctors would have reduced the cause-specific mortality for lung cancer almost as much in absolute terms as for cardiovascular disease.

Also note that the risk of death by lung cancer in nonsmokers is low—only 0.07 per 1000—causing RR to appear so high. This is one of the limitations of RR. Against this, the cardiovascular disease mortality is high even among nonsmokers, so that the RR is not so high.

Since AR and RR are functions of $p_1$ and $p_2$ only, $p_1$ and $p_2$ can be uniquely determined when AR and RR are known. Simple algebra yields

$$p_1 = \frac{AR \times RR}{RR - 1} \quad \text{and} \quad p_2 = \frac{AR}{RR - 1}.$$

### *AR in Matched Pairs*

**Matched pairs** is a setup where the same group of people are assessed twice, such as before and after an intervention. Two separate groups where individuals are one-to-one matched for the characteristics that can affect the outcome, except the antecedent under study, are also considered matched pairs. With a dichotomous antecedent and outcome, the data in matched pairs would appear as in Table A.13.

AR in the case of matched pairs (large *n*) is estimated as

$$\text{AR}_\text{M} = \frac{|b - c|}{n},$$

**TABLE A.13**
**Matched Pairs with Dichotomous Antecedent and Dichotomous Outcome: Prospective Study**

| | Partner 1: Antecedent Not Present (Not Exposed or Control) | | |
|---|---|---|---|
| **Partner 2: Antecedent Present (Exposed or Experiment)** | **Positive Outcome (Disease+)** | **Negative Outcome (Disease−)** | **Total** |
| Positive outcome (disease+) | *a* | *b* | ***a + b*** |
| Negative outcome (disease−) | *c* | *d* | ***c + d*** |
| **Total** | ***a + c*** | ***b + d*** | ***n = a + b + c + d*** |

where *b*, *c*, and *n* are the same as in Table A.13. For matched pairs, the hypothesis $H_0$: $\text{AR}_\text{M} = 0$ can be checked by the **McNemar test** when *n* is large. For CI, however, **standard error (SE)** is needed. For matched pairs, it is given by

$$\text{SE}(\text{AR}_\text{M}) = \sqrt{\frac{(b+c)n - (b-c)^2}{n^3}}.$$

Thus,

$$95\%\ \text{CI for AR}_\text{M}\text{: AR}_\text{M} \pm 1.96 * \text{SE}(\text{AR}_\text{M}).$$

These formulas are the same as for a difference in proportions in a matched pairs setup.

For the data on the common cold in Table A.14, $\text{AR}_\text{M} = (15 - 5)/50 = 0.20$. Thus, the therapy seems to increase the 1-week relief rate by 20%. At least, this is the finding from these 50 pairs. Also,

$$\text{SE}(\text{AR}_\text{M}) = \sqrt{\frac{(15+5)50 - (15-5)^2}{50^3}} = 0.08485.$$

Thus, the 95% CI for $\text{AR}_\text{M}$ is

$$(0.20 - 1.96 \times 0.08485,\ 0.20 + 1.96 \times 0.08485)$$

or

$$(0.03,\ 0.37).$$

**TABLE A.14**
**Trial for Therapy for Common Cold: Matched Pairs**

| | Without Therapy (Control Group) | | |
|---|---|---|---|
| **With Therapy (Experimental Group)** | **Relieved within 1 Week** | **Not Relieved within 1 Week** | **Total** |
| Relieved within 1 week | 22 | 15 | **37** |
| Not relieved within 1 week | 5 | 8 | **13** |
| **Total** | **27** | **23** | **50** |

The AR is 20% in this sample, but the CI says that the actual AR in the corresponding population could be as low as 3% or as high as 37%.

Since the CI does not contain 0, $H_0$: AR = 0 can be rejected at the 5% level. If the McNemar test with continuity correction is used for this hypothesis, then

$$\chi^2_M = \frac{\left(|15-5|-1\right)^2}{15+5} = 4.05.$$

This is more than the critical value 3.841 of chi-square at 1 df at a 5% level of significance. Thus, $H_0$: AR = 0 is rejected by this method also. The excess relief with the therapy is statistically significant.

Note the following regarding RR and AR:

- Risk in a statistical sense does not necessarily refer to an adverse outcome. In the data in Table A.14, the risk is for relief within 1 week, which is a positive feature. The term *protective effect* is used for such a factor.
- This section assumes large *n*, but the **binomial** methods can be used to find CI and to test $H_0$ on RR or AR when *n* is small.
- All the RR-related parameters can be estimated from retrospective or case–control studies by replacing RR with OR as an approximation under certain condition, as explained under the term **odds ratio**. This obviates the need to carry out expensive prospective studies. However, *AR does not have this feature.*
- Another advantage with RR opposed to AR is that RR often comes close to multiplying individual RRs when two independent factors act jointly in concert. This does not happen with AR.
- The same AR, such as a difference of 4% between 60% and 64%, has an entirely different implication than same difference, say, between 2% and 6%.
- RR generally represents the magnitude of the association and provides information that can be used in making a judgment on causality. Once a causality inference is drawn, AR assumes importance in delineating the public health importance of the exposure.
- Many times, AR gives more valid information to health managers than RR. This happens particularly when the disease is rare. Suppose oral cancer has a risk of 0.0002 in non-tobacco-chewing persons but has a risk of 0.015 in those who chewed tobacco for 10 years of more. Thus, RR = 0.015/0.0002 = 75, whereas AR = 0.015 − 0.0002 = 0.0148 (less than 1.5% only). Changing habits of tobacco chewing in this setup will not make much of a difference in the overall incidence of oral cancer, RR = 75 notwithstanding. This is all the more true if only a small percentage, such as 2%, of the population chews tobacco.

## attributable risk (AR) fraction, *see also* population attributable risk and population attributable fraction

Attributable risk (AR) measures the expected reduction in risk if the exposure factor is eliminated. A slight modification of AR is the AR fraction. This is the AR calculated as a proportion of the incidence in the exposed group, i.e.,

$$\text{AR fraction} = \frac{\pi_{11} - \pi_{12}}{\pi_{11}} = \left(1 - \frac{1}{\text{RR}}\right),$$

where $\pi_{11}$ is the risk in the exposed group and $\pi_{12}$ is the risk in the unexposed group. The numerator in this expression is the AR. In the alternative expression, RR is the relative risk. In short, the AR fraction is called the **attributable fraction**. It is the proportion contributed by the exposure to the elevated risk—thus, it measures the extent of preventability of the disease from eliminating a particular factor. This is also called a **prevented fraction** because this delineates the preventable incidence when the exposure is eliminated. Another name for the same quantity is **etiologic fraction**. When multiplied by 100, this becomes the AR percent. For example, if hypertension is responsible for 30% of the risk of myocardial infarction among adult males in your area, then AR fraction = 0.30. If you want to see this in action, see, for example, the work of Cosgrove et al. [1], who found after a systematic review of literature that only 20% of cases of type 2 diabetes can be attributed to depression in people.

Note that the AR fraction can be very high even when the AR is low. Suppose the risk of low birth weight in a developed country is 0.06 (6%). Of this 0.06, let 0.04 be due to poor nutrition of mothers. Then the AR fraction of low birth weight due to poor nutrition is 67%. Eliminating poor nutrition is expected to reduce the risk of low birth weight by 67%—from 6% to 4%—but 2% risk will remain.

1. Cosgrove MP, Sargeant LA, Griffin SJ. Does depression increase the risk of developing type 2 diabetes? *Occup Med (Lond)* 2008 Jan;58(1):7–14. http://occmed.oxfordjournals.org/content/58/1/7.long

## attributes

*Attribute* is the statistical name for any characteristic of the unit or subject of the study. The convention is to restrict this term to **qualitative** characteristics measured on **nominal** or **ordinal scales**. Thus, sex, disease, nature of diet, exercise, smoking, type of treatment, etc. are attributes, but creatinine level, hemoglobin level, blood pressure, etc. are not. The latter are on a **metric scale** but can be converted into attributes, such as hemoglobin level converted to no anemia, mild anemia, and severe anemia.

A response on each attribute can be divided into at least two categories—with attribute and without attribute. This is **dichotomous**. You can have **polytomous categories** of an attribute such as blood group, which can be one of O, A, B, and AB. These are **mutually exclusive and exhaustive categories** since each person has to have just one blood group and no other blood group is possible in the same person. However, an attribute such as a symptom may not have mutually exclusive and exhaustive categories. One person can have two or more symptoms, and a subject may have a symptom other than those listed. Observations for an attribute are in terms of the number of subjects (**frequencies**) who fall into different categories of an attribute. If two or more attributes are studied together, the observations are in terms of the number of subjects in cross-classification.

Frequencies in mutually exclusive and exhaustive categories of one or more attributes give rise to what is called a **contingency table**. A basic statistical method for studying association or a pattern in responses of one or more attributes in contingency tables is **chi-square**. But other methods such as **logistic regression** and

A

**log-linear models** are also used for studying attributes when the situation so demands.

In the context of **factor analysis**, where the objective is to discover inherent natural structure among variables, the observed variables are considered **surface attributes**, and the actual underlying factors, *internal attributes*. Factor analysis determines the number and nature of the underlying internal attributes, and the pattern of their influence on surface attributes.

## AUC, *see* area under the concentration curve (AUC curve)

## autocorrelation

Autocorrelation (also called **serial correlation**) is the ordinary **correlation** of values with their preceding values. Correlation of $x_t$ at time $t$ with its value $x_{t-k}$ at time $(t - k)$ is called autocorrelation of lag $k$. If there are $n$ time points so that there are a total of $n$ values, and if $k = 3$, autocorrelation of **lag** 3 will be obtained as the usual correlation between $(x_4, x_5, \ldots, x_n)$ and $(x_1, x_2, \ldots, x_{n-3})$. When a product–moment correlation coefficient is used, for sample values, this is given by

$$\text{autocorrelation of lag } k\text{: } r_k = \frac{\Sigma(x_t - \bar{x})(x_{t-k} - \bar{x})}{\sqrt{\Sigma(x_t - \bar{x})^2}\sqrt{\Sigma(x_{t-k} - \bar{x})^2}}.$$

The usual correlation requires two different series of values, but autocorrelation does not require two series. This is the correlation with its own past values. Inherent in the concept of autocorrelation is a **time series** setup where the subjects or units are measured serially at a large number of different points in time. For example, if a man is checking his prostate-specific antigen level at 3-month intervals, the value in June will depend to a great extent on what it was in March. Death rate due to kidney diseases in a country in 2015 will depend on what it was in previous years. They are autocorrelated. This correlation may remain even when the time factor is properly accounted for because other factors such as population characteristics may also change over this period but are not accounted for.

Beware that the correlation is not necessarily with immediate preceding values but can have what is called **lag** of more than 1. This lag is also called *periodicity*. For example, diseases such as dengue fever have a seasonal pattern since dengue fever depends on *aegypti* mosquito density, which rises in the rainy season in dengue-prone areas. When monthly data are available, it may be more useful to examine the correlation of incidence in a month with the value in the corresponding month a year ago—that is a lag of 12 months. The periodicity of this disease is 12 months. Autocorrelation is a tool to find repeating patterns. The absence of autocorrelation of any lag is an indication of randomness. That is, the observations, despite being in time series, can be considered random in the sense of being independent of time.

In the case of *stationary time series*, the variance of $x_t$ at time $t$ is the same as the variance of $x_{t-k}$ at another point in time $(t - k)$. Thus, the two values in the denominator in the calculation of autocorrelation should be nearly the same in this case, particularly when the time series is sufficiently long. This will make the formula simple, and some books do give this simple formula.

When the lag is not easily identifiable from extraneous considerations, the autocorrelation function, also called **correlogram**, can provide an answer. This plots the autocorrelation of various lags versus lag (Figure A.6) and is also called a correlogram. In Figure A.6, the autocorrelation of lag 2 is high, and autocorrelations of all other lags are small, indicating that the periodicity is 2. The autocorrelation is declining as the lag increases—meaning thereby that the values are related more to the recent past and not so much to the distant past.

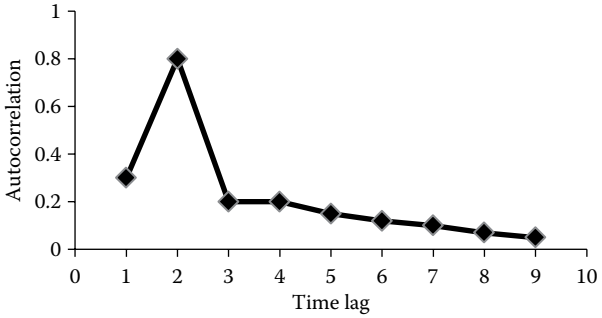



**FIGURE A.6** Autocorrelation function (correlogram).

The presence of autocorrelation violates independence, which is a requirement for many statistical methods. In effect, this tends to reduce the number of variables actually available for modeling since some values can be reasonably predicted by others. Thus, it is customary to check for autocorrelation before such methods are used. The most popular statistical method for this check is the **Durbin–Watson test**. When a change in value from the previous occasion has sensible meaning, this can be a strategy worth trying to find if these differences are uncorrelated. In that case, there is no need to worry about autocorrelation.

In the case of **regression**, for example, a snakily trend or any other nonrandom pattern of **residuals** is an indication that autocorrelation is present. A spurious trend and underestimation of uncertainties may occur when autocorrelation is present but ignored. For this reason, a whole set of separate methods are used for time series analysis with methods such as **autoregressive moving average (ARMA) models**. Similarly, **repeated measures ANOVA** is a method for serial measurements that is distinct from the usual **ANOVA**. Sometimes, **multivariate methods** are used that consider autocorrelated observations as one set.

Awaya and Nishimura [1] have provided a good example of how autocorrelation can be useful. They found that short-term changes in the number of Kawasaki disease cases in Japan have a significant positive correlation (they called it cross-correlation) with pollen releases 9–10 months ago. This suggests first that Kawasaki disease development follows pollen release and second that the lapsed period is somewhere around 9–10 months. The disease could also be the cumulative effect of pollen exposure for elapsed months.

1. Awaya A, Nishimura C. A combination of cross correlation and trend analyses reveals that Kawasaki disease is a pollen-induced delayed-type hyper-sensitivity disease. *Int J Environ Res Public Health* 2014;11(3):2628–41. http://www.mdpi.com/1660-4601/11/3/2628

## autoregressive models, *see* autoregressive moving average (ARMA) models

## autoregressive moving average (ARMA) models

Autoregressive moving average (ARMA) models are tools for analysis of the form and extent of the effect of previous values on the

current values in a sufficiently long **time series**. Generally, observations for not less than 50 time points are needed to use this model, and it is desirable to have more than 100. The ARMA model combines two methods—autoregression and moving averages.

### *Autoregression*

It is natural to expect in a time series setup that the observed values will have some kind of correlation with previous values—called **autocorrelation**. A step further is the form of this relationship—called **regression**. Autoregression is the regression on previous values. An elementary form of autoregression is $x_t = a_0 + a_1 x_{t-1} + e_t$, where, as you can see, the model presumes that the value $x_t$ at time $t$ is mostly determined linearly by the value $x_{t-1}$ at time $(t - 1)$. In this equation, $a_0$ is the **intercept**, $a_1$ is the **regression coefficient**, and $e_t$ is the error term. This equation is based on time **lag** 1, but you may want to investigate dependence on values with some other lag. In that case, $x_{t-1}$ is replaced by $x_{t-K}$ for lag $K$. This model can be generalized to say that the value at time $t$ depends on values on all the $K$ previous occasions. This can also be generalized to include nonlinear terms. For linear terms, the model becomes

$$\text{autoregressive model of order } K, \text{ AR}(K):$$
$$x_t = \alpha_0 + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \ldots + \alpha_K x_{t-K} + \varepsilon_t,$$

where $\varepsilon_t$ are the errors that are independent with each other and with $x$'s, and have fixed variance $\sigma^2$.

The first step in building up an autoregressive model is to identify the lag. This must be much less than the length of the series. You may have extraneous knowledge leading you to believe that a value is mostly determined by only the previous value, in which case the lag is 1; that it is mostly determined by the corresponding value 12 months ago, as in the case of a seasonal trend; or any such plausible reason for any other lag. If there are no a priori reasons for any particular lag, the lag can be statistically identified by plotting the autocorrelation function. Autocorrelation can be significant for two or more lags. In Figure A.6, under the topic **autocorrelation**, it is possible that the autocorrelation of lag 1 is also high in addition to the autocorrelation of lag 2. In that case, the autoregressive model just mentioned can legitimately have $K = 2$.

The next step after identifying the lag is to estimate the regression coefficients of the autoregressive model. These are the parameters of the model. The estimation can be done by the **least squares method** or by using what are called Yule–Walker equations, but the Burg method is considered better [1]. These are mathematically complex procedures that we are not discussing in this volume, but the names are provided so that you can make a judicious choice while running a software package. Keep in mind, though, that higher $K$ makes the model complex, although it may provide a good fit. Parsimony cannot be sacrificed for minor gain; choose the least $K$ that still provides a reasonably good fit. A model with the fewest **parameters** yet adequately describing the data is preferred.

Autoregressive models can be useful in many **longitudinal studies**, particularly when the time points are equally spaced. Once the values are found to be dependent on the previous values (significant autoregression), it can be argued that the same trend will continue, and future values can be partially predicted by the current and previous values. For example, autoregressive models reveal that cognitive declines precede and predict physical decline in Alzheimer disease [2]. In this application, there are two series—cognitive functions and physical functions—but the basic premise is the same, and autoregressive models were used.

### *ARMA Models*

As separately discussed, **moving averages** are the averages of the first $M$ values (from 1 to $M$) out of a total $n$ ($M \ll n$), then the next $M$ values (from 2 to $M + 1$), then the next $M$ values (from 3 to $M + 2$), etc. When calculated for an appropriate period $M$, called order $M$, moving averages can considerably smoothen the time trend by averaging out short-term fluctuations. Thus, moving averages serve two purposes—they help first in detecting a trend and second in recognizing a change in trend.

The moving average (MA) model in this context is defined in terms of the errors because it is already adjusted for the mean, and it is expressed as

$$\text{moving average model of order } M:$$
$$\text{MA}(M)\ x_t = \beta_0 + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \ldots + \beta_M \varepsilon_{t-M}.$$

Obviously, $M$ has to be much less than the length of the series denoted by $n$. When AR and MA models are combined, we get the following:

$$\text{ARMA model of order } (K, M): x_t = \Sigma_{k=1}^{K} \alpha_k x_{t-k} + \Sigma_{m=1}^{M} \beta_m \varepsilon_{t-m} + \varepsilon_t.$$

This ARMA model has come to be regarded as the most basic tool to model a time series. Through a complex process as described by Box et al. [3], some parts of which have been mentioned earlier, the parameters of this model are estimated, and the model is used for forecasting. Sometimes $\log x_t$ is modeled in place of $x_t$ itself, particularly when the variance increases as the values increase, since the log-transformation stabilizes the variance in this case.

As in the case of any model, an important step in ARMA is to validate the model. For this, the first step is to check the biological plausibility so that it looks justified. It should be able to serve the purpose for which it is built and should be feasible to adopt. The second step is to examine this statistically: not only should the model-generated values be close to the sample values, but also, the residuals should be random. Estimated parameters should also be statistically significant.

A step further is autoregressive integrated moving average (ARIMA) models. A moving average is integrated when the terms of the series are "differenced" to make it stationary by removing seasonality. For details, see Box et al. [3].


1. De Hoon MJL, van der Hagen THJJ, Schoonewelle H., van Dam H. Why Yule–Walker should not be used for autoregressive modeling. http://www-stat.wharton.upenn.edu/~steele/Courses/956/ResourceDetails/YWSourceFiles/WhyNotToUseYW.pdf, last accessed May 18, 2015.
2. Zahodne LB, Manly JJ, MacKay-Brandt A, Stern Y. Cognitive declines precede and predict functional declines in aging and Alzheimer's disease. *PLoS One* 2013 Sep 2;8(9):e73645. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3759461/
3. Box GEP, Jenkins GM, Reinsel GC, Ljung GM. *Time Series Analysis: Forecasting and Control*, Fifth Edition. Wiley, 2015.


## average deviation, *see* variation (measures of)

## average linkage method of clustering, *see also* cluster analysis

Average linkage is one of the several methods of **hierarchical** agglomerative clustering. Clustering is the process of putting similar

A

units in one group and dissimilar ones in another distinct groups. In the hierarchical agglomerative process, the two most similar units are put into one group at the first stage. This group is now considered as one entity. Now the distance of this entity from other units is compared with the other distances between various pairs of units. Again, the closest are joined together. This hierarchical agglomerative process goes on in stages, reducing the number of entities by one each time. This is a bottom-up approach. The process is continued until all units are clustered together as one big entity, although it can be stopped midway when the desired kind of clusters are obtained. Note that subsequent clusters completely contain previously formed clusters in this method.

It may not be immediately clear how to compute the distance between two entities containing, say, $I$ and $J$ units, respectively. Several methods are available, such as **single linkage, complete linkage**, and **centroid**. Different methods can give different results. Jain et al. [1] and several others have studied the relative merits and demerits of some of these methods. No specific guideline can be given, but the method of average linkage has been found to perform better in many situations in the sense that the clusters obtained by this method are more compact within and fairly distinct from one another.

The average linkage method uses the average of the distances between units belonging to different entities as the measure of distance between two entities. This is also called the unweighted pair-group method using arithmetic averages **(UPGMA) method of clustering**. Consider entity A comprising $I$ units ($a_1$, $a_2$, …, $a_I$) and entity B comprising $J$ units ($b_1$, $b_2$, …, $b_J$). Then

average linkage method:

$$\text{distance between entity A and entity B } d_{AB} = \frac{1}{I \times J}\Sigma_i \Sigma_j (d_{ij}),$$

where $d_{ij}$ is the distance between $a_i$ and $b_j$. In other words, find all the distances between units of the first entity and units of the second entity, and take the average. In Figure A.7, entity A has 17 units, and entity B has 6 units. Some distances between these units are shown but not all. The average of all these distances is the distance between entity A and entity B under this method. If entity A has $I$ = 5 units and entity B has $J$ = 4 units, this would be the average of 5 × 4 = 20 distances. This method can be modified to include distances between pairs belonging to the same entity also. That is, if entity A has $I$ = 5 units and entity B has $J$ = 4 units, the distances between pairs of all these 9 entities are considered. These will be 9 × 8/2 = 36 distances, and the average of these 36 distances will be considered as the distance between entity A and entity B. This modification is called *average linkage within groups*.

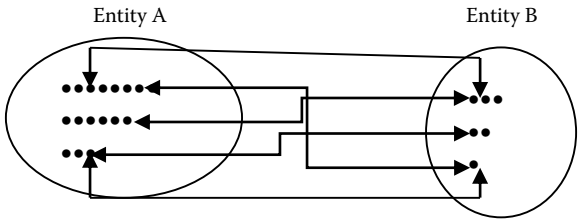



**FIGURE A.7** Average linkage distance between entity A and entity B is the average of distances between units.

The distance between units can be obtained by different methods, but the most popular is the square of the **Euclidean distance**. This is defined as $d_{ij} = (a_i - b_j)^2$ when $a_i$ and $b_j$ are univariate, and as $d_{ij} = \Sigma(a_i - b_j)^2$ when $a_i$ and $b_j$ are multivariate quantities (the sum is over elements of the multivariate quantity). For further details, see Everitt et al. [2].

1. Jain NC, lndrayan A, Goel LR. Monte Carlo comparison of six hierarchical clustering methods on random data. *Pattern Recog* 1986;19:95–9. http://www.sciencedirect.com/science/article/pii/0031320386900385
2. Everitt BS, Landau S, Leese M, Stahl D. *Cluster Analysis*, Fifth Edition. Wiley, 2011.

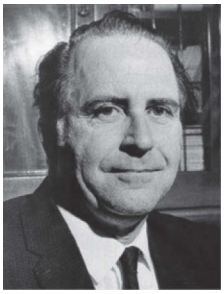
# B



## backward elimination, *see* stepwise methods

## Balaam design, *see also* crossover design

The Balaam design, first proposed by Leslie Balaam in 1968 [1], is a type of higher-order two-period **crossover design** for comparison of regimen A with regimen B in which subjects are randomly allocated to one of four sequences: AA, AB, BA, or BB. Sequence AA means that the subjects in this group receive regimen A in period 1 followed by regimen A again in period 2. Sequence AB means regimen A followed by regimen B, etc. Thus, all subjects receive a regimen two times. It has features of both parallel design and crossover design. The advantage of the Balaam design as opposed to the simple (AB/BA) crossover design is that it works when the effect of a treatment in the first period could be different from its effect in period 2. Most commonly, this happens when a carryover effect is present. Although still desirable, the washout period is less crucial in this design. However, the requirement of subjects in a Balaam design is twice that in the conventional AB/BA crossover design because now we have AA and BB groups as well.



Leslie Balaam
(Courtesy of the University of Sydney Archives.)

Berlin et al. [2] used a Balaam design for comparing the efficacy of a steroid nasal spray with an antihistaminic nasal spray in the treatment of perennial allergic rhinitis. Each of these nasal sprays was compared with placebo—thus, these in fact are two separate trials. The **difference-in-differences approach** was used to compare the two modes of treatment. Described here are essential features of one of these trials that would illustrate the use of the Balaam design.

Twenty patients with perennial allergic rhinitis were randomly allocated to four groups of five patients each. These groups received (i) a steroid nasal spray followed by a steroid nasal spray, (ii) a steroid nasal spray followed by placebo, (iii) placebo followed by a steroid nasal spray, and (iv) placebo followed by placebo, respectively. This completes the Balaam design for this trial. The crossover was done in the middle of the eighth week. Among the outcomes were severity of congestion, daytime sleepiness, and sleep. Each was scored on a scale of 0 to 4. The mean difference in these scores between the drug and the placebo was estimated.

The same process was followed for an antihistaminic spray on another group of 24 patients, and the mean difference in scores between spray and placebo were obtained. This difference was compared with the difference found with the steroid spray. The conclusion they reached is that the steroid spray was more effective than the antihistaminic spray in cases of perennial allergic rhinitis.

For the method of analysis of data from the Balaam design, see Jones and Kenward [3]. The method generally assumes that the carryover effect is determined by the first treatment irrespective of what the second treatment is. This may not hold true with some treatments.


1. Balaam LN. A two-period design with $t^2$ experimental units. *Biometrics* 1968;24:61–73. http://www.jstor.org/discover/10.2307/2528460?uid=3738256&uid=2&uid=4&sid=21104608564967
2. Berlin JM, Golden SJ, Teets S, Lehman EB, Lucas T, Craig TJ. Efficacy of a steroid nasal spray compared with an antihistamine nasal spray in the treatment of perennial allergic rhinitis. *J Am Osteopath Assoc* 2000 Jul;100(7 Suppl):S8–13. http://www.ncbi.nlm.nih.gov/pubmed/10948809
3. Jones B, Kenward MG. *Design and Analysis of Cross-Over Trials*, Third Edition. CRC Press, Boca Raton, FL, 2014.


## balanced and unbalanced designs

One of the important specifications in the **design of a study** is the sample size in different groups under study. If the number of subjects in each group is the same, the design is called balanced—if not, then it is an unbalanced design.

Consider a hypothetical example of a drug for improving hematocrit level in girls of age 10–12 years, 13–15 years, and 16–18 years in a developing country where this measure is generally low in this segment of the population. The sample size could be 30 in each group in a **one-way** setup. Since the number of subjects is the same in each group, this is a balanced design.

Sometimes, the situation demands an unbalanced design, such as illustrated in Table B.1. A study was undertaken by Wang et al. [1] on the effect of maternal smoking on birth weight in the United States. For our example, we vary this study somewhat and assume that some women, who were otherwise habitual smokers, gave up smoking off and on but not completely during pregnancy. Nonsmokers are excluded in this example. Suppose that such women were asked

**TABLE B.1**
**Average Birth Weight (kg) of Children Born to Women with Different Amounts and Durations of Smoking**

| Duration of Smoking in Pregnancy | Amount of Smoking | | | |
|---|---|---|---|---|
| | Mild | Moderate | Heavy | All |
| <18 weeks | 3.45 ($n$ = 15) | 3.42 ($n$ = 12) | 3.43 ($n$ = 7) | **3.44 ($n$ = 34)** |
| 18–31 weeks | 3.38 ($n$ = 8) | 3.40 ($n$ = 10) | 3.39 ($n$ = 6) | **3.39 ($n$ = 24)** |
| 32+ weeks | 3.35 ($n$ = 25) | 3.30 ($n$ = 23) | 3.18 ($n$ = 29) | **3.25 ($n$ = 17)** |
| **All** | **3.41 ($n$ = 28)** | **3.40 ($n$ = 25)** | **3.32 ($n$ = 22)** | **3.38 ($n$ = 75)** |

*Note:* Entries are average birth weight in kilograms.

at the time of their last antenatal visit just before birth about the duration of smoking (factor 1) and the amount of smoking (factor 2). Thus, this is a **two-way design**. At the end of the pregnancy, the researcher categorized duration of smoking as <18 weeks, 18–31 weeks, or ≥32 weeks. These are the three levels of factor 1. The amount of smoking was categorized as mild (1–9 cigarettes per day), moderate (10–19 cigarettes per day), or heavy (20+ cigarettes per day). Only days when there was at least some smoking are included in this calculation, and the amount of smoking is the average per smoking day. These are the three levels of factor 2. The **outcome** of interest (**response variable**) was birth weight of the children born to these women.

In Table B.1, there are, for instance, 15 women who smoked an average of 1–9 cigarettes per day for a total of less than 18 weeks during the entire pregnancy. The average birth weight of their babies was 3.45 kg. There are 8 women who smoked an average of 1–9 cigarettes per day for a total of 18–31 weeks, and the average birth weight of their babies was 3.38 kg, and so on. Because of unequal numbers in different groups, this design is unbalanced.

Analysis of data from an unbalanced design is more complicated than that from a balanced design for a number of reasons. Firstly, the computations are more complex, although that is not a limitation, as the software takes care of this easily. Secondly, an unbalanced design requires more careful consideration of the validity assumptions such as **homoscedasticity**, particularly when the group sizes are not large. Thirdly, the **power** of the study for detecting a specified difference is largest when the total number of subjects is equally divided among groups. Fourthly, the reason for unequal numbers may have to be explained. Thus, use an unbalanced design only when there is full justification for doing so.

1. Wang X, Tager IB, van Vunakis H, Speizer FE, Hanrahan JP. Maternal smoking during pregnancy, urine cotinine concentrations, and birth outcomes: A prospective cohort study. *Int J Epidemiol* 1997;26:978–88.http://ije.oxfordjournals.org/content/26/5/978.long

## Bangdiwala *B*, *see also* Cohen kappa

The Bangdiwala *B*-statistic is a measure of qualitative agreement between two independent raters, two observers, two methods, etc., when the same subjects are classified into the same *K* categories. The problem is frequent in health and medicine: for example, when two radiologists interpret the same x-ray as being definitely negative, probably negative, probably positive, or definitely positive for a particular lesion. **Cohen kappa** also has the same function, but it adjusts for chance agreement by frequencies, whereas Bangdiwala *B* adjusts for chance agreement by areas measured by marginal totals. This was proposed by Shrikant Bangdiwala in 1985 [1].

Shrikant Bangdiwala

**TABLE B.2**
**Assessment of Intrathecal Synthesis by Two Laboratories**

| | Laboratory 1 | | | |
|---|---|---|---|---|
| **Laboratory 2** | **Positive** | **Doubtful** | **Negative** | **Total** |
| Positive | 36 | 5 | 3 | **44** |
| Doubtful | 7 | 12 | 6 | **25** |
| Negative | 1 | 4 | 55 | **60** |
| **Total** | **44** | **21** | **64** | **129** |

$$\text{Bangdiwala } B = \frac{\Sigma O_{kk}^2}{\Sigma(O_{k\bullet}O_{\bullet k})},\ k = 1,2,\ldots,K;$$

where $O_{kk}$ is the cell frequency in the $k$th row and $k$th column (diagonal element), $O_{k\bullet}$ is the marginal total in the $k$th row, and $O_{\bullet k}$ is the marginal total in the $k$th column.

Cohen kappa is much more commonly used, but its value ranges from $-a/(1-a)$ to 1, where $a$ is the agreement expected by chance. The value of the *B*-statistic ranges from 0 for no agreement to 1 for perfect agreement, which makes *B* more easily interpretable.

Consider $n$ subjects or units divided into $K$ categories by two methods, for example, $n = 129$ suspected cases of multiple sclerosis assessed for intrathecal synthesis by two laboratories as positive, doubtful, or negative ($K = 3$ categories). For such data in Table B.2,

$$B = \frac{36^2 + 12^2 + 55^2}{44\times 44 + 25\times 21 + 60\times 64} = 4465/6301 = 0.71.$$

Cohen kappa for the same data is $\kappa = 0.68$. There is hardly any difference for these data, but 0.71 on a scale of 0 to 1 can be appreciated better than 0.68 on a scale of −0.61 to 1. The value −0.61 is the value of Cohen kappa for these data when there is no agreement. Munoz and Bangdiwala [2] have provided a comparison of Cohen kappa and Bangdiwala *B*, and highlighted how *B* can also be represented graphically (see **agreement charts**).

1. Bangdiwala S. A graphical test for observer agreement. *Proc 45th Int Statistical Institute Meeting*, Amsterdam, 1985:307–8.
2. Munoz SR, Bangdiwala SI. Interpretation of Kappa and *B* statistics measures of agreement. *J Appl Stat* 1997;24(1):105–12. http://www.tandfonline.com/doi/abs/10.1080/02664769723918#.UvD4ydKSxIA

## bar diagrams

The bar diagram is probably the most common form of visual representation of data and is indeed very versatile. If the data are **mean**, **rate**, or **ratio**, the bar may be the most appropriate diagram. It is especially suitable for categories on a **nominal** or **ordinal scale**, although it can be drawn for **metric categories** as well. The bar diagram is more suited to disjointed categories, but that is not a prerequisite. For example, age-specific fertility rate in age groups 15–19, 20–24, 25–29, etc. can be shown by a bar diagram despite age categories being contiguous. This is because the value to be depicted is a rate. Had this value been the number of subjects in each age group, the diagram would have been a **histogram**.

In a bar diagram, groups are represented on one axis, and the values in those groups, on the second axis. The length of the bar is proportional to the value in that group. The width of the bar is such that the bar diagram does not look awkward, but it has no meaning.

Generally, a suitable gap is kept between the bars. The width of the bars and the space between consecutive bars is usually kept the same unless one particular category is to be highlighted.

Figure B.1a represents the percentage distribution of palpable breast lumps by histopathological diagnosis in different age groups. This is called a *divided-bar diagram*. According to these data, as age advances, the relative incidence of carcinoma increases. If the total number of women with palpable breast lumps was highest in the age group 45–49 years, these bars would not show this information, since each bar represents the total cases in the respective age groups as 100%. If the number of cases of age 30–34 years is 20 and that of age 35–39 years is 80, both will have the same height in this type of diagram. The comparison in this graph is of percentages with different diagnoses across various age groups.

Figure B.1b is a *multiple-bar diagram* showing the mean and **standard deviation** (SD) of some lung function parameters in male workers in factories with different pollutants. Both between- and within-factory comparisons can be made with the help of this diagram.

The bars do not have to be vertical. Sometimes, horizontal bars give a better representation. In fact, some call diagrams with vertical bars *column diagrams* and those with horizontal bars *bar diagrams*. Figure B.1c is one such diagram with horizontal bars, simultaneously showing two pieces of information for some magisterial districts in South Africa [1]. One horizontal axis measures the number of cervical smears per 1000 women, and the other, the ratio of malignant to suspicious smears. Note how a bar diagram can quickly become complex.

Sometimes, it is difficult to properly interpret a bar diagram in the literature because the information provided is incomplete or the diagram is not properly drawn. Whenever feasible, each bar should be accompanied by the value it represents. For example, if a bar shows that 42% of cases reporting pain in the knees have arthritis, mention "42%" at the top of the bar or inside the bar so that the reader can immediately understand. The bar diagrams in Figure B.1 have this deficiency.

A special bar diagram used in demography is the **population pyramid**. This shows the number or percentage of males and females in different age groups in a population. Another is the **epidemic curve**, which shows the number of new cases arising in a population at different points in time. See these topics for details.

1. Bailie R, Bourne D. Surveillance for equity in cervical cytology screening. *Int J Epidemiol* 1996;25:46–52. http://ije.oxfordjournals.org/content/25/1/46.long

## Barnard test

It is customary to use the **Fisher exact test** for detecting any **association** in a **2 × 2 table** with small frequencies. This test assumes fixed marginal totals—both row totals as well as column totals. This restriction is valid for a situation where, for example, you have six persons with disease and six without disease, and the test is built up in a manner that it is constrained to give six positive (some may be without disease) and six negative (some may be with disease) results. Note that this is an unnatural restriction and would rarely

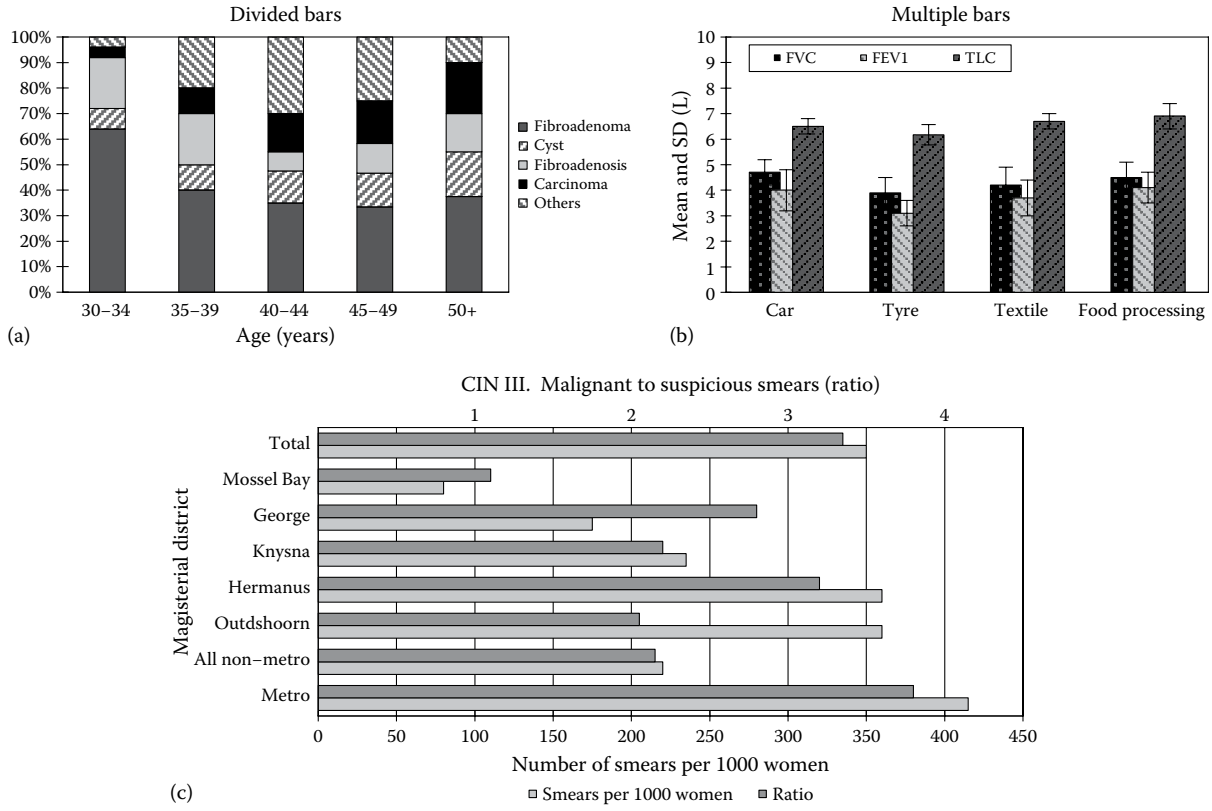



**FIGURE B.1** Different types of bar diagrams: (a) Palpable breast lump by histopathological diagnosis in different age groups; (b) mean and SD of lung functions in male workers in different types of factories; (c) ratio of malignant to suspicious smears and their screening incidence by magisterial district, Western Cape, South Africa. (From Bailie R, Bourne D. *Int J Epidemiol* 1996;25:46–52. http://ije.oxfordjournals.org/content/25/1/46.long.)

B

**TABLE B.3**
**Distribution of Diabetic and Nondiabetic Subjects by Obesity**

| | Diabetics | Nondiabetics | Total |
|---|---|---|---|
| Obese | 10 | 5 | **15** |
| Nonobese | 2 | 4 | **6** |
| **Total** | **12** | **9** | **21** |

hold in practical situations. This restriction makes the test overly conservative (i.e., fails to reject the **null hypothesis** when it should). To overcome this problem, another test, called the Barnard test [1], has been proposed, which requires fixed margins on one side and not both sides. This would be the case when two populations are being sampled as naturally occurs in a **case–control study** and in a **prospective study**. George Barnard proposed this in 1945, to the chagrin of Fisher. This test is more powerful but computationally difficult and is not popular yet. For a comparison between the Fisher exact test and the Barnard test, see Mehta and Senchaudhari [2]. They have shown that the *P*-value for the same data is 0.0641 by Fisher exact test and only 0.0341 by Barnard test. At a significance level of 0.05, the Fisher test does not reject the null, but the Barnard test does. This illustrates that the Barnard test is more powerful in detecting a difference.


George Barnard

The Barnard test uses the product of two **binomials** in place of the hypergeometric used by Fisher. Consider a sample of 12 diabetics and 9 nondiabetic subjects. Suppose 10 of 12 and 5 of 9, respectively, are obese. This could be written as a 2 × 2 table (Table B.3).

Note in this case that the column totals are fixed but not the row totals. Under the null hypothesis that the prevalence of obesity is the same $\pi$ in the two groups, the probability of this configuration by the product of two binomials (product **law of probability** can be used in this case since the two groups are independent) is

$$[^{12}C_{10}\,\pi^{10}(1-\pi)^{(12-10)}]*[^{9}C_{5}\pi^{5}(1-\pi)^{(9-5)}].$$

A similar probability can be obtained for all the configurations adverse to the null and favorable to the alternative hypothesis. The sum of all these probabilities will be the *P*-value for the Barnard test.


1. Barnard GA. A new test for 2 × 2 tables. *Nature* 1945;156 (3954):177. http://www.nature.com/nature/journal/v156/n3974/pdf/156783b0.pdf
2. Mehta CR, Senchaudhuri P. Conditional versus unconditional exact test for comparing two binomials. 2003. http://www.researchgate.net/publication/242179503_Conditional_versus_Unconditional_Exact_Tests_for_Comparing_Two_Binomials, last accessed May 23, 2015.


## Bartlett test

**Homoscedasticity** is the equality of variances in different groups. This requirement should be reasonably satisfied for **analysis of variance (ANOVA)** and some other parametric procedures. This can be checked graphically by the **box plot** for different groups. The conventional statistical test for checking homoscedasticity is the Bartlett test. This was proposed by Maurice Bartlett in 1937 [1].


Maurice Bartlett

The Bartlett test compares the weighted **arithmetic mean** of the sample variances with their weighted **geometric mean**. The geometric mean has the property that it is always less than the arithmetic mean and is equal only when all the values are equal. A large difference between the arithmetic mean and the geometric mean indicates that the values are different. A function of these two types of means is used as the **test criterion**, which follows an approximate **chi-square distribution** for large $n$ when the underlying distribution is Gaussian. The **degrees of freedom** are $K - 1$, where $K$ is the number of groups. The test should not be statistically significant for assuming homoscedasticity. The approximation is unsatisfactory if sample size in many of the $K$ groups is less than 5.

The validity of this test is heavily dependent on the distribution of the underlying **variable** being Gaussian and requires a large **sample size**. For this reason, many software packages now use the **Levene test**. The Levene test is based on the median and thus is more robust to any departure from Gaussianity.

Besides **ANOVA**, the Bartlett test can be used in a variety of other situations where equality of variances is a requirement. Its multivariate analog is available [2] for use with **Hotelling $T^2$** and **multivariate ANOVA** setups to test homogeneity of **dispersion matrices** across groups. This version of the test also requires **multivariate Gaussianity** of the data.

One of the requirements of a successful **principal components** analysis and **factor analysis** is that the **correlations** are sufficiently high among variables. Another Bartlett test [3] can be used to test the null hypothesis that the correlation matrix is proportionate to an identity. An identity matrix is the one in which the diagonal elements are unity (the correlation of a variable with itself is always 1) and the off-diagonal elements are zero (no correlation between variables). This is called the Bartlett test for **sphericity**. Such a test should reject the null hypothesis (say, $P < 0.05$) for factor analysis to be a useful procedure.


1. Bartlett MS. Properties of sufficiency and statistical tests. *Proc Royal Stat Soc*, Series A1937;160:268–82. http://www.jstor.org/discover/10.2307/96803?uid=3738256&uid=2&uid=4&sid=21103028921443
2. Bartlett MS. On the theoretical specification of sampling properties of autocorrelated time series. *Journal of Royal Statistical Society, Series B, Supplement* 1946;8:27–41. http://www.jstor.org/discover/10.2307/2983611
3. Bartlett MS. Tests of significance in factor analysis. *Br J Stat Psychol* 1950;3:77–85. http://onlinelibrary.wiley.com/doi/10.1111/j.2044-8317.1950.tb00285.x/abstract

## baseline, *see also* baseline equivalence, baseline hazard/risk

Baseline refers to the status of the subjects of a study before exposure to a potentially hazardous or beneficial maneuver. Postexposure values are compared with baseline values for assessing whether the exposure made any difference. When preexposure values are not available for a group, values in another group without exposure can play the role of baseline. In this case, for a valid comparison, the exposed and unexposed groups are chosen such that they are similar at baseline, particularly with regard to the factors that can influence the outcome. This is called *baseline matching*. This is one of the methods used to achieve **baseline equivalence**.

In **paired samples**, the change from preexposure to postexposure is the typical outcome of interest. Since baseline (pre) values can affect the amount of change, the percent change is sometimes calculated. This works well provided that the percent change is independent of the baseline. This limitation tends to be overlooked, and this can lead to spurious results, as illustrated in the following example.

Given here are hypothetical data on the rise or fall in hemoglobin (Hb) levels in a random sample of eight adolescent girls after iron supplementation for 2 weeks.

Rise in Hb level (g/dL): 0.4 0.7 −0.9 0.3 0.1 0.5 0.9 0.2

A one-sample ***t*-test** reveals that this rise, on average (mean rise = 0.275 g/dL), is not statistically **significant**. Although the rise in this example is calculated over the baseline values, a rise of 0.3 over, for example, 8.7 g/dL is considered the same in this example as the same rise over, say, 13.7 g/dL. In this sense, the rise does not consider the actual Hb level in the subjects before the start of the supplementation. Now examine the situation when the baseline levels are also considered. The following are the corresponding Hb levels (g/dL) of these eight girls before supplementation:

Baseline Hb level (g/dL): 10.1 8.5 13.8 9.7 12.4 9.0 8.6 12.1

This indicates that the girls with the higher baseline levels of Hb showed a smaller rise. In fact, a girl with an Hb level of 13.8 g/dL showed a fall. The relationship can be examined by running a **simple linear regression** of rise on the baseline level. This regression equation is

$$\text{rise} = 2.85 - 0.244\ (\text{baseline Hb}).$$

This shows that the rise, on average, is higher when the baseline values are lower and becomes negative in this case when the baseline Hb level exceeds 11.7 g/dL. Since a fall can be ruled out in the case of iron supplementation except for extraneous reasons, it might be concluded that iron supplementation in such girls is possibly not useful when Hb level is already 11.7 g/dL or higher. The lesson from this example is that consideration of baseline values can be important for valid inference.

Baseline is important in some other contexts as well. In the case of population studies, baseline information can be obtained by what is called **health situation analysis**. Baseline information on enrolled subjects in a clinical trial may indicate that the design needs to be modified with respect to doses, sample size, duration of follow-up, type of subjects (say, age 70+ to be excluded), etc. In a prospective study, some subjects generally drop out during the course of the investigation. Since baseline information on them would be available in this case, that can be used for adjustment of results.

It is usually preferable to undertake an **analysis of covariance (ANCOVA)** of the postintervention values with baseline values as a **covariate** rather than undertaking an analysis of changes from baseline. ANCOVA would adjust results for differentials at the baseline. For a convincing example, see Vickers [1].

1. Vickers AJ. Statistical reanalysis of four recent randomized trials of acupuncture for pain using analysis of covariance. *Clin J Pain* 2004;20:319–23. http://www.ncbi.nlm.nih.gov/pubmed/15322438

## baseline equivalence, *see also* baseline

The term *baseline equivalence* is generally used in the context of case–control studies, medical experiments, and clinical trials. It refers to the fact that the subjects in different treatment groups to start with are similar, and more or less ensures that any difference observed subsequently between groups can be legitimately ascribed to the exposure or intervention.

In a case–control setup, generally, the subjects with disease and controls without disease are compared with respect to past exposure. These two groups should be equivalent for all characteristics except the exposures under study. This is referred to as baseline matching.

Baseline equivalence is usually easily achieved in medical experiments because of the nature of the material. For example, an experiment on the shape of red blood cells at various ionic strengths will usually fulfill the baseline equivalence requirement because various groups of cells (including those in the control group) typically come from the same stock. It is because of this equivalence and standardized laboratory conditions that the experiments are able to provide compelling evidence of the presence or absence of a cause–effect type of relationship.

In a clinical trial setup, the difference in outcome between the test group and the control group can be legitimately ascribed to the intervention when the participants with and without intervention are equivalent to begin with. **Randomization** is a very potent tool to achieve such equivalence. This works well for trials on a large number of participants but can occasionally fail for small samples. If a trial on a large enough sample cannot be conducted, alternate strategies are as follows:

- Identify pairs of participants matched for baseline characteristics and **randomly allocate** one of each pair to the two **arms** (test and control) of the trial. However, this approach is applicable only if a group of subjects is readily available for randomization so that matched pairs can be selected. In practice, that is rarely the case as the subjects are generally recruited and randomized one at a time.
- Adopt **stratified randomization** to ensure that the subjects with similar baseline characteristics are randomized in nearly equal numbers to treatment groups.
- Use **minimization**, which determines allocation to treatment groups at the time of randomization of each individual subject by considering the baseline characteristics of all previously randomized subjects and the one now to be randomized. The minimization process then allocates the subject with nearly 100% probability to the treatment group, which results in the best baseline equivalence between the treatment groups.

It is quite common to use statistical tests to compare groups with respect to baseline characteristics to confirm baseline equivalence.

However, if allocation to groups has been at random, this is not usually appropriate or technically valid. Such tests determine the probability that observed between-group differences in baseline characteristics could have arisen by random allocation to the groups, so they are not really appropriate when it is already known that allocation was at random.

## baseline hazard/risk

**Hazard** is the instantaneous rate of an outcome at a particular point in time. This presumes that the outcome depends on time, for example, the hazard of death at age 70 years. While the hazard will vary with time, in most situations, it will also depend on other factors as well. For example, the hazard of death of a person at age 70 years when the person is male, was diagnosed with leukemia 2 years ago, and is thin is different from the hazard of death of a person who is female and thin with no other risk factor. Baseline hazard is the hazard in the absence of any risk factor. In this example, the hazard of death at age 70 years for a person who has no other risk factor is the baseline hazard for this age.

The most pronounced application of the baseline hazard is in the **Cox model**. This is given by

$$h(t) = h_0(t)*\exp(b_1x_1 + b_2x_2 + \ldots + b_Kx_K),$$

where $h(t)$ is the hazard of the outcome such as death at time $t$ and $h_0(t)$ is the baseline hazard at this time when all $x$'s are 0. These $x$'s are the risk factors, as explained in the preceding paragraph. In the context of the Cox model, they are called *covariates*. Note that $h_0(t)$ represents the effect of time alone without the covariates. The binary outcome in this is not necessarily survive/die and could be any end point reached or not reached. Interpretation of the baseline hazard $h_0(t)$ remains the same.

Another similar term is *baseline risk*. This has a slightly different interpretation. This is the risk in an unexposed group at a defined end point. Note that **risk** is obtained for the terminal event, whereas hazard can be computed for any point in time during the interregnum.

Baseline risk can play an important role in the interpretation of **relative risk** (RR). For example, if the 10-year risk of death with leukemia is 0.99, and with anemia, only 0.02, then the RR of death with leukemia is nearly (0.99/0/02 ≈) 50. On the other hand, if the comparison is with breast cancer cases, where the 10-year risk of death may be 0.60, then the RR is only 0.99/0.60 = 1.65. This is as high as it can get in this situation because no risk can exceed 1 (see the numerator of this RR). The value of RR is greatly influenced by the risk in the control group. Also, 1 out 2 is 50%, and 50 out of 100 is also 50%. Underlying numbers are important. These two aspects are many times forgotten while interpreting the value of RR.

## bathtub curve/distribution

A bathtub curve has a shape similar to that of a bathtub, with high values at two extremes and low values in the middle. It is difficult to imagine a medical situation where values at both the extremes are nearly equal to give it an exact bathtub shape, but some medical situations approximate this pattern. The plot of the risk of death in human beings over age is shaped nearly like a bathtub (Figure B.2a). The risk of death is relatively high in the earliest years of life, decreasing fairly soon to a minimum, remaining low for a substantial part of life, and then starting to climb again sometime around age 50 years, when people start to die rapidly as they grow older. Thus, it has three distinct phases—decreasing risk, nearly a constant risk, and steeply increasing risk.

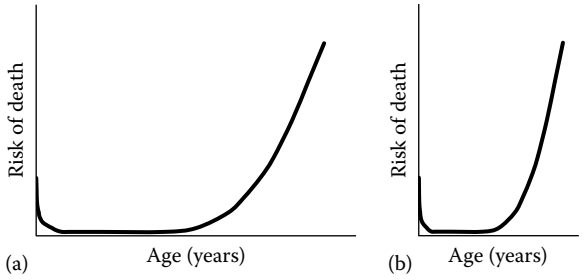



**FIGURE B.2** Bathtub curve for risk of death over age (a) with normal $x$-axis and (b) with shrunken $x$-axis.

The risk of many diseases or risk of death is high for unusually low values of most medical parameters as well as for exceedingly high values of these parameters. For example, low body mass index (BMI) is also harmful, although possibly not as much as very high BMI, while persons with values between, say, 20 and 27 kg/m$^2$ have low risk. Similarly, the risk of failure of, say, a kidney transplantation is high in the initial few days or weeks. If that period passes, the risk is low for quite some time but steeply increases after a few years as age advances. All these situations yield a bathtub-shaped curve.

If the $x$-axis (age in Figure B.2a) is shrunk, the same curve may look like a **J-shaped curve**. Figure B.2b is based on exactly the same data, but the $x$-axis is shrunk, giving it a J-shape. If the risk at the beginning of life is almost as high as at the end, this may look like a **U-shaped curve** with a shrunk $x$-axis, and like an actual bathtub for an expanded $x$-axis, the difference being that the flat part in the U-shaped curve in the middle is practically nonexistent or small, whereas this region is substantial in the bathtub curve.

Note that the examples we have provided are for the bathtub curve and not for bathtub distribution. This becomes a **distribution** when the $y$-axis is frequency or percentage of subjects in place of risk or rate. In the case of deaths over age, if you plot the number of deaths in a general population over age, you will have a bathtub shape to a point, say, up to the age 85 years, but after that, since the population is small, the *number of deaths* is also small. This gives rise to what we call a **smoke-pipe distribution**.

## Bayesian confidence interval, *see* Bayesian inference

## Bayesian inference, *see also* Bayes rule

Bayesian inference is a method to derive new knowledge on the basis of some kind of objective collation of the existing evidence. This looks like an elegant way to evolve but has stumbling blocks. Some believe that Bayes, who was associated with a church, devised his rule as an attempt to prove the existence of God [1]. We describe the Bayes methodology in the context of statistical inference.

Among the various ways in which the statistical inference approach can be divided, one is based on how the inferences are derived. One approach is called Bayesian, and the other is classical, called the *frequentist approach*. For example, when we find a **confidence interval (CI)**, the inference is based on what *repeated* CIs with similar samples will reveal. This is a frequentist approach, pioneered by Ronald Fisher, considered by many as the father of statistics. A 95% CI, when computed for, say, 200 random samples of the same size from the same population, is likely to contain the true value of the parameter 190 (95%) times. The value of the parameter is considered fixed. In contrast, in a Bayesian CI, the parameter value

is considered to have a **distribution**. This is called the prior distribution since it is based on prestudy knowledge. The sample values are used to "update" this *prior distribution* using the **Bayes rule**, and this updated distribution is called the *posterior distribution*. The CI is then based on this posterior distribution. In the case of testing of a hypothesis, Bayesian procedure is to compute the probability of the value of the parameter specified in the null hypothesis given in the sample values. In terms of **conditional probability**, this can be written as $P(\theta|x)$, where $\theta$ is the parameter, $x$ is the sample, and $P$ stands for probability. On the other hand, in a frequentist approach, our inference is based on $P(x|\theta)$.

Note that the Bayes rule is central to the Bayesian inference. This rule helps to convert one **conditional probability** $P(x|\theta)$ into its inverse $P(\theta|x)$ but requires knowledge of $P(\theta)$ and $P(x)$. These probabilities depend on the distribution of $\theta$ and the distribution of $x$. While the distribution of the sample values $x$ is easy to handle, the difficulty arises in postulating a prior distribution of $\theta$. This assumes that we know some properties of the parameter on the basis of which $P(\theta)$ can be postulated. For example, we may postulate on the basis of the previous experience that the chance of the next person in a renal clinic having a failed kidney is 1 in 15—that is, the clinic has been receiving, on average, 1 out of 15 cases with a failed kidney. This is our prior belief. If all of the next few cases turn out to have a failed kidney, our belief alters, reflecting the posterior probability. Such probabilities are commonly used in clinical practice. When a patient goes to a clinic with certain complaints, you mentally evaluate his/her chances of being affected by a particular disease on the basis of clinical findings. When the results of medical tests or radiological images are available, that belief may alter.

Kim and Kim [2] used a Bayesian inference approach in the empirical prediction of genomic susceptibilities for multiple cancer classes. They examined four prediction models and used their combined prediction of the susceptibility for each test individual using a Bayes approach.


1. Hooper M. Richard Price, Bayes' theorem, and God. *Significance* Feb 2013;36–9. http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2013.00638.x/epdf
2. Kim M, Kim SH. Empirical prediction of genomic susceptibilities for multiple cancer classes. *Proc Natl Acad Sci U S A*. 2014 Feb 4;111(5):1921–6. http://www.pnas.org/content/111/5/1921.full.pdf+html


## Bayes rule, *see also* Bayesian inference

The basic premise of Bayes is to change our mind as new evidence emerges. Thomas Bayes described this rule in the 1740s, but it was Laplace who described it independently in 1774 and gave it a modern mathematical form [1]. The rule can be understood easily in terms of the basic concepts of **probabilities**.

Thomas Bayes

The Bayes rule has widespread applications, but this section is restricted to its prominent use in medical sciences. Medical knowledge sometimes is such as to provide probabilities of the form $P$(complaints|disease), which is the probability of complaints for a given disease, whereas the probabilities actually required in diagnostic process are the probability of disease given the complaints, namely, $P$(disease|complaints). Both of these are conditional probabilities. One is the inverse of the other in the sense that the first is the probability of complaints conditioned on disease being known, and the second is the probability of disease conditioned on complaints being known. The Bayes rule is a tool that helps in finding such **inverse probabilities**.

For simplicity and for generalizability, denote the set of complaints by $C$ and the particular disease by $D$. When some additional information is available, $P(D|C)$ can be obtained from $P(C|D)$ by using the Bayes rule. For this conversion, two additional probabilities are required. The first is $P(D)$. This is the **prior probability** of the disease in the absence of any information and is the same as the prevalence of the disease in the subjects under investigation. This may be derived from, for example, the records of the health facility where this is to be applied or may be available from a survey. The second is $P(C)$. This is the **relative frequency** of the complaints in similar patients. Special efforts may be required to compute $P(C)$, but it is worth the effort given the difficulty in obtaining $P(D|C)$ directly. As it turns out, $P(C)$ can often be easily computed by an alternative method, given later in this section. Once these probabilities are available, the required inverse probability $P(D|C)$ can be computed from

$$\text{Bayes rule: } P(D \mid C) = \frac{P(C \mid D)P(D)}{P(C)}.$$

This gives the **posterior probability** after the complaints are known. The probabilities $P(C|D)$, $P(D)$, and $P(C)$ may be available in a research setup, but for everyday practice, an educated guess may have to be made on the basis of knowledge and experience, as just mentioned.

The Bayes rule is more useful when it is known that the set of complaints $C$ can occur only in $K$ **mutually exclusive and exhaustive categories**. If the group of complaints (abdominal pain, vomiting, and constipation) are considered to occur only in abdominal tuberculosis (abdTb) ($D_1$), amebiasis ($D_2$), and hepatitis ($D_3$) and no other disease, and if the diseases do not really overlap, then we have $K = 3$ mutually exclusive and exhaustive categories of $C$. If that is so, we can find the probability of abdTb given those complaints by using the following:

$$\text{Bayes rule (slightly expanded version): } P(D_1 \mid C) = \frac{P(C \mid D_1)P(D_1)}{P(C \mid D_1)P(D_1) + P(C \mid D_2)P(D_2) + P(C \mid D_3)P(D_3)}.$$

This version obviates the need to know $P(C)$—the relative frequency of complaints in this example, which could be difficult to obtain in many cases. The following example illustrates its use and also the caution required in interpreting an inverse probability.

Suppose only 1 in 2000 male adults in an area suffers from a coronary disease for which a screening tool is the electrocardiogram (ECG). When a person actually has the disease, suppose ECG is so good that it is positive 98% of the time, and it is also positive just 1% of the time in those who do not have the disease. What is the probability that an ECG-positive person actually has the disease?

Let $D_1$ = {person has the disease}, $D_2$ = {person does not have the disease}, and $C$ = {positive ECG}. Then, the preceding paragraph says that $P(D_1) = 1/2000 = 0.0005$, $P(D_2) = 1 - 0.0005 = 0.9995$, $P(C|D_1) = 0.98$, and $P(C|D_2) = 0.01$. Use the following to find $P(D_1|C)$.

Since $D_1$ and $D_2$ are mutually exclusive ($K = 2$ in this case), an expanded version of the Bayes rule gives

$$P(D_1 \mid C) = \frac{0.98 \times 0.0005}{0.98 \times 0.0005 + 0.01 \times 0.9995} = 0.047.$$

If the values are really as in this example, a very small $P(D_1|C)$ indicates that a positive ECG does not provide any clue to the disease. This may seem paradoxical in view of the 98% positivity of the ECG in coronary cases. The difficulty is in it being positive in 1% of the subjects without the disease. And this group is very large: 1999 out of 2000 in this screening exercise. Thus, most positive results arise from errors rather than from diseased cases. A tool with a much smaller error rate is required to be effective in such a situation.

In clinics, however, an ECG will be done only for those who are otherwise suspected to have a coronary problem. And in them, $P(D_1)$ would be high. $P(C|D_1)$ would also be high. Thus, $P(D_1|C)$ would be much higher in a clinic setup than the one obtained earlier in a screening example. ECG will not be a waste of resources in this setup.

1. McGrayne SB. *The Theory that would not Die*. Yale University Press, 2011.

## BCPE method, *see* the Box–Cox power exponential (BCPE) method

## bed–population ratio, *see* health infrastructure (indicators of)

## before–after design/study

In an experimental setup, when a separate group of units receives a control regimen, this is called a **parallel control** group. Some experiments are undertaken without a parallel control group. For example, a study measuring tail-flick latency in mice exposed to heat before and after administration of an analgesic does not have a parallel control group. Such a design is called a before–after design, and the experiment is called a before–after study. Since there is no parallel control group in this setup, this is sometimes referred to as an *uncontrolled design*, although this is rather misleading since each subject serves as its own control. Many call this a **quasi-experimental design** because of the lack of random allocation to a test and comparison group.

Instead of incorporating a separate control group, it is indeed sometimes prudent to assess the outcome before and after the intervention in the same unit. In a bacterial colony, the bacteria could be counted initially, provided a specific favorable environment such as sucrose in the medium, and counted again after specified time. This is a before–after experiment and the simplest form of a **repeated measures design**. In a **clinical trial**, one can measure oxidative stress enzyme levels (e.g., superoxide dismutase) in Parkinson disease, introduce an intervention to reduce this stress, and measure again. A parallel control group is not necessary in this setup, although some experiments may have a parallel group as well for observing the trend in them also after placebo.

Before–after experiments have smaller variability than parallel-group designs because the subjects for before–after measurements are the same. Such experiments also require fewer subjects to detect the same effect as compared with parallel control experiments.

A major problem with the before–after design is that it assumes that nothing except the intervention is changing the outcome. But the observed effect could be at least partially due to psychological factors that operate in a **placebo** group. There might also be some natural changes over time. The **Hawthorne effect** may also be working. Extremely low values at prestage could become slightly high, and extremely high values, slightly low because of the effect of **regression to the mean**. All these are inextricably mixed, and it is impossible with this type of design to separate the intervention effect from other effects. Lipsey and Wilson [1] identified 45 reviews on behavioral treatments that separately reported the pooled estimate for controlled and uncontrolled studies and found the observed effect for uncontrolled studies to be greater than that from studies with parallel controls. This explains why a before–after design is typically used when parallel controls are not feasible and randomization is difficult.

The analysis of data from a before–after study is generally done by considering the difference between *post-* and *pre-*values. Since both these values are subject to sampling fluctuation and measurement errors, the differences are kind of doubly exposed to these "ills." The better approach is to use pre-values as covariates and do an **analysis of covariance**. If a parallel control group is also subjected to before–after measurements, a commonsense approach to arrive at the causative relationship is the **difference-in-differences** method. The before–after difference in the test group is compared with the before–after difference in the control group. If this difference in differences is statistically significant, you can safely conclude that the test regimen has some effect. Realize, though, that this method also is approximate due to problems of double errors with the differences. One can hope that in the case of a difference in differences, these errors cancel out between the groups.

1. Lipsey MW, Wilson DB. The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *Am Psychol* 1993 Dec;48(12):1181–209. http://psycnet.apa.org/journals/amp/48/12/1181/

## Beherens–Fisher problem, *see* Welch test

## bell-shaped distribution, *see* Gaussian distribution

## Berkson bias

This bias occurs when a conclusion is drawn from a special group of subjects that do not represent the target population. Berkson bias is common in case–control studies in a hospital setting where the controls also are from the hospital and these controls show overrepresentation (or underrepresentation) of the risk factor under study. Overrepresentation can happen since hospital controls will usually have some disease and this disease may also be related to the risk factor under study.

Consider a simple example of smoking as a risk factor for lung cancer. If this study is done in a hospital setting using a case–control design, you might take patients with diseases other than cancer from the same hospital as controls. However, these patients may also have a preponderance of smokers, as smoking can have manifestations other than lung cancer; for example, they may have asthma. Despite being noncancer patients, asthma patients are not proper controls for studying smoking as a risk factor. Controls should be chosen such

that they have the same prevalence of the risk factor (in our example, smoking) as in the general population. Even cases of fracture may not be appropriate controls, since it is possible that smokers get more fractures than nonsmokers.


Joseph Berkson

The concept of Berkson bias was first introduced by Joseph Berkson in 1946 [1]. Details are available in Westreich [2].

Hernández-Díaz et al. [3] have discussed the paradox that among low-birth-weight (LBW) babies in the United States, infant mortality rate (IMR) was lower for children born to smokers compared with the children born to nonsmokers. This was not due to a beneficial effect of maternal smoking but possibly because LBW babies born to nonsmokers had other debilitating conditions such as birth defects that caused LBW and higher mortality. Another example is provided by Peroutka et al. [4], who concluded that certain migraine comorbidities reported in the literature may have resulted from Berkson bias as opposed to a shared pathophysiological variation in the C3 gene. These examples illustrate that Berkson bias is an important phenomenon and cannot be overlooked by studies that investigate cause and effect or etiology using hospital patients.


1. Berkson J. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bull* 1946;2:47–53. Reprinted *Int J Epidemiol* 2014;43:511–5. http://ije.oxfordjournals.org/content/early/2014/02/28/ije.dyu022.full
2. Westreich D. Berkson's bias, selection bias, and missing data. *Epidemiology* 2012 Jan;23(1):159–64. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3237868/
3. Hernández-Díaz S, Schisterman EF, Hernán MA. The birth weight "paradox" uncovered? *Am J Epidemiol* 2006 Dec 1;164(11):1115–20. http://aje.oxfordjournals.org/content/164/11/1115.long
4. Peroutka SJ, Price SC, Jones KW. The comorbid association of migraine with osteoarthritis and hypertension: Complement C3F and Berkson's bias. *Cephalalgia* 1997 Feb;17(1):23–6. http://cep.sagepub.com/content/17/1/23.long


## best subset method of choosing predictors

As the name implies, the best subset method is choosing a subset of predictors out of all possible subsets that best predicts or explains the outcome. The criterion for "best" is defined before the procedure begins—this could be highest **coefficient of determination**, highest **statistical significance**, least **misclassifications**, most convincing biological plausibility, or any other such criterion.

In a multivariable situation such as in **regression** and **discriminant function**, it is customary to start with a large number of possible predictors. Using all of them is one possibility, but that tends to make the model too complex for adoption. The estimates of the regression coefficients also become less efficient. Various statistical methods are used to select a few that are more relevant or significant. Suppose there are $K$ predictors ($x_1$, $x_2$, …, $x_K$) to begin with, and the outcome variable is $y$. Under the best subset method, the relationship of $y$ with each $x_k$, of $y$ with pairs $x_k$ and $x_{k*}$, of $y$ with each triplet of $x$'s, etc. are examined. That is, all possible combinations are tried. With $K$ predictors, this will amount to a total of $2^K$ combinations. If $K = 4$, there will be a total of $2^4 = 16$ combinations. This number will rise steeply as the number of candidate predictors increases. Fortunately, statistical packages can easily handle these calculations, and the subset of predictors that provide the best prediction can be identified.

In a **multiple linear regression** setup, for which the best subset method is commonly used, only the linear combinations are examined. However, it is possible to have, for example, $x_2 = x_1^2$ or $x_3 = x_1 * x_2$, or any such function, so long as they do not introduce **multicollinearity**. All other requirements such as **independence**, **Gaussianity**, and **homoscedasticity** must also be fulfilled.

## beta distribution

The beta distribution is a **distribution** with varying shapes depending on the value of its parameters (Figure B.3). This distribution has two **parameters** in the same way that the Gaussian distribution has two parameters. One is generally denoted by α and the other by β, although these are not universally accepted notations (**Gaussian distribution** has universally accepted notations for its parameters, namely, μ and σ). In the case of Gaussian distribution, μ denotes mean, and σ denotes standard deviation, but this is not so for α and β for beta distribution. But they also determine the shape of the distribution. Figure B.3 shows the shape of this distribution for different values of α and β. For example, the distribution is just a flat line (see **uniform distribution**) for α = 1 and β = 1 and is right skewed for α = 2 and β = 6 but an extreme shape for α = 2 and β = 0.1. It has a symmetric shape when α = β.

Because of the wide and varied shapes available for different values of α and β within the beta distribution, this distribution is sometimes used to model health outcomes. For example, in an interesting paper, Kalmijn et al. [1] used beta distribution to model the distribution of happiness score among respondents. They measured happiness on a scale of 0 to 10. Among its statistical usage is in exact **confidence interval** for a probability π in a **binomial** case, which is obtained by using a beta distribution.

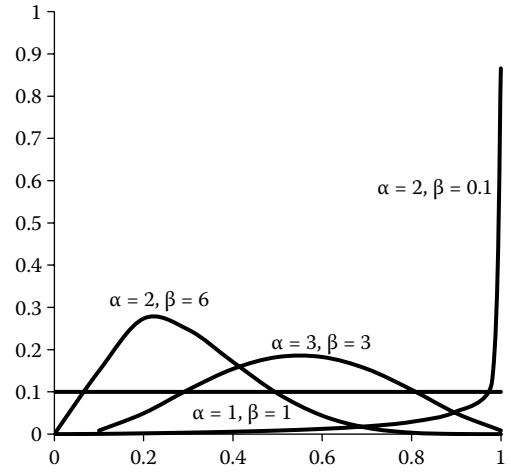



**FIGURE B.3** Shape of beta distribution for different values of the parameters α, β.

While trying to fit a beta distribution to a data set, the first step would be to **normalize** the values $y$ by using the transformation $x = (y - A)/(A - B)$, where (A, B) is the possible range of $y$ values. The possible range can be wider than the range you actually observed in your sample. These can be ignored. After normalization, all values of $x$ will be between 0 and 1. The next step is to calculate the mean and variance of your $x$ values. The estimates of $\alpha$ and $\beta$ are obtained by solving the equations

$$\text{mean} = \frac{\alpha}{\alpha+\beta} \text{ and } \text{variance} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

These equations can be solved easily for $\alpha$ and $\beta$ by using some high school algebra. Use the estimated values of $\alpha$ and $\beta$ to get the distribution with the help of appropriate software.

1. Kalmijn WM, Arends LR, Veenhoven R. Happiness Scale Interval Study: Methodological considerations. *Soc Indic Res* 2011 Jul;102(3):497–515. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3105228/pdf/11205_2010_Article_9688.pdf

## beta error, *see* Type II error

## biased estimate, *see* unbiased estimate

## biased sample

A sample is considered **unbiased** when it truly represents the features of the target population. If it does not, it is a biased sample. This is a frequent source of error in statistical conclusions. This can happen even when the selection is **random** or even when **random allocation** is made in experimental studies. A large sample tends to magnify this bias rather than control it. The following types of subjects can provide a biased sample in medical studies.

### *Survivors*

Consider the relationship of lung functions with age after the age of 50 years. It is well known that lung functions decline because of biological degeneration in old age, but the gradient differs from population to population. How does one find the exact effect of age in a particular population? Those with poor lung functions are likely to be in poor health and thus have less longevity. They may not be available in old age. Any estimation based on survivors is bound to be biased. It may not be easy to find the correct gradient in this case. Perhaps a more valid picture is obtained if lung function parameters of those dying at varying ages are available in records and are used for adjustment, but such data are not usually available. In the absence of such adjustment, lung functions among survivors are not a true reflection of the age-based gradient. The actual gradient could be a sharper decline. Similarly, a study based only on hospital cases will exclude those who die before admission. Serious cases, those residing in remote areas, and those who are too poor to afford hospitalization tend to be excluded. Similar bias occurs in a variety of other situations where the study is on prevalent cases rather than on incident cases.

The same sort of caution is required in some serial procedures in medicine. In some cancers, surgery is undertaken when radiotherapy fails and the patient remains operable at the conclusion of radiotherapy. Patients not responding to radiotherapy are not necessarily uniform in their grade of malignancy or age, nor in their resolve to fight the disease. Thus, additional caution is required at the time of drawing conclusions in such cases.

### *Volunteers*

Early **phases of clinical trials** are often undertaken on volunteers. They self-select themselves as the subjects. Volunteers tend to be very different from the general class of subjects. Many of them are either hopeless terminal cases or people with exceptional courage, either of which could affect the response. The results, thus, are not applicable to the general class of patients. Nevertheless, volunteer studies have a definite place in phase I trials as they do provide important clues on the toxicity of the regimen under test, the dose level that can be tolerated, and the potential for further testing of the modality.

Even for nonvolunteers, medical ethics requires that the subject's consent be taken for participation in a study. Real consent, after fully explaining the underlying uncertainties, is difficult to obtain without an inducement. Again, the subjects self-select themselves by providing or refusing consent, and some bias is likely. Some of this is eliminated by random allocation. This is feasible for **clinical trials** but not, for example, for cross-sectional **surveys**. Even in clinical trials, the generalizability suffers due to self-selection of subjects because of informed consent. Conclusions based on such studies can be fallacious for the general class of patients, although such subjects do provide evidence of efficacy of the regimen. Not many professionals appreciate this limitation imposed by consent-based selection, and they wonder later why their results cannot be reproduced in practical situations.

Similar problems, although of less severity, arise with surveys based on mailed questionnaires, subjects listed in a telephone directory, and Internet users. None of these sources are adequate for an unbiased sample of the general population. Because of the high nonresponse rate in many such surveys, the results are rarely valid even for the restricted class they represent.

The solution in all these cases is proper adjustment of the results, although this is not straightforward. The adjustment can be done only when relevant information is available on at least some subjects who are truly representative of the target population. In the absence of this, the alternative is to use the results only for the restricted population that these subjects represent. The third approach is to interpret the results as indicative of what *might* be occurring in the total population. In the case of nonresponse by some subjects after inclusion, you would most likely have their basic information, such as their demographics. If so, identify the type of subjects who dropped out, and use this information to adjust the results.

### *Clinic Subjects*

Clinic and hospital subjects also form a biased sample, as they tend to have a more severe form of disease and, in some countries, include mostly those who can afford these services. Mild cases tend to ignore their condition or self-treat themselves. An interesting example is migraine, which was once believed to be more common in the intelligent professional class [1]. An epidemiological study on a random sample could not substantiate such a relationship when the role of **confounding factors** was eliminated. Those in the more intelligent professional class perhaps seek medical assistance in the early phases of the disease and with greater frequency than their nonprofessional counterparts. Despite such a limitation, clinic-based studies do give important information on the presenting symptoms, their correspondence with laboratory and radiological findings, response

to various therapeutic procedures, prognostic features, etc. But the results are seldom applicable to the types of cases that do not show up in clinics. It is sometimes believed that consecutive cases coming to a clinic would be free from bias. While such cases could be true representatives of clinic subjects, the general bias in all clinic-based subjects still remains.

A similar bias occurs when the sample is restricted to people in employment. Such persons are healthier than the general population. For this reason, results based on employed people cannot be generalized to the other segments of the population.

A **nonrandom sample** tends to be biased anyway, but **random samples** can also be biased. For example, **systematic random sampling** can yield a biased sample if periodicity or trend is hidden in the subjects from 1 to *N*. As in the case of **simple random sampling**, by chance, this method can fail to give adequate representation to some specific groups of interest, particularly when the sample size is small A random sample, irrespective of the method used, tends to be unbiased when it is large. It is often mistakenly thought that random sampling necessarily produces a representative sample. Although that becomes increasingly likely as the sample size increases, the very nature of a random process implies that anything is equally likely—thus, even a very large random sample can be theoretically nonrepresentative.

1. Hill AB. *A Short Textbook of Medical Statistics*. The English Language Book Society, London, 1977: p. 261.

## bias in literature review

Many medical studies depend on what published literature says. Few realize that the literature itself could be biased—**publication bias** is prominent and well acknowledged. In addition, the list of references provided at the end of most research articles can also be biased. Researchers have a natural tendency to quote only that part of the literature that tends to support their findings—perhaps even at the cost of the credibility of the study. This is done to create an impression that their research is more persuasive. The next researcher then also tends to review the same literature and thus gets the same biased view. The opposite or indifferent view tends to be ignored or overlooked unless that also is a dominant view and is easily traceable. This bias can be avoided by including those articles that challenge your thesis. Research must remain a relentless search for truth instead of advancing a particular view. Thus, all views must be represented in the review of literature. Recent literature, as opposed to old literature, is likely to be more representative of the current knowledge.

Second is the language bias that could result in a biased review of literature. This occurs because the literature in languages other than that of the researcher is rarely included in the review. This literature can provide a new perspective but is mostly ignored.

Third, bias can occur when the literature search by the authors is based on certain specific terms. This search will find only a particular type of literature, and the search may therefore not be comprehensive or representative, something that is often unintentional. To avoid this bias, the search terms should be chosen such that divergent findings can be represented. The assistance of a specialist librarian experienced in literature search may be beneficial.

In **meta-analysis**, it is a regular practice to filter out articles that do not meet specified criteria. This is necessary also so that only credible and relevant articles are picked up. This would occur, for example, when articles following a particular methodology are included and others excluded. There is a distinct risk in this process that some articles with variant findings are not included. If so, the meta-analysis can give biased findings.



## bias in medical studies and their minimization, *see also* bias (statistical)

The results of a medical study become clouded when some bias is detected after the results are available. Therefore, it is important that all sources of bias are considered while conducting a study and all efforts are made to avoid or control them.

### *List of Biases in Medical Studies*

Since bias can occur at every step of research (planning, execution, analysis, and reporting), a large number of sources of bias exist in a medical research setup. The following list describes the most common and most important. These are not **mutually exclusive**. In fact, the overlap is substantial. Also, some of the biases in this list are collections of biases of a similar type. If all these are stated separately, the list may become unmanageable. Biases are described in brief here in the order in which studies are done, i.e., planning, execution, analysis, and reporting.

1. *Bias in concepts:* This bias occurs when there is a lack of clarity about the concepts to be used in the proposed study. Lack of clarity gives an opportunity to the investigators to use subjective interpretation that can vary from person to person. This bias can also occur when the logic used is faulty: sometimes, the premise of the logic itself can be incorrect. For example, it is generally believed that lower values of body mass index and blood pressure are better for health. In fact, very low values of these are also associated with increased morbidity and mortality. Ignoring stress as a factor in causation of a disease such as diabetes just because stress is so difficult to measure also comes under this kind of bias.
2. *Definition bias:* The subjects and medical condition under study should be sharply defined so that there is no room for ambiguity. For example, a study on tuberculosis cases should specify that these are be sputum positive, Mantoux positive, radiologically established, or some combination. A blurred definition is open to subjective interpretation, and this can affect the **validity** of the study.
3. *Bias in design:* This bias occurs when the **case** group and the **control** group are not equivalent at **baseline** and differentials in prognostic factors are not properly accounted for at the time of analysis. Design bias also depends on the structure of the study. For example, an **ecological study** has more chance of bias in results than a double-blind **randomized controlled trial** (see **bias pyramid**). **Random allocation** and large sample size tend to minimize this type of bias.
4. *Bias in the selection of subjects:* The subjects included in a study may not truly represent the target population. This can happen when the sample is not randomly selected or when the sample size is too small. In both these instances, the sample may fail to represent the entire spectrum of subjects in the target population. Studies on volunteers and clinic subjects always have this kind of bias. Indrayan [1] has discussed this at length. Selection bias can also occur because the serious cases have already died and are

not available with the same frequency as the mild cases (also called **survival bias**). For example, Bidzan et al. [2] had an initial group of 158 older patients with mild cognitive impairment, but conclusions were drawn on 52 who were available at follow-up 5 years later. This is obviously biased. A similar bias can occur in selection of cases with diseases that have highly variable incubation periods, such as AIDS. (See also *length bias*.). Bias due to self-selection or volunteers is obvious since those not consenting are excluded.

5. *Bias due to concomitant medication or concurrent disease:* Selected patients may suffer from another apparently unrelated condition, but their response might differ either because of the condition itself or because of medication given concurrently for that condition. This can be controlled by appropriate restrictions on concurrent disease in the **inclusion and exclusion criteria** and by strict restriction of concurrent medications where feasible.
6. *Instruction bias:* When there are no instructions or when unclear instructions are prepared for conducting a study, the investigators use discretion, and this can vary from person to person and from time to time. Modern clinical trial protocols and associated guidelines usually give very clear and detailed instructions regarding trial conduct.
7. *Length bias:* C**ase–control studies** are generally based on prevalent cases rather than incident cases. **Prevalence** is dominated by those who survive for a longer duration. And these patients may be qualitatively different from those who die early. Thus, the sample would include disproportionately more of those who are healthier and survive longer. The conclusions cannot be generalized to those who died earlier. The disease profile can also differ since there would be more cases in whom disease progression is slow. Those suffering from an aggressive form of the disease would be missed because of the rapid progression of the disease and early death.
8. *Bias in detection of cases:* Error can occur in diagnostic or screening criteria. For example, in a prostate cancer detection study, if prostate biopsies are not performed on men with normal results after screening, the true **sensitivity** and **specificity** of the test cannot be determined. Also, a laboratory investigation done properly in a hospital setting is less prone to error in the detection of cases compared to one carried out in a field setting. Detection bias also occurs when cases with mild disease do not report or are difficult to detect. If this is inadvertent, the results would be biased without anybody knowing that such a bias was present.
9. *Lead-time bias:* Not all cases are detected at the same stage of the disease. With regard to cancers, some may be detected by screening before they are clinically apparent, for example, by Pap smear, whereas some may not be detected until clinical manifestation of the disease starts appearing. In some cases, detection occurs when the disease is already at an advanced stage. But the follow-up is generally from the time of detection. This difference in "lead time" can cause systematic error in the results. These issues can be partly addressed by strictly defining inclusion criteria in terms of stage of disease at detection and/or by stratifying subjects according to the stage of disease at detection.
10. *Bias due to confounder:* This bias occurs when **confounders** are not dealt with adequately. If that happens, any difference or **association** cannot be fully ascribed to the **antecedent factors** under study. For this, identify the confounders after a thorough review of the literature and devise a strategy to deal with them.
11. *Bias due to epistemic factors:* Efforts can be made to control only those factors that are known. But there may be many unknown factors that could affect the results. These **epistemic** factors can bias the results in very unpredictable ways. There is nothing specific that can be done for the unknown factors, but randomization and adequate sample size seem helpful for minimizing this type of bias.
12. *Contamination in controls:* Control subjects are generally those that receive placebo or regular therapy. If these subjects are in their homes, it is difficult to know if they have received some other therapy that can affect their status as controls. In the prostate cancer detection project reported by Concato et al. [3], the control subjects are those who were under routine care. But some of these may have been screened outside the study and treated. Thus, their survival rate would not be sufficiently "pure" to be compared with the survival of those who were screened by the test procedures. In a field situation, contamination in a control group can occur if the control group is in close proximity to an unblinded test group and learns from the experience of the latter. In another situation, a neighboring area may not be the test area of the research, but some other program may be going on there that has a spillover effect on the control area.
13. *Berkson bias:* Comparison of hospital cases with hospital controls can be biased if exposure increases the chance of admission. This can lead to an overrepresentation of subjects with that exposure. Cases of injury in motor vehicle accidents can suffer from this kind of bias. See also **Berkson bias**.
14. *Bias in ascertainment or assessment:* This bias occurs in unblended studies when unknowingly or deliberately, the investigators are more thorough with cases than with controls. A similar problem can also occur when subjects belonging to a particular social group have records but others have to depend on recall. Sometimes, this is also called *information bias.*
15. *Interviewer bias or observer bias:* Interviewer bias occurs when one is able to elicit better responses from one group of patients (say, those who are better educated) relative to another (such as illiterates). Observer bias occurs when the observer unwittingly (or even intentionally) exercises more care about one type of response or measurement than others, for instance, those supporting a particular hypothesis versus those opposing the hypothesis. Observer bias can also occur if, for example, the observer is not fully alert when listening to Korotkoff sounds while measuring blood pressure or not able to properly rotate the endoscope to get an all-around view of, say, the duodenum in a suspected case of peptic ulcer.
16. *Instrument bias:* This occurs when the measuring instrument is not properly calibrated. For example, a weighing machine that does not show zero when resting will give an inaccurate measurement of weight. Another kind of instrument bias occurs when a device does not provide a complete picture of the target organ, thereby giving false or incomplete information. For example, an endoscope

might not reach the site of interest. Another example is **Likert scale** assessment, where +3 may be a more frequent response on a scale of −5 to +5 than +8 on a scale of 0 to 10, although both are same. A third kind of instrument bias occurs when an instrument is considered the **gold standard** because this is acknowledged as the best while forgetting that the best may still be imperfect. Errors can occur even with this "gold." Fourth is when **predictivity** of an instrument is used in a new setup without considering that different prevalence in this new setup would affect the predictivity.

17. *Hawthorne effect:* When subjects know that they are being observed or being investigated, they often alter their behavior and response. In fact, this is the basis for including a **placebo** group in a trial. The usual responses of subjects are not the same as when they are under a scanner. Such effects should not cause the results of a trial to be biased in favor of one treatment, as both the groups are under observation. See also **Hawthorne effect**.
18. *Recall bias:* There are two types of recall bias. One arises from better recall of recent events than those that occurred a long time ago. Also, serious episodes are easier to recall than mild episodes. The second type of recall bias occurs when cases suffering from a disease are able to recall events much more easily than healthy controls. A thorough probing can possibly help to reduce this bias.
19. *Response bias:* Cases with serious illness are likely to give more correct responses regarding history and current ailments compared with controls. This is not just because of recall but also because patients with serious illness tend to keep meticulous records. Some patients, such as those suffering from sexually transmitted diseases (STDs), may intentionally suppress sexual history and other information because of the stigma attached to these diseases. Injury history may be distorted to avoid legal consequences. If the subjects are able to exchange notes, their responses to questions might alter and, in some cases, might even become nearly uniform. An unexpected illness, death in the family, or any such drastic event may produce an extreme response. Response bias can also be regarded as a type of *information bias*.
20. *Bias due to protocol violation:* It is not uncommon in a clinical trial that some subjects do not receive the full intervention or the correct intervention, or some ineligible subjects are randomly allocated in error. This occurs when the study **protocol** is not faithfully followed, something that can bias the results.
21. *Repeat testing bias:* In a pretest–posttest situation, the subjects tend to remember some of the previous questions, and they may no longer commit previous errors in the posttest—thus doing better for reasons other than the intervention. The observer may acquire expertise to elicit the correct response on the second or third occasion. Conversely, fatigue may set in with repeat testing, which could alter the response. Moreover, many biological measurements have a strong tendency toward the mean (see **regression to the mean**): extremely high scorers tend to score lower in subsequent testing, and extremely low scorers tend to do better in a subsequent test, whereas midrange scores remain similar.
22. *Clustering bias:* This is related to repeat testing bias. When the subjects belong to an affinity group, they tend to give a similar response. For example, people in one profession, people living close together, family members, etc. are such affinity groups. When members of such groups are study subjects by design, use **design effect** to minimize the effect of clustering. If clustering occurs by chance, you may not even know that it is there, and the results would be biased.
23. *Midcourse bias:* Sometimes, after enrollment, subjects have to be excluded if they develop an unrelated condition such as an injury, or become so seriously ill that their continuation in the trial is no longer in their interest. If a new facility such as a health center is started or closed for the population being observed for a study, the response may alter. If two independent trials are going on in the same population, one may contaminate the other. An unexpected event such as a disease outbreak can alter the response of even those who are not affected.
24. *Self-improvement effect:* Many diseases are self-limiting. Improvement over time occurs irrespective of the intervention, and it may be partially or fully incorrectly ascribed to the intervention. Diseases such as arthritis and asthma have natural periods of remission that may look like the effect of therapy. The use of proper controls and random allocation may address this source of bias.
25. *Bias due to digit preference:* It is well known that most people have a special love for the digits 0 and 5. Measurements are more frequently recorded ending with these digits. A person aged 69 or 71 is likely to report his/her age as 70 years. Another manifestation of digit preference is in forming intervals for quantitative data. Blood glucose level categories would be commonly chosen as 70–79, 80–89, 90–99, etc., and not 64–71, 72–79, etc. If the digit 0 is preferred, 88, 89, 90, 91, and 92 can be recorded as 90. Thus, intervals such as 88–92, 93–97, and 98–102, may be better to ameliorate the effect of digit preference rather than the conventional 85–89, 90–94, 95–99, etc. See also **digit preference**.
26. *Bias due to nonresponse:* In most medical studies, particularly those requiring follow-up, some subjects refuse to cooperate, suffer an injury, die, or become untraceable. Nonrespondents have two types of effects on the results. First, they are generally different from those who respond, and their exclusion can lead to biased results. Second, nonresponse reduces the sample size and can result in substantial differences between the numbers in different groups, both of which can decrease the **power** of the study to detect specified differences or **associations**.
27. *Attrition bias:* The pattern of nonresponse can differ between subgroups in the sense that in one subgroup, more severe cases drop out, whereas in another group, mostly mild cases drop out. In a rheumatoid arthritis databank study, attrition during follow-up was high in patients of young age, who were less educated, and were non-whites [4]. Everything possible should be done to convince the subjects to respond.
28. *Bias in handling outliers:* No objective rule can define a value as an outlier other than that the value must be far away from the mainstream values. If the duration of hospital stay after a particular surgery is mostly between 6 and 10 days, some researchers would call 18 days an outlier and may exclude it on the suspicion of it being an incorrect recording or transcription, and some would consider it correct and include it in their calculation. In view of this subjective and essentially arbitrary definition of an outlier,

many would not exclude any extreme value, howsoever different it might be. Thus, the results would vary depending on a researcher's approach to detecting and handling outliers. Generally speaking, you should not exclude a suspected outlier from analysis unless there is convincing reason to label it as an outlier. When an analysis with such exclusions is performed, an analysis without any exclusions should also be normally performed.

29. *Recording bias:* At least two types of errors can occur in recording data. The first arises due to the inability to properly decipher the writing on case sheets, particularly since physicians are notorious for illegible writing. This can happen particularly with similar-looking digits such as 1 and 7, and 3 and 5. Thus, the entry of data may be in error. The second arises due to the carelessness of the investigator. A diastolic level of 87 can be wrongly recorded as 78, or a code 4 entered as 6 when remembered incorrectly. Some errors may also be typographical since wrongly pressing adjacent keys on the keyboard is not uncommon. Errors can also occur when data are manually transcribed from one document to another.
30. *Bias in analysis:* See **bias (statistical)**.
31. *Bias due to competing cofactors:* Some factors influence results **synergistic**ally or **antagonistic**ally when relevant cofactors are present. If this is not properly taken into account, the effect of an intervention can be underestimated or overestimated. It is therefore important to identify as many such factors as possible and take them into account during analysis.
32. *Prevalence–incidence bias:* This occurs when effects of risk factors on prevalence could be a function of the duration of the disease and can be mistaken for effects on disease occurrence when not properly accounted for.
33. *Interpretation bias:* This arises from the tendency among some researchers to interpret the results in favor of a particular hypothesis, ignoring the opposite evidence. This would be mostly unintentional but can also be intentional in rare cases.
34. *Reporting bias:* Researchers may have conscious or subconscious preconceptions and expectations of results, and this may result in writing reports that are not really consistent with the data available. For example, it is easy to suppress contradictory evidence by not talking about it.
35. *Bias in presentation of results:* Scales in graphs can be chosen such that a small change looks like a big change or vice versa [1]. Also, the researcher may merely state the inconvenient findings that contradict the main conclusion but not highlight them in the same way as done with the findings that support the main conclusion.

For more details, see Delgado-Rodriguez and Llorca [5].

### *Steps for Minimizing Bias in Medical Studies*

The purpose of describing various types of biases in so much detail is to create awareness of the need to avoid or at least minimize them. Everything possible should be done to keep them under control so that you do not have to give an explanation such as "hand of God" in a soccer game. In the context of a medical study, bias in results and conclusions is a function of the study design and its implementation. The following steps can be suggested to minimize bias in the results in a research setup. Not all steps apply to all the situations. Adopt the ones that apply to your setup. Details of some of these steps are given by Indrayan [1].

1. Develop an unbiased scientific temperament by realizing that you are in the occupation of a relentless search for truth.
2. Specify the problem and the objectives to the minutest detail.
3. Assess the validity of the identified **target population** and the groups to be included in the study in the context of objectives and the methodology. The inclusion and exclusion criteria should be precisely worded to address this problem.
4. Assess the validity of antecedents and outcomes for providing the correct answer to your questions.
5. Beware of epistemic **uncertainties** arising from the limitation of scientific knowledge. The use of double-blind randomized controlled trials goes a long way toward minimize this bias.
6. Evaluate the **reliability** and **validity** of the measurements required to assess the antecedents and outcomes, as well as of the other tools you plan to deploy.
7. Where appropriate, consider undertaking a **pilot study** and pretest the tools. Make changes as needed.
8. Try to identify all possible confounding factors and other sources of bias, and develop an appropriate design that can take care of most of these biases, if not all.
9. Choose a representative sample, preferably by a random method as and when possible.
10. Choose an adequate size of the sample in each group.
11. Rigorously train yourself and your coworkers in making correct assessments prior to the beginning of a study.
12. Use **matching**, **blinding**, **masking**, and **random allocation** as needed.
13. Monitor each stage of research, including periodic checking of data.
14. Make determined efforts to minimize nonresponse and partial response.
15. Double-check the data and rectify errors in recording, entries, etc.
16. Analyze the data with proper statistical methods. Use **standardized** or **adjusted rates** where needed, perform **stratified analysis**, or use **mathematical models** such as **regression** to take care of those confounders that could not be ruled out by design.
17. Interpret the results in an objective manner, based only on the evidence at hand.
18. Report only the **evidence-based** results, enthusiastically but dispassionately.
19. Exercise extreme care in drafting the report and keep comments or opinions clearly separate from evidence-based results.

Bias and other aspects of design can be very adequately taken care of if you imagine yourself presenting the results a couple of years hence to a critical but friendly audience [6]. Consider what your colleagues would question or advise at that time, and their reaction when you conclude that the results are significant or if you conclude that the results are not significant: Might there be alternative explanations of the results? Are there any confounding factors that have been missed? Or could **chance** or **sampling error** be an explanation? Such considerations will help you to develop a proper design and to conduct the study in a conscientious manner. A word of caution in the end. Do not try to control or eliminate all the sources of bias at the time of planning. This may affect the proper execution of the study and practical utility of results. Control the effect of some sources at the time of analysis. Others that remain can be acknowledged as limitations of the study. This is appreciated much more than the claim of completely unbiased results.

1. Indrayan A. *Medical Biostatistics*, Third Edition. CRC Press, Boca Raton, FL, 2012.
2. Bidzan L, Pachalska M, Bidzan M. Predictors of clinical outcome in MCI. *Med Sci Monit* 2007; 13:CR398–405. http://www.ncbi.nlm.nih.gov/pubmed/17767119
3. Concato J, Peduzzi P, Kamina A, Horwitz RI. A nested case–control study of the effectiveness of screening for prostate cancer: Research design. *J Clin Epidemiol* 2001;54:558–64. http://www.ncbi.nlm.nih.gov/pubmed/11377115
4. Krishnan E, Murtagh K, Bruce B, Cline D, Singh G, Fries JF. Attrition bias in rheumatoid arthritis databanks: A case study of 6346 patients in 11 databanks and 65,649 administrations of the Health Assessment Questionnaire. *J Rheumatol* 2004;31:1320–6. http://www.ncbi.nlm.nih.gov/pubmed/15229950
5. Delgado-Rodrıguez M, Llorca J. Bias. *J Epidemiol Community Health* 2004;58:635–41. http://jech.bmj.com/content/58/8/635.full
6. Elwood M. Forward projection—Using critical appraisal in the design of studies. *Int J Epidemiol* 2002;31:1071–3. http://ije.oxfordjournals.org/content/31/5/1071.full.pdf

## bias in publications, *see* publication bias

## bias pyramid, *see* evidence (levels of)

## bias (statistical), *see also* biased sample, unbiased estimator

Statistical bias occurs when an inappropriate statistical procedure is used. This could be in the calculation of sample size, sampling (see **biased sample**), statistical estimation (see **unbiased estimate**), model building, or any other method of statistical analysis. If an inappropriate method is used, the results may not reflect the true state of affairs and can lead to wrong conclusions.

Amongst many types of statistical bias, some are described elsewhere in this volume, as indicated in bold in the previous paragraph. What follows is a description of two other types of statistical bias.

1. *Bias in analysis:* This again can be of two types. The first occurs when gearing the analysis to support a particular **hypothesis**. When comparing *pre-* and *post*-values, for example, hemoglobin (Hb) levels before and after weekly supplementation of iron, the average increase may be so small that it will not be detected by a comparison of the means. But it may be detected when evaluated as a proportion of subjects with levels <10 g/dL before and after iron supplementation. This will happen when the rise in Hb level is small in many subjects but crosses the threshold of 10 g/dL. The second bias in analysis can arise due to differential interpretation of ***P*-values**. When $P = 0.055$, one researcher may refuse to say that it is significant at a 0.05 level, and another may say that it is marginally significant. Some researchers may change the **level of significance** from 5% to 10% after having seen the results, although the level of significance should be decided prior to the analysis.
2. *Bias due to lack of power:* **Statistical tests** are almost invariably used to check the **significance** of differences or **associations**. The statistical **power** of these tests to detect a specified difference or association depends to a large extent on the number of subjects included in the study and the variability of the data. If the study is conducted on a small sample, even a big difference cannot be detected, leading to a false-negative conclusion. When conducted on an appropriate number of subjects, the conclusion can change.



## bidirectional dependence

Two variables $x$ and $y$ are considered to be bidirectionally dependent when $y$ depends on $x$ and $x$ depends on $y$—for example, health depends on exercise, and exercise depends on health. The general statistical setup is that one or more variables are **dependent** on one or more **independent variables**. Statistical analysis of data in this setup is relatively easy since independent variables can be considered fixed in most situations, such as in **regression**. However, rarely, the dependence is bidirectional. Data with bidirectional dependence are not easy to analyze. For example, this may require a *two-stage regression* procedure. The details are too complex to be described in this book. Those interested may refer to Faries et al. [1].

1. Faries D, Leon AC, Haro JM, Obenchain RL. *Analysis of Observational Health Care Data Using SAS*. SAS Publishing, 2010.

## bihistogram

*Bihistogram* is the term used for a figure that has two **histograms** placed opposite to one another for comparison. Sometimes, they can be placed side by side also. Figure B.4 is a bihistogram where histograms for diastolic levels of blood pressure (BP) for hypertensive persons and controls are compared.

Histograms show the distribution of a quantitative variable whose values are divided into categories. For example, in Figure B.4, the diastolic BP categories (in mmHg) could be 70–74, 75–79, 80–84, etc. In the case of nonobese persons, the values start from 70 mmHg in this bihistogram, whereas they start from 85 mmHg for obese people. The peak (highest frequency) for obese people is at 100–104 mmHg but is at 85–89 mmHg for nonobese persons. The comparison shows that the two distributions are nearly the same except that there is a shift of nearly 15 mmHg for the obese group. Thus, the bihistogram provides an effective comparison of the distribution in the two groups.

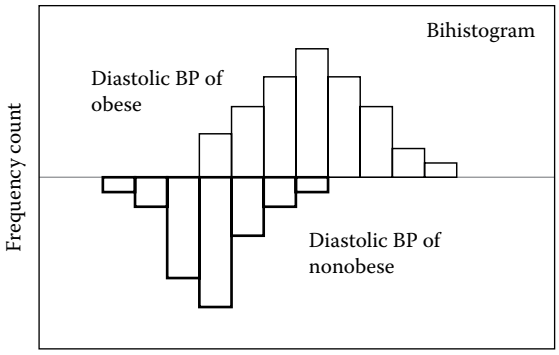



**FIGURE B.4** Bihistogram of diastolic BP in obese and nonobese subjects.

## bikini syndrome, *see* confidence intervals (the concept of)

B

## bimodal distribution

**Mode** is the most common value in a set of data—a definition that assumes that only one value can be the most common. Sometimes, you can have two or more modes. Consider the following data on duration of immobility in days in cases of acute polymyositis of the back in 38 women.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | 5 | 9 | 7 | 36 | 4 | 6 | 7 | 5 | 8 | 3 | 6 | 5 |
| 7 | 8 | 10 | 7 | 14 | 10 | 9 | 4 | 6 | 11 | 9 | 6 | 5 |
| 8 | 8 | 6 | 7 | 5 | 5 | 12 | 3 | 5 | 9 | 10 | 7 | |

Mode = 5 days and 7 days, occurring in 7 patients each

A distribution containing two modes such as in this example is called a bimodal distribution. In this example, the maximum frequencies are equal, but that is not a requirement for a distribution to be bimodal. Mode essentially means a peak in a distribution. You can have many peaks in one distribution—and thus many modes. When you plot a **frequency curve**, the values around the mode have less frequency, and the peak stands out. The age distribution of Hodgkin disease and leukemia is bimodal with one (smaller) peak around 20 years and the other (bigger) peak at around 60 years. Figure B.5 describes a similar bimodal distribution of Crohn disease. A bimodal distribution can also arise with a mixture of two separate distributions, such as of height when men and women are mixed. In a figure in the topic **normal range of medical parameters**, diseased and healthy cases are mixed, giving rise to two modes—one for healthy subjects and the other for nonhealthy subjects.

Sometimes, sample values appear to have two or more modes due to **sampling fluctuations** rather than an actual bimodal distribution. In most situations, a biological explanation will be available, as in the example we have cited. When there are two modes and they have unequal peaks as in Figure B.5, the value with a higher mode is called a *major mode*, and the other, a *minor mode*. When two or more modes are really present, mean and median may not be adequately representative of the **central value**. Such a distribution is not **Gaussian** (normal), and many statistical tests such as **Student *t*** and **ANOVA** *F* may fail. Instead, **nonparametric methods** should be used for data values that have bimodal distribution. Most bimodal, and for that matter, multimodal, distributions can be considered a mixture of two or more unimodal distributions as just described and can be analyzed accordingly.

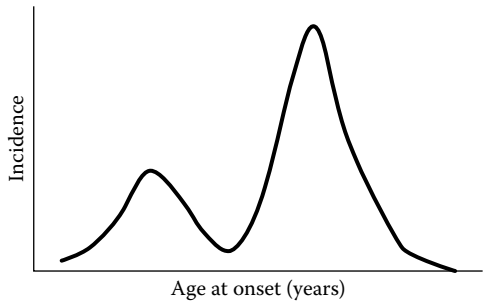



**FIGURE B.5** Age distribution of onset of Crohn disease: a bimodal distribution.

## binary dependent/outcome/predictor/response/variable

When the assessment of a characteristic is in terms of only two categories, such as yes/no, present/absent, or favorable/unfavorable, these are called **dichotomous categories.** The corresponding variable is called a binary variable. Recording gender as male or female is an obvious example. You can have a binary predictor or a binary **independent variable** such as in a **regression models**, or you can have a binary outcome, response, or **dependent variable**. The occurrence of malaria dependent partly on the density of female anopheles mosquitoes in an area and the diagnosis of primary biliary cirrhosis dependent on elevated serum alkaline phosphatase levels are examples in which the dependent variable (presence or absence of disease) is binary. In the context of outcome, such as cured/not cured, this is also called a **quantal response**. A **continuous variable** can also be made binary, such as with diastolic blood pressure <90 mmHg or ≥90 mmHg, although this amounts to loss of information.

A binary outcome in *n* independent trials (such as *n* subjects) leads to **binomial distribution**. In all such dichotomous situations, the binary variable *y* is given a value of 0 for a negative response or 1 for a positive response, with no other possibilities. The probability of a positive response is $P(y = 1)$ and is denoted by $\pi$ for population. This is estimated by the corresponding proportion in the sample, denoted by *p*.

**Logistic regression** is the most appropriate statistical method for binary response when the objective is to predict the response with a set of predictors or to explain it with a set of independent variables. In this case, the dependent variable is actually the proportion or probability of subjects with a positive response. Even though the **Cox model** is for **hazard**, the dependent variable in this case is also binary, such as survived/dead. **Discriminant functions** can also be used for binary outcomes. The concept of **number needed to treat (NNT)** is especially useful for comparing two treatment regimens with a binary response.

In the case of assessment of subjects on two binary variables, the results can be studied using a $2 \times 2$ table. This can be analyzed in a variety of ways, as mentioned under **two-by-two table**. There could be a $2 \times K$ table, where one variable is binary and the other has *K* categories. An example is distribution of male and female cases of glaucoma by blood group. Such tables are generally analyzed by using a **chi-square** test.

## binomial distribution/probability

The binomial distribution arises for "successes" in binary outcomes when (i) there are *n* independent trials (occurrence in one does not affect the outcome in any other trial) and (ii) the probability of success remains the same for each trial. For example, if 20 males of age 60 years and above are randomly selected, the number of males with an enlarged prostate out of 20 will have a binomial distribution. In this case, "success" is having an enlarged prostate, and *n* is 60. Note that a person with an enlarged prostate in the sample does not affect the chance of any other person in the sample of having or not having an enlarged prostate. The outcome in different persons is independent. Also, if all the persons in the sample are from the same milieu, the chance of one with an enlarged prostate is the same as that of anyone else in the sample.

A prominent application of binomial distribution is illustrated in the following example. Consider the 5-year survival among patients with cervical cancer. There is no noncancer or any other group for comparison. Two kinds of statistical questions can arise in this case.

(i) If the proportion surviving for 5 years among patients with cervical cancer is known to be 30%, what is the chance that at least 6 will survive for at least 5 years in a random sample of 10 patients? (ii) If the number surviving in a random sample of 20 patients is only 4, how likely is that the survival rate in the long run could be 30%? These two questions are, in fact, two ways of looking at the same statistical issue. The answer to these questions is obtained by the binomial distribution, as described next.

### *Binomial Distribution*

For simplicity, let us call the event of interest a success. Denote its probability by $\pi$. In the preceding example, the event of interest is survival for 5 years, and $\pi = 0.3$. It can be shown by the multiplication **law of probability** that the probability of $x$ successes in $n$ independent trials is given by

$$\text{binomial distribution: } P(x) = {}^{n}C_{x}\pi^{x}(1-\pi)^{n-x},$$

where ${}^{n}Cx = \dfrac{n!}{x!(n-x)!}$ and $x! = x(x-1)(x-2)\ldots 3.2.1$ (for example, ${}^{5}\mathrm{C}_{3} = \dfrac{5!}{3!(5-3)!} = \dfrac{5\times 4\times 3!}{3!\times(2\times 1)} = 10$), and $\pi$ is the probability of success in one trial. In our example, a trial corresponds to one patient. Patients behave independently in the sense that the survival of one patient does not affect the chance of survival of another. Also, the chance of survival for each patient should be the same. This would be so when the patients are homogeneous with respect to prognostic factors. When these conditions are fulfilled, this distribution can be used to answer the two questions earlier posed.

### *Binomial Probability*

In question (i) above, $n = 10$, and $\pi = 0.3$. The success in this case is survival. You need to find $P(x \geq 6)$. Because not more than 10 successes are possible in 10 patients and the successes are **mutually exclusive**, by addition law of probability,

$$\begin{aligned} P(x \geq 6) &= P(x=6) + \mathrm{P}(x=7) + \ldots + P(x=10) \\ &= {}^{10}C_{6}(0.3)^{6}(0.7)^{4} + {}^{10}C_{7}(0.3)^{7}(0.7)^{3} + \ldots + {}^{10}C_{10}(0.3)^{10}(0.7)^{0} \end{aligned}$$

and, from the equation mentioned earlier,

$$= 0.0368 + 0.0090 + 0.0014 + 0.0001 + 0.0000 = 0.047.$$

Thus, the chance that at least 6 will survive for 5 years in a sample of 10 patients is only 4.7%, which may seem surprisingly low.

Question (ii) is more appropriately answered by obtaining the probability of $x = 4$ or a more extreme value, i.e., $P(x \leq 4)$. Because of mutually exclusive values of $x$, for $\pi = 0.3$ and $n = 20$,

$$\begin{aligned} P(x \leq 4) &= P(x=0) + P(x=1) + P(x=2) + P(x=3) + P(x=4) \\ &= {}^{20}\mathrm{C}_{0}(0.3)^{0}(0.7)^{20} + {}^{20}\mathrm{C}_{1}(0.3)^{1}(0.7)^{19} + {}^{20}\mathrm{C}_{2}(0.3)^{2}(0.7)^{18} \\ &\quad + {}^{20}\mathrm{C}_{3}(0.3)^{3}(0.7)^{17} + {}^{20}\mathrm{C}_{4}(0.3)^{4}(0.7)^{16} \\ &= 0.0008 + 0.0068 + 0.0278 + 0.0716 + 0.1304 = 0.238. \end{aligned}$$

This probability is fairly high. Thus, it is not unlikely (probability is 0.238) that the survival rate in the long run is 30%.

Suppose for some reason that the interest is in finding the probability of at least two survivals out of $n = 10$. In notation, this is $P(x \geq 2)$. Calculating all the probabilities for $x = 2, 3, \ldots, 10$ by hand would take a lot of time and effort. In this situation, use the fact that $P(x \geq 2) = 1 - P(x \leq 1)$. Now this is to be computed only for $x = 0$ and $x = 1$. Similarly, for example, $P(x \leq 8) = 1 - P(x \geq 9)$.

### *Large* n*: Gaussian (Normal) Approximation to Binomial*

The calculation of the binomial probability can become complex and time consuming when $n$ is large and when it is to be computed for several different values of $x$. When the **Gaussian conditions** are satisfied, which is likely when $n$ is large, the binomial distribution can be approximated by a **Gaussian distribution**. This approximation arises from the **central limit theorem** as the binomial $x$ also is a summation type of variable—this time, the sum of 1's and 0's for success and failure, respectively. For Gaussian approximation of a binomial, the following are needed:

$$\text{mean of a binomial variable } x = n\pi, \text{ and } \mathrm{SD} = \sqrt{n\pi(1-\pi)}.$$

Thus, for large $n$,

$$z = \frac{x - n\pi}{\sqrt{n\pi(1-\pi)}}$$

is a standard Gaussian variate. When $n$ is large, this can be used to answer the same types of questions as posed earlier. This is illustrated in examples given as follows.

If the proportion surviving for at least 3 years among cases of cancer of the cervix is 60%, what is the chance that at least 40 will survive for 3 or more years in a random sample of 50 such patients?

For $n = 50$ and $\pi = 0.60$, mean number of survivors, $n\pi = 50 \times 0.60 = 30$, and

$$\mathrm{SD} = \sqrt{n\pi(1-\pi)} = \sqrt{50\times 0.60\times 0.40} = 3.464.$$

Since $n\pi \geq 8$ and $n(1-\pi) \geq 8$ also, the Gaussian conditions are satisfied, and we can use Gaussian approximation. With **continuity correction**, $P(x \geq 40) = P(x \geq 39.5)$. Thus,

$$\begin{aligned} P(x \geq 40) &= P\left(\frac{x - \text{mean}}{\mathrm{SD}} \geq \frac{39.5 - 30}{3.464}\right) \\ &= P(z \geq 2.74) \\ &= 0.0031 \text{ from Gaussian distribution} \end{aligned}$$

This low probability indicates that there is practically no chance that 40 or more patients will survive for at least 3 years in a sample of 50 when the survival rate is 60%. This might seem unbelievable, but it is true.

The probability just calculated is the approximate probability based on a Gaussian distribution. The exact binomial probability by binomial distribution for $x \geq 40$, when $n = 50$ and $\pi = 0.60$, is 0.0008.

## bioassays, *see also* parallel-line assays, slope-ratio assays, quantal assays

Assays are investigative procedures used for assessing the presence, amount, or activity of a substance. Most such assays are

B

based on chemical reactions between the test substance, the analyte, and various test reagents, where the outcome of the test is measured by several different methods. *Bioassays*, short for "biological assays," use the effect of the analyte on a living organism (animal, plant, bacteria, in vivo) or living tissues (e.g. cells, in vitro).

Conventional assays are experiments where we intentionally do something at different intensities to see what the effect is and compare it with the effect of a standard intervention at similar intensities. The effect of an intervention is technically called the response. In the case of assays, we apply the same intervention at different intensities in the hope that the response will be higher for the increased intensity. Many bioassays study such **dose–response relationships**. This particularly applies to poisons such as insecticides and pesticides. The larger the dose, the more insects will die. In this case, the response is death. In general, the response could be any other outcome. In the case of carcinogens, for example, an increased dose means a greater chance of cancer. The primary objective of assays is to estimate the **relative potency** of the test intervention compared with the standard intervention for achieving a particular outcome. Many times, the test preparation is a dilution or concentration of the standard preparation, leading to what is called *dilution assays*.

Statistically, there are basically two types of bioassays. **Quantal assay** is the term used when the response is binary, such as survive/die, relieved/not relieved, and cholesterol level high/within normal limits. These are particularly useful for estimating **median effective dose ($ED_{50}$)** or median lethal dose ($LD_{50}$) as the case may be. This is the dose at which 50% of the experimental units respond. The others are quantitative assays where the response is quantitative, such as time taken to respond, cholesterol level itself, and pain score. Death is a quantal response, but the time elapsed before death is quantitative. When the comparison is with a standard preparation, an assay with a quantitative response is called a quantitative assay.

Quantitative assays generally are parallel line or slope ratios depending on the pattern of response in the test preparation group and the standard preparation group. The basic setup in **parallel-line assays** is that the response in the test and the standard preparation differs by a fixed amount at each dose of the drug. In **slope-ratio assays**, the response is the same for both test and standard preparation at some baseline but increases (or decreases) at a faster rate for one preparation than the other as the dose increases. However, the trend in both should be linear. In slope-ratio assays, the estimate of the relative potency is given by the ratio of the slopes, hence the name. Quite often, the dose and the response are transformed, such as by logarithm to achieve linearity in the relationship. Such transformed dose and response are respectively called *dose metameter* and *response metameter*. When such transformations are used, the interpretation of the estimate of parameters such as relative potency needs special care. See the topic **logarithmic scale**.

Statistical methods for analyzing bioassays were initially described by Emmens [1] in his book published in 1948 but are generally attributed to Finney [2] because of his comprehensive book on this subject. See **parallel-line assays**, **slope-ratio assays**, and **quantal assays** for some details about analysis of data from different types of bioassays.

1. Emmens CW. *Principles of Biological Assay*. Chapman & Hall, 1948.
2. Finney DJ. *Statistical Method in Biological Assay*. Charles Griffin and Company, 1952.

## bioavailability, *see also* half-life of medications, area under the concentration curve

*Bioavailability* is the term used for the rate and extent to which an external substance is available at its intended organ or site in a biological system at different points in time after its administration. In the vast majority of cases, the concentration in bloodstream is measured in place of the site or organ, although exceptionally, concentration in other tissues or body fluids may be examined. This, in a way, measures the absorption or the concentration of the substance as a surrogate to its therapeutic potency. Bioavailability can be studied either in terms of amount, say in micrograms (or percentage of the amount ingested, such as 35%) or in terms of rate per unit of time, such as micrograms per hour. This can also be measured as micrograms per kilogram of body weight.

Bioavailability is considered an important pharmacological property of a substance and is assessed through **pharmacokinetic studies**, most commonly in phase I of a drug trial. This also is used to establish or refute bioequivalence of two substances (generic versus brand), two formulations (tablet versus capsule), two routes of administration (oral versus injectable), etc. As explained for **bioequivalence**, this equivalence is concluded when the bioavailability of the substances are same at different points in time. Bioequivalence has become a big issue since the 1980s, when generic drugs started competing with brand-name drugs, claiming that these provide the same response. Bioavailability is the anchor for bioequivalence studies.

Among various parameters used for assessing bioavailability, the most common is the plot of concentration versus time, called the concentration curve. The **area under the concentration curve (AUC curve)** provides a good measure of overall bioavailability of the drug, although there are deficiencies also with this measure, as explained under that topic. Among other important pharmacokinetic parameters used to assess bioavailability are **half-life**, maximum (peak) concentration ($C_{max}$), and time after administration to reach the peak value ($T_{max}$). See the topic **bioequivalence** in case you are interested to know more.

Usually, a **crossover design** is used to conduct comparative bioavailability studies. This means that a random half of the subjects receive the AB sequence, and the other half, the BA sequence. Each subject gets a washout period for the substance to exit the system either through metabolizing or through excretion. In case a carryover effect remains, one can think of other designs, such as the **Balaam design**, where the sequences administered are AA, AB, BA, and BB.

A large number of bioavailability studies are available in the literature. For example, Salem and Kuratko [1] compared bioavailability of two active substances in krill and fish oil, and also discussed some difficulties in such a comparison. Gaudreault et al. [2] evaluated the truncated area under the curve as a measure of the relative extent of bioavailability. This area is different from the area under the receiver operating characteristic (ROC) curve.

1. Salem N Jr, Kuratko CN. A reexamination of krill oil bioavailability studies. *Lipids Health Dis* 2014 Aug 26;13(1):137. http://www.lipidworld.com/content/pdf/1476-511X-13-137.pdf
2. Gaudreault J, Potvin D, Lavigne J, Lalonde RL. Truncated area under the curve as a measure of relative extent of bioavailability: Evaluation using experimental data and Monte Carlo simulations. *Pharm Res* 1998;15:1621–9. http://link.springer.com/article/10.1023/A%3A1011971620661#page-1

## bioequivalence, *see also* bioavailability, area under the concentration curve

In the context of treatment regimens, equivalence has two distinct dimensions—therapeutic equivalence and bioequivalence. Only the end point at a particular time point is considered for therapeutic equivalence, whereas bioequivalence considers the entire course of the regimen. When the course of the disease or the improvement pattern over a period of time is the same for two regimens, they are considered bioequivalent. **Bioavailability** is the main consideration that decides bioequivalence. Bioequivalence implies therapeutic equivalence, but the reverse is obviously not necessarily true.

Bioequivalence may be relevant in situations where the course of the disease or the mechanism working behind is important. A diuretic may be as effective in controlling blood pressure as a beta-blocker, but their mechanisms for reaching the same end point are different. In this case, they are therapeutically equivalent but not bioequivalent. In some specific patients, diuretics may be more effective than beta-blockers, and for that, the mechanism is important. Then the interest would be in bioequivalence. When the presentation (color, packaging, size, etc.) or mode of administration (daily/weekly, once a day/twice a day) or ingestion (oral/injectable, etc.) of a regimen is changed, the interest could be in either bioequivalence or therapeutic equivalence.

Bioequivalence requires **pharmacokinetic studies**, and the comparison may be in terms of peak concentration ($C_{max}$), time to reach its peak ($T_{max}$), **half-life**, **area under the concentration curve (AUC curve)**, etc. Regulatory agencies mostly consider only the AUC curve and $C_{max}$—$C_{max}$ being used as a measure of rate of delivery. $T_{max}$ is considered when there are specific claims of rapid delivery. The difficulty is that the area under the curve has limited physical meaning. When used in isolation, it may fail to provide an adequate assessment of the difference in trend, even on average, because it is neither specific nor sensitive to changes in patterns. Curves with markedly different trends can give the same area [see Figure A.5 in the topic **area under the concentration curve (AUC curve)**]. Also, as time passes, some patients drop out or get cured—thus, average response at different points in time is based on different $n$. This is ignored while studying AUC.

In a statistical test for bioequivalence of two pharmaceutical preparations, if the difference in concentration or any such parameter is less than a specified clinically unimportant margin, you can conclude that the groups are essentially equivalent even if the difference is statistically significant. Generally, the rule of 80/125 is followed for equivalence, which means that the difference between the two regimens should not exceed the 20% limit. Note that 20% of 125 is 25, which takes it back to 100, and 20% of 100 is 20, which reduces it to 80. This applies to confidence interval (CI). For example, the 90% CI for the ratio of these parameters should be between 0.80 and 1.25. Also note that most bioequivalence parameters are multiplicative instead of additive—thus, many times, log transformation is used. Thus, they are typically compared in their logarithmic form. This applies to concentration but does not apply to time-based parameters such as $T_{max}$ and half-life, since time is not multiplicative. For sample size and other details, see **equivalence tests**.

Let us reemphasize that a distinction should also be made between bioequivalence at the individual patient level and the average bioequivalence in groups of patients. If the two regimens under comparison produce different responses in individuals, this is ignored in average bioequivalence, whereas in individual bioequivalence, this interaction is an important consideration. Average bioequivalence would imply that either of these regimens can be prescribed to *newly* incoming patients. Individual bioequivalence would mean that the *existing* patient can be switched from one regimen to the other in the midst of the ongoing treatment. For example, drug switchability or interchangeability can be from brand name to generic while the patient is still under treatment, say for cost considerations.

Bioequivalence studies can be done on healthy subjects as well as for some regimens using the **crossover** strategy. In the case of drugs with a very long half-life, a parallel-group design may be used. Bioequivalence studies necessarily use a **repeated measures design**, because only then can the pharmacokinetics and the course of disease be studied. For statistical details of bioequivalence, consult Patterson and James [1].

1. Patterson S, James B. *Bioequivalence and Statistics in Clinical Pharmacology*. CRC Press, Boca Raton, FL, 2006.

## bioinformatics

Bioinformatics can be crudely understood as computer-based information system applications to biology. Most relevant of these for us are applications to health and disease in humans. Medical marvels, riding on bioinformatics, are poised to take center stage in our health care. The *Star Trek* fantasy of a doctor scanning the body with a handheld machine to get the particulars of a patient's condition looks to be within reach. Indrayan [1] visualized a pocket health monitor two decades ago, and now, there is a talk of a real scenario where your smartphone can turn into a personal doctor [2].

Bioinformatics is a rapidly evolving science facilitated by expanded computer efficiency. Computer technology has made it feasible to study the structure and characteristics of billions of molecules in various biological organisms, including human beings. The human genome itself is estimated to have 3 billion base pairs [3]. Bioinformatics is the science that tries to systematically organize the information about these billions of items so that useful information with regard to health and disease can be extracted. This also helps in identifying drugs that can be specifically targeted, relying on genomic sequencing.

A statistical component in bioinformatics is within- and between-subjects variations, which exist in all biological systems. Subjects tend to behave differently at different points in time, and different subjects have different profiles. This variation generates uncertainty, which can be studied only by statistical methods.

1. Indrayan A. A health monitor in your packet. *CSI Communications* 1996;20(1):8–9. http://www.researchgate.net/publication/249322843_A_health_monitor_in_your_pocket
2. Sabar A. Inside the technology that can turn your smartphone into a personal doctor: The fantastic tricorder device that "Bones" used to scan aliens on "Star Trek" is nearly at hand—In your cellphone. *Smithsonian Magazine* May 2014. http://www.smithsonianmag.com/innovation/inside-technology-can-turn-your-smartphone-personal-doctor-180951177/#m7zPr82rrs1wBUPd.99
3. What is bioinformatics? Bioplanet.com. http://www.bioplanet.com/what-is-bioinformatics/, last accessed July 21, 2015.

## biostatistics

Biostatistics comprises statistical methods that are used to manage some crucial aspects of uncertainties in the field of medicine and health, and the statistical methods in this are those that help to derive meaning from data by considering the underlying uncertainties and variations.

Medicine involves close interactions with the patient, and a large number of steps are usually taken before arriving at a treatment regimen. The patient's history is taken; measurements such as weight, blood pressure, and heart rate are recorded; physical examination is carried out; and a range of investigations are undertaken. In the course of these steps, a patient sometimes encounters many observers and many instruments. Variations among them contribute their share to the uncertainties seen in clinical practice. The assessment of diagnosis, treatment, and prognosis can all go wrong. **Medical uncertainties** are profound, and it is important to delineate them and contain their effect on our health care decisions.

Uncertainties are not something that doctors are necessarily used to handling. Sometimes, even communicating and understanding risks can be difficult. The role of **statistics** is precisely this. Statistics can be regarded as the *science of management of uncertainties in an empirical setup*—a tool to measure them and minimize their impact on decisions. A basic feature of statistical methods is that they enable us to draw conclusions about the whole from the study of a part—a sample. This provides a framework within which to think critically about evidence, and to understand and appreciate the concepts of risks and probability.

Another aspect is of errors in the data and missing values. Statistical methods also help in detecting and handling these aberrations that give rise to statistical lies [1], which statistics is notorious for.

Uncertainties are prominent in practically all medical situations and need to be properly managed. Management in any sphere is a complex process, more so if it concerns phenomena such as uncertainties. Management of uncertainty requires a science that understands randomness, instability, and variation. Biostatistics is the subject that deals specifically with these aspects. Although the *bio* part of *biostatistics* should stand for all biological sciences, it has become convention to apply the term *biostatistics* to statistical applications in only medical and health sciences.

Since the uncertainties are glaring, one wonders how medicine has been successful, sometimes very successful, in giving succor to mankind. The silver lining is that a trend can still be detected among these variations, and following this trend yields results within clinical tolerance in most cases. The term *clinical tolerance* signifies that the medical intervention may not necessarily restore the body system to its homeostatic level but tends to bring it closer to that level so that the patient feels better, almost cured. Also note the emphasis on "most cases." Positive results are not obtained in all cases, nor is this expected. But a large percentage of cases respond to medical intervention. Thus, the scenario is doubly probabilistic and underscores the prominent role of biostatistics in medical decision making.

Many clinicians deal with these uncertainties in their own subjective ways, and some are very successful. But most are not as skillful. To restore a semblance of science, methods are needed to measure these uncertainties, to evaluate their impact, and of course to keep their impact under control. Such a need is more conspicuous in a research setup than in everyday practice. All these aspects are primarily attributed to the domain of biostatistics. Biostatistical methods not only contribute to probabilistic thinking but also encourage critical and logical thinking. This helps to judge the believability of something that nobody has witnessed.

Biostatistical methods are becoming more complex by the day just as almost all other sciences. Many medical professionals find it difficult to catch up, and many consider biostatistics an unnecessary intrusion into their clinical acumen. But biostatistics seems to have firmly entrenched itself in medical research, if not in clinical practice, which gives the analogy to a joke that it is a wife that you cannot live with and cannot do without either. Another popular quote that says that biostatistics is used more for support than illumination just as a streetlight pole is for drunkard also illustrates how biostatistics is treated by some medical researchers. Biostatistics consultation should start at the beginning of a project. Consulting after the project is over, when the reviewers have found fault, is like a postmortem that can diagnose the cause of death but cannot revive the project.

Distinguish biostatistics from medical **data mining**, which relates to the existing data. Hospitals these days keep an electronic record of all patients they manage, particularly those admitted. You may collate these data, tabulate them, present them in graphs and diagrams, etc. This also is a perfectly valid statistical activity, although it has no chance element. After examining such data for thousands of patients, for example, undergoing gastrointestinal (GI) surgery, you may reach a conclusion that women of age 50–59 years are the ones who more commonly chose elective GI surgery in that hospital compared with any other age-sex groups. This may have some implication for the policy of the hospital, but this conclusion cannot be generalized to other hospitals, although one can surmise that this is so for other hospitals also catering to a similar population. Statistical inference such as estimation and test of hypothesis cannot be used in data mining unless this is considered as a **random sample** from a hypothetical hyperpopulation of patients coming to that hospital. If that assumption is made, you are back to the fold of biostatistics.

1. Fung K. The pending marriage of big data and statistics. *Significance* 2013;10(4):22–5. https://tagteam.harvard.edu/hub_feeds/1980/feed_items/261241

## biplot

A biplot is a specific type of graphical display of **multivariate** data. Although two or three variables are easy to depict, depiction of a larger number is difficult. How can mean plasma glucose level (both fasting and postprandial), weight, and blood pressure (systolic and diastolic) be shown for two groups of subjects in the same figure? Several types of biplots have been proposed that can show four or even five variables simultaneously. However, the methods are complex, and the figure so obtained can be difficult to interpret. One such biplot is shown in Figure B.6 for stroke patients. You will appreciate that such a figure, as of now, does not seem to serve the purpose of better perception and adequate cognition that a figure is supposed to serve. For further details of biplots, see Gabriel and Odoroff [1].

1. Gabriel KR, Odoroff CL. Biplots in biomedical research. *Stat Med* 1990;9:469–85. http://www.ncbi.nlm.nih.gov/pubmed/2349401
2. De Wit L, Molas M, Dejaeger E, De Weerdt W, Feys H, Jenni W, Lincoln N, Putman K, Schupp W, Lesaffre E. The use of a biplot in studying outcomes after stroke. *Neurorehabil Neural Repair* 2009 Oct;23(8):825–30. http://www.ncbi.nlm.nih.gov/pubmed/19498014

## birth and death registration

Almost all countries around the world require that each birth and death be registered with the local registrar. Generally, the local authority such as a municipality is assigned the function to register such events and issue a birth or death certificate as the case may be. This is called the **civil registration system**. Although some countries now have an online system for registering births and deaths through hospitals, many still require manual registrations. For example, in India, an act provides that each birth and death must be reported to the local registrar within 21 days.

A *birth certificate* will generally contain information on the place of birth, sex, age of the mother, and sometimes other details

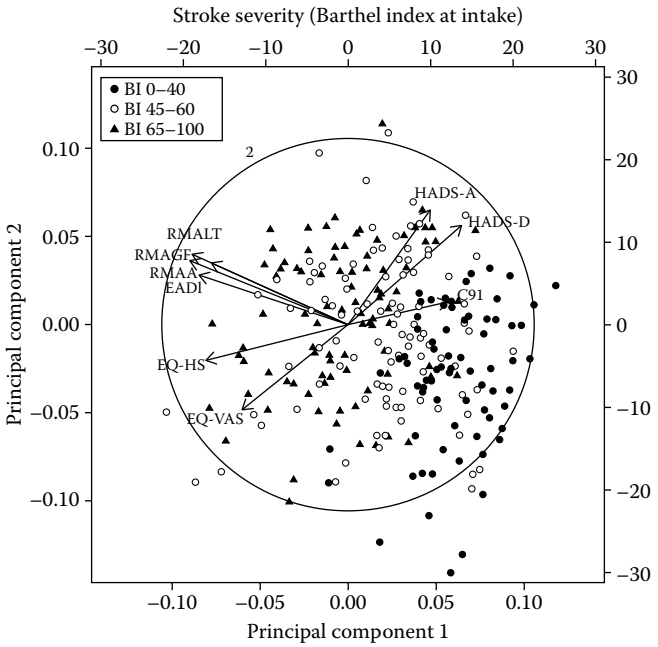



**FIGURE B.6** Biplot without CSI and with projected HADS-A and HADS-D vectors. BI—Barthel index; CSI—Caregiver Strain Index; EQ-HS—EuroQol health state; EQ-VAS—EuroQol visual analogue scale; HADS-A—Hospital Anxiety and Depression Scale–Anxiety; HADS-D—Hospital Anxiety and Depression Scale–Depression; RMAGF, RMALT, and RMAA—Rivermead Motor Assessment gross function–leg, trunk, and arm. (From De Wit L, Molas M, Dejaeger E, De Weerdt W, Feys H, Jenni W, Lincoln N, Putman K, Schupp W, Lesaffre E. The use of a biplot in studying outcomes after stroke. Neurorehabil Neural Repair 2009 Oct;23(8):825–30. http://www.ncbi.nlm.nih.gov/pubmed/19498014.)

(e.g., occupation) of the mother and/or father. A *death certificate* carries information on age, sex, place of death, and, most importantly, the **cause of death**. The cause should be medically certified. The recorded causes should ideally follow the **International Classification of Diseases (ICD)** system so that they are uniformly understood across the world, although this often does not happen. The World Health Organization (WHO) recommends that the causes be divided into immediate, underlying, and contributory.

Birth certificates provide a legal framework not just for health benefits but also for, say, school admission, juvenile justice, child labor, inheritance, etc. Similarly, death certificates also help in inheritance, property disputes, etc. Cause of death is particularly useful for targeted interventions against causes that are alarming or preventable, thus helping to reduce premature mortality.

In countries where a complete listing of the population is available, birth and death registrations can be used to instantly update records. These records can also be analyzed to ascertain whether more births or deaths are occurring, at what age of women in the case of birth, at what age in the case of death and with what cause, etc. Since age and cause of death are important indicators of health status and health requirements, birth and death registration serves a very useful statistical purpose.

A civil registration system may not able to capture all births and deaths, particularly in developing countries. The cause of death may or may not be medically certified. Such countries therefore try to collect data from other sources to fulfill their requirement for planning and evaluation of their health programs.

## birth cohort

A birth cohort is a group of people born in a certain specified period and who are periodically followed up for their health parameters since birth, sometimes from pregnancy. Thus, a birth cohort study is a **prospective study** with no randomization of the exposure. The objective of the follow-up is to keep complete track of various health events since birth and to study their relationship with one another and with other environmental and behavioral factors. Birth cohort studies are believed to be a reliable source for generating **cause–effect relationships** since information on many **antecedents** and **confounders** is available that can be taken into account when deriving net relationships with the outcomes. This can help policy makers.

A birth cohort can be chosen to be representative of births in a defined area. Follow-up frequency and period of follow-up is decided beforehand, and the outcome of interest is specified. Covariates under study are chosen and confounders identified. If the outcome of interest is rare, the cohort size obviously should be large so that enough events occur in the follow-up period for drawing a reliable conclusion.

When followed up until adulthood, birth cohorts provide a unique opportunity to study the impact of early life events, such as birth weight and childhood obesity, on subsequent adverse health events. There has been a considerable interest in such studies ever since the Barker hypothesis [1] was put forward. A large range of hypotheses can be investigated by birth cohort studies.

Birth cohort studies are not simple to implement, because they involve huge cost and efforts for long-term follow-up. In addition, such cohorts may be truncated because of migration, noncooperation, and occasional deaths. More challenging is the need to deal with time-related changes—new technology, new knowledge, and new methods could easily make earlier observations obsolete, and environmental, social, and physical changes could severely affect the results even if time-dependent covariate analysis is done. Birth cohort studies may not provide breakthrough results during the early phase, and this can cause frustration and despondence. Despite such problems, there has been a marked increase in birth cohort studies because of their inherent value in providing a comprehensive picture of events in real time.

With digitization of records since birth, these kinds of studies have become easy in many countries. When complete records are available, historical birth cohorts can be analyzed for events since birth. In one such study in Australia, Srinivasjois et al. [1] observed that those born at a shorter gestational age remain more vulnerable to subsequent hospitalization for a variety of causes up until 18 years of age compared with full-term infants. Magnusson et al. [2] studied the diet of a birth cohort at the age of 8 years in Sweden for development of allergic rhinitis and nonallergic rhinitis in the next 8 years, and concluded that regular consumption of oily fish and dietary long-chain n-3 polyunsaturated fatty acids (PUFAs) in childhood might decrease the risk of rhinitis. For a 20-year follow-up of a birth cohort in Australia for clinical cardiovascular risk during young adulthood in offspring of hypertensive pregnancies, see Davis et al. [3].

A good account of birth cohort studies has been provided by Lawlor et al. [4].

1. Srinivasjois R, Slimings C, Einarsdóttir K, Burgner D, Leonard H. Association of gestational age at birth with reasons for subsequent hospitalisation: 18 years of follow-up in a Western Australian population study. *PLoS One* 2015 Jun 26;10(6):e0130535. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4482718/

2. Magnusson J, Kull I, Westman M, Håkansson N, Wolk A, Melén E, Wickman M, Bergström A. Fish and polyunsaturated fat intake and development of allergic and nonallergic rhinitis. *J Allergy Clin Immunol* 2015 Jul 4. pii: S0091-6749(15)00772-1. http://www.ncbi.nlm.nih.gov/pubmed/26152316
3. Davis EF, Lewandowski AJ, Aye C, Williamson W, Boardman H, Huang RC, Mori TA, Newnham J, Beilin LJ, Leeson P. Clinical cardiovascular risk during young adulthood in offspring of hypertensive pregnancies: Insights from a 20-year prospective follow-up birth cohort. *BMJ Open* 2015 Jun 23;5(6):e008136. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4480003/
4. Lawlor DA, Andersen AM, Batty GD. Birth cohort studies: Past, present and future. *Int J Epidemiol* 2009 Aug;38(4):897–902. http://ije.oxfordjournals.org/content/38/4/897.long

## birth–death process, *see* stochastic processes

## birth order

Are you the first child born to your mother? Then your birth order is 1. As the name implies, the birth order of the eldest child is 1, that of the next is 2, etc., counting only the live births. Stillbirths and other pregnancy terminations such as abortions are excluded. Deaths occurring after birth do not alter the birth order of siblings.

Birth order can affect health in a variety of ways. The first child initially enjoys all the attention and resources, whereas the second and subsequent births generally get divided attention and divided resources. The effect of birth order on psychological development, intellectual capabilities, and personality traits is well known, but it affects physical health as well. Birth order affects the hemoglobin level of the woman, which can affect the birth weight, which in turn can affect a large number of health conditions. For example, birth order has been found to be associated with cardiovascular risk factors in young adulthood in Sweden [1] and posttraumatic stress disorder in the United Kingdom [2].

Birth order can have significance in many epidemiological studies. It is seen to affect height in Swedish men [3] and suicide rate in Norway [4]. Statistically, isolation of the effect of birth order is not always straightforward, because the outcome may have a multifactorial etiology. For example, Turner et al. [5] found a V-shaped effect of birth order on autism in multiplex families in the United States, where middle births are at high risk, and a linear effect in simplex families, where the risk increases with each additional birth. It is therefore necessary to exercise care when studying birth order as an antecedent for an outcome.

1. Jelenkovic A, Silventoinen K, Tynelius P, Myrskylä M, Rasmussen F. Association of birth order with cardiovascular disease risk factors in young adulthood: A study of one million Swedish men. *PLoS One* 2013 May 16;8(5):e63361. http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0063361
2. Green B, Griffiths EC. Birth order and post-traumatic stress disorder. *Psychol Health Med* 2013 Mar 11. http://www.tandfonline.com/doi/full/10.1080/13548506.2013.774432#abstract, last accessed July 18, 2015.
3. Myrskylä M, Silventoinen K, Jelenkovic A, Tynelius P, Rasmussen F. The association between height and birth order: Evidence from 652,518 Swedish men. *J Epidemiol Community Health* 2013 Jul;67(7):571–7. http://www.ncbi.nlm.nih.gov/pubmed/23645856
4. Bjørngaard JH, Bjerkeset O, Vatten L, Janszky I, Gunnell D, Romundstad P. Maternal age at child birth, birth order, and suicide at a young age: A sibling comparison. *Am J Epidemiol* 2013 Apr 1;177(7):638–44. http://aje.oxfordjournals.org/content/177/7/638.long
5. Turner T, Pihur V, Chakravarti A. Quantifying and modeling birth order effects in autism. *PLoS One* 2011;6(10):e26418. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3198479/

## birth rate, *see* fertility indicators

## birth weight ratio

This is defined as

$$\text{BWR} = \frac{\text{birth weight ratio:}\quad \text{actual birth weight}}{\text{expected birth weight for gestational age and gender}}.$$

This measures how a child's weight at birth is different from what is expected for his/her gestational age and gender. A statistical relationship (third-degree polynomial) is available to find the expected weight for gestational age and gender [1]. More crudely, the median birth weight for a specific gestational age can be used as the expected weight in the denominator.

The BWR has special significance for developing countries such as Bangladesh, where many premature births take place. The BWR is used as an index of small-for-gestational-age (SGA) births relating to children who could not grow normally during pregnancy and require special care. Generally accepted categorization of the BWR is as follows:

| | |
|---|---|
| Normal: | $\text{BWR} \geq 0.90$ |
| Mild SGA: | $0.75 \leq \text{BWR} < 0.90$ |
| Severe SGA: | $\text{BWR} < 0.75$ |

SGA is valuable in predicting several health outcomes, particularly during infancy. For example, Voskamp et al. [2] found in a nationwide study in the Netherlands that the BWR has better discriminating power than birth weight percentile for perinatal deaths. SGA can affect growth and development. It has been found to affect height but not age at menarche in Korea [3].

1. Cnathingius S, Haglund B, Kramer MS. Difference in late fetal death rate in association with determinants of small for gestational age fetuses: Population based cohort study. *BMJ* 1998; 316:1483–7. http://www.bmj.com/content/316/7143/1483.pdf%2Bhtml
2. Voskamp BJ, Kazemier BM, Schuit E, Mol BW, Buimer M, Pajkrt E, Ganzevoort W. Birth weight ratio as an alternative to birth weight percentile to express infant weight in research and clinical practice: A nationwide cohort study. *Obstet Gynecol Int* 2014;2014:749476. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4147261/
3. Shim YS, Park HK, Yang S, Hwang IT. Age at menarche and adult height in girls born small for gestational age. *Ann Pediatr Endocrinol Metab* 2013 Jun;18(2):76–80. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4027099/

## biserial correlation, *see* point-biserial correlation

## bivariate distributions

When two variables are considered together, this is called a bivariate setup. The **joint distribution** of the variables $x$ and $y$ is the pattern they jointly follow. This helps us to know the probability that the two

variables together take given values. For example, the joint distribution of albumin and globulin levels will indicate the chance that a random person has, say, albumin between 4 and 5 g/dL and globulin between 7 and 8 g/dL. This also indicates the distribution of each variable over all possible values of the other, called the **marginal distribution**. This can also tell us the distribution of one variable for specific values of the other, called the **conditional distribution**. Further explanation is as follows.

Simultaneous study of two variables helps to explore or determine the extent and nature of the relationship between them. For example, we all know how age and height are intimately related in children, and that the relationship is **nonlinear**. In this case, both variables are **continuous**, but one or both of them can be **discrete**. Blood group (O, A, B, AB) distributions in different races (e.g., Aryan, Caucasian, Black, and Mongoloid) is an example where both variables are discrete. At the extreme of this is a **binary** bivariate setup, such as the relationship between **antecedent** present/absent and **outcome** present/absent (see **two-by-two tables**).

The number of subjects with different values of the two variables together gives shape to a bivariate distribution. An example of bivariate distribution of systolic and diastolic level of blood pressure (BP) in women of age 40–44 years is in Table B.4. Values of diastolic and systolic levels are given in intervals for convenience, but that is not necessary. Note, for example, that women with diastolic BP of 60–64 mmHg most frequently have systolic BP between 120 and 129 mmHg (**mode** = 120–129 mmHg), whereas those with diastolic BP of 70–74 mmHg have modal systolic level = 130–139 mmHg. These findings come from conditional distribution of systolic BP for diastolic BP between 60 and 64 mmHg and conditional distribution of systolic BP for diastolic BP between 70 and 74 mmHg, respectively. Similarly, there is a distribution of diastolic level for each interval of systolic level. This is the essential feature of a bivariate distribution.

The distribution of all women by diastolic level is in the last column of Table B.4. This is the marginal distribution of diastolic level. Similarly, the last row is the marginal distribution of systolic level in these 600 women. In many applications, the term *marginal* is ignored, and the distribution among total subjects is simply called the **distribution** of the concerned variable.

As mentioned earlier, we can have a bivariate distribution when one or both variables are discrete. Distribution of birth weight of boys born at different birth orders will be a bivariate distribution where one variable (birth order) is discrete and the other (birth weight) is continuous. In this case, the distribution of all boys (irrespective of birth order) is the (marginal) distribution of birth weight in these newborns.

The concept of joint distribution, conditional distribution, and marginal distribution can be extended to a multivariate setup when the number of variables is three or more.

## bivariate Gaussian distribution

A bivariate Gaussian distribution, better known as bivariate normal distribution, is a special type of **bivariate distribution** that follows a bell-type shape (Figure B.7a). This figure has three axes. In the context of the systolic and diastolic BP example in Table B.4 (but with ungrouped data), the bottom two axes depict systolic ($x$) and diastolic levels ($y$), and the vertical axis depicts the number of subjects. Note the symmetry of this shape and equal standard deviations (SDs) of $x$ and $y$ in Figure B.7a. In this case, you will see the same shape no matter which side the plot is viewed from. The frequency is highest (peak of the distribution) when $x$ and $y$ values are at the middle of their ranges, and the decline of frequencies on either side of these values follows the same pattern. The peak is such that **kurtosis** is 0.

For contrasting with other bivariate distributions, see Figure B.7b, which is for data in Table B.4 with BP groups intact. Since this is based on sample values, the shape is not smooth, and there are bumps and corners. This does not look like a bivariate Gaussian distribution, as the distribution has a longer tail on one side than the other side. Also, it can be seen that the variability of the systolic level is higher relative to that of the diastolic level. When this occurs in a bivariate Gaussian distribution, the shape is not exactly symmetrical, as in Figure B.7c, where the length of the bell at the base is more than its width because of unequal SDs. The shape depends not only on SDs but also on the **correlation** between $x$ and $y$.

## black-box approach

Using computer programs without properly understanding the underlying procedure is sometimes called a black-box approach. Users generally know what goes in and what comes out of a computer package, but few understand what happens in between. This

**TABLE B.4**
**Bivariate Distribution of Systolic and Diastolic Level of Blood Pressure (BP) in 600 Healthy Females of Age 40–44 Years**

| | Systolic BP (mmHg) | | | | | | | Total (Marginal distribution of diastolic BP) |
|---|---|---|---|---|---|---|---|---|
| **Diastolic BP (mmHg)** | **100–109** | **110–119** | **120–129** | **130–139** | **140–149** | **150–159** | **160–169** | |
| 60–64 | 1 | 2 | 5 | 2 | 2 | 0 | 0 | **12** |
| 65–69 | 4 | 8 | 12 | 4 | 1 | 0 | 0 | **29** |
| 70–74 | 1 | 7 | 18 | 33 | 12 | 8 | 2 | **81** |
| 75–79 | 3 | 12 | 15 | 58 | 22 | 18 | 4 | **132** |
| 80–84 | 2 | 23 | 20 | 45 | 71 | 32 | 7 | **200** |
| 85–89 | 0 | 2 | 5 | 15 | 54 | 26 | 5 | **107** |
| 90–94 | 0 | 0 | 1 | 3 | 6 | 20 | 9 | **39** |
| **Total (marginal distribution of systolic BP)** | **11** | **54** | **76** | **160** | **168** | **104** | **27** | **600** |

*Note:* Shaded-in row is the conditional distribution of systolic level when the diastolic level is between 60 and 64 mmHg; shaded-in column is the conditional distribution of the diastolic level when the systolic level is between 110 and 119 mmHg.

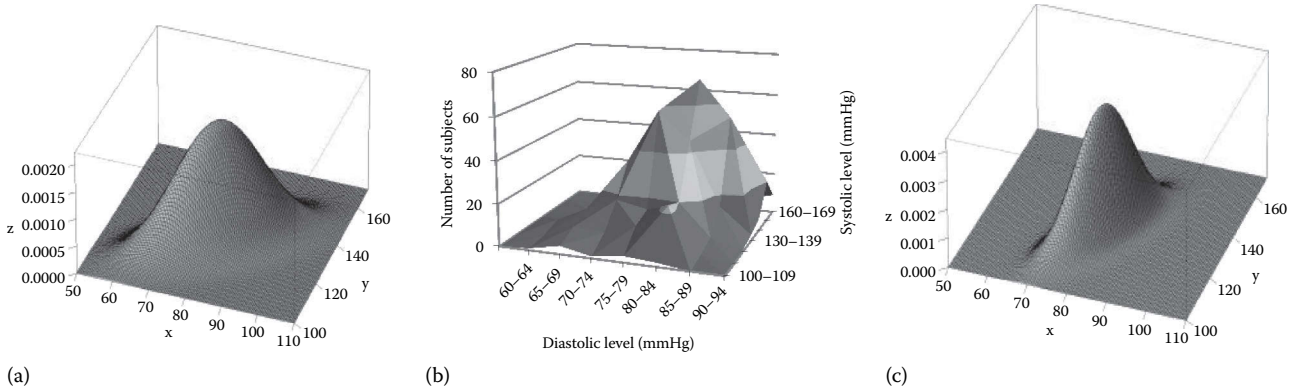



**FIGURE B.7** (a) Bivariate Gaussian distribution with equal standard deviation of $x$ and $y$; (b) bivariate distribution of data in Table B.4 and (c) bivariate Gaussian distribution with unequal standard deviations.

approach is frequently used for computational problems in biostatistics by those who find it difficult to spend time to understand the procedure, or when they possibly lack expertise.

As statistical methods are becoming increasingly complex by the day, use of a black-box approach is inevitably increasing. Many health professionals and some statisticians find the intricate mathematics behind these methods too complex to understand, and resort immediately to the use of computer software to get a solution to their problems. Most statistical software packages are not yet intelligent enough to check the underlying assumptions and appropriateness of the method for the problem at hand. Even if software packages are intelligent for the specified method, users sometimes are not fully aware of how to use them. Further difficulties arise in proper interpretation of the computer output, particularly since statistical software tends to give large output as a default. Some researchers use this output in their report without full appreciation of the implications. It is difficult to ascertain from the finished report whether or not the results and conclusions are based on proper reading of the computer output. Thus, some research reports remain of doubtful value.

A black-box approach can be used for any calculation for which a computer package is available, but it is frequently used for **predictive models**, particularly in **data mining**. Kuhn and Johnson [1] have given an example of a model to "[i]dentify patients who will be admitted to a hospital within the next year, using historical data," to illustrate how challenging unfolding a black-box algorithm could be.

1. Kuhn M, Johnson K. Who's afraid of the big black box? Statisticians' vital role in big data and predictive modeling. *Significance* July 2014;11(3);35–7. http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2014.00753.x/abstract

## Bland–Altman method of agreement, *see also* agreement assessment (overall)

The Bland–Altman method is the most widely used method of assessing **agreement** between the same quantitative measurement taken on the same subjects by two methods, two sites, two observers, etc. The paper describing this was originally published in a statistical journal (*The Statistician*) in 1983 [1], and it was repeated in a revised form for a medical audience in *The Lancet* [2] in 1986 with a reverse sequence of authorship. From then on, this paper has become one of the most cited in medical literature.

Martin Bland

The Bland–Altman method of agreement is for finding whether the two measurements taken on the same subjects agree or not. Thus, this investigates the presence or otherwise of an $x = y$ relationship for each individual. If the two measurements agree, all points will fall on an $x = y$ line. This is different from regression since regression is for averages. For details, see **agreement assessment (overall)**. Some other methods of assessing agreement are discussed under that topic.

The two measurements will rarely agree completely. Differences will arise, and the question is whether the differences are within the tolerance range or not. In the Bland–Altman method, the differences $d = (x - y)$ in the values obtained by the two methods or observers under comparison are examined. If these differences are randomly distributed around 0 and none of the differences is large, the agreement is considered good. The method was initially devised for measurements where neither of the two is a gold standard. This method considers $x$ to be as good or as bad as $y$—it is symmetric. Modifications have now appeared that allow it to be used when the comparison is with the gold [3].

A graphical approach for the setup in which both measurements can be in error is to plot $d$ versus $(x + y)/2$. This has become known as the *Bland–Altman plot*. A flat line around 0 is indicative of good agreement. Depending upon which is labeled $x$ and which $y$, an upward trend indicates that $x$ is generally more than $y$, and a downward trend, that $y$ is more than $x$. If $x$ is a gold standard against which $y$ values are to be assessed for agreement, plot $(y - x)$ versus $x$ [3].

A commonsense approach is to consider agreement as reasonably good if, say, 95% of these differences fall within the prespecified clinically tolerable range and the other 5% are also not too far. Statistically, note that when the two methods or two observers are actually measuring the same variable, then the difference $d$ is mostly the measurement error. Such errors are known to follow a **Gaussian** (normal) **distribution** irrespective of the distribution of the original measurements. Thus, the distribution of $d$ in most cases would be Gaussian even when the

distribution of $x$ or of $y$ or both is far from Gaussian. Then the limits $\bar{d} \pm 1.96 s_d$ are likely to cover differences in nearly 95% of subjects, where $s_d$ is the standard deviation (SD) of the differences. The literature describes them as the **limits of agreement**, but they are actually *limits of disagreement* because they delineate the extent of disagreement. If these limits are within clinical tolerance in the sense that a difference of that magnitude does not alter the management of the subjects, then one method could be replaced by the other. The mean difference $\bar{d}$ is the **bias** between the two sets of measurements, and $s_d$ measures the magnitude of random error. For further details, see Bland and Altman [2]. They also discuss how error in estimating $\bar{d} \pm 1.96 s_d$ can affect the limits and how to calculate limits of disagreement in the case of **repeated measures**.

The Bland–Altman approach has its limitations. Suppose a method consistently gives a level 0.5 mg/dL higher than the other method. The **correlation coefficient** between these two methods would be a perfect 1.0. Correlation fails to detect systematic bias. This well-known artifact incidentally also highlights the limitation of the limits-of-disagreement approach. The difference between measurements by two methods is always +0.5 mg/dL—thus, the SD of difference is 0. The limits of disagreement in this case are (+0.5, +0.5). This, in fact, is just one value and not limits. A naive argument could be that these "limits" are within clinical tolerance and thus, the agreement is good. To detect this kind of fallacy, plot the differences against the mean of paired values. This trend can immediately reveal this kind of systematic bias. Secondly, the method considers absolute difference and does not work well when the difference between two measurements under investigation is in proportion to the values—a larger difference for large values. For this setup, Indrayan [4] has proposed a modification (see **agreement assessment**), although the Bland–Altman method can also be used after taking logarithms.

As an example, consider the following data of Chawla et al. [5] on systolic blood pressure (BP) readings derived from the plethysmographic waveform of a pulse oximeter, which could be useful in a pulseless disease such as Takayasu syndrome. The readings were obtained (i) at the disappearance of the waveform on the pulse oximeter on gradual inflation of the cuff and (ii) at the reappearance on gradual deflation. In addition, BP was measured in a conventional manner by monitoring the Korotkoff sounds. The study was done on 100 healthy volunteers. The readings at disappearance of the waveform were observed to be generally higher, and at reappearance, generally lower. Thus, the average (AVRG) of the two is considered a suitable value for investigating the agreement with the Korotkoff readings. The results are in Table B.5. The scatter and the line of equality are shown in Figure B.8.

**TABLE B.5**
**Results of Agreement Analysis of AVRG with Korotkoff Method of Measuring Systolic Level of Blood Pressure in Healthy Subjects**

| | AVRG | | Korotkoff |
|---|---|---|---|
| Mean systolic BP (mmHg) | 115.1 | | 115.5 |
| SD (mmHg) | 13.4 | | 13.2 |
| Mean difference (mmHg) | | −0.4 | |
| *P*-value for paired *t* | | >0.50 | |
| Correlation coefficient ($r$) | | 0.87 | |
| SD of difference, $s_d$ (mmHg) | | 6.7 | |
| Limits of disagreement (mmHg) | | (−13.5,12.7) | |
| **Intraclass correlation coefficient** ($r_I$) | | 0.87 | |

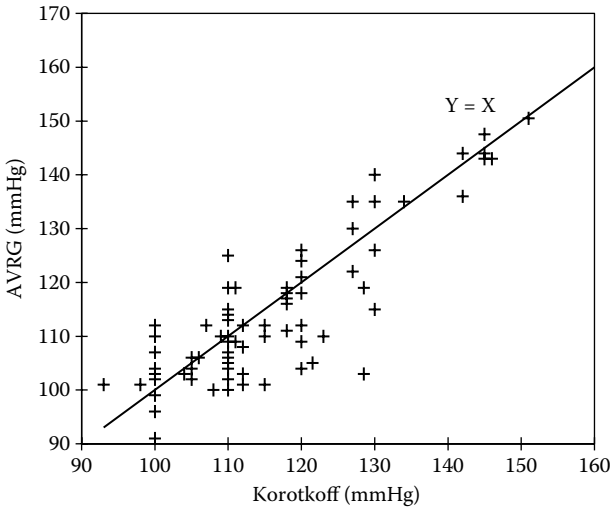



**FIGURE B.8** Scatter of the pulse oximeter and Korotkoff readings (AVRG = average of readings at disappearance and reappearance of a waveform).

Despite the means being nearly equal and the correlation coefficient high, the limits of disagreement show that a difference of nearly 13 mmHg can arise between the two readings on either side (average of pulse oximetry readings can give either less or more than the Korotkoff readings). These limits are further subject to sampling fluctuation [2], and the actual difference in individual cases can be higher. It is for the clinician to decide whether a difference of such magnitude is tolerable. If it is, then the agreement can be considered good, and pulse oximetry readings can be used as a substitute for Korotkoff readings; otherwise, they should not be regarded as interchangeable. Thus, the final decision is clinical rather than statistical when this procedure is used.


1. Altman DG, Bland JM. Measurement in medicine: The analysis of method comparison studies. *The Statistician* 1983;32:307–17. http://www.jstor.org/discover/10.2307/2987937?uid=3738256&uid=2&uid=4&sid=21103402993057
2. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986 i:307–10. http://www-users.york.ac.uk/~mb55/meas/ba.pdf
3. Krouwer JS. Why Bland–Altman plots should use X, not (Y+X)/2 when X is a reference method. *Stat Med* 2008 Feb 28;27(5):778–80. http://www.ncbi.nlm.nih.gov/pubmed/17907247
4. Indrayan A. *Medical Biostatistics*, Third Edition. CRC Press, Boca Raton, FL, 2012.
5. Chawla R, Kumarvel V, Girdhar KK, Sethi AK, lndrayan A, Bhattacharya A. Can pulse oximetry be used to measure systolic blood pressure? *Anesth Analg* 1992;74:196–200. http://www.ncbi.nlm.nih.gov/pubmed/1731537


## blinding, masking, and concealment of allocation

Blinding in health and medicine is withholding information from those involved. This can be used in a variety of research setups, such as blinded reviews of the articles submitted for publication so the reviewers do not know who has conducted the study and where, and can provide their unbiased view. In case–control studies, for example, blinding the observers, where feasible, can help in obtaining unbiased data. The purpose is to look fair to those who are being assessed and to provide unbiased data as much as possible. Our concern in this section is with blinding in a clinical trial setup, where

fairness and unbiased data have special significance. Related topics of masking and concealment of allocation are also discussed in this section.

In a **clinical trial** setup, sometimes, the exposure or the accompanying baseline information and sometimes the outcome are not correctly assessed. This can occur due to bias of the observer or of the recording clerk, who may classify a subject into a particular category of interest supporting one's individual hypothesis. Observer **bias** can also occur if the subjects with disease are evaluated more intensively and more carefully than those without disease. In addition, there is a tendency for the subjects to respond differently depending on whether they are in the treatment group or in the control group. They may show some psychological effect when actually, none is present, or they may conceal a real effect. They may want to switch from one group to another depending upon which group is showing better results. This can disturb the trial. To control all these biases, three precautions are taken. The first is called blinding of the subjects and the assessors, the second is called masking of the regimen, and the third is concealment of allocation. These are closely related but different procedures, although often confused with one another.

Blinding is not revealing to those involved in the trial which subject is receiving the test regimen and which the control regimen. Thus, this term refers to the subjects and the assessors and not the regimen. The corresponding term for making the regimen apparently indistinguishable is called masking. Concealment of allocation means that the person doing the allocation does not know what the next allocation is going to be. This is also described in this entry. Masking and concealment of allocation are needed for effective blinding. The details are as follows.

## *Single, Double, and Triple Blinding*

Trials with no blinding are called **open trials** or *open-label trials*. In this type of trial, everybody knows who is receiving which treatment. Experience suggests that unblinded trials are more likely to falsely show a benefit of the active regimen than blinded trials. Some subjects may show improvement or deterioration unrelated to the true effect of the treatment if they know to which group they belong. Participants in the control group may feel discriminated against if the allocation is open. Also, patients who know that they are receiving a new regimen may either exhibit increased anxiety or have favorable expectations. In addition, subjects are more likely to seek adjunct intervention in an open trial and more likely to dropout. If the treating clinician is aware of the upcoming allocation, he/she could exclude patients he/she considers unsuitable for that treatment or include patients considered especially suitable. This can introduce bias. During the course of the trial, a clinician who knows the group assignment, is more likely to administer cointervention and more likely to adjust dose. All these potential sources of bias can be avoided if subjects and the observers are blinded.

*Single blinding* refers to subjects not told about the treatment allocation. This eliminates the possibility of participants psychologically changing their response when they know that they are in a particular group. Blinding works as an insulation against such biases as the subjects are generally more committed when told in advance that they can get any of the regimens under trial and give consent to participate. Single blinding occurs automatically if the subjects are unconscious or anesthetized when a treatment is given or a procedure undertaken, such as using or implanting a device, or changing the policy of intensive care. The term *single blinding* can also be used when the assessors are blinded but not the subjects. When this is done, this is specified upfront, or else the usual single blinding is assumed, where the subjects are blinded and not the assessors. Although any single blinding is far from ideal, it may be unavoidable if some features or effects of treatments are apparent to trial personnel and observers but not to the subjects, or vice versa.

If the assessor knows what treatment a particular subject is receiving, this may affect the way the questions are asked, investigations done, or interpretations made. Thus, it is desirable that the assessor also is kept blind. This removes possible bias of the physicians or nurses involved in patient assessment—or at least mitigates any subconscious influence on the outcome assessment. Unblinded assessments can also be unbiased when done by credible people, but blinding improves the confidence in the results without much additional cost. When observers, assessors, and other study-related personal are blinded, in addition to the subjects, this is called *double blinding*. Such a precaution is an important criterion for validity of the results of a trial. A double-blind **randomized controlled trial** is considered a gold standard for assessing the **efficacy** and safety of a new regimen or any other intervention.

Sometimes, even the data analyst has preconceptions, expectations, and prejudices. He/she might be interested in particular findings and can gear the analysis and interpretation accordingly. Most such biases at the data management stage occur in the coding of adverse events when manually done or in the interpretation of handwriting when the case sheets are handwritten. If the data analyst is aware of the treatment allocation, he/she could be selective in whether to seek clarification when something is not clear and may impose his/her own interpretation. To avoid this, the analyst can also be kept blind about the codes. The codes are broken only after the data analysis is complete. This makes the trial *triple blind*. However, make sure blinding the data analyst is not jeopardizing the analysis and presentation. Ii is better to specify in advance what analysis would be done and what tables would be prepared.

In a blinded trial, other than during the phase of analysis, treatment allocations should not be available to anyone, even in the form of treatment A and treatment B. It is also important that nobody involved in the trial knows which patients are receiving the same treatment. Blinding can be implemented by involving a third party who keeps the record that a subject has received treatment or control. Rigid coding systems, such as code X for the treatment and code Y for the placebo, should be avoided because breaking the code for one patient breaks it for the rest of the trial. This can easily happen in a medical care setup, where laboratory investigations or side effects can identify that a subject has received the active regimen or placebo.

Details of how blinding is actually intended to be implemented should be fully specified in the protocol, and how it was actually implemented should be stated when reporting a trial. Merely stating that blinding was done is not enough. In fact, such details should be given in the protocol itself.

## *Difficulties in Blinding*

Blinding can be difficult and sometimes not achievable. One problem is masking, as described later in this section, and the second problem is its feasibility. For assessing outcomes such as quality of life, readmissions, and falls after hip surgery, blinding is just not possible if one maneuver is keeping patients in the hospital for a specified number of days, and the other is early discharge and home rehabilitation. In most surgical interventions, the control has to be another kind of surgery and not a placebo. A sham surgery may be unethical many times because it exposes a patient to surgical risks. In either case, it is extremely difficult to enforce blinding in a surgical trial. The patient can be kept blind after proper consent, but the surgeon definitely

knows. However, a mechanism can possibly be developed wherein all assessments subsequent to the operations are done by another surgeon who does not know and cannot decipher whether the patient belongs to the test surgery or the control surgery group.

If the effects of one of the treatments are such as to be apparent (e.g., marked facial flushing after taking a drug or developing different symptoms after different types of surgery), it is almost impossible to maintain the blindness. A similar problem arises if, for example, two different durations of hospitalizations are being compared.

Blinding is rarely feasible in studies requiring long follow-up. At some stage, something happens that can break the code. It is very difficult to enforce blinding in field trials also except in some typical situations. Special efforts may be needed. For example, Lwegaba [1] reports a single-blinded field trial on educational material for tobacco prevention. The trial was conducted on school students, and the schools were distantly located so that contact between students did not occur and the blinding could be maintained.

Some trials keep a provision of **interim appraisals**. This has the potential to unblind an otherwise blinded trial. Unblinding is necessary to assess whether the treatment arm is giving sufficient evidence of efficacy or of futility relative to the control or for sample reestimation. Efforts are made that this unblinding remain confined to an unconcerned group, such as the Data Safety and Monitoring Board, and does not extend to the participants and the assessors, and the blinding for them remains in effect till such time that the trial is completed.

Morality issues are attached to blinding because the information is withheld from the participants, on the one hand, who would be keen to know what they are getting in case an unusual side effect appears or unusual recovery occurs, and on the other hand, from the doctor, who may not be able to take remedial measures if anything happens to the detriment of the participant. Nonetheless, this is considered acceptable so long as the subjects are fully informed about blindness and they provide consent. Also, breaking the code must be axiomatic if the care providers consider it necessary in the interest of the patients.

There are logistical problems in ensuring that the blinding remains effective. We have already mentioned side effects that can easily indicate which subject is on an active drug and who on placebo. Masking the regimens to at least look similar, as discussed in a subsequent paragraph, is a challenge in many practical situations. A substantial improvement in some patients and not others may also give indication as to who is in which group.

For more details of blinding, see any good book on clinical trials, such as that by Friedman et al. [2].

## Masking

The term *blinding* applies to the human beings involved in a project, and *masking*, to the regimens and the procedures. An obvious prerequisite for maintaining blinding is that the treatments must not be distinguishable. Regimens and procedures in a trial must be disguised to look similar to the subjects and to the assessors as much as possible. This can be easily illustrated for drugs where the active drug under trial and the placebo should be exactly alike in appearance (color, size, weight, packaging, smell, etc.) and possibly in taste. Note that the treating physician would know the treatment because he/she is giving the treatment unless the masking of the regimen is done and the regimen, such as tablets, are in envelopes for administration. For masking, both the regimens have to be administered equally often. If one is once a day and the other twice a day, the former has to be given a placebo a second time each day to look like a similar treatment. This looks easy in this example, but consider one regimen requiring six tablets a day and the other just two. You can have four placebos, but the subjects may find six too cumbersome to swallow and may decide to take three. All these may be placebos. This may be difficult to track. Thus, masking is not always easy. Consider beforehand what is feasible. A classic example is two anesthetic agents, one of which is to be given at 30 mL and the other at 60 mL. Dilution in this case is not advisable. For masking in this situation, a placebo (say, saline) of 60 mL to the first group and 30 mL to the other groups of the same appearance is given. The order is randomized. This might look strange, as both the groups are getting placebo. This becomes necessary in this case to mask the regimen. This is called the *double dummy technique*.

In the case of hospitalized patients, the two groups should undergo the same rituals in terms of diet, laboratory and radiological investigations, transfer from intensive care to routine care, etc. The term used for such procedures is *sham* (just like in *sham surgery*), whereas the equivalent term for drugs is *placebo*. This is applicable to the single-blind (subject is blind) setup since in a double-blind trial, no one knows the allocation, and all will inevitably be treated identically.

## Concealment of Allocation

Allocation concealment is when the person allocating the treatments in a clinical trial does not know what treatment the *next* person is going to get. The allocation of the subjects to the treatments under trial is generally done with the help of opaque sealed envelopes that contain the allocation. They are opened only after the subject's name and other identification is written on the envelope so that the treatment cannot be changed. Concealment means that the envelopes are in random sequence and the serial on the envelope is not able to reveal what treatment it contains. A third party keeps all the records of the random sequence. This prevents bias of the allocating person in choosing which subject will get what treatment. Using a pharmacy as the third party is common for concealment of allocation in the case of drug trials.

Do not confuse concealment of allocation with **blinding**. Concealment seeks to protect the sequence of the assignment, whereas blindness is not knowing who is getting what. The question of sequence before actual allocation does not come in the case of blinding, but it comes in the case of concealment of allocation. Concealment is always feasible, but blinding is not always feasible. Blinding protects the sequence *after* allocation and is a safeguard for response and ascertainment bias, which occur after the treatment is administered. When the allocation is not concealed, even random assignment can be subverted. This could lead to selective withdrawals before the treatment starts. Details and an example of how all this was implemented in a trial on misoprostol are given by Piaggio et al. [3], although they seem to have mixed up **masking** with blinding.


1. Lwegaba A. Field trial to test and evaluate primary tobacco prevention methods in clusters of elementary schools in Barbados. *West Indian Med J* 2005;54:283–91. http://caribbean.scielo.org/pdf/wimj/v54n5/v54n5a03.pdf
2. Friedman LM, Furberg CD, DeMets DL. *Fundamentals of Clinical Trials*, Fourth Edition. Springer, 2010.
3. Piaggio G, Elbourne D, Schulz KF, Villar J, Pinol AP, Gülmezoglu AM; WHO Research Group to evaluate Misoprostol in the Management of the Third Stage of Labour. The reporting of methods for reducing and detecting bias: An example from the WHO Misoprostol Third Stage of Labour equivalence randomised controlled trial. *BMC Med Res Methodol* 2003 Oct 3;3:19. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC280677/

## block, cluster, and stratified randomization, and minimization, *see also* random allocation

B

**Randomization** is random allocation of subjects to the arms of a trial. This is done primarily to achieve **baseline equivalence** of the groups under trial. Beside simple individual randomization, this can be done by the method of block randomization, cluster randomization, and stratified randomization in certain specific situations or by using more complex techniques such as minimization and adaptive randomization.

### *Block Randomization*

Block randomization is one of the most common methods of randomization. This requires that subjects are divided into $M$ blocks of size $2n/M$ each, where $n$ is the stipulated size of each of the two groups. In the case of more than two groups, the block size must be a multiple of the number of groups. For two groups, the block size can be 4 or 6 or 8, but not 5 or 7. If you have enrolled a total of 80 subjects, you can make 20 blocks of 4 subjects each. Within each block, allocate two subjects at random to group 1 and the other two to group 2. Thus, you can have one of the following allocations.

(1,1,2,2), (1,2,1,2), (1,2,2,1), (2,2,1,1), (2,1,1,2), (2,1,2,1).

One of these blocks is chosen at random for the first four subjects and allocated to group 1 or group 2 accordingly. Then proceed to the second block of four subjects, and so on until all subjects are randomized. This method is called block randomization.

An advantage of block randomization is the possibility of one group becoming full before the other is ruled out. But the difficulty is that you know that the fourth subject after the first three going to groups 1,1,2 must go to group 2. Thus, blinding is difficult. For this, several random block sizes are advocated that are concealed from the investigators.

### *Cluster Randomization*

In place of randomization of individual subjects, it is sometimes convenient to randomize groups. This is also used when randomization of individuals is not feasible. For example, you may want to assign residents of an entire village to a particular mosquito control strategy for malaria and another village to another strategy. The data are still collected at an individual level. This is called cluster randomization.

Individual randomization is advocated in a setup where the response of an individual is independent of that of another individual, whether belonging to the same village, family, or any other cluster. For example, kidney transplant patients in a hospital may be undergoing a similar preoperative and postoperative protocol, as opposed to the patients in another hospital, and the responses of patients within each hospital will be correlated with one another. In this case, individual randomization will not work, and cluster randomization can provide more reliable results. However, cluster randomization does not have the same statistical efficiency as individual randomization has. It has low statistical power due to a clustering effect. The sample size has to be larger to compensate for this loss. This may not be a big problem, because chunks of subjects are selected in each cluster so that a few clusters may give a big sample of individuals. The problem of statistical analysis of a cluster-randomized trial is more challenging as this becomes complicated because the **clustering effect** has to be eliminated.

In cluster randomization, generally, only a small number of clusters are randomized. This may still give a large number of individuals, but subjects within each cluster are likely to have similar baseline characteristics that affect the response. Few clusters implies an increased chance of imbalance in the test and the control group. Both groups should be thoroughly examined for baseline equivalence; otherwise, the researchers should keep track of the imbalances at the time of statistical analysis to ensure that this does not affect the results.

### *Stratified Randomization*

Stratified randomization seeks to ensure that the subjects with important covariates are rationally distributed amongst the groups—thereby reducing the possibility of baseline imbalance. This can also be done for anticipated imbalance in the responses. If your study is on a wonder dose that controls blood sugar level for 1 month, and if you know that the effect could be different in males of age <50 years compared to females of age ≥50 years, you may want to divide the enrolled subjects as <50M, ≥50M, <50F, and ≥50F, so that each of these strata is adequately represented, and then divide them equally into group 1 receiving the test drug and group 2 receiving the control regimen. Such stratification is useful when the effect is expected to vary across strata and helps in better interpretation. Some groups may be of special interest and need to be overrepresented. This stratification will not only help to ensure an adequate number of subjects in each stratum but also help in finding which specific group benefits most from the regimen. This conclusion will have as much reliability as afforded by the sample size in each stratum.

Since many medical responses are sex specific, sex-stratified randomization is common. Golomb et al. [1] used this to study statin effect on aggression in the United States. Buhule et al. [2] observed that stratified randomization by obesity, age group, and census region controls better for batch effects in 450K methylation analysis in the United States.

Nonetheless, stratified randomization is not practical if there are many relevant covariates even if a large sample is available. Thus, it is restricted to the most important or strongly related covariates.

### *Minimization*

Minimization uses a computer-based process to determine randomized allocation to treatment groups in real time at the time of randomization of each individual subject by considering baseline characteristics of all previously randomized subjects and the one to be randomized. This process tends to allocate the subject to the group that results in the best baseline equivalence between the groups, while usually, at the same time, also considering equality of allocations to each treatment. Pure minimization is deterministic, which can predict in advance the group to be allocated, but in the method we are discussing, a chance element is introduced. A system that generates this process can be quite complicated. For details, see Saghaei [3].

### *Adaptive Randomization*

Adaptive randomization uses a similar computer-based randomization system as does minimization, but rather than just determining random allocation to treatment in order to maximize baseline equivalence or equality of size of treatment groups, it dynamically adjusts the ratio of allocation to treatment groups in light of emerging efficacy results. This technique may, for example, be used to reduce the number of subjects exposed to an ineffective treatment (such as placebo) should the emerging results show a large difference between groups. Thall and Nguyen [4] have discussed this type of randomization for improving utility-based dose finding with bivariate ordinal outcomes.

1. Golomb BA, Dimsdale JE, Koslik HJ et al. Statin effects on aggression: Results from the UCSD statin study, a randomized control trial. *PLoS One* 2015 Jul 1;10(7):e0124451. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0124451
2. Buhule OD, Minster RL, Hawley NL, Medvedovic M, Sun G, Viali S, Deka R, McGarvey ST, Weeks DE. Stratified randomization controls better for batch effects in 450K methylation analysis: A cautionary tale. *Front Genet* 2014 Oct 13;5:354. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4195366/
3. Saghaei M. An Overview of randomization and minimization programs for randomized clinical trials. *J Med Signals Sens* 2011 Jan–Apr;1(1):55–61. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3317766/
4. Thall PF, Nguyen HQ. Adaptive randomization to improve utility-based dose-finding with bivariate ordinal outcomes. *J Biopharm Stat* 2012;22(4):785–801. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3385658/

## BMI, *see* body mass index

## body mass index, *see also* obesity (measures of)

**Obesity** has been found to be associated with risk of diseases such as hypertension, atherosclerosis, gallbladder disease, and diabetes. It is now standard practice in clinics to assess obesity and give advice accordingly. Ideally, it should be assessed by the amount of fat present in the body, but that is difficult to determine. Weight relative to height is considered a good surrogate. The following indicator is commonly used:

$$\text{body mass index: BMI} = \frac{\text{weight in kilograms}}{(\text{height in meters})^2}$$

Adolphe Quetelet

The name *body mass index* was given by Keys et al. in 1972 [2], although the original name for this index is **Quetelet index** after the Belgian scientist Adolphe Quetelet, who first suggested it in 1832 [3]. The unit of BMI is kg/m$^2$. A World Health Organization (WHO) expert committee [4] has given a nomogram that computes BMI for a given height and weight. Another nomogram is in Figure B.9. This reads the value of BMI for a given height and weight and also tells whether this is low, normal, high, or extremely high.

BMI has the same basic form as the general **ponderal index**. The exponent of the denominator of the ponderal index is 3.0 in the case of infants but progressively declines as age advances. It finally settles down to 2.0 for adults. Thus, we have height$^3$ in the denominator for infants and height$^2$ in the denominator for adults.

Statistically, persons with higher BMI tend to exhibit greater variability in BMI than those with lower BMI. Thus, care is needed in using methods (such as **ANOVA**) that require **homogeneity of variances**.

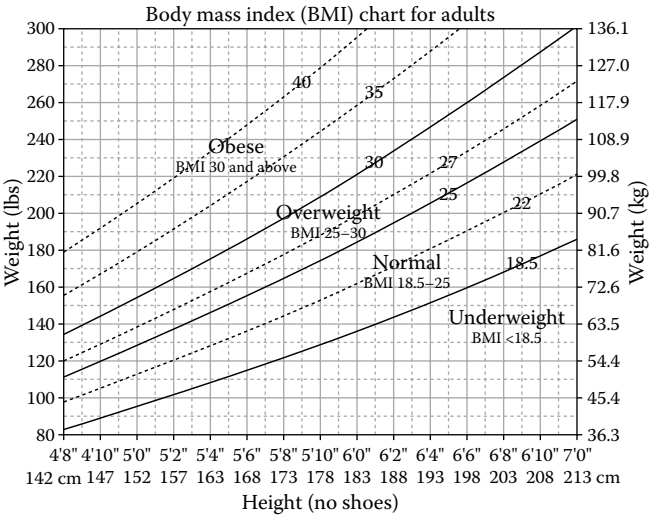



**FIGURE B.9** A nomogram for BMI from height and weight of adults. (From Vertex42. BMI Chart (Body Mass Index). http://www.vertex42.com/ExcelTemplates/bmi-chart.html.)

In the case of self-reporting, it has been observed that people tend to report less weight but greater height. This can make a substantial difference in the value of BMI. Thus, be careful in using BMI based on self-reported height and weight.

BMI is mainly determined by heredity, physical activity, and carbohydrate intake. It is considered age-sex independent in adults, and nearly the same thresholds can be used for females as for males without much error, although it tends to be slightly less in females. However, it is lower in children. For adults, generally, a person with BMI < 16kg/m$^2$ is usually categorized as severely underweight, 16–18.5 as thin, 18.5–25 as normal, 25–30 as overweight, 30–35 as obese, and 35+ as morbidly obese. These categories can vary depending on the context. Individuals with a large body frame and muscle mass, such as athletes, may be wrongly classified as obese. For Asian adults, the WHO has given lower thresholds [5].

Recent evidence suggests that BMI < 16 kg/m$^2$ in adults is more of a risk for early mortality than BMI > 25 kg/m$^2$. A disaggregated analysis by Welch et al. [6] has indicated that BMI = 26 or 27 has nearly the same risk as BMI = 25 kg/m$^2$—thus, these values are no longer considered to indicate overweight. This shows that care must be taken in interpreting BMIs.

BMI has been found to have an association with many health conditions. Besides the usual heart-related parameters, it has also been found to be negatively associated with ovarian cancer risk—for increasing quartiles of BMI, the odds ratio exhibited a decreasing trend [7]. Children with low birth weight but with rapid increase in BMI after the age of 1 year have been found to be at greater risk of coronary heart disease later in life, particularly male children [8]. Obese women can have a longer duration of labor in childbirth.

In US boys, BMI is least (median, 15.3 kg/m$^2$) at age 6 years, and in girls, at age 5 years (median, 15.2 kg/m$^2$). This rises to 23.0 kg/m$^2$ at age 20 years in boys and to 21.7 kg/m$^2$ in girls at this age [9]. Thus, BMI is not age independent in children. For internationally applicable cutoffs for well-to-do children, see Cole et al. [10]. Anderson et al. [11] have given a number of nomograms that more effectively take account of variables referred to previously, such as age-specific and gender-specific height and weight.

1. Vertex42. BMI Chart (Body Mass Index). http://www.vertex42.com/ExcelTemplates/bmi-chart.html
2. Keys A, Fidanza F, Karvonen MJ, Kimura N, Taylor HL. Indices of relative weight and obesity. *J Chron Dis* 1972;25:329–43. Quoted by Apell SP, Wahlsten O, Gawlitza H. Body mass index—A physics perspective. http://arxiv.org/ftp/arxiv/papers/1109/1109.0296.pdf
3. Quetelet A. Recherchessur le poids de l'homme aux different ages. *Nouveaux Memoire de l'Academie Royale des Scienceset Belles-Lettres de Bruxelles* 1832: t. VII. Quoted by Apell SP, Wahlsten O, Gawlitza H. Body mass index—A physics perspective http://arxiv.org/ftp/arxiv/papers/1109/1109.0296.pdf)
4. WHO. Physical status: The use and interpretation of anthropometry. *Tech Rep Ser No*. 854, 1995:434. http://helid.digicollection.org/en/d/Jh0211e/
5. WHO Expert Consultation. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies. *Lancet* 2004;363:157–63. http://www.who.int/nutrition/publications/bmi_asia_strategies.pdf
6. Welch HG, Schwartz LM, Woloshin S. The exaggerated relations between diet, body weight and mortality: The case for a categorical data approach. *CMAJ* March 29, 2005;172(7): 891–5. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC554875/
7. Lukanova A, Toniolo P, Lundin E et al. Body mass index in relation to ovarian cancer: A multi-centre nested case–control study. *Int J Cancer* 2002;99:603–8. http://cebp.aacrjournals.org/content/14/5/1307.long
8. Eriksson JG, Forsen T, Tuomilehto J, Osmond J, Barker DJ. Early growth and coronary heart disease in later life: Longitudinal study. *BMJ* 2001;322:949–53. http://www.bmj.com/content/322/7292/949.pdf%2Bhtml
9. Centers for Disease Prevention and Control. *Growth Charts*. http://www.cdc.gov/growthcharts/
10. Cole TJ, Flegal KM, Nicholls D, Jackson AA. Body mass index cut offs to define thinness in children and adolescents: International survey. *BMJ* 2007;335:194–201. http://www.bmj.com/content/335/7612/194.pdf%2Bhtml
11. Anderson RH, Baker EJ, Penny DJ, Redington AN, Rigby ML, Wernovsky G. *Paediatric Cardiology*, Third Edition. Elsevier, 2009.


## Bonferroni procedure/test, *see also* multiple comparisons

Sometimes, several statistical tests are done on the same data, and a combined conclusion is drawn. If each test is undertaken at, say, a 5% **level of significance**, the total probability of **Type I error** in all tests together is markedly increased. In other words, the more such tests are done, the more likely that at least some will produce a significant result just by chance when no real effect is present. This situation specifically arises in analyses of variance where the means of several groups are compared with each other, termed **multiple comparisons**. Many methods have been proposed to ensure that the probability of **Type I error** does not exceed the predetermined level, and the Bonferroni procedure is one of them.

Bonferroni is the simplest method to ensure that the total probability $\alpha$ of a Type I error does not exceed the desired level in case of multiple comparisons. Under this procedure, a difference is considered **significant** only if the corresponding ***P*-value** is less than $\alpha/H$, where $H$ is the number of comparisons. If there are four groups and *all* pair-wise comparisons are required, then $H = 6$. Then a difference would be considered significant at the 5% level if $P < 0.05/6$, that is, if $P < 0.0083$ for that difference.

The Bonferroni procedure is conservative in the sense that the actual total probability of Type I error will be much less than $\alpha$. This means that there is an additional chance that some differences are actually significant but pronounced not significant. As explained under **tests of hypotheses (philosophy of)**, this is not a major limitation in an empirical sense for any statistical test. The advantage of the Bonferroni procedure is that $H$ can be only as much as the number of comparisons of interest. If there are four groups and the interest is only in comparing group 1 with group 2, 2 with 3, and 3 with 4 (and not, for example, in comparing group l with group 3), then $H = 3$. A small $H$ improves the efficiency of this procedure.

If there are many **predictors**, as in a **regression** setup, an alpha level of 0.05 for each individual predictor may be too high because all tests are based on the same data. Exercise caution and see if you would like to use a Bonferroni type of procedure to adjust the **significance level** in such cases.

## bootstrap procedure, *see* resampling

## Bowker test, *see* the McNemar–Bowker test

## box-and-whiskers plot

As a part of exploratory data analysis, the box-and-whiskers plot was first suggested by John Tukey in 1977 [1]. This is also called a *box plot*. A box is made around the value of the **median** with two divisions. The lower end of the box represents the first **quartile** $Q_1$, and the upper end, the third quartile $Q_3$ (Figure B.10). The height of the box, thus, is the **interquartile range** (IQR), covering the middle 50% of the values. The larger the height of this box, the greater the spread of the values. The width of the box is arbitrary. Vertical lines are drawn from the lower end of the box ($Q_1$) to the "minimum value," and from the upper end of the box ($Q_3$) to the "maximum value." These look like whiskers. Values more than three interquartile distances away from the median are considered clear outliers and marked with an asterisk, and values more than 1.5 × IQR (called *inner fence*) but less than 3.0 × IQR (called *outer fence*) are considered mild outliers and marked by 0. The minimum and maximum values used to draw whiskers exclude these mild and clear outliers. Note that a single graph shows so many values The diagram can be drawn vertically or horizontally.

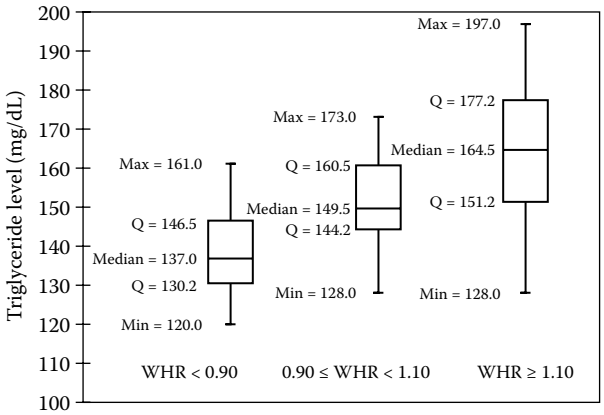



**FIGURE B.10** Box-and-whiskers plots for the data of three WHR groups.

Box-and-whiskers plots for the triglyceride levels (TGLs) in different waist–hip ratio (WHR) categories for hypothetical data are shown in Figure B.10. The values of median, $Q_1$, and $Q_3$ as well as the minimum and maximum values are also shown. In this case, there are no outliers. The figure on the whole displays the location of the median and the dispersion of the data as well as the skewness. The plot for three WHR groups in Figure B.10 shows that the dispersion of TGL is fairly spread out (big box) and is symmetric when WHR ≥ 1.1 but relatively compact (short box) and very skewed (the distance between the median and $Q_1$ is very different from the distance between the median and $Q_3$) when 0.9 ≤ WHR < 1.1. Rising box-and-whiskers plots in this figure also show that TGL rises as WHR increases.

1. Champkin, J. Timeline of statistics *Significance* 2013;10;23–6. http://www.statslife.org.uk/history-of-stats-science/1190-the-timeline-of-statistics

## Box–Cox power exponential (BCPE) method, *see also* LMS method

Lambda-mu-sigma (**LMS**) of Cole and Green [1] and the Box–Cox power exponential (BCPE) of Rigby and Stasinopoulos [2] are popular methods to obtain smoothened centile curves. These are commonly used to obtain percentile curves for various growth parameters in children, such as by the World Health Organization (WHO) Multicentre Growth Reference Study Group [3], and growth reference curves for school-aged children and adolescents [4]. The application extends to any setup where centiles are estimated for different time points. For example, these methods have been used to obtain reference values of differences between TW3-C radius, ulna and short bone (RUS) and TW3-C carpal bone ages of children in China [5]; to obtain centile charts for placental weight for singleton deliveries in Aberdeen, United Kingdom [6]; and for normal values of aortic dimensions, distensibility, and pulse wave velocity in children and young adults in Germany [7].

A main feature of these plots is that percentiles appear to follow a smooth curve over age. Two major statistical issues are involved when plotting these curves. The first is to correctly find different percentiles at each age, and the second is to achieve smoothness of the percentile curves over age. Note that the first problem is to obtain various percentiles at age 1 year, at 2 years, at 3 years, etc. The second problem is to ensure that each of these percentiles forms a smooth curve when plotted versus age. These involve the following problems.

Experience suggests that the **distribution** of many medical measurements is not **Gaussian** (normal) but, rather, tends to be **skewed** and with disturbed **kurtosis**. The LMS method is primarily for correcting skewness, and it does not handle kurtosis, whereas the BCPE method handles both skewness and kurtosis. To understand the BCPE method, it is necessary to first understand the LMS method.

Let $y$ denote the measurement under consideration, such as weight and arm circumference. Consider a distribution of $y$ that has either positive or negative kurtosis. In this case LMS-adjusted $z_{LMS}$ will not give you the correct $z$-score. Additional adjustment is needed for kurtosis. This adjustment is not in terms of changing the transformation but in terms of considering that the distribution of $z_{LMS}$ is not **standard normal** but, rather, another complex distribution called the BCPE distribution. This is where BCPE comes in. Besides LMS ($\lambda$, $\mu$, $\sigma$; see **LMS method**), this distribution has another parameter, denoted by $\tau$. However, the skewness parameter $\lambda$ under the LMS method is now denoted by $\nu$ since $\lambda$ would be the notation for **Box–Cox power transformation** of the $x$ variable.

The notation $x$ is for a variable such as age in the case of weight, while $y$ is the notation for weight in this case. Weight centiles are required for each value of $x$. BCPE distribution reduces to LMS when $\tau = 2$. Thus, the LMS method is a special case of the BCPE method. You might as well use the BCPE method straightaway. This will give $\tau \approx 2$ if the kurtosis is already 0. The values of $\mu$, $\sigma$, $\nu$, and $\tau$ are estimated with the help of special software in a manner wherein the transformed $z$ has a standard Gaussian distribution. The BCPE method is also called the **LMSP method**, where P stands for *power*. The $p$th percentile for BCPE distribution is obtained by using an inverse function, but it does not have an explicit expression of the type that exists for the LMS method. Software help will be required. The software generally used for this purpose is the *Generalized Additive Model for Location, Scale, and Shape* (GAMLSS) developed by Stasinopolous and Rigby [8]. This software is available at http://www.gamlss.org. Indrayan [9] has provided full details of this method in a nonmathematical language.

1. Cole TJ, Green PJ. Smoothing reference centile curves: The LMS method and penalized likelihood. *Stat Med* 1992 Jul;11(10):1305–19. http://www.ncbi.nlm.nih.gov/pubmed/1518992
2. Rigby RA, Stasinopoulos DM. Smooth centile curves for skew and kurtotic data modelled using the Box–Cox power exponential distribution. *Stat Med* 2004 Oct 15;23(19):3053–76. http://www.ncbi.nlm.nih.gov/pubmed/15351960
3. WHO Multicentre Growth Reference Study Group. WHO Child Growth Standards based on length/height, weight and age. *Acta Paediatr Suppl* 2006 Apr;450:76–85. http://www.ncbi.nlm.nih.gov/pubmed/16817681
4. de Onis M, Onyango AW, Borghi E, Siyam A, Nishida C, Siekmann J. Development of a WHO growth reference for school-aged children and adolescents. *Bull World Health Organ* 2007 Sep;85(9):660–7. http://www.who.int/bulletin/volumes/85/9/07-043497/en/
5. Zhang SY, Liu LJ, Han YS, Liu G, Ma ZG, Shen XZ, Xu RL, Hua JQ. Reference values of differences between TW3-C RUS and TW3-C carpal bone ages of children from five cities of China. *Zhonghua Er Ke Za Zhi* 2008 Nov;46(11):851–5. [Chinese] http://www.ncbi.nlm.nih.gov/pubmed/19099904
6. Wallace JM, Bhattacharya S, Horgan GW. Gestational age, gender and parity specific centile charts for placental weight for singleton deliveries in Aberdeen, UK. *Placenta* 2013 Mar;34(3):269–74. http://www.ncbi.nlm.nih.gov/pubmed/23332414
7. Voges I, Jerosch-Herold M, Hedderich J, Pardun E, Hart C, Gabbert DD, Hansen JH, Petko C, Kramer HH, Rickers C. Normal values of aortic dimensions, distensibility, and pulse wave velocity in children and young adults: A cross-sectional study. *J Cardiovasc Magn Reson* 2012 Nov 14;14:77. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3514112/
8. Stasinopoulos DM, Rigby RA. Generalized additive models for location, scale and shape (GAMLSS) in R. *J Statistical Software* 2007;23(7):1–46. http://www.jstatsoft.org/v23/i07
9. Indrayan A. Demystifying LMS and BCPE methods of centile estimation for growth and other health parameters. *Indian Pediatr* 2014 Jan;51(1):37–43. http://www.indianpediatrics.net/jan2014/jan-37-43.htm

## Box–Cox power transformations, *see also* power transformations

This is a family of transformations, at their simplest of the form $z = y^{\lambda}$, which are used to reduce asymmetry in data. Many commonly used transformations such as logarithm and square root are specific cases of the Box–Cox power transformation.

B

A square-root transformation ($z = y^{0.5}$) tends to correct mild positive skewness because this transformation shrinks values (when values are >1), but importantly, higher values shrink more than lower values. This is because $\sqrt{(4)} = 2$ (reduces to one-half) and $\sqrt{(25)} = 5$ (reduces to one-fifth). Thus, the right tail shrinks faster toward the mean than does the left tail. On the other hand, if the distribution is negatively skewed with a longer left tail, the transformation $y^{2.0}$ will do just the reverse. It will stretch the right tail much more than the left tail—thereby tending to correct left skewness. Both these transformations are of the type $y^{\lambda}$. The value of power $\lambda$ depends on the type of skewness and the extent of skewness. This is called **power transformation** and is a potent tool to correct skewness. This is the initial form of the Box–Cox power transformation [1] and can be extended.

Extension of this approach involves normalization of the values prior to the transformation. To achieve this, divide each $y$ by its **central value** $\mu$ and calculate $y/\mu$, and the power transformation is applied to these values. This central value could be the mean or median or any other value. You may be aware that the mean is not a good representative central value in the case of a skewed distribution—thus, other options remain for exploration. For example, if the weight of 4-year-old girls has mode = 13.6 kg, we may divide each weight by this value. Then values less than 13.6 would become less than 1.0, and values more than 13.6 would become more than 1.0. When square-root transformation is applied to these transformed values, for example, $\sqrt{(0.2)} = 0.447$ (increases by a factor of more than 2) and $\sqrt{(0.9)} = 0.949$ (increases very slightly). The effect is that the left tail of the distribution stretches where the transformed values are less than 1.0. On the positive side of the mode, for example, $\sqrt{(1.4)} = 1.18$ (reduces by 0.22) and $\sqrt{(2.4)} = 1.55$ (reduces by 0.85)—that is, the values on the positive side shrink, but larger values shrink more, thus correcting right skewness. Similarly, square transformation (or any power more than 1.0) tends to correct left skewness. The advantage of using $y/\mu$ is that it transforms the values to less than 1 on one side and more than 1 on the other side of $\mu$, and then transformation makes more sense. Note that square-root transformation is $(y/\mu)^{0.5}$ and square transformation is $(y/\mu)^2$. In general, the power transformation is $(y/\mu)^{\lambda}$, and the value of $\lambda$ depends on the type of skewness and extent of skewness as before. $\lambda < 1$ is for correcting right skewness and $\lambda > 1$ for left skewness. $\lambda = 1$ is no correction. If the distribution is already Gaussian, no correction is required, and then $\lambda = 1$.

Considering this, an extended form of power transformation is the following:

$$z_{\text{LMS}} = \frac{1}{\sigma_L \lambda}\left[\left(\frac{y}{\mu}\right)^{\lambda} - 1\right] \text{ for } y, \mu, \sigma_L, \text{ and } \lambda > 0,$$

where $\sigma_L$ is as explained shortly. This is the advanced form of the Box–Cox power transformation. The original measurements, such as weight in our example, may have any skewed distribution with a single mode; the distribution of $z_{\text{LMS}}$ with this transformation will be standard normal; and this will give the correct ***z*-score** for calculating the **percentiles** provided that the kurtosis is already 0. Note the involvement of lambda ($\lambda$), mu ($\mu$), and sigma ($\sigma_L$), making it an **LMS method**. Instead of lambda, mu, and sigma, it is customary to use L, M, and S.

The rationale of $(y/\mu)^{\lambda}$ is already explained, and $\sigma_L$ is in the denominator just as in $z = (y - \mu)/\sigma$. But in LMS, $\sigma_l$ is the **coefficient of variation** $\sigma/\mu$. Note that when $\sigma_L = \sigma/\mu$ and $\lambda = 1$, $z_{\text{LMS}}$ in the aforementioned equation reduces the usual $z$-score, namely, $(y - \text{mean})/\text{SD}$. This reinforces that $\lambda = 1$ means no correction for skewness.

For $\lambda = 0$, $z_{\text{LMS}}$ in this equation becomes 0/0 and is replaced by its mathematical equivalent, $[(1/\sigma_L)^*\ln(y/\mu)]$. In this case, this becomes **log transformation**. For such a transformation, negative values of $\lambda$ are ruled out, and negative values of $y$ or $\mu$ are also excluded since these can make $y/\mu$ negative, whose root (such as square root) does not exist. All medical parameters of the type we are discussing have positive values. For example, weight can never be negative. When the values are negative, for example, for a change from pretreatment to posttreatment or a difference between, say, right and left measurements, this transformation would work only after adding slightly more than the minimum difference so that all values are more than 0. If minimum difference is –3, add 3.1 to all the differences and then use this transformation.

Indrayan [2] has provided an easy nonmathematical explanation of this method.


1. Box GEP, Cox DR. An analysis of transformations. *J Royal Statistical Soc Series B* 1964;26(2):211–52. http://www.jstor.org/discover/10.2307/2984418?sid=21105162011671&uid=4&uid=2
2. Indrayan A. Demystifying LMS and BCPE methods of centile estimation for growth and other health parameters. *Indian Pediatr* 2014 Jan;51(1):37–43. http://www.indianpediatrics.net/jan2014/jan-37-43.htm


## Box *M* test

In the case of **repeated measures ANOVA** and **multivariate ANOVA (MANOVA)**, among other things, you need to confirm homogeneity of covariance matrices (also called the **dispersion matrix**) across groups when there are two or more groups, particularly when the sample sizes are different in different groups. To check homogeneity, the Box *M* test [1] is used. The basic requirement for a Box *M* test is **multivariate Gaussian** (normal) **distribution** of the values under consideration. The sample size in each group should be large (say >20), and the number of groups should not be too many (≤5). It appears that this test is rarely used in medical literature.

George Box

The Box *M* test is naturally based on the sample dispersion matrices and the pooled dispersion matrix under the null hypothesis. A statistical software package will do this easily and convert this to an *F*-test for the final result. When significant (say, $P < 0.05$), this might indicate that the dispersion matrices are unequal but also could be due to the fact that distributions are not Gaussian. One way to get around this problem in the case of large samples is to equalize the sample sizes in different groups by, say, randomly dropping some observations with a larger sample. The requirement of homogeneity of dispersions is not that great for equal sample sizes. But the better method for checking the homogeneity of dispersion matrices is the multivariate analog of the **Levene test**. This test works better, as this is based on the median. If the dispersions are not homogeneous, use a **Pillai trace** in place of the **Wilk lambda** for testing equality of means in a multivariate setup.

1. Box GEP. A general distribution theory for a class of likelihood criteria. *Biometrika,* 1949;36,317–46. http://www.jstor.org/discover/10.2307/2332671?uid=3738256&uid=2&uid=4&sid=21103403296627

## Box–Pierce and Ljung–Box *Q* test

The Box–Pierce and Ljung–Box tests are two versions of the same test, which is used to check that a series of values have any autocorrelation. If you are not conversant on the concept of **autocorrelation**, see this topic in this volume. As an example, in a monthly series of values with seasonal fluctuations (such as cases of dengue fever each month), the values in the month of September in different years are likely to be correlated. Hence, in a monthly series, this has autocorrelation of lag 12.

For the Box–Pierce and Ljung–Box tests, you have to decide the lag of autocorrelation you suspect and would like to check. This will be based on your knowledge about the series, such as lag 12 for seasonal variations.

If you decide lag $K = 5$, all autocorrelations of lag $K \le 5$ will be checked by this test. Denote these autocorrelations by $r_k$ ($k = 1, 2, 3, 4, 5$). In general, $k = 1, 2, \ldots, K$. Now,

$$\text{Box–Pierce test: } Q = n\sum_k r_k^2 ,$$

where the sum is over $k = 1, 2, \ldots, K$, and $n$ is the total number of values in the series. This criterion is treated as a **chi-square** test with $K$ degrees of freedom (df's). If not significant at the specified level of significance, all autocorrelations of lag $K \le 5$ can be considered not significant. This is taken to provide evidence that the series is random. This test was first proposed in 1978 [1].

The Box–Pierce test is a simplified version of

$$\text{Ljung–Box test: } Q = n(n+2)\sum_k \frac{r_k^2}{n-K} .$$

This also is checked with chi-square with $K$ df's [2].

1. Box GEP, Pierce D. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *J Amer Statistical Assoc* 1970;65:1509–26. http://www.stat.purdue.edu/~mlevins/STAT598K_2012/Box_Pierce_1970.pdf
2. Ljung G, Box GEP. On a measure of lack of fit in time series models. *Biometrika*. 1978;66:67–72. http://stat.wharton.upenn.edu/~steele/Courses/956/Resource/TestingNormality/LjungBox.pdf, last accessed September 8, 2015.

## Breslow–Day test

Sometimes, data are collected or broken into different **strata**. For example, you may have a history of hypertension (yes/no) for people with and without glaucoma, and this is divided by age (<40 years, 40–59, 60–79, and 80+) and sex (M/F). This will make a total of eight strata. Suppose you have the **odds ratio** (OR) of hypertension for each strata, but you are also interested in **pooled odds ratio** for all strata combined. This pooling is done after proper adjustment of strata sizes. However, the pooling is not valid if one or more ORs is very different from the others. They must be homogenous. For example, this homogeneity is required for the **Mantel–Haenszal test**. The Breslow–Day test is used to test the **homogeneity** of ORs across strata. Since OR is a measure of association, this test can also be used to check whether association between factors A and B is homogenous at different levels of factor C. The criterion used for this test has an approximate chi-square distribution with $(K - 1)$ degrees of freedom (df's), where $K$ is the number of strata. This is when each stratum has a $2 \times 2$ table. If the order of the contingency tables is higher, **log-linear models** can be explored.

Most statistical software will provide the result of the Breslow–Day test. This is based on a complex calculation requiring pooled OR by the Mantel–Haenszel method. For the Breslow–Day test to be valid, the sample size should be relatively large in each stratum, and at least 80% of the expected cell counts should be greater than 5. Note that this is a stricter sample size requirement than the requirement for the Mantel–Haenszel test itself, where this requirement applies to the overall sample size and not for each stratum. Tarone [1] derived an adjustment factor that is subtracted from the Breslow–Day statistic, resulting in an asymptotically chi-squared statistic and relaxing the stricter sample size requirement. This correction is also available in most statistical software. For further details of the Breslow–Day test, see Breslow and Day [2].

Woo et al. [3] used the Breslow–Day test to determine whether hypercholesterolemia and apolipoprotein E polymorphisms affect the risk of intracerebral hemorrhage by statin use. Ko et al. [4] used this test to study ALPK1 genetic regulation and risk in relation to gout.

1. Tarone, RE. On heterogeneity tests based on efficient scores, *Biometrika* 1985;72:91–5. http://www.jstor.org/stable/2336337?seq=1#page_scan_tab_contents
2. Breslow NE, Day NE. *Statistical Methods in Cancer Research: Vol. I—The Analysis of Case–Control Studies*. International Agency for Research on Cancer, Lyon, 1980.
3. Woo D, Deka R, Falcone GJ, Flaherty ML, Haverbusch M, Martini SR, Greenberg SM, Ayres AM et al. Apolipoprotein E, statins, and risk of intracerebral hemorrhage. *Stroke* 2013 Nov;44(11):3013–7. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3873717/
4. Ko AM, Tu HP, Liu TT, Chang JG, Yuo CY, Chiang SL, Chang SJ et al. ALPK1 genetic regulation and risk in relation to gout. *Int J Epidemiol* 2013 Apr;42(2):466–74. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3695596/

## Breslow test

A commonly used test for comparing survival pattern in two groups, say under different treatment regimens, is the **log-rank test**. This test gives the same importance to different time points even when one time point has more subjects than the others. Sometimes, time points with a higher number of subjects may need to be given more weight since those may have better **reliability**. In that case, another method, such as the Breslow test [1], can be used for comparing survival curves. This is also known as the *Gehan–Breslow test* and **generalized Wilcoxon test**. For details, see Hosmer et al. [2].


Norman Breslow

B

Since the Breslow test gives more weight to the time points with a higher number of subjects, initial time points where the number of subjects at risk is higher tend to determine the statistical significance. As time passes, subjects die or drop out, and the numbers become smaller. In practice, the difference between survival at initial stages in the two groups under comparison probably would be negligible, and it is only at later stages that differences would become pronounced. In this situation, the Breslow test may lead to incorrect conclusions. As a middle path between the log-rank and Breslow tests, some prefer the **Tarone–Ware test** [3], which gives weight proportional to $\sqrt{n_t}$, where $n_t$ is the number at risk at time point $t$.


1. Breslow NE. Covariance analysis of censored survival data. *Biometrics* 1974;30:89–99. http://www.jstor.org/discover/10.2307/2529620?uid=3738256&uid=2&uid=4&sid=21103028795773
2. Hosmer DW, Lemeshow S, May S. *Applied Survival Analysis: Regression Modeling of Time to Event Data,* Second Edition. Wiley Interscience, 2008.
3. Tarone RE, Ware J. On distribution-free tests for equality of survival distributions. *Biometrika* 1977;64:156–60. http://www.jstor.org/discover/10.2307/2335790?uid=3738256&uid=2&uid=4&sid=21103028795773


## Broca index

The Broca index is probably the simplest measure of excess or insufficient weight in otherwise healthy adults. According to this index, normal weight in kilograms is height in centimeters minus 100.

Broca index for normal weight (kg) = height (cm) − 100.

If a person's height is 172 cm, his/her normal weight should be 72 kg. This can be considered a formula for calculating *ideal weight* in adults. You can see that this is fairly simplistic but works reasonably well for both men and women for general assessment. A suggested modification for women is to subtract 105 from height in cm instead of 100. However, the Broca index may be too simple and not appropriate for specialized applications such as medication dosing.

Paul Broca
(Archives Serge NICOLAS, Paris Descartes University, with permission.)

The formula was first proposed by a French surgeon, Paul Broca, and appears in *Memoires d'anthropologie*, Paris, 1871.

## Brown–Forsythe test

The Brown–Forsythe test is an alternative to the usual ANOVA $F$-test when the requirement of homogeneity of variances is not satisfied.

The most commonly used test for comparing means of three or more groups is the ANOVA ***F*-test**. The requirements for the validity of this test are, among others, that the underlying distribution of the variable under consideration is **Gaussian** (normal) and that the variances in different groups are **homogeneous**. Having equal sample sizes in different groups is an additional help. However, the $F$-test is **robust** in the sense that it works reasonably well for minor violation of these requirements. The requirement of Gaussianity can be further relaxed when the sample size is large because of the **central limit theorem**.

Except when the data are highly skewed or when most group sizes are less than 10, the **Welch test** performs well for testing equality of means in situations of unequal variances and unequal $n$'s. Group sizes between 6 and 10 are your call. When group sizes are really small such as $n < 6$ for most groups, and you have unequal $n$'s and unequal variances, the Brown–Forsythe test [1] is used. This test uses the median in its calculation instead of the mean—thus, it is less sensitive to departures from assumptions. The mathematical procedure for both is complex. The names are given so that you can make an appropriate choice while running a statistical package.

There is another test with the same name but used for testing homogeneity of variances [2] in situations where the **Levene test** is not valid.


1. Brown MB, Forsythe AB. The small sample behavior of some statistics which test the equality of several means. *Technometrics* 1974;16:129–32. http://www.jstor.org/discover/10.2307/1267501?uid=3738256&uid=2&uid=4&sid=21103403296627
2. Brown MB, Forsythe AB. Robust tests for the equality of variances. *J Amer Statistical Assoc* 1974;69:364–7. http://www.jstor.org/discover/10.2307/2285659?uid=3738256&uid=2&uid=4&sid=21103403296627


## bubble charts

A bubble chart is an extension of a **scatter diagram** where a third dimension is added that signifies size in some sense. A scatter diagram is a plot of points $y$ versus $x$, where both are quantitative but can be either positive or negative or zero. In a bubble chart, these points become bubbles of the size in the third dimension—thus, size cannot be negative. A bubble chart is more effective when the size can be interpreted as an outcome of $x$ and $y$. However, the bubbles do not give a good indication of the actual value represented by each bubble.

Consider the percentage of deaths due to cancer at different ages in the years 1970–2010 in a particular country. In Figure B.11, age at death is on the $y$-axis and year on the $x$-axis. The percentage of death is represented by the size of the bubble. The figure shows that

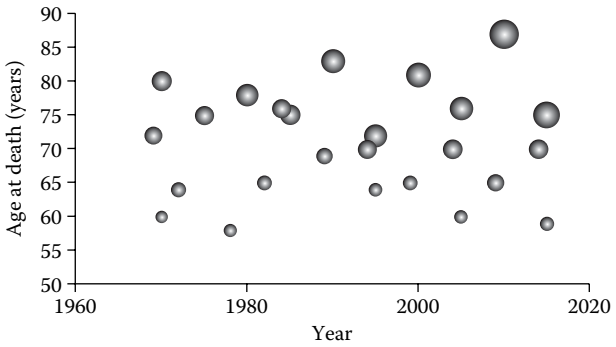



**FIGURE B.11** Percentage of deaths due to cancer in a country at different ages from 55 to 90 years in the years 1970–2015.

the percentage of deaths due to cancer has increased in this country not just as age is increasing but also as time has progressed. The increase in percentage of deaths due to cancer is more pronounced in recent years than in previous years.

More dimensions can be added, although that will make the chart more complex and difficult to understand. For example, you can have hollow bubbles for males and solid ones for females so that the trend in both the sexes can be studied. Further dimensions can be represented by different colors of bubbles. Each of the bubbles can even be a pie diagram to show percentages due to various causes. For an actual example of the use of a bubble chart, see the work of Chien et al. [1], who advocate bubble charts to enhance adherence to quality-of-care guidelines for colorectal cancer patients.

1. Chien TW, Lin YF, Chang CH, Tsai MT, Uen YH. Using a bubble chart to enhance adherence to quality-of-care guidelines for colorectal cancer patients. *Eur J Cancer Care (Engl)* 2012 Nov;21(6):712–21. http://www.ncbi.nlm.nih.gov/pubmed/22335545

## burden of disease, *see also* disability-adjusted life years (DALYs)

Although initially coined for a specific measure of ill health in a population, *burden of disease* is now a generic term for how severely a specific population is affected by one or a group of diseases. An easy and possibly most commonly used measure of burden of a particular disease is the **age-specific** and overall **mortality** from that disease. For example, according to the recent **global burden of disease** study [1], noncommunicable diseases (NCDs) account for more than two-thirds of all deaths across the world, and they have increased over the 40 years from 1970 to 2010. This can be interpreted as the burden of NCDs in the world. One can also say that nearly 80% of the total burden of all diseases is borne by low- and middle-income countries [2]. Next, burden can be assessed in terms of morbidity caused by the disease. This can be assessed in terms of **incidence** and **prevalence**. Duration and severity of sickness can also be included in this assessment. The higher the morbidity, the greater is the burden. Although the scenario is changing, developing countries have greater burden from communicable diseases than from NCDs. The situation in the developed world is the reverse.

Perhaps the most appropriate use of the term is when the impacts of both mortality and morbidity are combined. This was initiated by the World Bank in the early 1990s for almost all countries of the world and is called global burden of disease (GBD) [3]. This was put forward in a report jointly produced by Harvard University and the World Health Organization (WHO) [4]. The metric used to measure the burden of disease in this report is **disability-adjusted life years** (DALYs) lost due to a particular disease or a group of diseases. This metric combines different indicators of morbidity as well as mortality, takes into account the age of the affected person, and discounts future years lost to the present in case of death. The latter two adjustments proved controversial and have now been removed in the latest report, this time by the Institute of Health Metrics and Evaluation, called GBD 2010. This also changed from an incidence perspective to prevalence perspective, i.e., the morbidity is assessed on the basis of its existence instead of its new appearance. **Disability weights** have been revised, generally downward, and more disease conditions have been added.

The GBD 2010 is the most comprehensive study of its kind, producing comparative metrics for 291 different causes of premature death and disability across 187 countries, 20 age groups, and both sexes for three time periods: 1990, 2005, and 2010. The study also estimated 67 potentially preventable causes of ill health, or risk factors, such as smoking, high blood pressure, and household air pollution [1]. In addition, the WHO has released their estimates for the year 2012. According to these estimates, age-standardized DALYs lost in Afghanistan was 68,970 per 100,000 population in 2012 compared with, for example, 17,696 DALYs lost in Australia [5]. Note the big difference.

The burden of diseases in almost all countries of the world is steadily but surely declining, with the contribution of NCDs increasing and that of other diseases declining. The increase of NCDs is mostly due to ageing of populations across the world and partly contributed to by a more adverse lifestyle. Communicable diseases are declining due to control measures such as immunizations and improvement in environmental conditions. The contribution of accidents and traumas seems to be fairly stable over time and across countries.

### *Salient Features of the Burden-of-Disease Methodology*

The burden-of-disease methodology is intricate and cannot be fully described here. For a complete methodology for computing burden of disease by the WHO, see Ref. [6]. For details of the metric used to assess burden of disease, see the topic **disability-adjusted life years (DALYs)**. Other salient features that can help in the understanding the essentials are as follows.

The burden-of-disease methodology divides all causes of mortality and morbidity into three broad groups: group 1 (communicable, maternal, perinatal, and nutritional conditions); group 2 (NCDs); and group 3 (injuries). Each of these is divided into groups and subgroups. For example, skin diseases are subdivided into psoriasis, acne, eczema, scabies, urticaria, etc. These are all mutually exclusive and exhaustive categories so that there is no duplication and nothing is left out. Age-sex-specific morbidity (including severity) and mortality data on all these conditions are obtained from national health statistics, surveys, research studies, etc., and they are systematically compiled and collated to provide national estimates. Many times, models are developed that can provide reasonable estimates on the basis of **surrogates**. Separate efforts are made to estimate the disability weights for each severity condition of each disease. For countries with little or no data, data from neighboring countries with similar conditions are extrapolated. Special attention is paid to the **epidemiological consistency** of various estimates and their **validity**. These estimates are used to calculate the DALYs lost as estimates of the burden of disease in each country for each condition of ill health in different age-sex groups. Despite huge efforts involved, there is always uncertainty, and **uncertainty intervals** are also computed. This is clearly an enormous exercise, but the good news is that the burden of disease is estimated with tremendous success by a large team of professionals. Subnational efforts are made by the respective countries.

1. Institute of Health Metrics and Evaluation. *The Global Burden of Disease: Generating Evidence, Guiding Policy*. Institute of Health Metrics and Evaluation and University of Washington, 2013. http://www.healthdata.org/policy-report/global-burden-disease-generating-evidence-guiding-policy, last accessed July 23, 2015
2. World Health Organization. *Global Status Report on Non-Communicable Diseases 2010*. Geneva, 2011: p. vii. http://www.who.int/nmh/publications/ncd_report2010/en/
3. World Bank. *World Development Report 1993: Investing in Health*. Oxford University Press, 1993; pp. 290–1. http://files.dcp2.org/pdf/WorldDevelopmentReport1993.pdf
4. Murray JL, Lopez AD (eds.). *Global Burden of Disease: A Comprehensive Assessment of Mortality and Disability from Diseases, Injuries, and Risk Factors in 1990 and Projected to 2020*. Harvard University Press, 1996.

5. World Health Organization. *Burden of Disease*. http://gamapserver.who.int/gho/interactive_charts/mbd/as_daly_rates/atlas.html, last accessed July 13, 2015.
6. World Health Organization. *WHO Methods and Data Sources for Global Burden of Disease Estimates 2000–2011*. WHO, 2013. http://www.who.int/healthinfo/statistics/GlobalDALYmethods_2000_2011.pdf?ua=1

B

## butterfly effect

An enormous effect on the outcome resulting from very minor changes in input values is called the butterfly effect. This can happen due to a ripple effect of chain reactions as in the chaos theory. For example, Dorn [1] has mentioned the following: "Regulated microRNA expression levels, as during a tissue stress response, induce small changes in initial conditions that are sequentially amplified through secondary and higher-order interactions, producing a systemic ripple effect that ultimately invokes disproportionately large (and frequently unanticipated) phenotypes."

The term *butterfly effect* is believed to have been first used by Lorenz in the context of weather forecasting when he observed tremendous changes in the forecast when the input was 0.506 instead of the more accurate figure of 0.506127 [2]. Lorenz's own paper [3] was presented in a conference with no reference to these values, but he is credited to have coined this term. In biostatistics, this can happen in modeling when a slight variation in the value of one input variable substantially changes the expected outcome. This can happen when the variable has high power or is an exponent. For example, $x^8 = 1{,}280{,}631$ for $x = 5.80$ and 1,298,402 for $x = 5.81$—a difference of 0.01 in the value of $x$, which is not even 0.2% of $x$, has made a difference of 17,771, which is nearly 1.4%, 7 times of 0.2%. Similarly, $2^x = 65{,}536$ for $x = 16$ but $2^x = 131{,}072$ for $x = 17$, i.e., if $x$ is increased by 1, $2^x$ becomes double. Figure B.12 illustrates this for cancer. A small difference in the initial value has changed the entire course of the disease. During carcinogenesis, a minor change in one host cell out of millions can have a cascading effect with serious consequences in the course of time [2].

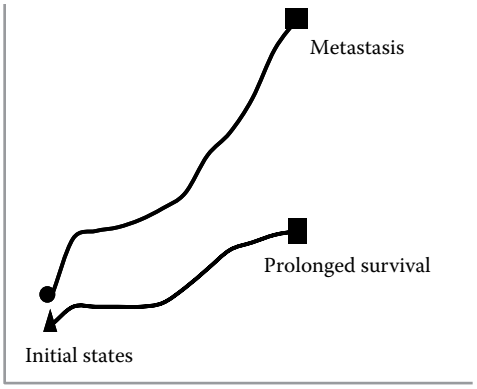



**FIGURE B.12** Butterfly effect in cancer.

A similar issue can arise when a model involves a small proportional difference between two large numbers. If, for example, a calculation involves $d = (x - y)$, if $x$ is 10,010 and $y$ is 10,000, then $d$ will be 10. However, if $y$ decreased by just 1%, to 9900, then $d$ would increase to 110, a massive increase of 1100%. Because of a possible butterfly effect, statistical models involving functions such as powers and exponents, or small proportional differences between large numbers, need special care. In particular, the input value should be fairly accurate to give correct answers.

1. Dorn GW 2nd. MicroRNAs and the butterfly effect. *Cell Cycle*. 2013 Mar 1;12(5):707–8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3610711/pdf/cc-12-707.pdf
2. Dey P. Butterfly effect and cancer. *Indian J Pathol Microbiol* 2011;54(2):435–6. http://www.ijpmonline.org/article.asp?issn=0377-4929;year=2011;volume=54;issue=2;spage=435;epage=436;aulast=Dey, last accessed July 22, 2015.
3. Lorenz EL. Predictability: Does the flap of a butterfly's wings in Brazil set off a tornado in Texas? *American Association of Advancement of Science 139th Meeting, 1972*. http://eaps4.mit.edu/research/Lorenz/Butterfly_1972.pdf

# C



## calibration

Calibration is the process of adjusting the setting of a tool according to a defined standard so that the readings provided by this tool have increased validity, i.e., they can be believed to be correct. This process is intimately related to quality control; it increases the confidence in the performance of the tool. A new tool must always be calibrated anyway, but an old tool also is calibrated when it is suspected to have "rusted" due to repeated use, misuse, or long nonuse. This clearly applies to laboratory and clinical instruments such as a biochemical analyzer and a sphygmomanometer. A questionnaire, say for assessing quality of life, also needs calibration when translated to another language so that it provides the same response as the original standard would in the identical setting.

For statistical measurements, calibration is the process of adjusting $x$ against a standard $y$. Simply stated, the first stage involves measuring both $x$ and $y$ for a sample of $n$ subjects and using these to develop a relationship between $x$ and $y$. In the second stage, this relationship is exploited to predict $y$ for other subjects by using the observed $x$. Thus, the standard values along with the observed values must be available from the first stage. Statisticians prefer that this relation be a **regression** equation, but not of $y$ on $x$ as usually done but of $x$ on $y$. This is because $y$ is standard and known in this setup. The most elementary regression of $x$ on $y$ is $x = a + by$, where $a$ and $b$ are estimated as in **simple linear regression**. Once this equation is obtained to your satisfaction in the sense that the **residuals** are negligible and randomly distributed around zero, then

$$\text{the calibrated value of } x = \frac{\text{observed value of } x - a}{b}.$$

This is an inverse of the regression equation and is frequently called an *inverse regression*. In a complex form, the regression equation may involve terms such as square, reciprocal, and logarithm. The inverse relationship in that case would not be linear but a curve, called the *calibration curve*.

Mossavar-Rahmani et al. [1] describe calibration of a food frequency questionnaire reporting energy intake using water and urinary nitrogen as biomarkers of objective measures of total energy expenditure. The latter was the standard they used. The statistical method they used is much more complex than just described, but the example illustrates a situation where calibration can be useful. Food frequency questionnaire is known to provide flawed estimates of energy intake, and the reporting can be calibrated to give a more correct estimate of the energy intake.

The method can be extended to calibration between two tools, both of which can be in error [2]. Neither is standard, but both can be standardized to give the same value for the same subjects. The objective in this setup is to increase the likelihood that both tools give the same or nearly the same readings.


1. Mossavar-Rahmani Y, Tinker LF, Huang Y, Neuhouser ML, McCann SE, Seguin RA, Vitolins MZ, Curb JD, Prentice RL. Factors relating to eating style, social desirability, body image and eating meals at home increase the precision of calibration equations correcting self-report measures of diet using recovery biomarkers: Findings from the Women's Health Initiative. *Nutr J* 2013 May 16;12:63. http://www.nutritionj.com/content/pdf/1475-2891-12-63.pdf
2. Osborne C. Statistical calibration: A review. *Int Stat Rev* 1991;59: 309–36. http://www.jstor.org/discover/10.2307/1403690?uid=3738256&uid=2&uid=4&sid=21103337240357


## canonical correlation

Canonical correlation is the correlation between one set of variables and another set of variables. Thus, this is a **multivariate** method. For example, you may have wondered if liver and kidney functions have any correlation, and if yes, how much. In this case, one set is liver function tests, say, levels of albumin, bilirubin, prothrombin time, aspartate transaminase, and alkaline phosphatase, and the second set comprises kidney function tests such as serum creatinine, creatinine clearance, blood urea nitrogen, and glomerular filtration rate. In this example, liver function is a five-measurement entity and kidney function is a four-measurement entity. How do you find the correlation between sets of measurements? The answer is canonical correlation.

The ordinary (Pearsonian) **correlation coefficient** assesses the degree of (linear) correlation between two measurements, and the **multiple correlation** coefficient assesses the degree of correlation between one measurement and a set of two or more measurements. Multiple correlation, in fact, is between one variable and the linear combination of the set of other variables. As an extension, the canonical correlation is between the linear combination of one set of variables and the linear combination of the other set of variables. Consider a set of $J$ variables $(x_1, x_2, \ldots, x_J)$ and a second set of $K$ variables $(y_1, y_2, \ldots, y_K)$. Then the canonical correlation is the correlation between $t_1 = a_1x_1 + a_2x_2 + \ldots + a_Jx_J$ and $t_2 = b_1y_1 + b_2y_2 + \ldots + b_Ky_K$. This was first proposed by Harold Hotelling in 1936 [1].

As in multiple correlations, these linear combinations are chosen in a manner that the correlation between these combinations is maximum. In our example, this means the correlation between a linear combination of the five liver function values and a linear combination of four kidney function values is the canonical correlation where linear combinations are those that maximize the correlation. If this correlation is more than, say, 0.7, you can say that liver and kidney functions are closely related, and if this is less than, say, 0.3, you can say that these functions have weak relation. Remember though, as in all Pearsonian correlations, that these are for linear relationships. Quadratic or other kinds of nonlinear relations are not properly assessed by this correlation. It suffers from the same deficiencies as listed for **correlation**.

C

An example of the use of a canonical correlation is between scores of the 11 subscales of the pattern element scale (PES) and 6 cognition scores in dementia patients in China [2]. The authors reported this to be 0.507 and statistically significant ($P < 0.001$). They concluded that PES, which is a relatively brief tool, is helpful in discriminating pattern elements in dementia. This conclusion is a bit too farfetched but illustrates how canonical correlation can be used.

The linear combinations $t_1$ and $t_2$ that provide maximum correlation are called *canonical variates*. These variates can be used as univariate entities and analyzed as usual for inference. However, they may lack substantive biologic meaning. Though extension is seldom done, canonical correlation does not stop with a single relationship between the sets of variables. You can find other pairs of canonical variates that are independent of the previous pairs, but they will provide lower correlation. For further details, see Ref. [3].


1. Hotelling H. Relations between two sets of variates. *Biometrika* 1936;28:321–77. http://www.jstor.org/discover/10.2307/2333955?uid=3738256&uid=2&uid=4&sid=21103363979653
2. Shi J, Tian J, Long Z, Liu X, Wei M, Ni J, Liu J et al. The pattern element scale: A brief tool of traditional medical subtyping for dementia. *Evid Based Complement Alternat Med* 2013;2013:460562. http://www.hindawi.com/journals/ecam/2013/460562/
3. Hair JF, Jr., Anderson RE, Tatham RL, Black WC. Canonical correlation analysis. Adapted from Chapter 8, *Multivariate Data Analysis*, Fifth Edition. Prentice Hall, 1998.


## capture–recapture method

Capture–recapture is a simple method to estimate the total size of a population by a particular kind of **resampling**. The methodology was originally devised for obtaining counts of animals in a closed area. Consider a sample of 48 deer ($n_1 = 48$) that are captured, marked, and then released back to their area for mixing up with other animals. They mix with other deer. Subsequently, a second sample of 35 deer ($n_2 = 35$) is captured from the same area. If 16 ($m = 16$) of them are found to have the mark, an estimate of the total number of deer is obtained by inflating by a factor of 35/16. Thus, an estimate of the total count $N$ is $48 \times 35/16 = 105$. A statistically better (**unbiased**) estimate in the long run is obtained when 1 is added to each of these numbers and finally 1 is subtracted. Thus,

$$\text{capture} - \text{recapture estimate } N = \frac{(n_1+1)(n_2+1)}{m+1} - 1.$$

where $n_1$ is the size of the first sample and $n_2$ of the second sample, and $m$ is the number of cases found to appear in both the samples. Note that it requires only basic arithmetic. For our example, this gives $N = (49 \times 36/17) - 1 = 102$, which may be more realistic than the value 105 arrived earlier.

The capture–recapture methodology was first applied for estimating the size of the human population in the 1940s [1]. But there is evidence that Laplace first used essentially the same method to estimate the population of France. This publication appeared in 1783. He used the count of newborn babies and population/births ratio of certain areas to come up with an estimate of the population [2].

The capture–recapture method allows us to estimate the size of the population not just where direct census is impossible or impractical but also at a much cheaper cost of census even if possible. This helps to estimate the total number of cases when their incomplete count is available from two or more *independent* sources. In health and medicine, these sources could be hospital records, physicians in private practice, death certificates, or any other such list of cases. The count of duplicate cases, which appear in more than one list, can help to substantially improve the estimate of the total cases. Hook and Regal [3] argued that the capture–recapture methodology can improve prevalence estimates even for apparently exhaustive surveys. It can also be used to estimate the size of hidden population such as the number of drug users in a district.

The methodology was adopted by McCarty et al. [4] to assess the prevalence of childhood diabetes in Madrid, Spain. A population-based registry identified 432 cases through hospital inpatient records. Another source was the Spanish Diabetes Association, which recorded 138 cases. It was found on matching that 119 cases were common to the two sources. Thus, an estimate of the total number of cases of childhood diabetes is

$$\frac{(432+1)(138+1)}{119+1} - 1 = 501.$$

The results can be converted to prevalence rates as shown in Table C.1.

The cases in either source are undercounts, but the duplicates helped to come up with an estimate of the total cases as well as an improved prevalence rate. This also helps to de-emphasize the need for all registers to be complete.

Note the following for capture–recapture methodology:

- The methodology assumes that there are no intermediary additions or deletions. If there are any, the estimate may have to be revised accordingly.
- It is necessary that the captured subjects "move around freely" and are homogeneously mixed when "released." The two sources should be independent of one another. This is not fulfilled in a situation where, for example, some cases seen by the first source tend to be referred to the second. In such cases, the capture–recapture estimate may be too low. In the case of cancer deaths, if the two sources are death certificates and hospital discharge records, more severe cases are more likely to appear in both the lists—they are not independent.
- Like all estimates, the estimate of the total count by the capture–recapture methodology is subject to sampling fluctuation. Methods are available to find the standard error of this estimate and to construct a 95% confidence interval. If interested, see the work of Laporte et al. [5]. For further reading, see Krebs [6].

**TABLE C.1**
**Capture–Recapture Estimate of Childhood Diabetes in Madrid**

| | Hospital Records | Diabetes Association | Common | Capture–Recapture Estimate |
|---|---|---|---|---|
| Number of cases | 432 | 138 | 119 | 501 |
| Prevalence rate per 100,000 population (total population 4.4 million) | 9.8 | 3.1 | – | 11.4 |
| Extent of undercount (as percent of the last column) | 14% | 72% | – | 0% |

1. Tilling K. Capture–recapture methods—Useful or misleading? *Int J Epidemiol* 2001;30:12–4. http://ije.oxfordjournals.org/content/30/1/12.full.pdf+html
2. Amoros J. Recapturing Laplace. *Significance* 2014;11(3 July):38–9. http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2014.00754.x/abstract
3. Hook EB, Regal RR. Capture–recapture methods in epidemiology: Methods and limitations. *Epidemiol Rev* 1995;17:243–64. http://epirev.oxfordjournals.org/content/17/2/243.full.pdf
4. McCarty DJ, Tull ES, Moy CS, Kwoh CK, LaPorte RE. Ascertainment corrected rates: Applications of capture–recapture methods. *Int J Epidemiol* 1993;22:559–65. http://ije.oxfordjournals.org/content/22/3/559.full.pdf
5. Laporte RE, Tull ES, McCarty D. Monitoring the incidence of myocardial infarctions: Applications of capture–mark–recapture technology. *Int J Epidemiol* 1992;21:258–62. http://ije.oxfordjournals.org/content/21/2/258.full.pdf
6. Krebs CJ. *Ecology: The Experimental Analysis of Distribution and Abundance*, Sixth Edition. Benjamin Cummings, 2009.

## carryover effect, *see* crossover designs/trials

## cartogram

A cartogram is a map that depicts the size of various areas in proportion to the magnitude of the problem in hand. A cartogram of 35 ECG leads is used in precordial mapping for diagnosis of combined myocardial ventricular hypertrophy. The concern in this volume is mostly with statistical mapping and not medical mapping as for ECG. Sometimes it is convenient to show the area of a city or a country in proportion to the population or the number of cases occurring in that city or country. The actual boundaries are approximated, although efforts are made to retain the shape. An example of such a cartogram is Figure C.1, which shows the number of persons living with AIDS in 2007 in each state of the United States [1]. Note how Florida is oversized and Washington state is drastically undersized.

Lovett et al. [2] have discussed cartogram as a health service quality improvement tool in London, and Kotviski and Barbola Ide [3] have prepared one for incidence of scorpion stings in the city of Ponta Grossa in Paraná State of Brazil.

1. CodieMaps. *Using a Cartogram to Map AIDS in the USA, 2011*. https://codiemaps.wordpress.com/tag/cartogram/, last accessed April 15, 2015.
2. Lovett DA, Poots AJ, Clements JT, Green SA, Samarasundera E, Bell D. Using geographical information systems and cartograms as a health service quality improvement tool. *Spat Spatiotemporal Epidemiol* 2014 Jul;10:67–74. http://www.sciencedirect.com/science/article/pii/S1877584514000306
3. Kotviski BM, Barbola Ide F. [Spatial distribution of scorpion stings in Ponta Grossa, Paraná State, Brazil]. [Article in Portuguese] *Cad Saude Publica* 2013 Sep;29(9):1843–58. http://www.scielosp.org/scielo.php?script=sci_arttext&pid=S0102-311X2013001300023&lng=en&nrm=iso&tlng=en

## case–control studies, *see* retrospective studies

## case-fatality rate

The case-fatality rate (CFR) is the rate at which affected persons die. In terms of percentage, this can be stated as

$$\text{case-fatality rate (CFR)} = \frac{\text{deaths among those affected}}{\text{number of individuals affected}} * 100.$$

The CFR represents the severity of a disease in terms of the associated mortality. Thus, this measures the most important aspect of

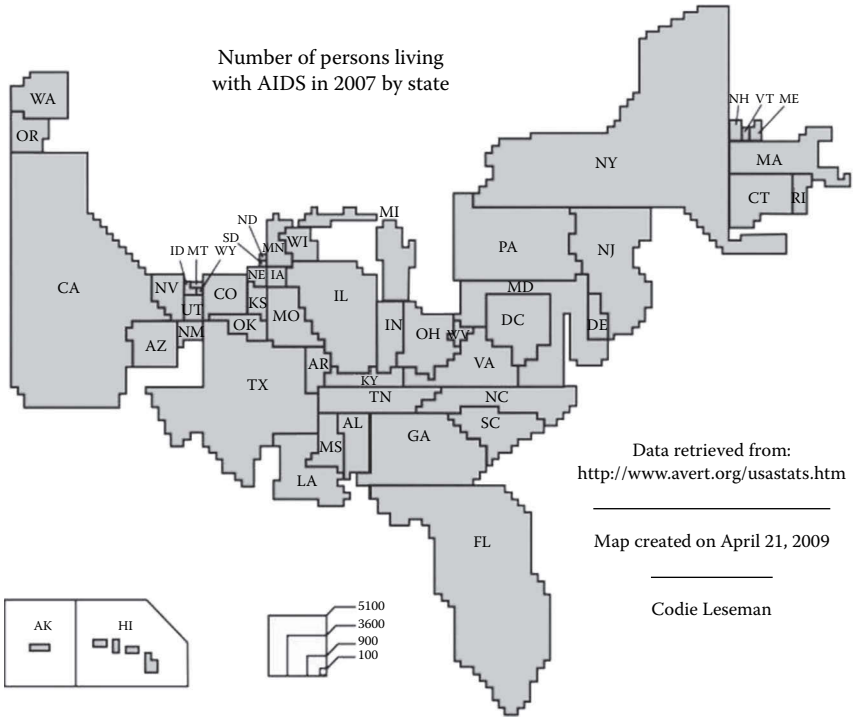



**FIGURE C.1** Cartogram: number of persons living with AIDS in 2007 by state in the United States. (From CodieMaps. *Using a Cartogram to Map AIDS in the USA, 2011*. https://codiemaps.wordpress.com/tag/cartogram/, last accessed April 15, 2015. With permission.)

C

prognosis that is important for both the patient and the doctor. The CFR is high for diseases such as rabies and tetanus, and low for diseases such as typhoid and influenza. Case fatality can be considered to measure the virulence of a disease and can help in planning control measures. The CFR is intimately related to duration, for example, 1-year case fatality of leukemia is very different from a 5-year case fatality. For this reason, the CFR is customarily used for diseases of short duration such as peritonitis.

Since everybody dies anyway, case fatality must have a qualifier for duration within which death would be counted. But this is hardly ever done. Case fatality rate is mostly used in hospital setups where some events are bound to be fatal during hospital stay despite best care. But some fatalities may be avoided. Case fatality per 1000 patients admitted can be an indicator of the quality of care in a hospital, particularly when calculated separately for each disease or condition. It may reflect the ability to manage critical patients with different diseases. There is a rider though. Because some hospitals are specialized to treat serious and terminal cases, interhospital comparison on the basis of gross fatality may not be valid. The same is true for different units of a hospital. However, trend within a hospital can certainly be monitored over time using this indicator. Mortality within 48 h in many cases is determined by the condition of the patient at the time of admission instead of the quality of medical care. When mortality within 48 h is excluded, this rate is called the *net fatality rate* or *net death rate*. This is relatively more valid for comparison between units or between hospitals.

The CFR in a hospital should be interpreted with caution because **Simpson paradox** may operate. This is illustrated by the data in Table C.2. In this example, case fatality in patients in two hospitals is the same on aggregate (19.8% and 20.0%), but glaring differences emerge when it is broken down by age and severity of illness of the patients. If one looks at the top row for the total alone in this table, it would seem that the two hospitals have nearly the same case-fatality rates. The breakdown by age in the next two rows shows that both the hospitals have nearly the same case-fatality rates in each age group also. But further breakdown by severity of illness in the subsequent rows shows that hospital A has less fatality in every group. Why then are the rates in the aggregate the same? The reason is that hospital A is catering much more to serious cases and to older cases. These have higher fatality although less than in hospital B for such patients. The breakdown clearly establishes better performance of hospital A. This was otherwise concealed by the aggregate rate in the top three rows of the table. Li et al. [1] give a similar example of CFR in the context of H5N1 avian influenza.

**TABLE C.2**
**Case-Fatality Rate (CFR) in Two Hospitals by Age of Patients and Severity of Disease**

| Severity and Age | Hospital A | | | Hospital B | | |
|---|---|---|---|---|---|---|
| | ***n*** | **Deaths** | **CFR (%)** | ***n*** | **Deaths** | **CFR (%)** |
| *Total* | *800* | *158* | *19.8* | *600* | *120* | *20.0* |
| –60 years | 280 | 38 | 13.6 | 340 | 48 | 14.1 |
| 60+ years | 520 | 120 | 23.1 | 260 | 72 | 27.7 |
| *Mild* | *200* | *28* | *14.0* | *350* | *62* | *17.7* |
| –60 years | 60 | 5 | 8.3 | 190 | 21 | 11.1 |
| 60+ years | 140 | 23 | 16.4 | 160 | 41 | 25.6 |
| *Moderate* | *300* | *51* | *17.0* | *200* | *41* | *20.5* |
| –60 years | 120 | 13 | 10.8 | 120 | 19 | 15.8 |
| 60+ years | 180 | 38 | 21.1 | 80 | 22 | 27.5 |
| *Serious* | *300* | *79* | *26.3* | *50* | *17* | *34.0* |
| –60 years | 100 | 20 | 21.0 | 30 | 8 | 26.7 |
| 60+ years | 200 | 59 | 29.5 | 20 | 9 | 45.0 |

The reverse of what is seen in this example can also happen. There might be a large difference in aggregate mortality rates in two hospitals falsely showing that the quality of care in one hospital is better than in the other. This might vanish when adjustment for severity of cases is made. Silber and Rosenbaum [2] have discussed this aspect in detail.

An improvement of 1.5/1000 in case fatality in the case of stage IV cancer has a different implication than a similar improvement in cases of peritonitis. Such gains may look statistically the same but have different social implications. You should be careful in interpreting such improvements. Put value to the outcome and interpret accordingly.

Biostatistically, case fatality, along with remission, is an important consideration in **epidemiologic consistency** of **prevalence** and **incidence** of a condition. CFR, expressed per case, is a proportion and can be treated as a **binomial** variable for the purpose of finding the confidence interval and for test of hypothesis. For most diseases, this rate is small, and the same precautions are required as for any small $p$ in a binomial setup. For example, this may be unreliable for small samples. For a sufficiently large sample, **Gaussian** approximation can be used for statistical inferences.


1. Li FCK, Choi BCK, Sly T, Pak AWP. Finding the real case-fatality rate of H5N1 avian influenza. *J Epidemiol Comm Health* 2008;62(6):555–9. http://jech.bmj.com/content/62/6/555.abstract
2. Silber JH, Rosenbaum PR. A spurious correlation between hospital mortality and complication rates: The importance of severity adjustment. *Med Care* 1997;35(10 Suppl.):OS77–92. http://www.jstor.org/discover/10.2307/3767251?uid=3738256&uid=2&uid=4&sid=21103304694767


## case reports

The detailed documentation of an unusual case is called a case report. The details may include the demographic particulars of the patient, etiology of the condition, presenting signs and symptoms, treatment history, disease-related experiences of the patient, timeline, any complications, photographs, laboratory and radiological investigation reports, diagnosis, current treatment, outcome, and prognostic implications. The emphasis is generally on what is unusual about this case or what makes it so unique, and then what lessons can be drawn. Related references of the literature are provided in a case report. An informed written consent is taken from the patient or his/her relative, yet the reporting is done in a manner in which the identity of the patient remains undisclosed.

New information given in a case report can provide learning with regard to exceptions to the rules—thus making the medical fraternity aware of the unusual events that can possibly happen. Statistically, the role of case reports is only to provide instances of outliers. Such outliers can disturb the main theme of the findings. These can help to prepare exclusion criteria for subjects in a research study. Such criteria are specified at the time of writing a research **protocol**. If an unanticipated unusual case appears during the course of the study, this can be excluded from analysis and described separately as a case report.

## cases (selection of)

Although the term *case* is generic, a person with disease or a condition of interest is customarily called a case. The question of the selection of cases is relevant mostly for case–control studies because,

for these studies, random selection is not all that important, and the selection of cases can substantially affect the results. Case–control studies are discussed under the topic **retrospective studies**.

The source of cases with disease or any other outcome of interest can be hospital inpatients, patients seen as outpatients, cases identified in a survey, cases listed in records of a health facility, etc. Hospital-based studies may be biased for, say, subjects from upper socioeconomic class, whose nutritional status may be different, or for any other factor influencing hospital intake. Population-based studies can become very expensive. Each, however, has relevance in specific situations.

A case could be either newly diagnosed (incident) or an already existing (prevalent) subject with disease. Inclusion of prevalent cases, particularly for chronic disease, can easily increase the sample size, but care is required at the time of interpretation of results because the factors determining the duration of disease can be important. Note that duration decides that a case is prevalent or not at a particular point in time. Mild cases or those physically strong may survive longer and add to the prevalence. At the same time, cases surviving for long are more likely to have recall lapse. Newly diagnosed cases do not have any such problem and are thus preferable for research.

Selection bias is said to have occurred when the study group has a different composition with regard to etiologic factors such as heredity, age, gender, nutrition status, and addictions, compared with the composition in the target population. Studies on volunteers or on clinic subjects almost invariably suffer from such a bias. A method of selection that has a random component is considered insulation against such bias. But such selection fails to take cognizance of special bias that can result from sources such as improper definition of the population. In a study on causes of psychiatric illness in old age, if the subjects are those who are single and of age 70 years or above at the time of enrollment, the bias occurs because some with severe illness may have already expired before attaining the age of 70 years. Also, in the case of analytical studies such as in the case–control setup, random selection does not help much since the emphasis is on comparing apples with apples to come to a fair conclusion on association or causation [1]. It is important to properly sort relevant factors from irrelevant factors. Nonrandom selection of cases (and controls) can restrict the applicability of results to a wider population but can enable valid conclusion on antecedent–outcome relationship. Sometimes this is more important than generalization. External validity may have to be sacrificed to achieve internal validity. In experiments and clinical trials, the focus is on causal inference, and that is achieved more by baseline equivalence of cases and controls rather than by random selection from the population. Random allocation takes care of statistical need of probabilistic base for inference.

An important consideration in the selection of subjects for any study is the feasibility of obtaining data on them. This means that the subjects must be approachable and cooperative. For a case–control study, accurate and complete information must be available on the subjects so that they can be correctly classified into exposed and nonexposed groups, and the effect of other characteristics on the outcome can be properly assessed.

1. Rothman KJ, Gallacher JEJ, Hatch EE. Why representativeness should be avoided. *Int J Epidemiol* 2013;42:1012–4. http://ije.oxfordjournals.org/content/42/4/1012.full.pdf+html

## case series/studies

A case series is obviously a series of cases of specific type and is an extension of **case reports**. The cases share some common feature(s) to be part of a series. For example, they may be suffering from the same disease, or they may be undergoing the same regimen of treatment. In medical research, a case series signifies a **descriptive study**, although here no specific design for selection of cases is formulated. Neither a provision of control nor any protocol for allocation of treatment regimens is made. No cause–effect, not even associations, can be concluded by a case series, but they are capable of generating a hypothesis for planning an **analytical study**.

There is a question regarding the minimum number of cases required for it to be called a series. An analysis of case series articles appearing in the PubMed database published in 2009 found that the median number of cases reported per series was 7 and the range was 1–6432 [1]. The authors suggest that a case series should contain no less than four cases. If they are so few, the cases are individually described, but if they are many, the results can be summarized in terms of tables and figures with means, proportions, and rates. Inclusion and exclusion criteria should be specified in any case.

Popular among the case series is the first official documentation in 1981 of five young gay men with pneumocystis *Pneumonia carinii* in Los Angeles that were later identified to have HIV [2].

### *Case Studies*

A case study is the detailed study of an event and its origin, process, and consequences, with the purpose of deriving learning for wider application. Case studies have been found to be effective tools for teaching and learning various aspects of subjects such as their management with informatics technology. The primary advantage with this methodology is its flexibility, as case studies do not have to follow a fixed mold and can be done in a format suitable to study the event. For example, a case study can include results from several studies on that kind of event.

Medical case studies are sometimes confused with case reports. Indeed, the terminology has mixed use in the literature. A case report is for a single patient, whereas a case study is for a particular event. A case study of sudden rise in deaths from motor vehicle accidents within the precincts of a particular city would draw information from many sources and would try to reach to a reliable conclusion regarding the factors responsible for such a rise. This would not be so in a case report.

Biostatistics case studies can be on topics such as survival analysis and diagnostic evaluation of medical tools. A case study on survival analysis may discuss survival analysis methods used by various workers in different setups. The focus can be on the logical basis of the inferential methods used in different studies including validation or violation of the requirements of those methods. Alternative scenarios with modification in the data and alternative methods can be discussed, and the effect on conclusions can be indicated. As already stated, no specific format of case studies can be suggested to fit all.

1. Abu-Zidan FM, Abbas AK, Hefny AF. Clinical "case series": A concept analysis. *Afr Health Sci* 2012 December; 12(4):557–62. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3598300/
2. Avert. *History of HIV & AIDS in the U.S.* http://www.avert.org/history-hiv-aids-us.htm, last accessed April 23, 2015.

## casewise, pairwise, and listwise deletion

The question of deletion of cases arises in a research setup when part of the values is missing or wrongly recorded for some subjects in a dataset. Missing values cannot be used in any analysis unless we impute them. Thus, they are excluded. We illustrate various types

of deletion with the help of data in Table C.3. Out of, say, 200 cases of liver disease, only those with missing data are shown. Two cases with nothing missing are also shown for illustration. All others with complete data are not shown.

While calculating something like mean and standard deviation (SD), you can exclude case numbers 79 and 164 for total bilirubin level, case numbers 14, 127, 164, and 183 for AST level, and case numbers 127, 164, and 191 for ALT level. These are the cases for which that particular information is missing. This is called **listwise deletion**. The description in the literature varies. Realize though that under this deletion, different means and SDs would be based on different cases—thus, they would not be strictly comparable.

For calculating the correlation between the AST level and the ALT level, you will have to exclude case numbers 14, 127, 164, 183, and 191 for which any of the two or both are missing. This is called **pairwise deletion**. For correlation between total bilirubin level and AST level, excluded cases will be numbers 14, 79, 127, 164, and 183. Again the two correlations are not based on the same subjects—thus, some comparability is lost. For running polytomous **logistic regression** with disease as dependent and total bilirubin and ALT levels as regressors, not only case numbers 127 and 164 will be excluded since they do not have vital information on their disease but also case numbers 79, 127, 164, and 191 because at least one of the data on regressors is missing. This also is an extension of pairwise deletion. Under this deletion, if regressors are AST and ALT levels, case numbers 14, 127, 164, 183, and 191 will be excluded, but case number 79 will be included. Thus, the two logistic regressions will have different cases. If all three are regressors, all cases listed in Table C.3 except numbers 92 and 132 will be excluded.

In order to base all the calculations on the same subjects, cases with any one piece or more information missing are excluded. This will exclude six out of the eight cases listed in Table C.3 for calculation of mean/SD, correlations, as well as logistic regression. This is called **casewise deletion** since the entire case is deleted as though that case is not in the sample. Statistically, this is the most desirable kind of deletion but has the potential to exclude a large number of cases since one or the other piece of information may be missing for many cases. This reduces the sample size and affects the statistical **power** to detect the kind of result you are hoping for. Also, the cases with incomplete information may be of specific type, and their exclusion can disturb the representativeness of the subjects, rendering the findings biased.

Missing data always present difficulties whatever you do. No deletion is better than the other unless the missing values are randomly distributed and the sample size is large enough to remain robust despite casewise deletions. The other option is **imputation**, although that also has its own problems. Thus, everything possible should be done to get correct and complete data on as many subjects as possible. Deletions and imputations are the last resort.

**TABLE C.3**
**Missing Data**

| Case No. | Total Bilirubin Level (mg/dL) | AST Level (U/L) | ALT Level (U/L) | Disease |
|---|---|---|---|---|
| 14 | 0.7 | Missing | 30 | Hepatitis |
| 79 | Missing | 32 | 51 | Cirrhosis |
| 92 | 2.1 | 25 | 58 | Malignancy |
| 127 | 1.7 | Missing | Missing | Missing |
| 132 | 1.9 | 43 | 47 | Malignancy |
| 164 | Missing | Missing | Missing | Missing |
| 183 | 2.2 | Missing | 35 | Hepatitis |
| 191 | 2.9 | 61 | Missing | Hepatitis |

## categorical data (analysis of)—overall

Conventionally, data on qualitative variables are called categorical, but data on those quantitative variables that have few categories are also considered categorical. Thus, data on age of adults divided into 20–29 years, 30–39 years, etc., are categorical. This is discussed in detail under the topic **variables**. Theoretically, age divided into 1-year intervals is also categorical but is considered as good as continuous, and categories are ignored for analysis. From the statistics viewpoint, it is not much important whether the data are categorical or not. What is more important is that the focus of conclusions is on proportions, percentages, or frequency of subjects falling into different categories, or on mean and SD of the values. The analysis of categorical data implies that the conclusions will be based on what proportion of cases fall into different categories. In health and medicine, this kind of inference is extensively drawn. Even the quantitative values such as hemoglobin levels are many times categorized into anemia being present or absent because of the convenience in understanding such categories.

If there is just one variable with two categories such as yes/no or present/absent (binary or dichotomous variable), this is analyzed by using binomial distribution that includes its Gaussian approximation for a large sample. This is presented under the topic **binomial distribution/probability**. If this variable has more than two categories (**polytomous**), the interest would be usually to know whether or not they fall into a specified pattern. This almost invariably happens with **nominal** categories and is discussed under the topic **goodness-of-fit**. For large samples, this uses **chi-square** as the criterion for testing the hypothesis on specific pattern of frequencies. Small samples require considering **multinomial distribution**.

The basic statistical method of analysis of categorical data is indeed chi-square for testing of hypotheses regarding not only goodness-of-fit but also many other types of problems. This is used to test the hypothesis regarding the association between two categorical variables when the data are cross-classified into an $R \times C$ **contingency table**. These methods are presented under **chi-square for larger two-way tables** in this volume. A similar method is applicable to three-way tables. **Log-linear models** can be used to understand the pattern of frequencies in different cells in two- or three-way tables. All these methods require large samples.

A special case of an $R \times C$ table is the **two-by-two table**. For small samples, a $2 \times 2$ table is analyzed by the Fisher exact test. Such tables are extensively used to calculate odds ratio and relative risk. These are also used in crossover trials and equivalence tests with binary response, and in assessing sensitivity and specificity and predictivities of medical tests. A $2 \times 2$ table also arises in case of matched pairs for which the **McNemar test** is used for large samples and the **exact test** for small samples. All these methods are discussed separately under the respective topics. For assessing the degree of association, various methods are discussed in this volume for **association between dichotomous categories** and for **association between polytomous characteristics**. For dichotomy in repeated measures, see the **Cochran *Q* test**.

The methods indicated so far are valid when the categories have no order or, if they have order, it can be ignored. For ordinal categories, there is a separate **chi-square for trend in proportions**. The degree of **association between polytomous characteristics** can also be assessed. **Nonparametric methods** based on ranks are also commonly used for analyzing data with ordinal categories.

When the dependent in a regression setup is categorical, logistic regression is the method of choice. This is among the most popular methods of data analysis in biostatistics. Full length books are available on this topic, but the essentials are presented under the topic **logistic regression**. Logistic regression can be used for ordinal categories also.

We have discussed a large number of methods of analysis of categorical data in this book, but the discussion is still not comprehensive. For more details, consult Stokes et al. [1].

1. Stokes ME, Davis CS, Koch GG. *Categorical Data Analysis Using SAS*, Third Edition. SAS Institute, 2012.

## categorical variables, *see* variables

## categories of data values

Categories naturally arise for many medical measurements, while for some others, they are created. Natural categories occur for **nominal** variables such as blood group, occupation, site of injury, site of cancer, type of disease, and complaints. They are called nominal categories. They do not have any natural order—none is more or higher than the other. This can be dichotomous with two categories (sex: male and female) or can be polytomous as in the examples just cited.

Categories for **ordinal variables** are also mostly perceived as natural. Examples are severity of disease categories such as none, mild, moderate, serious, and critical; opinion from most unfavorable to most favorable; obesity as thin, normal, overweight, obese, and morbidly obese; and test result as negative, equivocal, probably positive, and definitely positive. These are called ordinal categories and arise mostly because we do not have a suitable measurement of the underlying continuum. The main reason for nonavailability of a metric scale is that many of these characteristics are multifactorial. For example, disease severity depends on signs and symptoms, and measurements such as blood pressure and plasma glucose levels, radiological assessment, part affected, etc. To pool them together on a metric scale is a challenge. Sometimes ordinal categories are used for convenience and easy communicability. Obesity is an example that can be exactly measured by body mass index, but some workers may like to consider other features also such as waist measurement and skinfold thickness for categorizing obesity. If so, the categories just mentioned may be more convenient. It is easy to categorize burns as severe on visual inspection alone, whereas a large number of assessments may be needed to say that the extent of burns is 83%.

Sometimes a device to measure a characteristic is easily available but is not adopted because such a level of accuracy is not needed. Smoking can be measured as the number of cigarettes smoked, but categories such as none, light, moderate, and heavy seem to serve the purpose sufficiently well in many clinical situations. Age can be measured in terms of years, but categorization into child, adult, and old may be adequate in some situations. Note that ordinal categories also have names and thus can be called *ordinal nominal* in that sense. However, this is different from ordinal metric categories as discussed later in this section.

The basic advantage of using ordinal categories rather than metric measurements is convenience in eliciting, recording, and reporting. The use of any sophisticated device is avoided. Ordinal categories are often easier to comprehend than metric categories. In the process, however, valuable accurate information is lost, and the analysis of data is rendered less efficient. Metric measurements are amenable to a host of mathematical manipulations that are not possible with ordinal measurements. Thus, hard measurements such as blood pressure level are preferred instead of grade of hypertension, and prostate volume instead of grade of enlargement. If the efforts required for hard measurements are enormous such as in measuring the size of the brain, or when no metric scale is available, use ordinal categories. However, beware of anomalies in some ordinal categories. What is mild for one person may be moderate for you. Very rarely these terms are strictly defined.

Quite often, numerals are associated with ordinal categories such as 0 for none, 1 for mild, 2 for moderate, and 3 for serious. These numbers are then subjected to all sorts of algebraic calculations. Such calculations are valid only when the moderate degree is considered two times the mild degree and the serious degree is considered three times the mild. These numerals also assume that the difference between mild and no disease is the same as that between serious and moderate disease. In practice, this may not be so. Thus, caution is required in assigning numerals to ordinal categories and in drawing conclusions when based on calculations involving such numbers. Note that these are not codes or scores.

### *Categories for Metric Measurements*

Exact measurements on the metric scale are indeed statistically preferable to ordinal categories. The irony is that sometimes circumstances force grouping of metric data into categories even after exact data are obtained. The weight of a woman may be recorded to the nearest kilogram but may have to be categorized into 5 kg intervals such as (40–44), (45–49), etc. Data reported in this manner are called **grouped data**; the process is commonly referred to as categorizing continuous variables, and the groups are called **class intervals**. The reasons for doing this may be one or more of the following:

- Consider a dataset containing systolic BP levels of 1200 persons. The only effective way to present these in a report is by using groups in mmHg such as (100–109), (110–119), (120–129), etc., and stating the number of subjects in each such group. This saves space and at the same time makes the data more intelligible. Storage of grouped data may take only about half a page or 1 kB of space, whereas storage of 1200 individual values may take four pages or 8 kB of space. Such grouping also makes the data more sensible, while 1200 ungrouped values may be difficult to comprehend.

  Groups such as (0–4) and (5–9) for age assume that age is noted in terms of *completed* years or age last birthday. The interval (5–9) actually means 5 to less than 10 years and can also be written as (5–10) years. It is customary in such statistical grouping that the upper end of the interval is considered to belong to the next interval. Mathematically, these are written as [0–5), [5–10), etc., but this kind of exact notation is seldom used in practice. Wherever the intervals are continuous in this book, the convention of (0–4), (5–9), etc., is followed.
- It is well known that the end digit is predominantly 0 or 5 in many data values. This happens either because of approximation done by subjects themselves at the time of inquiry, such as stating one's age as 45 years instead of the more exact 44, or because of the observer's bias such as in recording a systolic 130 mmHg instead of the exact 132. Intervals (105–114), (115–124), (125–134), etc., or (108–112), (113–117), (118–122), etc., would dilute the effect of such digit preference. In another setting, suppose waist and hip sizes are measured without sufficient care

and could be in error of up to 5 mm. Grouping of waist–hip ratio in intervals (0.7–0.8), (0.8–0.9), (0.9–1.0), etc., would minimize the effect of such errors, and the purpose of assessing central obesity could still be adequately achieved despite errors, provided they are minor.

The preceding two reasons are valid for grouping at the stage of reporting or analysis, but sometimes even the recording is done in a grouped form. This is done for the following reasons:

- Eliciting a woman's age and anybody's income is sometimes considered a lack of courtesy. Some people prefer to keep such information confidential. Stating them in a grouped form may be more acceptable. The exact value remains confidential, yet data available are in a usable form.
- Many clinicians are accustomed to think in terms of anemia being present or absent and its degree as mild, moderate, or severe in place of exact hemoglobin or hematocrit values. Thus, they sometimes prefer grouped values. Two or more measurements can also be simultaneously considered in this kind of grouping. Categorization of growth of a child into excessive, normal, retarded, and dismal depends not only on height and weight but also on the age of reaching different milestones of development. Such multifactorial grouping is sometimes more relevant for the practice of medicine.
- In an experiment on lethal dose of a drug in mice, it is much easier to observe each morning and record the number of dead mice than to keep a continuous watch and note the exact time of death. In this case, the survival time would be available in 24 h categories. Serum glucose level is measured in units of 5 mg/dL because the analyzer in some cases is so calibrated. Thus, 5 mg/dL categories are inadvertently formed. Greater accuracy may be redundant in this case. If better accuracy is needed, cost and efforts may substantially increase.

Whenever data are available in an exact form, statistical analysis should be done using the exact data. Grouping in this case renders analysis less efficient in the sense that some important features of the data may fail to emerge. Calculations for metric categories are generally done using the midpoint of the categories. This unrealistically considers that all subjects with values in a category have the same value as the midpoint. Statistical inference becomes less efficient as the methods assume that the values in grouped data are flat within each interval. This is against the factual position since discontinuity is imposed by categorization across interval boundaries. The interval 160–169 of systolic BP forgets that 168 mmHg is more than 162 mmHg, and the difference is more than between 168 and 170, which now belongs to separate categories. Cutpoints of intervals are mostly arbitrary, and different cutpoints can give different results. In some cases, though, this loss can be compensated by increasing the sample size. A larger sample size helps to capture a better spectrum of values even when data are grouped. In fact, in cases in which grouped data can be rapidly obtained at a substantial saving, more reliable results can be obtained by investigating a larger sample within the same cost. In such a situation, grouped data on metric measurements can be rightly advocated. However, it is important that the number of groups and the width of intervals are carefully chosen so that the essential features of data are not compromised and the relevance is not lost. As a side note, metric categories also have order and thus are ordinal—ordinal metric to be more specific.

At the extreme is dichotomizing a continuous variable. An example is dividing medical measurements as normal and not normal. Whereas this could be a clinical necessity to decide whether or not to start the treatment, analysis of dichotomized continuous data entails severe loss of data information. For example, in a regression analysis, to develop a prognostic model for patients with primary biliary cirrhosis, bilirubin as a continuous explanatory variable explained 31% more of the variability in the data than when bilirubin was dichotomized at the median [1].

### *Statistical Features of Categories*

Statistical methods of analysis depend on statistical features of the categories. Most basic categorization is in two categories: yes/no, male/female, disease/no disease, survived/died, treated/not treated, etc. These are called **dichotomous categories**. The opposite of this are **polytomous categories** in which a variable has more than two categories. Blood group, site of lesion, complaints, and occupation are examples of variables with polytomous categories. Ordinal categories also are always polytomous since the concept of grading mostly exists only for three or more nominal categories (see the next paragraph). For analysis of categorical data of different types, see **categorical data (analysis of)**.

Metric categories can be dichotomous or polytomous. Diastolic blood pressure (BP) <90 mmHg and ≥90 mmHg are dichotomous. In this setup, we can have order (one is higher than the other)—pointing to the limitation of our statement that ordinal categories exist only for three or more categories. For diastolic BP, polytomous metric categories in mmHg can be 60–69, 70–79, etc.

Except for complaints, all categories mentioned in this section are **mutually exclusive and exhaustive**. Data on such categories are generally expressed in the form of **contingency tables**. On the contrary, complaints of a patient coming to a clinic are neither mutually exclusive nor exhaustive. If the complaints categories are pain, vomiting, constipation, and feeling of tiredness, a patient can have two or more complaints at the same time. This is called multiple response. Neither of these are exhaustive since a patient can have a complaint other than the four listed. Statistical analysis of such data requires special care as mentioned for **multiple responses**.

1. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: A bad idea. *Stat Med* 2006;25:127–41. http://www-psychology.concordia.ca/fac/kline/734/royston.pdf

## causal chain/pathway

Causal chain and causal pathway are synonyms for the series of events that lead sequentially one to another, finally giving rise to the outcome of our interest. The causal chain starting from DNA to RNA to proteins leading finally to biological functions is well known. In the context of a disease, a causal chain is the etiological pathway—how the disease started, what course it took and why, and why this is in the present condition. A **causal diagram**, as described for myocardial infarction under the next topic, is a visual description of the chain of events.

Pecoraro et al. [1] have found that minor trauma, cutaneous ulceration, and wound-healing failure could be the causal sequence for 72% of the cases in their series of diabetic patients with limb amputation. Thus, practical interventions that cause limb loss in such cases can be devised. As illustrated by Li and Lu [2], biostatistical study of the causal chain of how diseases respond to drugs can lead to new

uses of the existing drugs. An understanding of the ways in which xenobiotic substances perturb biological systems and lead to adverse outcomes can help to devise tools to avoid such outcomes [3].

Experiments are generally more effective in establishing a causal chain than **observational studies**, but these may be very expensive and time consuming. Biostatistical **models** can help to study causal chains in a quantitative manner and provide them much needed exactitude.


1. Pecoraro RE, Reiber GE, Burgess EM. Pathways to diabetic limb amputation: Basis for prevention. *Diabetes Care* 1990 May;13(5):513–21. http://www.ncbi.nlm.nih.gov/pubmed/2351029
2. Li J, Lu Z. Pathway-based drug repositioning using causal inference. *BMC Bioinformatics* 2013;14 Suppl 16:S3. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3853312/
3. Sturla SJ, Boobis AR, FitzGerald RE et al. Systems toxicology: From basic research to risk assessment. *Chem Res Toxicol* 2014 Mar 17;27(3):314–29. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3964730/


## causal diagrams

Causal diagrams depict the causal pathway between an antecedent and an outcome using nodes and arrows that are drawn with due consideration of the intermediaries and confounders. Thus, they represent a web of causation. Such diagrams are commonly used in epidemiologic research. Causal diagrams are explained by Greenland et al. [1] and recently discussed for matched designs by Mansournia et al. [2]. These diagrams are also called **etiology diagrams**.

One such diagram is in Figure C.2. This tries to describe the causation of myocardial infarction (MI). The process leading to MI is intricate, and the exact web of causation is still not fully known. This is just our hypothesis and is untested. The arrows show the direction of the effect from the cause. Nodes are the factors. If this figure is to be believed, ageing, personality profile, lifestyle, and genetic predisposition are independent of one another, and they together or alone give rise to the factors that ultimately cause MI. They are the *ancestors* in this web of causation; stress and anxiety is the *child* of the personality profile, while smoking is the child of emotional disturbance, stress and anxiety, and lifestyle. You can see that this causal diagram ignores other factors such as peer pressure and home environment as the cause of smoking. These possibly are represented by lifestyle in this diagram. This diagram proposes that hypertension is the child of genetic predisposition, changes in the arterial walls, and emotional disturbances, and has ageing and personality profile as ancestors. Smoking is in the causal pathway of MI but is not a direct contributor—its contribution is through hyperlipidemia and atherosclerosis. If this diagram is to be believed, lifestyle does not cause hypertension but is mediated through lack of physical exercise and obesity. Many would not agree with this proposition. Despite such complexity in the diagram, this still is possibly not a complete representation of the web of causation of MI. Perhaps any causal diagram is a simple version of the complex process—and thus fits into the definition of a **model**. In this diagram, there is no two-headed (bidirectional) arrow, but this can be used in a causal diagram when X affects Y and Y affects X. In Figure C.2, hypertension can sometimes lead to change in arterial walls in place of age-dependent changes to the walls of arteries leading to hypertension.

The difficulty with causal diagrams is that they trace causal path without specifying the quantitative effect. For example, Figure C.2 does not say how much is the effect of stress and anxiety on smoking and how much is of lifestyle. Also, the shape of the relationship is not depicted. Yet a properly drawn diagram is useful in visualizing the web of causation and to identify where to attack to prevent the disease, or to reduce the risk, or even what etiological factors to assess while examining a case with the disease.


1. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999 Jan;10(1):37–48. http://www.biostat.harvard.edu/robins/publications/causaldia.pdf
2. Mansournia MA, Hernán MA, Greenland S. Matched designs and causal diagrams. *Int J Epidemiol* 2013;42:860–9. http://ije.oxfordjournals.org/content/42/3/860.short


## causal inference, *see* cause–effect relationship

## cause–effect relationship

Cause–effect relationship is said to exist between an antecedent and an outcome when a change in the value of the antecedent necessarily causes a change in the outcome. By far, the most compelling

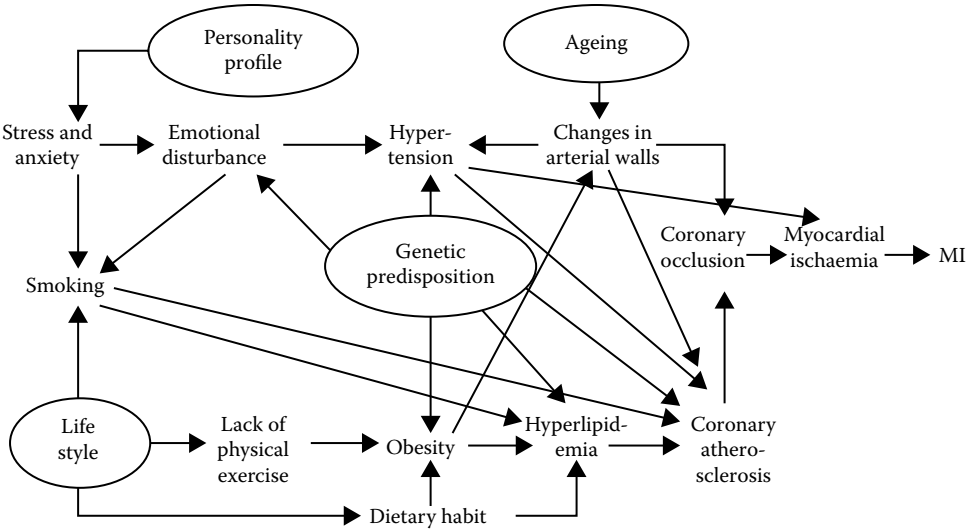



**FIGURE C.2** Hypothetical causal diagram for myocardial infarction (MI).

C

evidence of cause–effect is provided by **experiments** where an intervention is intentionally introduced to see how the outcome changes. "All" other factors are controlled by design or eliminated by analysis. This includes **clinical trials**. However, some unknown factors can provide a distorted picture. The effect could be random, at least in part, while we see it as causality. Under certain strict conditions, **observational studies** too can provide evidence of cause–effect. We examine both in the following paragraphs.

### *Causal Inference in Clinical Trials*

It is generally believed that properly conducted experiments—whether in the laboratory or in the clinic—provide compelling evidence for or against any cause–effect relationship between the antecedent and the outcome. The antecedent could be the regimen under trial, and the outcome may be any that indicates efficacy. However, there are limitations: compelling evidence—yes; indisputable evidence—no.

An antecedent is not considered a cause unless it has a plausible biological explanation for the effect. However, a trial can be conducted without full knowledge of the biological relationship, and this relationship can be explored later after the trial is over. In that case, however, the relationship is interpreted as a mere association and not cause–effect until such time that biological plausibility is established.

A perfectly valid trial in terms of randomization, control, and blindness may still provide results that are difficult to apply elsewhere. First, the selected subjects may not be representative of the target population; second, the conditions under which a trial is done may be too restrictive; and third, different trials may give different results. These differences propel synthesis techniques such as meta-analysis that allow a more reliable conclusion regarding causality [1].

**Randomized controlled trials** are one of the many ways, perhaps the most dependable, to establish the properties of a regimen. But there might also be alternate methods of arriving at credible answers. They all should match. In addition, all such trials are plagued with epistemic uncertainties—unrecognized factors may affect the outcome. A positive aspect of varying results of different trials on the same regimen is that they raise awareness about the uncertain domain and bring humility to our endeavors.

### *Correlation versus Cause–Effect Relationship in Observational Studies*

Some professionals feel that a randomized controlled trial is the only mode to provide scientific evidence of a cause–effect relationship. A moment's reflection will convince you that this is not so. Most accept that cigarette smoking *causes* lung cancer, but no trial has ever been conducted. The controlled experiments do provide direct evidence, but evidence from observational studies can be equally compelling when **confounders** are really under control and the results replicate in a variety of settings.

Keep in mind though that an association or a correlation does not indicate cause–effect. In health measurements, association can arise because of a large number of intervening factors. It is rarely a cause–effect type of relationship unless established by a carefully designed study that rules out the possibility of a significant role of any confounding factor. McLaughlin et al. [2] report from epidemiological studies that women in the United States with cosmetic breast implants have two- to threefold risk of suicides compared to the general population. This is statistically significant. However, this by itself does not support a cause–effect relationship. They speculate that psychopathological condition of women may be a confounding factor since this condition can contribute to the need for a breast implant as well as suicide. They also remark that any study aiming to investigate this relationship for cause–effect should also rule out year of birth, family history of psychiatric admission, etc., as factors. More important threats to cause–effect inference are not the known confounders but the unknown ones.

A strong correlation between heights of siblings in a family exists not because one is the cause of the other but because both are affected by parental height. Similarly, correlation between visual acuity and vital capacity in subjects of age 50 years and above is not of the cause–effect type but arises because both are products of the same degeneration process. No one expects vision to improve if vital capacity is improved by some therapy. The maternal mortality rate declined in Mexico between 1960 and 2010, and the proportional mortality from coronary diseases increased. These too have a strong negative correlation, but they do not exhibit a cause and effect type of relationship. Counterfactuals provide useful armory to refute causality.

An unusual confounding factor may provide useful information in some cases. Persons with depressed mood may have an elevated risk of lung cancer because of a third intervening factor—smoking [3]. Depressiveness seems to modify the effect of smoking on lung cancer either by a biologic mechanism or by affecting smoking behavior. This example illustrates how a seemingly nonsense correlation can sometimes lead to a plausible hypothesis. Another example is the parental age gap influencing the sex ratio of the firstborn children [4]. At the same time, our existing knowledge tells us that no drug can change the blood group. There is no need of empirical evidence to confirm this.

### *Criteria for a Correlation to Indicate Cause–Effect*

For an association or a correlation to indicate a cause and effect, the following points should be carefully examined:

- Association should be present not only in the prevalence but also in the incidence. This should continue to be statistically significant when the effect of other confounding factors is eliminated by means of a method such as multiple regression. If feasible, confirm this by an experiment. This is possible only for potentially beneficial regimens and not for potentially harmful regimens.
- Association or correlation must be consistent across various groups, periods, and geographical areas. Wherever or whenever the postulated cause is present, the effect must also be present. If an association is found between lipoprotein(a) and incidence of coronary artery disease (CAD) in Canada, Singapore, England, and South Africa, which has persisted since this protein was first discovered, then this is likely to be a cause of CAD, although this could be one of the many causes.
- There must be a dose–response kind of relationship in the sense that if the cause is present in higher amounts or in greater intensity, then the chance of effect should also be high. The more one smokes, the higher the risk of lung cancer becomes.
- The relationship should also be specific, i.e., if the postulated cause is absent, then the effect should also be absent. However, make a distinction between a necessary cause and a sufficient cause. A necessary cause is an exposure that is *required* for a particular outcome to occur—without this, the outcome would not occur, but the outcome may

not occur even when the necessary cause is present. A sufficient cause is an exposure that will produce the outcome. The outcome may occur due to other causes without the sufficient cause. For example, infection is necessary for typhoid to occur but not sufficient.

- The relationship must be biologically plausible, which means an explanation linking the two in a causal way is available. For example, cigarette smoke has cotinine that easily affects lung cells and causes aberrant cell behavior. Thus, a biological explanation is available.
- It is generally stated for a cause–effect relationship that the degree of association or the correlation should be high. This may be so when the factor under investigation is a dominant cause. The correlation between parental levels and children's level of cholesterol is not high, but this does not exclude genetic influence as a cause of raised cholesterol level in children whose parents have high level. When a correlation, albeit low, is consistently present, a causal relationship can still be inferred, but that would be one of the many causes. It should, however, be statistically significant so that **sampling fluctuation** is adequately ruled out as a likely explanation. In addition, also note that the correlation coefficient can be small just because a small range of values has been observed.

The causal role of elevated serum cholesterol in atherosclerotic CAD, for example, fulfills all these criteria. A consistent association has been found in prevalence as well as in incidence in a large number of studies carried out in populations with different backgrounds. A gradient in risk of the disease has been seen with rising cholesterol level. In controlled clinical trials, cholesterol-lowering drugs and diet therapy have been shown to result in a reduction of coronary events. Angiographic regression of coronary lesions has also been demonstrated. Biological plausibility has been established from animal and human studies that have demonstrated cholesterol deposition in atheromatous plaques.

### *Other Considerations in Cause–Effect Relationships*

Distinction may be made between a necessary cause and a sufficient cause as mentioned earlier in this section. Sexual intercourse is necessary for initiating pregnancy in natural course, but it is not sufficient. In fact, the correlation between the number of intercourses and the number of pregnancies, even without barriers, is negligible.

In this era of diseases with multifactorial etiology, one particular factor may not be sufficient to cause the disease. The presence or absence of one or more of the others may also be needed. Thus, a factor could be a contributing cause in the sense that it is a predisposing, enabling, precipitating, or reinforcing factor. For a detailed discussion of causal relationships in medicine, see Elwood [5]. Also, do not be carried away by the following popular perception: guns kill people! Guns are instrumental, but it is people who kill people. On the other hand, earthquakes do not kill people—it is the collapse of the brittle buildings that kills people.

The above discussion assumes that the role of chance and bias has been minimized, if not eliminated. Chance due to sampling fluctuation is adequately ruled out by demonstrating statistical significance. **Bias** can occur due to a host of factors such as the selection process, observational methods, differential recall of past events, and suppression of information. Measures such as standardization of methods, training, randomization, and matching can be used to minimize the bias. The cause–effect hypothesis is strengthened when all alternative explanations are also studied and shown to be not tenable. A larger sample size does help in increasing confidence, but its role in causal analysis beyond statistical significance is marginal.


1. Dimick JB, Diener-West M, Lipsett PA. Negative results of randomized clinical trials published in the surgical literature: Equivalency or error? *Arch Surg* 2001;136:796–800. http://archsurg.jamanetwork.com/article.aspx?articleid=391750
2. McLaughlin JK, Lipworth L, Tarone RE. Suicide among women with cosmetic breast implants: A review of the epidemiologic evidence. *J Long Term Eff Med Implants* 2003;13:445–50. http://www.iei.us/JLT1306(205)-Author-Proof-3.pdf
3. Knekt P, Raitasalo R, Heliovaara M et al. Elevated lung cancer risk among persons with depressed mood. *Am J Epidemiol* 1996;144:1096–103. http://aje.oxfordjournals.org/content/144/12/1096.full.pdf
4. Hakko H, Rasanen P, Jarvelin M, Tiihonen J. Parental age-gap and child sex ratio—Fact or fiction? *Int J Epidemiol* 1998;27:929–30. http://ije.oxfordjournals.org/content/27/5/929.full.pdf
5. Elwood JM. *Causal Relationship in Medicine: A Practical System for Critical Appraisal*. Oxford University Press, 1992.


## cause of death

According to the World Health Organization (WHO) [1], cause of death is the disease or injury that initiated the train of morbid events leading directly to death, or the circumstances of accident or violence that produced the fatal injury. Thus, this is the underlying cause of death and not the immediate cause of death. For example, for injuries, the cause could be road accident, fall, self-harm (e.g., suicide), interpersonal violence (e.g., homicide), collective violence (e.g., war), etc. It is long realized that the assessment of cause of death has a substantial subjective element. To minimize this, an **International Classification of Diseases (ICD)** is available that is periodically revised (generally every 10 years) to incorporate new knowledge.

In countries where facilities are available, cause of death is recorded in **death certificates**. This certification is nearly complete in countries where each death is attended by a qualified medical professional. In other countries, mostly developing nations, the method of verbal autopsy is adopted that requires interviewing the immediate family about the conditions leading to death. Besides the demographic details about the informant and the deceased, the certificate contains minimal information on the immediate and underlying cause as illustrated in Figure C.3.

According to the WHO estimates [3], nearly 56 million deaths occurred globally in the year 2012, of which 23.0% were due to infectious and parasitic diseases, 67.8% due to noncommunicable diseases, and 9.2% due to injuries. With aging of the population almost everywhere, the noncommunicable diseases are increasingly occupying the pie, and the role of infectious and parasitic diseases is shrinking.

Cause of death statistics raise sensitive issues about allocation of resources for control of diseases and about socioeconomic conditions that promote one cause at the expense of another. In terms of percentage, since total deaths are 100%, one may argue that myocardial infarction is better as a cause of death or cancer is better [4]. All causes cannot be controlled since death is inevitable.


1. WHO. *Mortality*. http://www.who.int/topics/mortality/en/, last accessed April 12, 2015.
2. WHO. *Medical Certification of the Cause of Death: Instructions for Physicians on Use of International Form of Medical Certificate of Cause of Death*. World Health Organization, 1979: p. 7. http://apps.who.int/iris/bitstream/10665/40557/1/9241560622.pdf

Medical certificate of cause of death

| | Cause of death | Approximate interval between onset and death |
|---|---|---|
| I<br>*Disease or condition directly leading to death** | (*a*).....<br>due to (or as a consequence of) | ..... |
| *Antecedent causes*<br>Morbid conditions, if any, giving rise to the above cause, Stating the underlying condition last | (*b*).....<br>due to (or as a consequence of)<br>(*c*)..... | .....<br>..... |
| II<br>*Other significant conditions* Contributing to death, but not related to the disease or condition causing it | - - - - - - - - - - - - - - - - - -<br>- - - - - - - - - - - - - - - - - - | .....<br>..... |

*This does not mean the mode of dying e.g., heart failure, asthenia, etc. It means the disease, injury, or complication which caused death

**FIGURE C.3** Minimal certificate of cause of death as prescribed by the World Health Organization. (From WHO. *Medical Certification of the Cause of Death: Instructions for Physicians on Use of International Form of Medical Certificate of Cause of Death*. World Health Organization, 1979: p. 7. http://apps.who.int/iris/bitstream/10665/40557/1/9241560622.pdf.)


3. WHO. *Health Statistics and Information Systems; Estimates for 2000–2012: Cause-Specific Mortality.* http://www.who.int/healthinfo/global_burden_disease/estimates/en/index1.html
4. Indrayan A. Can I choose the cause of my death? *BMJ* April 21, 2001;322(7292):1003. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1120113/


## cause-specific death rate, *see* death rates

## cell frequency

A cell in a **contingency table** arises when the subjects are cross-classified into some categories. If the subjects are being divided by their blood group and sex, there would be eight cells in this contingency table—four for males and another four for females. The term *cell* is generally used for such cross-classifications by two or more attributes, but it can be used for **one-way tables** also. Cell frequency is the number of subjects falling into a particular cell.

If two variables under study are serum creatinine level and whether kidney disease is present or not, the number of subjects in your sample with level 1.2–1.4 mg/dL and with no disease is the cell frequency for this cell. The term can be loosely used for one-way classifications also, such as for the number of subjects with serum creatinine level 1.2–1.4 mg/dL, but is more appropriate when dealing with cross-classification. An example is in Table C.4 on survival of cases in different age groups undergoing gastrointestinal surgeries in a hospital over a 5-year period. In this table, the cell frequencies are the number of cases as mentioned in different cells. The entire grid is shown so that you can see the cells.

In Table C.4 are what are called *observed cell frequencies*. These are the numbers actually found in the study. The other term is *expected cell frequency*. This is used for the frequency you expect when the two or more characteristics under study are not associated. See **association and its degree** if you want to know about association. In Table C.4, a total of 126 out of 2321 patients died. This is 5.4%. If deaths have nothing to do with age, they will be uniformly distributed across all age groups at this rate. Observed deaths, for example, in age group 80–89 years is 4/31 = 12.9%. If age and deaths are not related, this also should be the same 5.4% as in the total. Thus, if deaths and age are not associated, the expected cell frequency for age group 80–89 years is 5.4% of 31, i.e., 5.4 × 31/100 = 1.67. Little arithmetic may tell you that this can also be written as 31 × 126/2321. In the

**TABLE C.4**
**Survival of Cases of Different Age Groups Undergoing Gastrointestinal Surgery in a Hospital over a 5-Year Period**

| Age Group (Years) | Survival | | Total |
|---|---|---|---|
| | Yes | No | |
| 0–9 | 30 | 1 | **31** |
| 10–19 | 82 | 4 | **86** |
| 20–29 | 259 | 17 | **276** |
| 30–39 | 343 | 12 | **355** |
| 40–49 | 415 | 12 | **427** |
| 50–59 | 460 | 31 | **491** |
| 60–69 | 407 | 26 | **433** |
| 70–79 | 170 | 19 | **189** |
| 80–89 | 27 | 4 | **31** |
| 90–99 | 2 | 0 | **2** |
| **Total** | **2195** | **126** | **2321** |

numerator is the product of the two marginal totals corresponding to the cell, and in the denominator is the grand total. The general formula for the $(i, j)$th cell in a two-way table is as follows. Rows and columns are counted only for the cells where frequencies appear. In Table C.4, the column containing the age groups and totals, and the last row containing the totals are not counted for demarcating the $(i, j)$th cell.

$$\text{expected cell frequency: } E_{ij} = \frac{R_i \times C_j}{G},$$

where $R_i$ is the corresponding row total, $C_j$ is the corresponding column total, and $G$ is the grand total. Thus, the observed cell frequency in the third row and first column is $O_{31} = 259$, and the expected cell frequency for no association is $E_{31} = \frac{276 \times 2195}{2321} = 261.0$. Such observed and expected cell frequencies are required for the calculation of **chi-square** for testing the hypothesis regarding association between two variables. This test uses the premise that the expected and observed frequencies in any cell should not be very different when there is no association.

It is not uncommon in medical data that some cells in a contingency table have no frequency or zero frequency. It is important for the purpose of analysis to distinguish between observed zeroes and structural zeroes. An observed zero is one where some frequency could have occurred but happens to be zero in the sample. This is not much of a problem and is treated just like any other small frequency. A structural zero occurs when it is just not possible to have any subject in the cell.

You will need these basic concepts to understand medical literature, when trying to analyze data from your own clinic, and when trying to do research. Cell frequencies play a vital role in the validity of a chi-square test as just mentioned. This is a large sample test and requires that at least 80% of the cells must have an expected frequency of at least 5, and none should have an extremely small frequency such as 0 or 1. If this happens, this can make chi-square invalid for these data. Merging of cells is the answer in such cases.

## censoring (of observations)

Censoring is used for incomplete values, particularly for durations. Duration is the time elapsed from a defined entry point to the defined exit point. This is also referred to as time-to-event where event is the exit. Duration is a difficult variable because it requires that each subject is observed for as long as it does not exit. If you are interested in the duration of survival of children born with thalassemia, some may survive for 80 years. It can be very expensive to continue the study for such a long period. We censor it somewhere and say that we will observe it for 10 years and no longer. Duration of survival for people alive at the cutoff time would never be known except that they survived for at least 10 years in this example. The method of **survival analysis** is geared to meet this contingency, which can rarely be handled in any other manner.

Censoring can also occur for other reasons, e.g., the person moves out, dies due to an unrelated cause, refuses to cooperate after some follow-up, etc. The primary cause of such incomplete observation, however, is, as just stated, that the end-point event does not occur during the follow-up period; the study is terminated after a fixed time.

The kind of censoring mentioned in the preceding example on thalassemia is called *right censoring*. In this case, the survival in censored cases is known to be greater than the specified duration, but the exact duration is not known because the event had not occurred when the study stopped. Right censoring is the most common type of censoring for duration so much so that the word right is dropped and such values are referred to as just censored. This kind of censoring is illustrated in Figure C.4a. A hollow square in this figure represents right censoring. In cases D, F, and H, this occurred because the study has ended, whereas in case E, it is due to loss of the subject for follow-up. Rearranging the durations from minimum to maximum leads to the situation as shown in Figure C.4b.

In *left censoring*, the beginning time is not known, though the end time is known. This also is an incomplete segment, but this kind of censoring is rare in survival studies. This is shown for subject A in Figure C.4a.

The third is *interval censoring*, which occurs when the time-to-event is known to have occurred within a specified time interval, but the exact time is not known. This happens when a continuous vigil on all subjects cannot be kept because it is impractical or too expensive. If you are undertaking periodic visits to inspect a process, the exact time of occurrence will not be known but only that it has occurred sometime since the previous visit. If a group of diabetic persons are being assessed quarterly for the development of retinopathy beginning with the diagnosis of diabetes, the elapsed duration will be known only in quarterly intervals. The exact day or week will not be known.

For the methods of survival analysis, the censoring scheme should have nothing to do with the future survival or time-to-event. This means censoring should not provide any tilted information regarding duration in that group. This condition is generally fulfilled in right-censored values arising from the termination of the

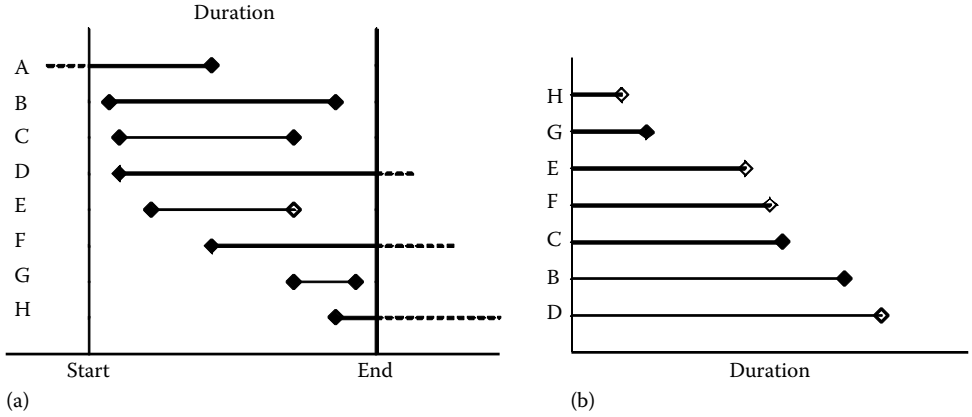



**FIGURE C.4** Censored and complete values (a) before ordering and (b) after ordering.

follow-up. But caution is required in those lost to follow-up for other reasons. In the case of prostate cancer, if a man is not traceable after a certain period, this should not be because he has moved to his family in the last moments of life or has dropped out of the study because he felt he has recovered. The implicit assumption in survival analysis is that the prospects of survival of censored cases are the same as those of uncensored cases. You should be convinced that this condition is met before using the methods of survival analysis.

## census

In contrast to sampling, census is the complete enumeration. The term is usually used for population counts that take place once in 10 years in many countries around the world, but it can be used for other statistical **populations** as well.

Population census looks like the first common statistical activity undertaken even in the ancient world. It may have been first carried out by Babylonians around 3800 BC [1]. The objective is to assess the needs of the population such as food, education, and health, which can be interpreted by the number of people in different age–sex groups in different areas. It can tell us several other interesting facts about the population. For example, the 2011 census in the United Kingdom revealed that 9% of people living as a couple were in an inter-ethnic relationship in England and Wales, which is up 2 percentage points from 2001 [2].

A census can be taken by several methods. In England, Wales, Scotland, and Northern Ireland, the householder receives a questionnaire in the post, completes it, and either submits it online or sends it back in the post. In some other countries, such as Ireland, a large field force visits every household in the land to deliver, and then collect, census questionnaires. The countries of Eastern Europe generally carry out interview censuses, where the enumerators collect and record the information on the householder's behalf [3]. Some countries, such as Finland, update their population register online by records of births, deaths, migration, etc. They can get the complete count and demographic details of the population anytime.

For a manual census of general population, an enumerator goes from house to house and visits other institutions such as hostels where people generally reside, and records elementary demographic details of each person such as age, sex, and occupation. In India, for example, the count is adjusted to be accurate for sunrise on March 1 of the first year of the decade. The last one was done in the year 2011. This included homeless people who sleep on pavements, railway stations, etc. Rural and urban areas are divided into census blocks for enumeration purposes for administrative convenience and to ensure that nobody is missed. The 2011 census in India also elicited information on type of drinking water and type of toilets in the households, besides age, sex, occupation, and housing of the people.

Generally, a short form is completed for each household. For example, for the U.S. census in 2010, the form contained only 10 questions. A sample of households may complete a more detailed questionnaire. Census data provide useful information for planning health services for different segments of population, as well as for assessing the impact of the health services. The data provide an accurate denominator for many rates such as birth rate and death rate.

We can think of a census of out- and in-patients for a hospital, although the term census is not used for this population. Because of computerization in most parts of the world, this is now easily done. The cases can be analyzed not just for their demographics (and compared with the demographics of the catchment area to identify the kind of people who more commonly come for different departments of the hospital) but also for the diseases from which they are suffering. Many hospitals may have a database of millions of patients accumulated over time, particularly if they form a consortium. Analysis of such a huge database for meaningful messages has become an important requirement as evidenced by rising science of data analytics and data mining.


1. Grajalez CG. Great moments in statistics: Ancient censuses. *Significance* 2013;10(6):21. http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2013.00706.x/epdf
2. UK Census. *What Does 2011 Census Tell Us about Inter-Ethnic Relationships?* http://www.ons.gov.uk/ons/rel/census/2011-census-analysis/what-does-the-2011-census-tell-us-about-inter-ethnic-relationships/sty-relationships.html
3. UK Census. *Method of Census Taking.* http://www.ons.gov.uk/ons/guide-method/census/2011/how-our-census-works/about-censuses/methods-of-census-taking/index.html


## centiles, *see* percentiles and percentile curves

## central limit theorem (CLT)

Considered as the most beautiful result in statistics, the central limit theorem (CLT), in its easily understandable form, states that irrespective of the statistical **distribution** of the variable $x$ (called the *underlying distribution*), the mean $\bar{x}$ of values in independently and randomly drawn sample subjects tends to follow a **Gaussian distribution** in repeated samples as the sample size increases. We have stated this in a blanket form, which seems practically true, but there are minor caveats such as that $x$ must have finite variance (theoretically there are distributions with infinite variance). There are also generalizations of the CLT, but those are not important for most medical applications. The CLT was first formulated by Laplace in 1812, but its utility was realized at the end of the nineteenth century when Aleksandr Lyapunov gave the first rigorous proof [1]. Polya first used the name central limit theorem in 1920.

Aleksandr Lyapunov

Figure C.5a shows a likely statistical distribution of years of survival of cancer patients. This is highly skewed. If we take several samples of size 10 and calculate average survival duration of the patients in each sample, the **sampling distribution** of mean would be nearly as shown in Figure C.5b. This is much more symmetric than the distribution of the original values. If we take several samples of size 40 and calculate the mean of each, the distribution of average survival time would be almost Gaussian as shown in Figure C.5c.

The theorem basically is for all those summaries that use the linear combination of sample values. It is applicable to sample mean since the numerator of $\bar{x} = \sum x/n$ is a sum. The CLT says that the distribution of the linear combination of a large number of values tends to become Gaussian. Thus, the theorem also applies to variance where the numerator $\sum(x-\bar{x})^2$ is also a sum, only that it

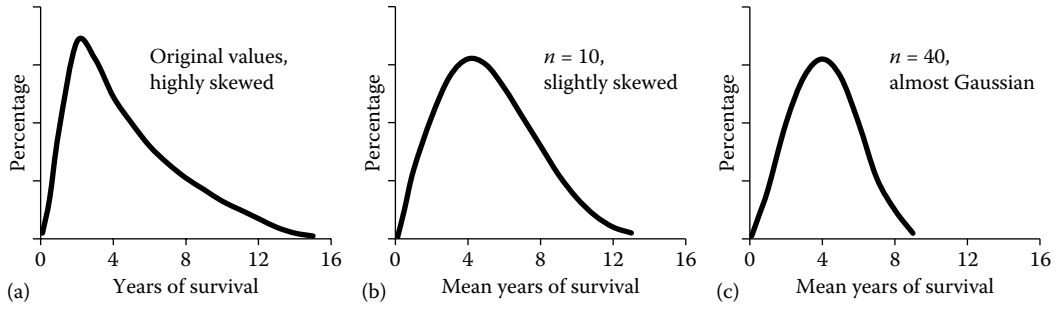



**FIGURE C.5** (a) Highly skewed distribution of survival time of cancer patients and the distribution of its mean with increasing sample size; (b) $n = 10$; (c) $n = 40$.

requires a larger $n$ for such complicated sums. This also applies to sample proportion since this proportion $p = \Sigma x/n$, where $x = 1$ for characteristic present and $x = 0$ otherwise. The numerator of the **regression coefficient** as well as of the **correlation coefficient** is also a sum of cross-products—thus, CLT is applicable to them also. Relative risk and odds ratio are ratios, and their logarithm is eligible to be under CLT; however, for these, a relatively larger sample size is required. All these tend to have a Gaussian distribution for large $n$. On the other hand, summaries such as median, mode, and percentiles are not based on sums or linear combinations—thus, CLT is not applicable to them. In general, they cannot be assumed to follow a Gaussian distribution for large $n$.

This theorem implies that the distribution of a variable such as survival duration may be highly **skewed**, but the distribution of *mean* survival duration becomes Gaussian when it is based on a large number of randomly selected subjects. This result places the Gaussian distribution at the center of the statistical methods. Skewed and other types of distributions such as **bathtub** and **Weibull** are difficult to handle, and this difficulty can be circumvented by increasing the sample size and considering the sample mean as the summary of choice for inference. At the same time, this theorem also underscores the need to have a large sample particularly when the underlying distribution is far from Gaussian and also that the samples must be drawn randomly.

How large is large for the CLT to be applicable? This depends on how different is the underlying distribution from Gaussian. For a slightly skewed distribution, perhaps a sample of size 10 is enough for the CLT to provide valid results, particularly for mean. This size could work for other non-Gaussian but *symmetric* distributions, e.g., the mean of values from the **uniform distribution** quickly follows the CLT. For a moderately skewed distribution, a sample of size 30 may be enough. For a highly skewed distribution such as that of tumor markers, you may need a sample of at least 100. Such a large sample size may be required also for a variable that follows a bathtub type of distribution. For most quantitative medical variables, however, a sample of size at least 30 does the job reasonably well for mean.

Aside from the fact that the CLT is focused on the Gaussian distribution of sample mean, it also obviates the need to depend on something like sample median for inference on skewed distributions if a sufficiently large sample is available. The sample median has problems with regard to its distribution and consequently in working out a confidence interval and in testing of hypothesis. These are easily worked out when the inference is based on sample mean. For example, it is well known that the **standard error (SE)** of the sample mean is $\sigma/\sqrt{n}$. The SE of other sample summaries such as median, mode, and percentiles is not commonly known. Thus, the CLT serves a twin purpose—(i) for enabling a Gaussian distribution and (ii) for removing dependence on median-like statistically difficult measures. However, for one sample, median or mode can be preferred in specific situations, particularly for small samples from non-Gaussian distributions.

1. Dunbar SR. Topics in probability theory and stochastic processes. http://www.math.unl.edu/~sdunbar1/ProbabilityTheory/Lessons/BernoulliTrials/DeMoivreLaplaceCLT/demoivrelaplaceclt.xml

## central values (understanding and which one to use), *see also* mean, median, and mode (calculation of)

Central value is the generic term for all those measures that can be considered to be representative of a dataset or of a **distribution**. As an exercise, consider the following simple data on duration of survival (years) of 14 patients after detection of esophagus cancer:

3, 7, 10, 5, 4, 4, 29, 5, 6, 3, 8, 4, 6, 7

If you are asked to choose a central value that could represent these durations, what would you choose? The natural answer is either mean or median or mode. Depending on the nature of values, one of these is considered as the most representative central value. Mean, median, and mode are also called **measures of central tendency** as they *tend* to lie somewhere in the center of the data.

The important question is, "Which is the best measure of central value for a given set of data?" For this, it is essential to first understand mean, median, and mode with correct perspective. While the details of their calculations in grouped and ungrouped data are given under the topic **mean**, **median**, **and mode**, the following is a brief description that can help to understand these measures. We subsequently give guidelines about which measure to choose and when.

### *Understanding Mean, Median, and Mode*

As just stated, a value nearly in the center, or commonly observed, can be considered as a representative of the central value. The arithmetic mean, commonly called an average, is a very popular measure of central value. If $x_1, x_2, \ldots, x_n$ are $n$ observations in our sample, the sample mean is $\bar{x} = \sum x_i/n$. The median is the middle value, which is obtained as the $\left(\frac{n+1}{2}\right)$th value if $n$ is odd after arranging in ascending order, and the average of $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th if $n$ is even. The mode is the most common value.

The following are the data on cases of acute polymyositis of the back in 38 women.

C

Duration of immobility (days):

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | 5 | 9 | 7 | 36 | 4 | 6 | 7 | 5 | 8 | 3 | 6 | 5 |
| 7 | 8 | 10 | 7 | 14 | 10 | 9 | 4 | 6 | 11 | 9 | 6 | 5 |
| 8 | 8 | 6 | 7 | 5 | 5 | 12 | 3 | 5 | 9 | 10 | 7 | |

Mean = 7.9 days
Median = Average of the 19th and 20th values after rearranging in ascending order = (7 + 7)/2 = 7 days
Mode = 5 days and 7 days, occurring in 7 patients each

A distribution containing two modes such as in this example is called a **bimodal distribution**. Mean is denoted by $\bar{x}$ for sample and by μ for population. No such universally accepted notations are available for median and mode.

A visual display can help to understand the real meaning of mean, median, and mode. Figure C.6 is a simple plot of points explaining the meaning of these measures. This displays the duration of hospital stay (days) after a surgery for 30 patients (numeric data not shown). The mean is the center that balances the beam on which the points are plotted. This balancing is a very useful feature of this central value. The median is such that half of the dots are less than or equal to the median. Although it is not so clear in Figure C.6 because of the **discrete** data, half of the values will be more than the median and the other half less. This may or may not happen with the mean. The mode clearly refers to the peak.

Figure C.7 depicts the location of the mean, median, and mode in symmetric, right-skewed, and left-skewed distributions (see **skewness**). Whereas the three coincide in the symmetric unimodal distribution, they change the order as the shape changes from right-skewed to left-skewed. In the left-skewed distribution, the value of mean is least, followed by median and then mode. These happen to be in the same order as they appear in the dictionary. The reverse happens in the case of the right-skewed distribution. The median is closer to the mean, but the gap between the median and mode is relatively large just as in a dictionary. Thus, values of mean, median, and mode give a fairly good idea of the skewness of the data values.

### *Which Central Value to Use and When?*

If you want to tell a patient one duration of immobility generally seen in cases of acute polymyositis of the back, would you use the mean, median, or mode? The mean is the popular choice because it is simple to calculate and easy to understand. It certainly should be preferred except in cases in which it can mislead. Add to the data in our example two more cases with durations 32 and 47 days. There is already a case of immobility of 36 days. The mean now becomes 9.5 days. Because 31 out of 40 patients are immobile for 9 days or less, 9.5 days cannot be a representative central value. The mean is vitiated due to **outliers**. Outliers are unusually high or unusually low values that do not go along with the other values. Whenever such values are present, the mean is not a good choice. In such instances, use the median because it is not affected by unusually high or unusually low values. For these 40 durations, the median is the average of the 20th and 21st values in ascending order. This again is (7 + 7)/2 = 7 days.

Use the mode when the interest is specifically in the most common value or the "typical" value. Consider the incubation period of cholera from exposure to the appearance of symptoms. This generally ranges from 1 to 4 days, but the most common (mode) is 2 days. This mode has special significance in this case because it can help to estimate a peak occurrence if the exposure is from a common source. Thus, hospitals can be geared up to handle the cases. In our example on duration of immobility, there are two modes—5 and 7 days—each seen in nearly one-fifth of the patients. *The mode can be multiple, but the mean and median are always unique*. The practical utility of the mode is sometimes lost when it is more than one except when it is genuinely bimodal. In the case of **nominal data**, the mode is an appropriate central value.

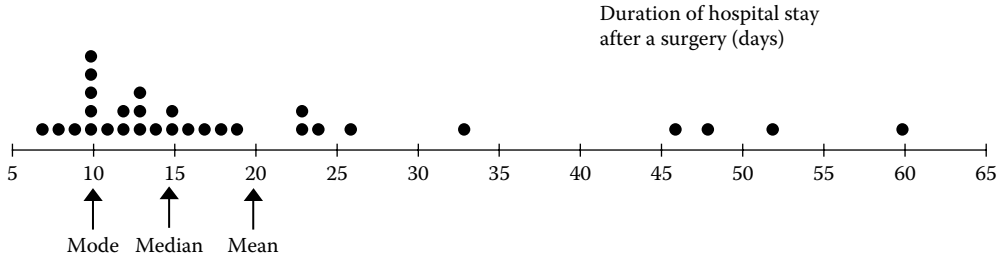



**FIGURE C.6** Schematic of mean, median, and mode.

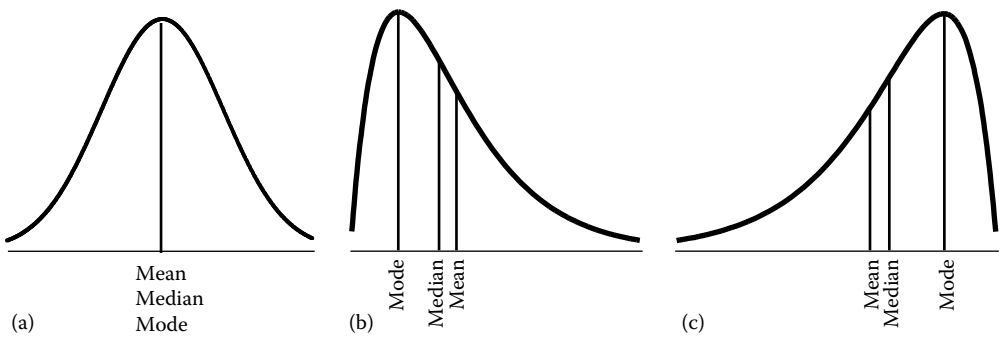



**FIGURE C.7** Location of mean, median, and mode in (a) symmetric, (b) right-skewed, and (c) left-skewed distributions.

Our guideline for choosing an appropriate central value is as follows: Always prefer mean to represent the central value except when (i) outliers are present (use the median then) or (ii) the interest is specifically in the most common value (use the mode then). If multiple modes are present, use the mean. In a large number of situations in medicine and health, the opportunity to choose may not arise because the three central values tend to be the same. Nonetheless, use the following precautions:

- These guidelines are for representing values in one sample, and not applicable to statistical inference such as confidence intervals and tests of hypothesis where mean outscores due to **central limit theorem**.
- The mean and other measures of centrality are sometimes misused. One misuse occurs when they are cited without the accompanying variability of the underlying values. The mean of the fasting plasma glucose level could be 105 mg/dL when the range is 90–110 mg/dL and also when the range is 60–200 mg/dL. The same means in these two cases have entirely different implications. Mean must always be accompanied by a measure of **variation**.
- Mean can be ridiculous when outliers are present. If 8 out of 10 persons in a sample do not take alcohol and 2 persons take 120 mL each per day, it is unwise to say that the average intake per person is 24 mL. Similarly, the average number of legs per person in this world is less than 2 as some do not have one or both legs. Mean, median, and mode are much more meaningful when they are based on relatively homogenous groups. For example, the mean survival period of assorted patients admitted to a hospital could be an inappropriate measure of efficiency since there would be cases of hernia with minimal risk of death and cases of peritonitis with grave prognosis. The survival period could be widely different in these cases.
- The number of values on which a mean is based must always be stated. The mean of values in two or three subjects has little meaning. The mean of 5 subjects may be the same as that of 200 subjects, but the two means have very different reliability.
- The other common misuse occurs when a mean or median is applied to one particular case in place of a group. If the median survival time after detection of a malignancy is 3 years, it does not imply that a person with this malignancy is likely to survive for 3 years. Even when the term "likely" is added, the statement does not become valid. The only correct statement is that nearly half survive for 3 years or less and the other half for more than 3 years.
- Arriving at a conclusion in medicine based on mean alone without considering other correlates could be misleading. If an antihypertensive drug is able to reduce diastolic blood pressure by 10 mmHg on average after 1 week of use by patients with hypertension, other considerations such as side effects, cost, and convenience of intake cannot be ignored altogether while evaluating the usefulness of the drug.
- Most scientific applications would be careful about the misuses just mentioned, but popular media such as television channels and newspapers sometimes pick up average values and report them without sufficient care about the variability attached to them. Thus, the public is not properly informed. This can cause all sorts of problems regarding the perception of a disease or in understanding the gravity of a health problem.
- In the case of binary data coded as 0 for absence and 1 for presence, the mean is the proportion of the subjects with presence of the characteristic.

For calculation of mean, median, and mode in different setups, such as in grouped data, see **mean**, **median**, **mode** (**calculation of**).

C

## centroid method of clustering, *see also* cluster analysis

Centroid is one of the several methods of **hierarchical** agglomerative clustering where the distance between two clusters is measured by the distance between their means (centroids in case of multivariate data). Clustering is the process of putting similar units in one group and dissimilar units in distinct groups. In the hierarchical agglomerative process, the two most similar units are put into one group at the first stage. This group is now considered as one entity. Now the distance of this entity from other units is compared with the other distances between various pairs of units. Again, the closest are joined together. This hierarchical agglomerative process goes on in stages, reducing the number of entities by one each time. This is a bottom-up approach. The process is continued until all units are clustered together as one big entity, although it can be stopped midway when the desired kind of clusters is obtained. Note that in the hierarchical method, subsequent clusters completely contain previously formed clusters.

It may not be immediately clear how to compute the distance between two entities containing, say, $I$ and $J$ units, respectively. Several methods are available such as **single linkage**, **complete linkage**, and **average linkage**. Different methods can give different results. Consider entity A comprising $I$ units ($a_1$, $a_2$,…, $a_I$) and entity B comprising $J$ units ($b_1$, $b_2$,…, $b_J$). Then, in the centroid method, the distance between entity A and entity B is equal to the distance between the mean of units in entity A and the mean of units in entity B. In other words, find the mean $\bar{a}$ of all units of entity A and the mean $\bar{b}$ of all units in entity B, and then compute the distance between $\bar{a}$ and $\bar{b}$. The distance can be obtained by different methods, but the most popular is the square of the **Euclidean distance**. In this case, this is defined as $d_{AB} = (\bar{a} - \bar{b})^2$ when $\bar{a}$ and $\bar{b}$ are univariate, and as $d_{AB} = \Sigma(\bar{a} - \bar{b})^2$ when $\bar{a}$ and $\bar{b}$ are multivariate quantities (the sum is over the elements of the multivariate quantity). These are obtained after **standardization** (by subtracting mean and dividing by the standard deviation) so that large values of any particular variable do not disproportionately affect the distance. Since it is based on means, the distance calculated by the centroid method is not much affected by outliers compared with other methods (such as complete linkage and single linkage), but this may not be sufficiently good to obtain internally homogeneous and externally isolated groups.

The centroid distance between them is the distance between $\bar{a}$ and $\bar{b}$. This method was originally proposed by Sokal and Michener in 1958 [1]. For further details of the centroid method, see Everitt et al. [2]. This method is also called the **group average method**. Since in calculating this distance we have disregarded that entity A has more units and entity B has less units, this is called the **unweighted group average method** (**UPGMA**). In Figure C.8, entity A has 17 units and entity B has 6 units. When this is weighted by the number of units, this becomes the weighted method.

The group average method is seen as a compromise between single linkage and complete linkage methods. In single linkage, the distance between entities is considered to be the same as between the closest units of the two clusters (nearest neighbor), and in complete

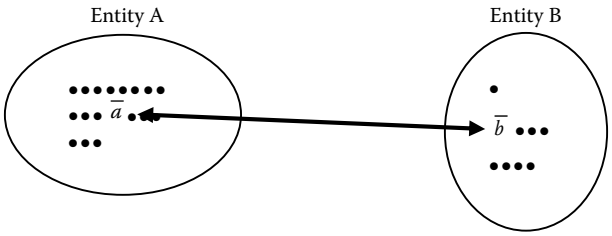



**FIGURE C.8** Centroid distance between entity A and entity B.

linkage the distance between the farthest units (farthest neighbor). Group average uses the distance between the entity averages.

We could not locate relevant examples in health and medicine where the group average method alone has been used. It has been mostly used in conjunction with other clustering methods. Feher and Schmidt [3] found group average the best out of many methods they tried for clustering molecular conformations.


1. Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull* 1958;38:1409–38. https://archive.org/details/cbarchive_33927_astatisticalmethodforevaluatin1902
2. Everitt BS, Landau S, Leese M, Stahl D. *Cluster Analysis*, Fifth Edition. Wiley, 2011.
3. Feher M, Schmidt JM. Metric and multidimensional scaling: Efficient tools for clustering molecular conformations. *J Chem Inf Comput Sci* 2001 Mar–Apr;41(2):346–53. http://pubs.acs.org/doi/abs/10.1021/ci000112%2B


## champaign glass effect, *see* health inequality

## chance, *see also* uncertainties

Anything unknown or too complex that makes it beyond human comprehension is termed chance. When we flip a coin, what determines that it will show head or tail? It probably depends on the size and weight of the coin, its position at the time of flipping, and the force with which flipping is done. Yet, can you flip it repeatedly to turn head each time? Such is the power of chance. When travelling in a vehicle, it is nearly impossible under normal circumstances to predict whether an accident will occur, and whether or not this will be fatal. Do we fully know why some women get breast cancer and others do not? **Risk factors** are called as such as they have a chance element, and all risk factors put together remain incapable of prediction with certainty. Because of the prominent chance element in all this, the prediction also is in terms of chance, statistically called **probability**.

Probability is by far the most common measure of chance. It quantifies and gives us a tool to differentiate one chance with the other. If the chance of death of a critically injured patient is 20% and of a typhoid patient is 0.4%, we know that the chance of death of a critically injured patient is not just steeply higher than that of a typhoid patient but is 50 times. This comparative evaluation of chance is hardly possible in any other way.

Chance is central to statistical thinking. If there is no chance, statistics does not apply. This is the basic difference between classical mathematics and statistics. The omnipresence of chance in health and medicine also explains why statistics pervades so much in medical sciences. This is what has given rise to a full-fledged science of **biostatistics**. No medical variable can be exactly predicted with certainty. There is always a need to attach probability to any medical prediction. On the contrary, for example, physical forces and their reactions can be exactly predicted. The role of chance is either absent or practically negligible in those sciences.

Several other examples can be cited. Ageing is a natural process, but some feel its effects more than others, to a varying degree. Some smokers develop lung cancer; others do not. Part of all such variation can be traced to factors such as personality traits, lifestyle, nutritional status, and genetic predisposition. However, these known factors fail to explain the entire variation. Two patients apparently similar, not just with regard to the disease condition but also for all other known factors, can respond differently to the same treatment regimen. Even susceptibility levels sometimes fail to account for all the variation. Thus, chance remains a factor. In some situations, these chance factors could be very prominent contributors to uncertainties, and in some situations, they can be minor.

## chance node, *see* decision analysis/tree

## chaotic measurements

Colloquially, chaos is a disorderly state, and chaotic measurements are the assessment of variables with haphazard values. Note that haphazardness is not random. Mathematically though, chaos has an enormous effect on outcome due to minor variation in initial or input values. Thus, this is intimately related to **butterfly effect** and is a consequence of extreme sensitivity of the outcome to the initial conditions. Chaotic systems limit our ability to predict even the deterministic phenomenon, least the stochastic phenomenon we encounter in health and medicine.

Sometimes chaotic values are confused with random variations. There is a clear distinction between the two. Chaotic values show a steep change over time or over stages with only a slight disturbance in initial values or input values, but they can still be largely predicted and modeled. This implies that we can say with considerable assurance what outcome is going to result when initial values are slightly different. This is mostly deterministic such as doubling at each passing stage. This cannot be done with random variations, although these variations too in statistics are modeled to follow a Gaussian pattern in the long run. This implies that random variations remain close to zero or near the amount of bias most of the time for fixed inputs. The chaotic values, on the other hand, can be very high, which in statistics are termed outliers.

For an application of chaotic measurements in modeling epileptic brain, see Fiasché et al. [1].


1. Fiasché M, Schliebs S, Nobili L. Integrating neural networks and chaotic measurements for modelling epileptic brain. *Lecture Notes in Computer Science* 2012;7552:653–60. http://link.springer.com/chapter/10.1007%2F978-3-642-33269-2_82


## charts (statistical)

Charts are organization of knowledge into some kind of figure for visual depiction. Statistical charts can be graphical or textual. For **graphs**, see this topic. In addition to those described there, we have also described **agreement charts**, **control charts**, **growth charts**, and **pedigree charts** under these topics. All these are factually diagrams. A more appropriate use of the term chart is for a figure that organizes *textual* information into boxes for depicting linkages of concepts, events, or activities. One such chart is in Figure C.9 on conventional diabetes screening. Specifically, this is called a *schematic chart*. Notice how it summarizes the steps to be taken at each phase of screening.

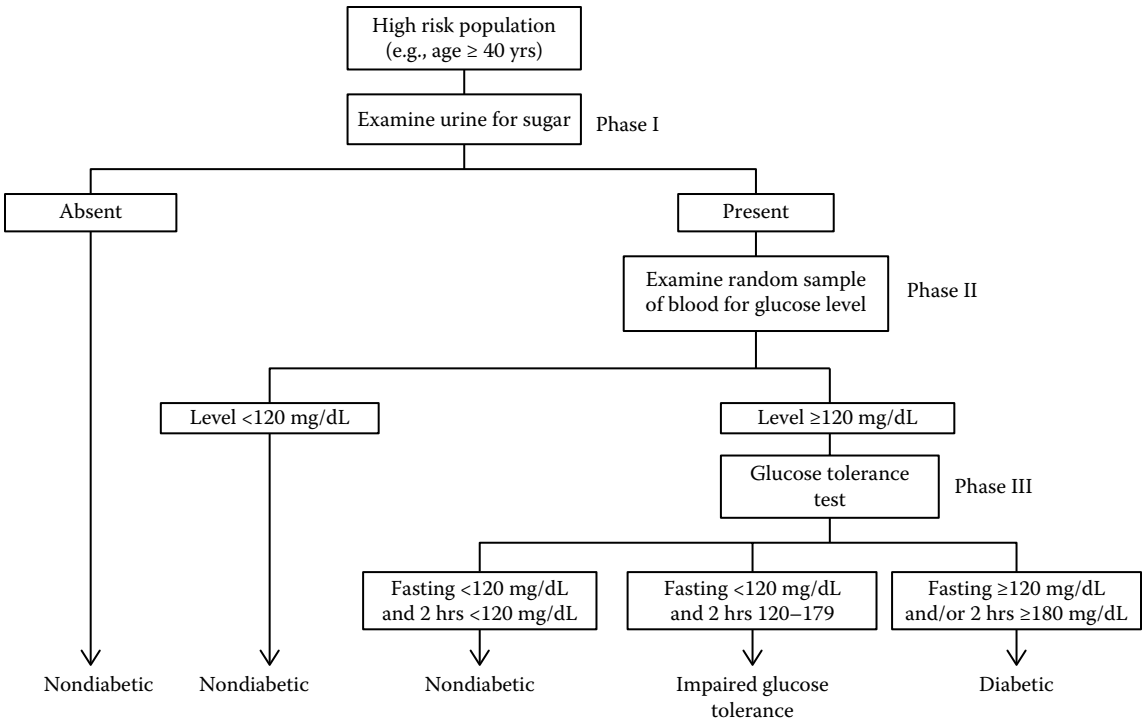



**FIGURE C.9** Scheme of multiphasic screening for diabetes mellitus.

There are other types of charts in health and medicine. Some of these are the Snellen chart, the retrospective chart, the medication chart, and the color chart. These are not statistical charts and are not in our purview.

## child mortality rate, *see* mortality rates

## chi-square—overall

A technical brief of chi-square distribution appears toward the end of this section. In health and medicine, we are mostly concerned about its application and we start with the applied aspects.

Chi-square is a versatile statistical method for the testing of several types of **hypothesis** relating to count of units (frequencies) in different groups. The most elementary of these is for finding whether or not frequencies in different groups follow a specified pattern. This is called a test for **goodness-of-fit**. Incidentally, this is also an easy portal to explain the essentials of the method of chi-square as follows.

Consider the data in the top row of Table C.5. This is a one-way table that gives blood group pattern of a sample of 150 cases of HIV. The problem we address is to find whether HIV positivity has any **association** with blood group pattern. This is the same as finding whether any blood group is more commonly seen in HIV cases. This can be tested only with comparison of the blood group pattern in the population from which these cases have come. Suppose the pattern in the population for blood groups O, A, B, and AB is 6:5:8:1. Since 6 + 5 + 8 + 1 = 20, this hypothesis in terms of probabilities is

$$H_0: \pi_1 = 6/20 = 0.30, \pi_2 = 5/20 = 0.25, \pi_3 = 8/20 = 0.40, \text{ and } \pi_4 = 1/20 = 0.05.$$

Note that the sum of these probabilities is 1.00.

**TABLE C.5**
**Calculation of Chi-Square for Testing the Null Hypothesis Regarding the Specified Pattern in HIV Cases**

| | Blood Group | | | | |
|---|---|---|---|---|---|
| | **O** | **A** | **B** | **AB** | **Total** |
| Observed frequency ($O_k$) | 57 | 36 | 51 | 6 | **150** |
| Expected frequency under $H_0$ ($E_k$) | 45.0 | 37.5 | 60.0 | 7.5 | **150.0** |
| $O_k - E_k$ | 12.0 | –1.5 | –9.0 | –1.5 | **0** |
| $(O_k - E_k)^2/E_k$ | 3.20 | 0.06 | 1.35 | 0.30 | **4.91 = $\chi^2$** |

Denote the observed frequencies in the four blood groups in the sample by $O_1$, $O_2$, $O_3$, and $O_4$, respectively. That is, $O_1 = 57$, $O_2 = 36$, $O_3 = 51$, and $O_4 = 6$. If $H_0$ is really true, then *expected frequencies* (see **cell frequency**), denoted by $E$'s, would be in the ratio specified in the hypothesis, i.e., $E_k = n\pi_k$ ($k$ = 1, 2, 3, 4). For $n$ = 150, $E_1$ = 150 × 0.3 = 45. Similarly, $E_2 = 37.5$, $E_3 = 60$, and $E_4 = 7.5$. A large difference between $O$'s and $E$'s would suggest that the observed pattern is different from that stipulated in the null hypothesis. This would be evidence against $H_0$ and in favor of $H_1$. It seems that the examination of the differences $(O_k - E_k)$ for different $k$ would be helpful. Since the total of the expected frequencies has to be the same 150 as that of the observed frequencies, it is imperative that some of these differences are negative and some positive, and the sum $\Sigma(O_k - E_k)$ would be always 0. As in the case of deviations $(x_i - \bar{x})$ for calculating SD, the square of these differences gets rid of the negative sign. This gives $(O_k - E_k)^2$. The magnitude of these squares is the key to the plausibility of $H_0$. But a difference of 1.5 over the expected 7.5 in blood group AB has a different meaning than the same difference over the expected 37.5 in blood group A. The former difference is one-fifth

of the corresponding expected frequency, whereas the latter is not even one-twentieth. Thus, the squared differences should be viewed in relation to the expected frequencies. The quantity $[(O_k - E_k)^2/E_k]$ becomes relatively free of the differentials in the magnitude of the expected frequencies in different groups and helps to give nearly equal weight to the groups. In place of taking the average of these quantities, this time, obtain the sum $\Sigma[(O_k - E_k)^2/E_k]$. This quantity is based entirely on frequencies and thus is unit free. This obviates the need to take the square root as is done at the time of calculating SD. To indicate that the quantity is a square, the sum is called chi-square ($\chi^2$). This is the **test criterion** in this case. Thus,

$$\text{chi-square (one-way table): } \chi^2 = \sum \frac{(O_k - E_k)^2}{E_k};\ k = 1,2,\ldots,K,$$

where $K$ is the total number of cells in the **contingency table**. In our example, $K = 4$.

Note that $E$'s are obtained assuming that $H_0$ is true. Thus, the value of $\chi^2$ in the formula given in this equation is under $H_0$. When $H_0$ is true, the difference between $O_k$ and $E_k$, i.e., $(O_k - E_k)$, should be small, and consequently the value of $\chi^2$ should also be small. In other words, a large value of $\chi^2$ is unlikely if $H_0$ is true. If the sample gives a large $\chi^2$, it provides evidence against $H_0$.

The ***P*-value** in this case is the probability of occurrence of the value of the criterion as extreme as or more extreme than obtained for the sample data. This requires distribution of the criterion under $H_0$. Such a distribution of $\chi^2$ is known. The exact shape of the distribution varies according to what is called the degrees of freedom (df). The df, in turn, depends mostly on the number of cells $K$. The concept of **degrees of freedom** is explained separately. The software automatically finds the df also and provides the ***P*-value**. If this $P$-value is less than the predetermined level of significance $\alpha$, reject the null; otherwise, do not reject the null.

In the case of the data in Table C.5, the df = $K - 1 = 4 - 1 = 3$. The calculations are presented in subsequent rows of this table. These show $\chi^2 = 4.91$. If a computer software package is used, it will automatically compare the calculated value of $\chi^2$ with its known distribution for 3 df and give $P = 0.178$. Otherwise, because 4.91 is less than the critical value 7.815 of chi-square for 3 df at 5% level of significance, the $P$-value is more than 0.05. Thus, the value 4.91 of $\chi^2$ obtained for these data is not all that unlikely when $H_0$ is true. That is, the frequencies observed in different blood groups in the example are not very inconsistent with the $H_0$. The sample values do not provide sufficient evidence against $H_0$, and it cannot be rejected. A preponderance of any blood group in cases of HIV cannot be concluded on the basis of this sample.

This inference is drawn despite an apparently clear excess of blood group O (57 subjects versus an expected 45) in this sample of HIV cases. This is because such a frequency *pattern* is not very unlikely to occur when the sample comes from the general population where the blood group ratios are as given in the null hypothesis. For further analysis, see **partitioning of chi-square**.

### Cautions in Using Chi-Square Test

The chi-square test does not require the frequency pattern to be Gaussian nor does it require any other specific pattern. Thus, the *chi-square test is a distribution-free procedure*. The following cautions are still needed:

- The use of chi-square for categorical data is well established, but $\chi^2$ itself is a **continuous variable**. Theoretically, it is based on an approximation. This approximation works fine when the expected frequency in any cell is not less than 5. When the number of categories $K$ is large, there can be a small relaxation. A rule of thumb is as follows: not more than one-fifth of categories (i.e., $K$/5 cells) should have $E_k < 5$, and almost none should be less than 1. When small frequencies are expected in many cells under $H_0$, either because of a small sample or because of very small $\pi$ in some cells, an exact **multinomial test** should be used.
- It is necessary to realize that chi-square is calculated from the actual frequencies in the cells. Percentages cannot be used.
- The chi-square test is basically a two-tailed test. Statistical significance in this case implies only presence of some difference from $H_0$, and it can seldom be labeled positive or negative. If the observed frequency is less than the expected frequency in one cell, it has to be more in one or more of the other cells because the total for both the observed and the expected frequencies is the same. In our example, the observed frequency is more for blood group O but less for the other three blood groups. Thus, the alternative hypothesis $H_1$ is two-sided. Some other tests, particularly when there are only two cells (**binary variable**), can have a one-sided alternative.
- The $\chi^2$ criterion is the sum $\Sigma[(O_k - E_k)^2/E_k]$. This would be large even if one particular difference $(O_k - E_k)$ is large. Thus, rejecting $H_0$ tells us only that there is at least one cell where the observed frequency is substantially different from the expected under the null hypothesis. It does not say where. On the other hand, if a large difference is present in only one cell, this can be masked by small differences in the other cells. This is what might be happening for blood group O in our example. For a more focused inference, further analysis by partitioning of chi-square may be helpful.

### Chi-Square Distribution

Mathematically, chi-square with $\nu$ df is defined as the sum of squares of $v$ independent standard Gaussian variates. As explained under that topic, if a variable $x_i$ has a **Gaussian distribution** with mean $\mu_i$ and variance $\sigma_i^2$, the standard Gaussian variate is $(x_i - \mu_i)/\sigma_i$. Thus, when $x_i$'s are independent,

$$\text{chi-square: } \chi^2 = \sum_{i=1}^{v} \frac{(x_i - \mu_i)^2}{\sigma_i^2}.$$

The degrees of freedom depend on the number of *independent* variates. When there is a restriction, such as that the total of observed frequencies is the same as the total of expected frequencies, the number of independent variates decreases by the number of restrictions. A different distribution of $\chi^2$ for different df is analogous to a different distribution of diastolic blood pressure in different age groups. Shapes of the distribution for some specific df's (denoted by $n$ in this figure) are given in Figure C.10. This gives you an idea how different df's can provide very different shapes and consequently different $P$-values. Also note that the shape of the chi-square distribution quickly looks like that of a Gaussian distribution even for 10 df.

Our description of chi-square is by way of an introduction to this extremely versatile procedure. As mentioned earlier, the chi-square test is used in a large number of setups. We have divided them into

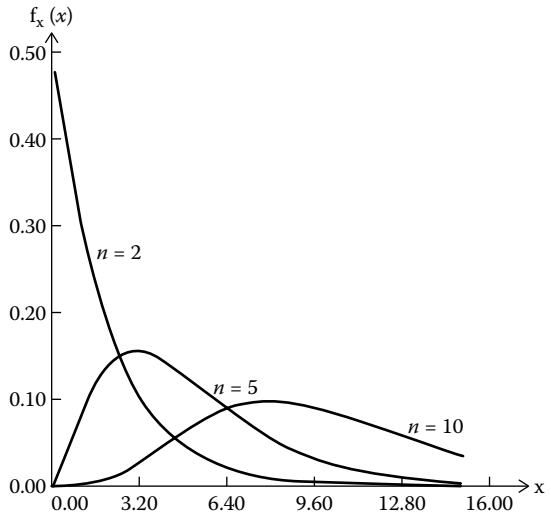



**FIGURE C.10** Shape of the distribution of chi-square for specified df's (denoted by *n* in this figure). (From Dudewicz EJ. *Introduction to Statistics and Probability.* Holt, Rinehart and Winston, 1976: p. 457.)

small segments to retain focus and to keep the text manageable. For details, see the next few topics.

1. Dudewicz EJ. *Introduction to Statistics and Probability.* Holt, Rinehart and Winston, 1976: p. 457.

## chi-square test for 2 × 2 tables, *see also* chi-square test for odds ratio and relative risk

If you are not familiar with the basics, see **chi-square—overall**. Among a large number of applications of chi-square, this section is restricted to two dichotomous variables, i.e., both characteristics observed as present or absent. One dichotomous variable could identify the presence or absence of the characteristic of interest, and the second variable could identify groups such as with and without disease, with disease A and with disease B, male and female, young and old, or any other such groups. These groups are independent, and the setup is essentially **bivariate**. Since both variables have two categories, this gives rise to a 2 × 2 table. This can also arise in a variety of other situations. One common situation is in calculating the **odds ratio** and the **relative risk**. The kind of contingency table arising in such cases is shown in Table C.6. $O_{rc}$ is the observed frequency in the $(r, c)$th cell ($r = 1, 2$; $c = 1, 2$), and inside the parentheses in each cell in the table are the corresponding probabilities. The dot in the subscript is for the corresponding total. The table is shown as though the relationship between antecedent and outcome is under study, but in fact, these can be any two dichotomous categories. This is also known as a *fourfold table*. There is a numerical example later in this section that provides a concrete view of this kind of table.

**TABLE C.6**
**General Structure of a 2 × 2 Contingency Table**

| Variable 2 (Outcome) | Variable 1 (Antecedent) Present | Absent | Total |
|---|---|---|---|
| Present | $O_{11}$ ($\pi_{11}$) | $O_{12}$ ($\pi_{12}$) | $\boldsymbol{O_{1\bullet}(\pi_{1\bullet})}$ |
| Absent | $O_{21}$ ($\pi_{21}$) | $O_{22}$ ($\pi_{22}$) | $\boldsymbol{O_{2\bullet}(\pi_{2\bullet})}$ |
| **Total** | $\boldsymbol{O_{\bullet 1}(\pi_{\bullet 1})}$ | $\boldsymbol{O_{\bullet 2}(\pi_{\bullet 2})}$ | $\boldsymbol{n}$ |

Even though the null hypotheses of no **association in prospective**, **retrospective**, and **cross-sectional studies** are different as explained under that topic, and consequently the interpretation is different, it can be shown that the test criterion is the same for all of them. Under any of those three $H_0$'s, the expected frequency in the $(r, c)$th cell, when the samples are independent, is given by

$$E_{rc} = O_{r\bullet} * O_{\bullet c};\ r, c = 1, 2.$$

Consider, for example, a study where $H_0$: $\pi_{11} = \pi_{21}$. This hypothesis implies in the context of Table C.6 that the proportion of subjects with antecedent in the two outcome categories should be the same. Each of these proportions would be the same as in the two categories combined. A similar statement should also be true for subjects without antecedent. This gives

$$(E_{11}/O_{1\bullet}) = (E_{21}/O_{2\bullet}) = (O_{\bullet 1}/n)$$

and

$$(E_{12}/O_{1\bullet}) = (E_{22}/O_{2\bullet}) = (O_{\bullet 2}/n)$$

or

$$E_{11} = \frac{O_{1\bullet}O_{\bullet 1}}{n},\quad E_{21} = \frac{O_{2\bullet}O_{\bullet 1}}{n},\quad E_{12} = \frac{O_{1\bullet}O_{\bullet 2}}{n},\quad E_{22} = \frac{O_{2\bullet}O_{\bullet 2}}{n}.$$

These are exactly the same as stated in the formula given in the topic **cell frequency**. The hypothesis of homogeneity in prospective and retrospective studies can also be shown to lead to the same formula.

Now the test criterion is

$$\text{chi-square for } 2 \times 2 \text{ table: } \chi^2 = \sum_{rc} \frac{(O_{rc} - E_{rc})^2}{E_{rc}};\quad r, c = 1, 2.$$

The justification is the same as explained for **chi-square**, and the applicability also requires each expected cell frequency to be at least five. In a 2 × 2 table, df = 1. There is freedom to choose the frequency arbitrarily in only one cell. The others are automatically decided because the row and column totals are considered fixed as illustrated in the following example. The test procedure is to calculate $\chi^2$ and find the probability $P$ of obtaining this or a higher value. A small value of $P$, as usual, is the evidence against $H_0$. If the $P$-value is sufficiently small, less than the predetermined level of significance, reject $H_0$; otherwise do not reject it. The whole procedure is illustrated in the following example.

Let the interest be in finding whether or not the prevalence of anemia in women is related to their parity status. Parity status is divided into two categories: two or less and three or more. Suppose the observed prevalence in a cross-sectional survey of 100 randomly selected women from a specified population is as given in Table C.7. In these data, 14 × 100/60 = 23% of women of parity ≤2 have anemia versus 16 × 100/40 = 40% women of parity ≥3. Is this association really present in the population?

The null hypothesis in this cross-sectional study would be that parity status has nothing to do with anemia status. This is the hypothesis of independence. That is, $H_0$: $\pi_{rc} = \pi_{r\bullet} * \pi_{\bullet c}$. If parity status has nothing to do with anemia status, then the ratio of anemics in both parity groups would be the same. That is, 30% of 60 women with parity ≤2 and 30% of 40 women with parity ≥3 should be anemic.

**TABLE C.7**
**Anemia and Parity in a Cross-Sectional Study of 100 Women**

| | Observed in the Survey | | | Expected under $H_0$ | |
|---|---|---|---|---|---|
| **Anemia** | **Parity ≤2** | **Parity ≥3** | **Total** | **Parity ≤2** | **Parity ≥3** |
| Present | 14 | 16 | **30** | 18 | 12 |
| Absent | 46 | 24 | **70** | 42 | 28 |
| **Total** | **60** | **40** | **100** | **60** | **40** |

Such expected frequencies under $H_0$ are also given in Table C.7. These can be verified to follow the formula given earlier. Thus,

$$\chi^2 = \frac{(14-18)^2}{18} + \frac{(16-12)^2}{12} + \frac{(46-42)^2}{42} + \frac{(24-28)^2}{28} = 3.17.$$

Only one cell frequency in a 2×2 table can be freely determined since the totals are considered fixed. If the number of nonanemics in the group with parity ≥ 3 is 18, then the number of anemics in this group has to be 22 (so that the total with parity ≥ 3 remains 40), that of anemics of parity ≤ 2 has to be 8, and that of nonanemics in this group has to be 52. Only that will keep the row and column totals unaltered.

A computer-based statistical package gives $P(\chi^2 \geq 3.17) = 0.075$. If the level of significance is $\alpha = 0.05$, this $P$-value is not sufficiently small. Thus, the $H_0$ cannot be rejected. The evidence in this sample of 100 women is not sufficient to conclude that the prevalence of anemia in women is related to parity status. The initial assumption ($H_0$) of no relation can be conceded in the absence of sufficient evidence against it.

There are several topics in this volume that are intimately related to this section. The same hypothesis can also be tested with the ***z*-test** for proportions. The $z$-test can also be used to find whether or not a predetermined medically significant difference is present between the two groups.

Sometimes the **Yates correction for continuity** is advised for this chi-square test for 2×2 tables. As already mentioned, the test described in this section is for large samples and requires that no expected cell frequency is less than five. For small samples, see the **Fisher exact test**. For paired data, see the **McNemar test**.

## chi-square test for larger two-way contingency tables

The preceding section is on **chi-square test for 2×2 tables**. This means that both the characteristics (or variables) are **dichotomous**. But a large number of variables are not dichotomous. Blood group has four categories. Subjects may be categorized for smoking as those who have never smoked, ex-smokers, mild smokers, and heavy smokers. All **polytomous** categories can be dichotomized, such as blood group into B and non-B groups and smoking into yes and no; but quite often, such dichotomy fails to serve the purpose. In some situations, such as in assessing the trend in proportion of sexually mature girls (with regard to, say, breast stage) of age 14 years with different grades of anemia, it is better to have as many grades of anemia as is feasible. Thus, $R \times C$ contingency tables are not uncommon, where the number of rows in the table is $R$ and the number of columns is $C$—either or both of them being more than 2.

The inference generally needed from such tables is whether the two variables are associated. If an association is found, further analysis can be done to measure the degree of **association**, to ascertain the presence of trend, if any, or to find which particular cell or cells in the contingency table are contributing to the relationship. The method for finding the presence or absence of association is basically the same for $R \times C$ tables as for 2×2 tables, although some generalization is needed. This section first discusses $2 \times C$ (or $R \times 2$) tables and then goes on to discuss $R \times C$ tables. The concern in this section is with a setup where both the variables are qualitative. If one of them is quantitative, examine whether methods such as **logistic regression** can be used. As in the case of 2×2 tables, the presumption throughout this section is that $n$ is large and additionally that the frequency in at least 80% of the cells is 5 or more. For small frequencies, the **Fisher exact test** can be extended to $R \times C$ tables. For more on this, see Mehta and Patel [1].

### *One Dichotomous and the Other Polytomous Variable (2×C Table)*

Consider an investigation on the relationship between dioxin dosage and enlarged prostate. Dioxins are by-products of combustion and other processes. They persist at lower levels virtually everywhere—in air, water, and soil. They are known to be disrupters that can offbalance the endocrine system. Children are particularly vulnerable. This can affect behavior, immune function, neurological development, and gender development. In an experiment, a group of 100 mouse fetuses are randomly divided into four equal groups and are exposed to none, low, medium, and heavy doses of a dioxin-like chemical. These dosages are in relative terms, but even a heavy dose would be a microdose. This experiment is along the lines reported by vom Saal et al. [2]. The numbers of fetuses that developed enlarged prostate are given in Table C.8. Some fetuses were lost and could not be observed.

The **null hypothesis** for rejection is that there is no effect of differential dose of dioxin on the incidence of enlarged prostate. If this $H_0$ is true, the enlarged prostate would be divided among the dose groups according to the total number of fetuses in each group. This means that 21 enlarged prostate cases should be in the ratio 23:25:25:20. This total is 93. The expected frequencies under $H_0$ therefore are 21 × 23/93, 21 × 25/93, 21 × 25/93, and 21 × 20/93, or 5.19, 5.65, 5.65, and 4.52, respectively. If the expected frequency in the $r$th row and the $c$th column is denoted by $E_{rc}$, it should be clear that

$$E_{rc} = O_{r\bullet} * O_{\bullet c}/n; \; r = 1, 2, \ldots, R; \; c = 1, 2, \ldots, C,$$

where $O_{r\bullet}$ is the total of the $r$th row, $O_{\bullet c}$ is the total of the $c$th column, and $n$ is the grand total. In this example, $R = 2$, $C = 4$, $O_{1\bullet} = 21$, $O_{2\bullet} = 72$, $O_{\bullet 1} = 23$, $O_{\bullet 2} = 25$, $O_{\bullet 3} = 25$, $O_{\bullet 4} = 20$, and $n = 93$. From the formula just given, $E_{11} = 5.19$, $E_{12} = 5.65$, $E_{13} = 5.65$, and $E_{14} = 4.52$. These are the same as obtained before. Also, $E_{21} = 17.81$, $E_{22} = 19.35$, $E_{23} = 19.35$, and $E_{24} = 15.48$. Now a test criterion is required.

**TABLE C.8**
**Dioxin Dosage and Enlarged Prostate in Mice**

| Enlarged Prostate | Dosage | | | | |
|---|---|---|---|---|---|
| | **None** | **Low** | **Medium** | **Heavy** | **Total** |
| Yes | 2 | 5 | 6 | 8 | **21** |
| No | 21 | 20 | 19 | 12 | **72** |
| **Total** | **23** | **25** | **25** | **20** | **93** |

### *Chi-Square for* R×C *Table*

Following the argument similar to that explained for **chi-square—overall**, the test criterion for an $R \times C$ table is

$$\text{chi-square for } R \times C \text{ table: } \chi^2 = \sum_{rc} \frac{(O_{rc} - E_{rc})^2}{E_{rc}};\ r = 1, 2, \ldots, R;\ c = 1, 2, \ldots, C,$$

where $E_{rc}$ is as already mentioned. The present section is restricted to the case in which $R = 2$ but $C > 2$; however, there is no such restriction for the above criterion. A large value of $\chi^2$ would indicate that the observed frequencies are very different from those expected under the null hypothesis and thus would provide evidence against the null. For it to be statistically significant, the $\chi^2$ value should be as large as is very unlikely to occur under $H_0$. As usual, it is considered very unlikely when the chances are less than the level of significance, say, 0.05. Statistical packages would readily give this *P*-value, so that a conclusion can be immediately drawn. This would again depend on the df. On the lines discussed under **degrees of freedom**, it can be seen for an $R \times C$ table that df = $(R - 1)(C - 1)$. In Table C.8, only $(2 - 1)(4 - 1) = 3$ cells out of 8 can be freely chosen because the row totals and column totals cannot be disturbed. This is the df in this example.

For the data in Table C.8, by the formula just given,

$$\chi^2 = \frac{(2-5.19)^2}{5.19} + \frac{(5-5.65)^2}{5.65} + \cdots + \frac{(12-15.48)^2}{15.48} = 6.13.$$

A statistical package gives $P = 0.105$ for 3 df. This shows that the chance of data coming from a population for which $H_0$ is true is not sufficiently small. The plausibility of $H_0$ is not adequately ruled out. Thus, $H_0$ cannot be rejected. The conclusion is that the dose level of the chemical does not significantly affect the proportion with enlarged prostate in these data.

The chi-square criterion in the above equation considers each dose level in this example on a nominal scale and is oblivious of its ordinal character. If the gradient or the trend is the concern, proceed as discussed under **chi-square test for trend in proportions and deviation from trend**. If there are repeated measures of the same subjects, use the **Cochran *Q* test**. For details, see Agresti [3].

### *Two Polytomous Variables*

Now we extend the method to $R \times C$ tables where both $R$ and $C$ are more than two. The objective is to find whether one qualitative variable is related to or affected by the other. For example, the interest may be in the relationship of extent of smoking (none, mild, moderate, heavy) to **social classification** of people. Or it may be on the influence of nutritional status (good, fair, poor) of pregnant women on maternal complications. Both variables are polytomous in these examples.

The basic method for finding whether or not the two qualitative variables with multiple categories are associated continues to be the same chi-square as given above provided the condition of large $n$ is met. The following example illustrates the method.

Douglas et al. [4] investigated the seasonality of sudden infant death syndrome (SIDS) by age at death in the United Kingdom. A total of 13,990 such deaths were observed. They provided data for each calendar month of death and each year of age. A summary is given in Table C.9. This table has $R = 4$ rows and $C = 3$ columns excluding totals. The total $n$ is very large, and no expected cell frequency under $H_0$ would be less than 5. Thus, this is a very appropriate case for applying the chi-square test. Statistically, this is a cross-sectional study in which 13,990 sudden deaths are divided by age at death and calendar month of death. The null hypothesis is that age at death of infants was not associated with the calendar month of death. The expected frequency under this $H_0$ can be obtained for each cell by the formula given earlier and the chi-square value by the above-mentioned formula or by statistical software. A software package obtained $\chi^2 = 40.46$ for these data. This has $(4 - 1)(3 - 1) = 6$ df. The package gives $P = 0.000$ (it should be stated as $P < 0.001$). This is far less than the conventional 0.05. Thus, $H_0$ was rejected, and it was concluded that age at death was indeed associated with the calendar month of death. Perusal of the data indicates that the deaths were proportionately more in January to April for those who died after 5 months of age. The biological implication of this result is not clear.

**TABLE C.9**
**Age and Calendar Month of Sudden Infant Deaths**

| Age at Death (Months) | Calendar Month of Death | | | |
|---|---|---|---|---|
| | Jan–Apr | May–Aug | Sep–Dec | Total |
| <2 | 831 | 490 | 745 | **2066** |
| 2–4 | 3163 | 1833 | 3022 | **8018** |
| 5–8 | 1457 | 750 | 1104 | **3311** |
| 9–12 | 283 | 140 | 172 | **595** |
| **Total** | **5734** | **3213** | **5043** | **13,990** |

Chi-square for $R \times C$ tables is valid for retrospective and prospective studies as much as for cross-sectional studies. Thus, the method remains the same for the three types of design. However, the nature of the null hypothesis, and hence of the conclusion, changes. As explained under **association in prospective, retrospective, and cross-sectional studies**, the null hypothesis in the case of a cross-sectional study is of independence of the two variables. If this is rejected, an association is concluded. In the case of retrospective and prospective studies, the null hypothesis is of a similar pattern of proportions in different groups. This is the hypothesis of homogeneity.

For completeness, it should be reiterated for $R \times C$ tables that the method of chi-square is again suitable only for large $n$. Strictly speaking, no expected cell frequency $E_{rc}$ should be less than five. As stated earlier, if there is a large number of cells, then possibly some relaxation can be made, but not more than 20% of expected cell frequencies should be less than five and none should be very small, say less than one. When the numbers are really small, you may have to collapse rows or columns or both. In an extreme case, the table may have to be collapsed to 2×2 and use the **Fisher exact test** if cell frequencies continue to be small. Collapsing should be done in a manner that biological relevance is not lost.

The chi-square test mentioned above evaluates only whether an association is present or not. The association may be present in specific cells, but the number of subjects in the other cells can mask it. This can be unmasked by a suitable **partitioning** of the table. Another point of interest may be the degree of association between two qualitative variables. For this, see **association between polytomous characteristics (degree of)**. For association in matched pairs with many categories ($C \times C$ tables), see the **McNemar–Bowker test**. The association in larger tables can also be studied by **log-linear models**.

1. Mehta CR, Patel NR. A network algorithm for performing Fisher's exact test in ($r \times c$) contingency tables. *J Am Stat Assoc* 1983;78:427–34. http://www.jstor.org/discover/10.2307/2288652?uid=3738256&uid=2&uid=4&sid=21103501109977

2. vom Saal FS, Timms BG, Montano MM et al. Prostate enlargement in mice due to fetal exposure to low doses of estradiol or diethylstilbestrol and opposite effects at high doses. *Proc Natl Acad Sci USA* 1997;94:2056–61. http://www.pnas.org/content/94/5/2056.full
3. Agresti A. *Analysis of Ordinal Categorical Data*. Wiley, 1984.
4. Douglas AS, Helms PJ, Jolliffe IT. Seasonality of sudden infant death syndrome (SIDS) by age at death. *Acta Paediatr* 1998;87:1033–38. http://onlinelibrary.wiley.com/doi/10.1111/j.1651-2227.1998.tb01409.x/abstract

C

## chi-square test for goodness-of-fit, *see* goodness-of-fit

## chi-square test for odds ratio and relative risk, *see also* chi-square test for 2 × 2 tables

Those not familiar with the **odds ratio (OR)** and **relative risk (RR)** may like to view these topics first. Also see **test of hypothesis** for apprising yourself about statistical tests. The chi-square test is used to test the null hypothesis ($H_0$) that OR = 1 or RR = 1 depending on whether the study is retrospective (from the outcome to the antecedent) or prospective (from the antecedent to the outcome). These two null hypotheses say that there is no relationship between the antecedent and the outcome. We explain these tests for OR and RR separately in the following paragraphs. These are basically same as the **chi-square test for 2 × 2 tables** for two independent samples.

### *Chi-Square Test for OR*

The $H_0$ in this case almost invariably is that OR = 1. This says that the presence of the antecedent is as common in cases as in controls. Since OR = RR if the outcome is rare, $H_0$: OR = 1 also says that the presence or absence of the antecedent does not influence the outcome in situations where the outcome is rare. A simple statement that takes care of both the directions of the relationship is that there is no association under $H_0$ between the antecedent and the outcome. The alternative could be one-sided ($H_1$: OR < 1 or $H_1$: OR > 1) or two-sided ($H_1$: OR ≠ 1). The latter is applicable when there is no a priori assurance that the relationship could be one-sided. The two-sided hypothesis is tested by classical **chi-square**. For the one-sided alternative, use ***z*-test**.

Suppose that 80 promiscuous males who were recently found to be HIV positive and 160 promiscuous controls (two controls per case) of the same age and socioeconomic–cultural–ethnic background are asked about their sexual behavior in terms of average frequency of extramarital contacts per month during the last 2 years. The data obtained are in Table C.10.

**TABLE C.10**
**Extramarital Contacts in HIV+ Subjects and Controls**

| | Average Extramarital Contacts/Month | | |
|---|---|---|---|
| **Group** | **One or More** | **Less Than One** | **Total** |
| HIV+ | 38 | 42 | **80** |
| Controls | 51 | 109 | **160** |
| **Total** | **89** | **151** | **240** |

The column totals in this case are 89 and 151, respectively, and $n = 240$. Thus, for chi-square, the expected **cell frequency** $E_{ij}$ in the $(i, j)$th cell is given by

$$E_{11} = \frac{80 \times 89}{240} = 29.67, \quad E_{12} = \frac{80 \times 151}{240} = 50.33,$$

$$E_{21} = \frac{160 \times 89}{240} = 59.33, \quad E_{22} = \frac{160 \times 151}{240} = 100.67.$$

Therefore, $\chi^2 = \Sigma[(O - E)^2/E]$ gives

$$\chi^2 = \frac{(8.33)^2}{29.67} + \frac{(8.33)^2}{50.33} + \frac{(8.33)^2}{59.33} + \frac{(8.33)^2}{100.67} = 5.58.$$

A computer package gives for 1 df $P(\chi^2 \geq 5.58) = 0.018$. The *P*-value is exceedingly small. There is an exceedingly small chance that this sample has come from a population where $H_0$: OR = 1 is true. Reject $H_0$ and conclude that OR ≠ 1. HIV positivity is associated with one or more average extramarital contacts per month in promiscuous males. In this case, OR = (38 × 109)/(51 × 42) = 1.93. This shows that the odds of one or more extramarital contacts per month are nearly twice as much in HIV-positive males as in the controls.

### *Chi-Square Test for RR*

This is similar to the test for OR just described. The null hypothesis is RR = 1 (no relationship). Consider the data in Table C.11 on lower tract illness in boys born to women of age <26 and ≥26 years.

For the data in Table C.11, the expected frequencies under $H_0$ can be calculated as usual, and chi-square can be calculated by the formula given for OR. This gives

$$\chi^2 = \frac{(48 - 37.29)^2}{37.29} + \frac{(117 - 127.71)^2}{127.71} + \frac{(65 - 75.71)^2}{75.71} + \frac{(270 - 259.29)^2}{259.29} = 5.93.$$

A statistical software package gives $P = 0.015$. The null hypothesis is rejected. Conclude that RR ≠ 1.

Note that chi-square is a two-tailed test and so the conclusion too is two-sided. If there is an a priori reason to ensure that RR would be more than 1, then $H_1$: RR > 1. It may be prudent then to use ***z*-test**.

## chi-square test for three-way contingency tables

If you are not already familiar with chi-square for two-way tables, familiarize yourself with the details described under the **chi-square**

**TABLE C.11**
**Maternal Age and Lower Respiratory Tract Illness (LRI) in Infant Boys**

| | LRI | | |
|---|---|---|---|
| **Maternal Age (Years)** | **Yes** | **No** | **Total** |
| <26 | 48 | 117 | **165** |
| ≥26 | 65 | 270 | **335** |

**test for larger two-way contingency tables**. These methods are extended in this section for three-way tables.

A three-way contingency table arises when the classification of the subjects is done with respect to three variables. Thus, this is a genuine multivariate categorical data setup—a step more than the bivariate setup in two-way contingency tables. The variables could be either on a nominal or on an ordinal scale. If any of them is on a metric scale, then it should be either discrete with a small number of values or continuous divided into a small number of broad categories for the methods of this section. This means that a variable such as aspartate aminotransferase (AST), formerly called serum glutamic-oxaloacetic transaminase (SGOT), should have categories such as (in U/L) 0–49, 50–99, and 100+ at least for constructing a contingency table and not small categories like 0–9, 10–19, 20–29, etc. In any case, the categories should be **mutually exclusive and exhaustive**. The interest must be in proportions of subjects in different categories and not the mean level.

Table C.12 is an example of a three-way table where the variables are visual acuity (VA), age group, and gender. Besides row and column, the third dimension is called *layer*. The numbers of rows, columns, and layers can be denoted by $R$, $C$, and $L$, respectively. In Table C.12, $R = 5$, $C = 3$, and $L = 2$. We have chosen to consider VA categories as main columns and gender as a subclassification (layer). You may wish to switch the labels, and that would be equally valid. The order of this table is $5 \times 3 \times 2$. The body of the table has 30 cells. The columns and rows with totals are not included in this count.

Denote the probability that a random person from the target population falls in the $(r, c, l)$th cell by $\pi_{rcl}$ ($r = 1, 2, \ldots, R$; $c = 1, 2, \ldots, C$; $l = 1, 2, \ldots, L$), and the observed frequency in the cell by $O_{rcl}$. The methods described in the present section are valid for large $n$ and need $E_{rcl} \geq 5$ in at least 80% of the cells, where $E$ is the expected cell frequency under the null hypothesis ($H_0$). One method for obtaining $E$'s for three-way tables will be described shortly. The case for small $n$ is too complex. In any case, it is seldom prudent to classify a small number of subjects simultaneously by three characteristics. Each of these characteristics will have a minimum of two categories, and thus, there are at least $(2 \times 2 \times 2 =) 8$ cells in a three-way contingency table.

The analysis of three-way contingency tables described in this section is not necessarily a comparison of three groups. One group of subjects can be cross-classified by levels of two variables. For example, Table C.12 can be considered to cross-classify subjects in three VA groups by age and gender.

As in the case of two-way tables, the null hypothesis in a three-way table could be that of homogeneity of different types, or of independence, depending upon individual variables being factors or responses. In Table C.12, if a sample of 520 males and 480 females was chosen and their age and VA elicited, then gender is a factor and the other two (VA and age) are responses. In this case, $H_0$ could be that age and VA have the same pattern of relationship in males as in females. If a sample of 1000 subjects is chosen and they are cross-classified by age, gender, and VA, then all three are responses, and $H_0$ is that of independence of the variables. Luckily, no matter what the type of $H_0$ is, the calculation proceeds along the same lines. Only the interpretation differs. The calculation is in terms of the usual $\chi^2 = \Sigma(O - E)^2/E$, although the calculations of $E$'s are not so straightforward in this case. These are illustrated in the following example.

Consider a survey carried out in Brazil among female family planning clients to study the profile of women who approve of sterilization [1]. The profile was studied in terms of the age of the women, the number of living children (LC), and the age of the youngest child. The data obtained on 1250 approvers are given in Table C.13.

**TABLE C.13**
**Women Approving Sterilization in a Survey in Brazil**

| | Observed Frequencies | | | |
|---|---|---|---|---|
| **Age of Woman and LC[a]** | **Age of the Youngest Child (Years)** | | | |
| | **–1** | **1–4** | **5+** | **Total** |
| | **Age ≤ 30 Years** | | | |
| LC < 3 | 37 | 95 | 22 | **154** |
| LC = 3 | 63 | 58 | 12 | **133** |
| LC > 3 | 77 | 47 | 3 | **127** |
| **Total 1** | **177** | **200** | **37** | ***414*** |
| | **Age > 30 Years** | | | |
| LC < 3 | 40 | 91 | 176 | **307** |
| LC = 3 | 57 | 65 | 79 | **201** |
| LC > 3 | 136 | 105 | 87 | **328** |
| **Total 2** | **233** | **261** | **342** | ***836*** |
| | **All Ages** | | | |
| LC < 3 | 77 | 186 | 198 | **461** |
| LC = 3 | 120 | 123 | 91 | **334** |
| LC > 3 | 213 | 152 | 90 | **455** |
| **Total** | ***410*** | ***461*** | ***379*** | ***1250*** |

*Source:* Adapted from Table 3 of Lassner KJ et al., *Studies Fam Plan* 1986;17:188–98. http://www.jstor.org/discover/10.2307/1966936?uid=3738256&uid=2&uid=4&sid=21103501420287. With permission.

[a] LC, number of living children.

**TABLE C.12**
**Distribution of 1000 Subjects Coming to a Cataract Clinic by Age, Gender, and Visual Acuity (VA) in the Worse Eye**

| Age Group | VA ≥ 6/60 | | | 6/60 < VA ≤ 1/60 | | | VA < 1/60 | | | Total | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **(Years)** | **M** | **F** | **P** | **M** | **F** | **P** | **M** | **F** | **P** | **M** | **F** | **P** |
| –49 | 11 | 8 | 19 | 37 | 32 | 69 | 12 | 10 | 22 | **60** | **50** | **110** |
| 50–59 | 18 | 21 | 39 | 69 | 73 | 142 | 13 | 16 | 29 | **100** | **110** | **210** |
| 60–69 | 25 | 21 | 46 | 183 | 142 | 325 | 42 | 47 | 89 | **250** | **210** | **460** |
| 70–79 | 10 | 11 | 21 | 54 | 44 | 98 | 26 | 25 | 51 | **90** | **80** | **170** |
| 80+ | 3 | 4 | 7 | 9 | 14 | 23 | 8 | 12 | 20 | **20** | **30** | **50** |
| **Total** | **67** | **65** | **132** | **352** | **305** | **657** | **101** | **110** | **211** | **520** | **480** | **1000** |

C

This is a cross-sectional study and all the three variables are responses. Thus, the appropriate $H_0$ is of independence, which in this case says that sterilization approvers are not concentrated in any specific combination of the categories of age, LC, and age of the youngest child.

Consider LC to be in rows so that $R = 3$, age of the youngest child in columns so that $C = 3$, and the age of woman in layers so that $L = 2$. The corresponding indexing subscripts are $r$, $c$, and $l$, respectively. The bottom part of the table is a two-way table with age collapsed. Total 1 and Total 2 give another two-way table with LC collapsed. The last row is a one-way table with age of women and LC collapsed.

Following a procedure similar to the one used for two-way tables, the expected frequencies for some cells under $H_0$ are calculated next. These are obtained by multiplication of the corresponding row, column, and layer totals divided by the grand total as illustrated in the following:

$$E_{111} = 1250 \times \frac{461}{1250} \times \frac{410}{1250} \times \frac{414}{1250} = 50.08 \text{ (observed} = 37),$$

$$E_{211} = 1250 \times \frac{334}{1250} \times \frac{410}{1250} \times \frac{414}{1250} = 36.28 \text{ (observed} = 63),$$

$$E_{311} = 1250 \times \frac{455}{1250} \times \frac{410}{1250} \times \frac{414}{1250} = 49.43 \text{ (observed} = 77), \text{ and}$$

$$E_{112} = 1250 \times \frac{461}{1250} \times \frac{410}{1250} \times \frac{836}{1250} = 101.13 \text{ (observed} = 40).$$

These frequencies are for the first four cells (excluding Total 1) of the first column in the table. The numerators are the corresponding marginal totals. All other expected frequencies can be calculated in a similar manner. These and their totals are arranged in Table C.14.

**TABLE C.14**
**Expected Frequencies of Women in the Example When They Are Not Affected by Age of Woman, Number of Living Children (LC), and Age of the Youngest Child**

| | Expected Frequencies | | | |
|---|---|---|---|---|
| | Age of the Youngest Child (Years) | | | |
| **Age of Woman and LC** | **–1** | **1–4** | **5+** | **Total** |
| | **Age ≤ 30 Years** | | | |
| LC < 3 | 50.08 | 56.31 | 46.29 | **152.68** |
| LC = 3 | 36.28 | 40.80 | 33.54 | **110.62** |
| LC > 3 | 49.43 | 55.58 | 45.69 | **150.70** |
| **Total 1** | **135.79** | **152.69** | **125.52** | ***414*** |
| | **Age > 30 Years** | | | |
| LC < 3 | 101.13 | 113.71 | 93.48 | **308.32** |
| LC = 3 | 73.27 | 82.38 | 67.73 | **223.38** |
| LC > 3 | 99.81 | 112.23 | 92.26 | **304.30** |
| **Total 2** | **274.21** | **308.32** | **253.47** | ***836*** |
| | **All Ages** | | | |
| LC < 3 | 151.21 | 170.02 | 139.77 | **461** |
| LC = 3 | 109.55 | 123.18 | 101.27 | **334** |
| LC > 3 | 149.24 | 167.81 | 137.95 | **455** |
| **Total** | ***410*** | ***461*** | ***379*** | ***1250*** |

Note that the totals given in bold italics match the corresponding observed totals in Table C.13. The calculation of $\chi^2$ is done as usual with the following formula:

$$\text{three-way table: } \chi^2 = \sum_{rcl} \frac{(O_{rcl} - E_{rcl})^2}{E_{rcl}};$$

$$r = 1, 2, \ldots R;\ c = 1, 2, \ldots C;\ l = 1, 2, \ldots, L.$$

Under $H_0$, this follows a chi-square distribution with df = $(R - 1)(C - 1)(L - 1)$. In this example, df = $(3 - 1)(3 - 1)(2 - 1) = 4$ and $\chi^2 = 277.47$. A software package gives $P << 0.05$. Thus, the hypothesis of independence is rejected. Conclude that at least two of the three variables are associated among the family planning approvers.

The detailed calculations are not shown. The purpose is to illustrate the *type* of calculations required in a three-way setup. Analysis of such tables of higher dimensions is indeed complex, but this should not be a worrying factor as long as valid conclusions can be drawn. Statistical software packages do such calculations easily.

Chi-square in the above formula gives an overall test. If significant, it indicates only that an association is present *somewhere*. To find exactly where this association is, one approach is partitioning as discussed under **partitioning of chi-square** for a one-way table. The second and more versatile approach is using **log-linear models**. When one of the variables can be considered dependent as in the case of prospective and retrospective studies, use **logistic regression**. This can handle a large number of variables simultaneously.

1. Lassner KJ, Janowitz B, Rodrigues CMB. Sterilisation approval and follow-through in Brazil. *Studies Fam Plan* 1986;17:188–98. http://www.jstor.org/discover/10.2307/1966936?uid=3738256&uid=2&uid=4&sid=21103501420287

## chi-square test for trend in proportions and deviation from trend

A version of **chi-square** is available that can be used for testing the presence or absence of an increasing or decreasing trend in proportions in **ordinal** categories. Before we give the details of the procedure followed for this test, consider the following example.

The data in Table C.15 are from an experiment on mice exposed to different dosages of dioxin to find the effect on developing enlarged prostate. It is easy to see from the first three rows of Table C.15 that the proportion with enlarged prostate increases with dose of dioxin. But that is the observation in this sample. Is there a substantial likelihood that the trend will persist in repeated samples?

**TABLE C.15**
**Dioxin-Treated Fetal Mice with Enlarged Prostate and Some Calculations for Chi-Square for Trend**

| | Dosage Score ($x_k$) | | | | |
|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3** | **Total** |
| Number of mice ($n_k$) | 23 | 25 | 25 | 20 | **93** |
| With enlarged prostate ($O_{1k}$) | 2 | 5 | 6 | 8 | **21** |
| Proportion with enlarged prostate ($p_k = O_{1k}/n_k$) | 0.09 | 0.20 | 0.24 | 0.40 | **0.2258** |
| $O_{1k}x_k$ | 0 | 5 | 12 | 24 | **41** |
| $n_k x_k$ | 0 | 25 | 50 | 60 | **135** |
| $n_k x_k^2$ | 0 | 25 | 100 | 180 | **305** |

A study of trend in proportions is relatively simple when the ordinal categories can be assigned a valid **score**. For the data in Table C.15, these scores for dosages could be 0, 1, 2, and 3 for none, low, medium, and heavy doses, respectively. Note that such linear scores have an inbuilt assumption that the difference between the effect of the no-dose and low-dose categories is the same as that between the heavy- and medium-dose categories, the effect of a heavy dose is three times that of a mild dose, etc. These scores are considered as usual metric quantities amenable to algebraic manipulations. This may not be exactly true in many situations. Yet, such scores seem to work reasonably well as an approximation in most practical situations.

One can also think of nonlinear scores for assessing trends. Condom use among patients with sexually transmitted diseases can be categorized as never, sometimes, often, and almost always. Spouse infection percentage may follow a trend in this case depending on the frequency of condom use. The scores to the regularity of condom use can be given as 0 for never, 1 for sometimes, 3 for often, and 6 for always. A scoring that adequately expresses the intensity of categories is not easy to devise, but some methods of determining scores are discussed under **scoring systems** in this volume. Results for trend would differ depending upon what scores are used.

Some additional calculations shown in Table C.15 are needed to compute the chi-square for the trend in proportions. The proportion with an enlarged prostate in this example steadily rises from 0.09 for no dosage of a dioxin-like chemical to 0.40 for a heavy dosage. Among various methods available to find the statistical significance of this trend, one method is to calculate

$$\text{chi-square for trend: } \chi^2_{\text{trend}} = \frac{(\sum O_{1k}x_k - O_{1\cdot}\bar{x})^2}{p(1-p)(\sum n_k x_k^2 - n\bar{x}^2)};\ k = 1, 2, \ldots, K,$$

where $O_{1\cdot} = \Sigma O_{1k}$, $n = \Sigma n_k$, $p = O_{1\cdot}/n$, and $\bar{x} = \Sigma n_k x_k/n$. The notation $x_k$ is for the dosage score. The criterion in this expression follows a chi-square distribution with only 1 df. A test can be performed as usual by finding the *P*-value. Sometimes this is called the **Cochran test for linear trend**.

In the case of Table C.15, $n = 93$, $O_{1\cdot} = 21$, $\bar{x} = 135/93 = 1.4516$, $p = 21/93 = 0.2258$, $\Sigma O_{1k}x_k = 41$, and $\Sigma n_k x_k^2 = 305$. Thus,

$$\chi^2_{\text{trend}} = \frac{(41 - 21\times 1.4516)^2}{0.2258\times 0.7742\ (305 - 93\times 1.4516^2)}$$
$$= \frac{110.5947}{19.0610}$$
$$= 5.80.$$

A computer package gives $P = 0.016$ for this value of chi-square at 1 df. Since this is less than 0.05, reject the null hypothesis of no trend at 5% **level of significance** and conclude that a trend in proportions is present. Note the following:

- When the scores are equally spaced as in Table C.15 ($x_k$ = 0, 1, 2, and 3), the chi-square in the criterion in the above equation tests linearity. Thus, the test performed above is for assessing *linear trend*. With only four categories in Table C.15, it may not be a good idea to investigate a curve. But the presence of a curve can also be tested by suitably modifying the scores. In this example, these could be changed to 0, 1, 4, and 9 (square of $x_k$) or any other plausible values for investigating a trend other than linear.
- The overall chi-square (as calculated by the usual method) value for the data in Table C.15 is 6.13. The difference between this and the value of $\chi^2_{\text{trend}}$ is the chi-square for deviation from the trend, i.e.,

$$\chi^2_{\text{deviation},(K-2)\text{df}} = \chi^2_{\text{overall}(K-1)\text{df}} - \chi^2_{\text{trend, 1 df}},$$

  where the left-hand side is the chi-square for deviation from the trend. A large value of this chi-square would indicate that a trend other than that studied is significant. In that case, other kinds of trend can be tried. For the data in Table C.15,

$$\chi^2_{\text{deviation,2df}} = 6.13 - 5.80 = 0.33.$$

  This is not significant ($P > 0.05$). The trend, in this case linear, seems adequate, and there is no statistical need to study any other kind of trend.
- As for any chi-square, this procedure is valid for only large $n$. In fact, all $n_k$'s should be reasonably large. Also, many $p_k$'s should not be close to 0 or close to 1.
- In this example, the chi-square for trend is significant, although the overall chi-square is not. This can happen because the test for trend has greater power to detect trend than the overall chi-square test. It is important to realize that the chi-square for trend and the general chi-square test the same null hypothesis, namely, equal proportions in different categories, but they are designed to detect different types of alternatives. The $\chi^2_{\text{trend}}$ is specially designed for the alternative that a trend exists in proportions, whereas the usual chi-square is an overall test for any type of association.
- Ordinal data are sometimes perceived to arise from categorical metric data. For investigating the association between maternal drinking and congenital malformations, Grauband and Knor [1] categorized the average number of alcohol drinks per day as 0, <1.0, 1.0–2.9, 3.0–5.9, ≥6.0. Since the categories are metric, midpoints seem to be the most appropriate scores for this kind of analysis. You should try to resist the temptation to use equally spaced scores such as 0, 1, 2, 3, and 4 in such cases since the drink categories are not following such linear pattern.

1. Grauband BI, Knor EL. Choice of column scores for testing independence in ordered 2×*K* contingency tables. *Biometrics* 1987;43:471–6. http://www.jstor.org/discover/10.2307/2531828?uid=3738256&uid=2&uid=4&sid=21103501680347

## choroplethic map, *see also* thematic map, spot map

Choroplethic map is a geographical map where areas are shaded according to the grading of the indicator under study. Areas with similar values get the same shade, and the shading becomes darker as the rate increases. This kind of map is good for an indicator that can be considered evenly distributed over all units of the area with the same shade. One such map is shown in Figure C.11 that depicts under-five mortality rates (U5MR) in the countries of the Asia-Pacific region for the year 2000. In this map, for example, all states of India are assumed to have the same under-five mortality. Australia has very low mortality and thus is lighter in color, and Laos and Cambodia have very high mortality and thus get the darkest color.

The selection of class intervals is critical for choroplethic mapping as different class intervals and different number of categories

C


**FIGURE C.11** Choroplethic map of under-five mortality rate in countries of the Asia-Pacific region.

can provide a different picture. In Figure C.11, the categories of U5MR are four (very low, low, high, and very high), but the cutoffs are not mentioned. Quite often, the categories of indicator values in choroplethic maps are arbitrarily chosen of equal width, and the number of categories also remains arbitrary. This introduces considerable subjectivity in the cognition and perception obtained from the maps. A discussion on this aspect is given by Indrayan and Kumar [1], who advocate that the categories should be natural, as dictated by the data, in place of arbitrary choices. These natural categories can be identified by consensus in the groups obtained by various methods of **cluster analysis**, and the picture thus obtained can be substantially different from the one based on equal-width categories. The four categories (excluding "data missing") shown in Figure C.11 are based on consideration of the statistical proximity in rates based on one method of cluster analysis but not on consensus across methods.

For further details of choroplethic maps, see Slocum et al. [2].

1. Indrayan A, Kumar R. Statistical choropleth cartography in epidemiology. *Int J Epidemiol* 1996;25:181–9. http://ije.oxfordjournals.org/content/25/1/181.full.pdf?origin=publication_detail
2. Slocum TA, McMaster RB, Kessler FC, Howard HH. *Thematic Cartography and Geovisualization*, Third Edition. Prentice Hall, 2008.

## civil registration system, *see* birth and death registration

## circular sampling, *see* systematic sampling

## classification, *see also* discriminant analysis/functions, cluster analysis

Classification is a generic term used for putting units into categories so that similar ones go into the same category and those that differ go into another category. When the categories are not known and are to be established, this is done through **cluster analysis**. In statistics, the term *classification* applies when the categories are already known. Statistical classification is assigning a unit to its most likely category out of a few possible known options. This presupposes that the broad category is known but the specific category is not known and is surmised on the basis of whatever ancillary information is available on that unit. For example, we may know that the patient has a liver disease but do not know which liver disease—it could be malignancy, cirrhosis, or hepatitis. On the basis of clinical and laboratory parameters, the patient is classified into one of the three possible categories and treated accordingly.

You can see that the statistical classification is based on probabilities. The unit is assigned to that category that is most likely, and this likelihood is obtained on the basis of the ancillary information.

Classification requires that a rule be established that can evaluate the probability of the unit belonging to each of the possible categories. This is called a *classification rule* or a *classification model* or *classification criteria*. The general procedure for this is to study the ancillary information of a sample of $n$ units whose classification is already known. Use this to develop a relationship between the probability of each category and the ancillary information. Check that this rule is working well not just on the sample used to develop it but also on another sample. This is done by preparing, what is called, a *classification table* (an example is in Table C.16). Its **robustness** for minor variations in the ancillary information is also ensured.

**TABLE C.16**
**Classification Table for Dichotomous Categories**

| Classified by the Model | Observed in the Sample | | |
|---|---|---|---|
| | Disease Present | Disease Absent | Total |
| Disease present | 65 | 5 | **70** |
| Disease absent | 35 | 95 | **130** |
| **Total** | **100** | **100** | **200** |

Only after such **validation** can this rule be used to classify the subjects whose categories are not known. Despite full care, many classification models fail to perform well in practice because possibly some crucial information was not considered at the time of developing the model, or there were other unsuspecting mishaps such as **confounding** or **multicollinearity** among the data used as ancillary information, or because the information on the category or the ancillary factors for some units was not correct.

Out of many possible **algorithms**, the easiest is when there are only two possible categories. These could be such as the person has a particular disease or not, the person is likely to survive for at least 5 years or not, a bone found in an excavation belongs to a male or a female, a medical test is sufficiently good for detecting a condition or not, or any other yes/no or positive/negative type of categories. The most common method for analyzing the relationship of dichotomous categories with other factors is logistic regression. Details are given under the topic **logistic regression**, but briefly, in a logistic regression, the probability of one of the dichotomous categories is considered as the dependent variable and the ancillary information is considered as independent variables. This is run for $n$ units whose category is already known and a logistic model is obtained. As a first step, this model is tried out on the base units that were used to develop the model. A unit is classified as positive if its model-predicted probability exceeds 0.5; otherwise, it is classified as negative. If this model performs well (say, it is able to correctly classify at least 90% units—called *classification accuracy*), the model is considered good.

The classification rule is tried on another sample where also the classification is already known. Note that the classification accuracy can be evaluated in this case since their category is already known. It is only after such external validation that a model such as logistic is adopted for classification of future units.

The results in Table C.16 are for a case–control study of 100 cases and 100 controls. Logistic regression was run for some variables containing the ancillary information. The classification accuracy of this model is (65 + 95)/200 = 80%, as these many were correctly classified. This is not bad. However, also consider that the model has been able to correctly classify 95 of 100 negative subjects but only 65 of 100 positive subjects. Thus, the model is mostly good in identifying subjects with no disease but performs poorly in correctly classifying subjects with the disease. The overall classification accuracy may mask specific performance of the classification model. Just be careful about such anomalies.

For multiple categories, the method of discriminant function is commonly used for classification. The details again are under the topic **discriminant analysis**, but briefly, mostly a set of linear combinations of the variables in the ancillary information is obtained in this method that has the maximum classification accuracy. If there are $K$ categories, a total of $(K - 1)$ combinations would be required under this method. These would divide $n$ subjects into $K$ **mutually exclusive and exhaustive categories**. The same combinations can be used to classify future units provided that the model is validated and robust.

The term classification is also used for nonstatistical categorization in health and medicine, for example, social classification of people based on income, education, and occupation, and classification of subjects in a clinical trial into those with no disease (controls), mild disease, moderate disease, serious disease, and critical disease. Such classifications have no probability element and are thus called nonstatistical.

## classification and regression trees

Classification and regression trees are a method that can help in predicting one of the many possible categories of a subject. The **dependent variable** must be **categorical** for this method. The emphasis in this method is on simplification of a complex process without compromising the accuracy of prediction. It determines a set of logical *if–then* conditions instead of linear equation for predicting the category of a subject. Such statements are straightforward, easily understood, and intuitively appealing. The method is nonparametric and does not need Gaussian distribution of the underlying variables; it does not need linearity either. The tree method is particularly suitable for data exploration where a priori information is scanty, and where biological reasons are not immediately available for classification. The method can sometimes reveal simple relationships between just a few variables that can go unnoticed by other methods.

Consider predicting whether a subject with multiple injuries (in motor vehicle accident, earthquake, etc.) is going to be a survivor with no disability, is going to be a survivor with some disability, or is not going to survive at all. The dimension of injuries can be observed as organs affected, severity of injury, number of fractures, etc. The individual's own characteristics such as age, sex, muscular strength, and bone density may be the other determining factors for the outcome. Emergency help (time elapsed since injury), available medical facilities, and expertise of the medical care providers are also important. In place of saying that one factor will contribute 5%, another 10%, etc., would it not be nice to say that if the head is injured, the age is less than 60 years, and medical help is excellent but available 2 h after the injury, then the person is most likely to survive but will have residual disability? Classification and regression tree has precisely this function.

The tree algorithm, as the name suggests, devises split nodes that generate branches by using if–then rules. The adequacy of prediction is tested at each step. A fresh split node is not needed when the preceding node does not improve the accuracy of prediction. It is possible that a tree so created still fails to correctly classify an adequate percentage of subjects. As in all other prediction methods, this failure would indicate that the predictors chosen for this purpose are not of the right kind. Also, if the tree becomes very large with many branches, it loses much of its operational utility.

The tree method is complex and we are describing only the essential features. For details, see Breiman et al. [1]. Briefly, the algorithm begins by identifying a **predictor** that splits the total subjects into two groups such that the subjects are similar within groups and dissimilar across groups, and the classification is correct for the largest percentage of subjects. The groups can be such that they have the strongest association with the response categories. If a predictor is quantitative, various cutoff points are tried, and the one providing the least within-groups **sum of squares** is chosen. For a qualitative predictor with $K$ categories, all $(2^K - 1)$ possible splits are tried. Then the process is repeated at second and subsequent steps with new predictors. The process stops when it does not significantly add to the accuracy of classification.

It is customary in classification and regression trees to define a "loss function" for misclassification. Some misclassifications have more severe implications than others, and for them, a larger loss is defined. The loss may be in terms of financial cost, in terms of inconvenience to the health agencies and the person concerned, or simply statistical with regard to the increase in **least squares** or increase in **proportional reduction in error**.

There is some similarity between classification trees and **discriminant analysis**. In the latter case, all predictors are considered together to classify a subject, whereas in the trees, one or a few predictors are considered in the first step, another one or more in the second step, yet another set in the third step, etc. More than three steps make the process too complex and generally not considered appropriate for classification trees.

Téllez-Gabriel et al. [2] used classification and regression tree analysis to establish two groups of patients of head and neck

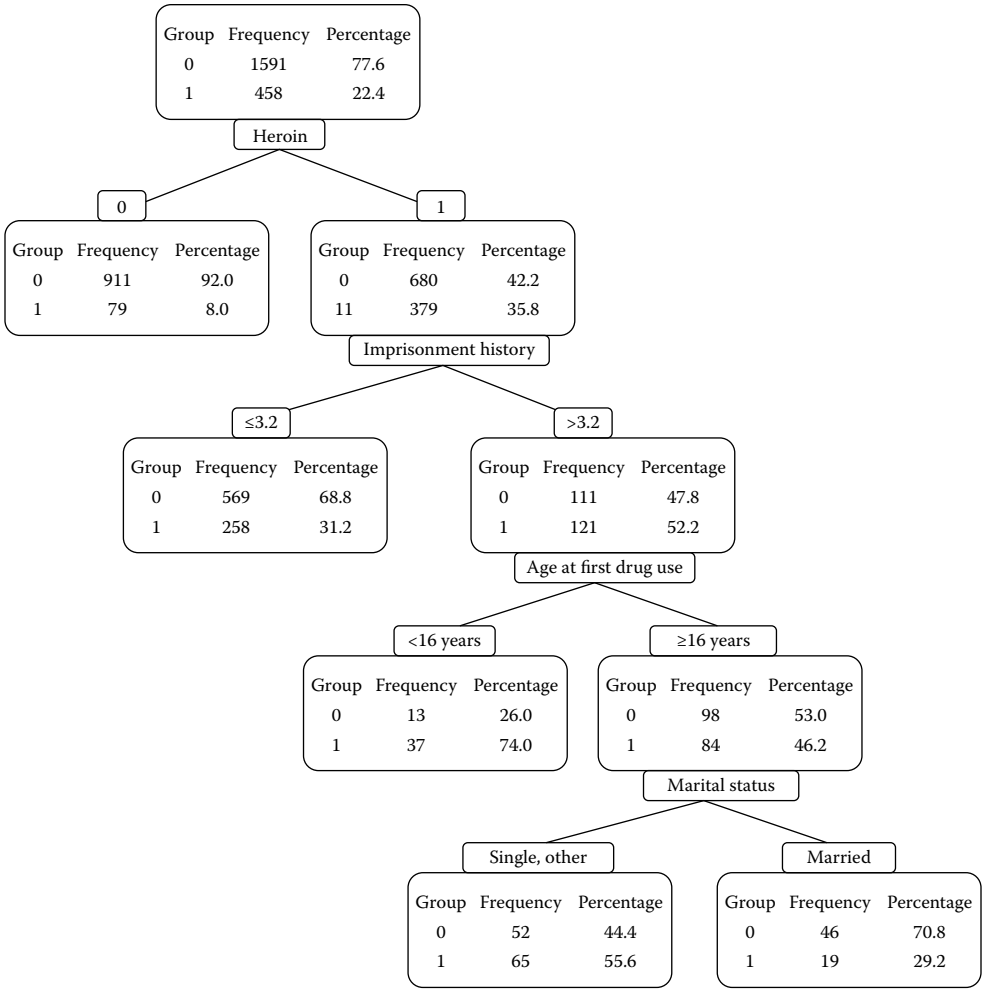



**FIGURE C.12** Classification and regression tree on training data (Code 1: prisoners with the history of drug injection). Note: At each node, the group with the highest percentage was considered as the predicting group. (From Rastegari A et al., *Addict Health* 2013 Winter–Spring;5(1–2):7–15. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3905563/. With permission.)

squamous carcinoma according to their RAB25 mRNA level and their risk of death. Rastegari et al. [3] have provided a simple example of use of classification and regression trees for factors influencing drug injection history among prisoners in Iran (Figure C.12). This is not an ideal example but illustrates the application.


1. Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.
2. Téllez-Gabriel M, Arroyo-Solera I, León X, Gallardo A, López M, Céspedes MV, Casanova I et al. High RAB25 expression is associated with good clinical outcome in patients with locally advanced head and neck squamous cell carcinoma. *Cancer Med* 2013 Dec;2(6):950–63. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3892400/
3. Rastegari A, Haghdoost AA, Baneshi MR. Factors influencing drug injection history among prisoners: A comparison between classification and regression trees and logistic regression analysis. *Addict Health* 2013 Winter–Spring;5(1–2):7–15. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3905563/


## class intervals

Class intervals divide the range of possible values of a variable on a **metric scale** into a small number of **categories** in order to facilitate the understanding and communication of the essentials, particularly the distribution, of a dataset. For example, for diastolic level (mmHg) of blood pressure (BP), you can have the number of people (frequency) with BP in class intervals 70–74, 75–79, 80–84, etc. This is also called grouping of data. Each interval has a lower limit and an upper limit, and the difference is called the width of the interval. In our example, each interval has width 5 mmHg, although the difference between the upper and lower limit is 4. This is because BP is measured in terms of integers, and the interval 70–74, for example, has values 70, 71, 72, 73, and 74, which are five values.

An inbuilt problem with the data reported in class intervals, such as cited for BP, is that the values 70 and 74 fall in the same interval, whereas values 74 and 75 fall in different intervals despite a relatively small difference. When you decide to use intervals, you automatically decide to live with this anomaly. However, a method such as **cluster analysis** can be used to discover natural cut points so that values within each interval have some kind of **affinity** compared with values in different intervals. For details, see Indrayan and Kumar [1], who have discussed it in the context of **choroplethic map**.

Formation of class intervals looks innocuous but requires extreme care. Many questions need answering before this can be done: how many intervals are appropriate, what should be the width of each

interval, should the width of all intervals be equal, whether any gap such as between 74 and 75 mmHg in our BP example should be allowed, etc. Common sense does play an important role in answering these questions, but the following guidelines may be helpful.

The question of class intervals arises only when you have a reasonably large sample. In case you have only six values, for example, there is no need to use class intervals since all these six values can be exactly stated. A rule of thumb is that you should have at least 10 values, preferably 20 or more values, before thinking of class intervals. It is customary to divide the range of values from minimum to maximum into at least four intervals for small $n$ and up to 12 intervals for large $n$. Equal intervals are preferred, but they are not a prerequisite. Thus, for diastolic BP of 15 persons that range from 68 to 90 mmHg, the intervals could be 68–73, 74–79, 80–85, and 86–91 mmHg. But for a sample of 80 persons, shorter intervals can be considered. If BP of 2000 persons is available, the intervals cannot be too short since the number of intervals should not exceed 12. Having a large number of intervals defeats the purpose of parsimony that class intervals are supposed to provide.

Same width of the intervals helps in interpretation. One can easily find which values are more common and which values are rather extreme. But exceptions can be made. Age intervals such as 0–1 year, 1–2 years, etc., for growth parameters in children are unacceptable since height and weight change so much from 0 to 1 month, 1 to 2 months, etc. For growth, monthly intervals may be needed up to, say, 6 months of age and then possibly quarterly intervals up to the age of 1 year. For health outcomes in general population, age intervals 0–1, 1–4, and 5–14 years are generally considered good. These do not have equal width.

The age intervals we just mentioned apparently have a gap, such as between 4 and 5 years. If the age is 4 years and 8 months on the date of interview, where will that child be placed? The convention in medical sciences is to record age in terms of completed years after the age of 1 year. Thus, a child of age 4 years and 8 months has completed 4 years but not 5 years. He or she will belong to the interval 0–4 years. Actually there is no gap between the intervals. Similarly, BP is measured in integers, although theoretically it is possible to have diastolic BP = 84.7 mmHg. We did not consider this possibility earlier in this section since such accuracy is not needed for BP measurement. Thus, there is no gap between, say, 80–84 and 85–89 mmHg intervals.

Another problem with the class intervals is our preference for categories beginning with 0 or 5. Hardly ever would one like to have intervals of width 7 or 8. Depending on the scale, the intervals would have width 0.1, 0.5. 1.0. 5, 10, etc. This arises from what is called **digit preference**. This causes no problem as long as the reporting and recording of values is accurate. However, often age 72 years is reported as 70 years and BP 137 mmHg is reported as 135. If that happens, you may like to consider different intervals such as 68–72, 73–77, etc. These would ameliorate the adverse effect of digit preference.

A further word of caution is in order. Class intervals should be used only for reporting of data and not for computations. Computations, such as for mean and standard deviation, have to needlessly assume that all the values in the interval are equal to its midpoint. For example, if you have 17 persons with BPs in the interval 80–84 mmHg, the calculations will be done as though all 17 values are 82 mmHg. This may be far too much of an approximation in some cases. Moreover, if exact values are available, there is no need to resort to any approximation. In some situations, though, the values are available in class intervals only. The age of women may be easy to elicit in the range, say, 20–29 years because some feel reluctant to give exact age. Similarly, for experiment on mice where they rapidly die, you may be able to visit the laboratory once every morning and note how many died during the last day. In this case, the exact time of death will not be available.

1. Indrayan A, Kumar R. Statistical choropleth cartography in epidemiology. *Int J Epidemiol* 1996;25:181–9. http://ije.oxfordjournals.org/content/25/1/181.full.pdf+html

## clinical equipoise, *see* equipoises

## clinical equivalence, *see* equivalence (types of) in clinical trials

## clinically important difference, *see* medically important effect (the concept of)

## clinical tolerance range, *see also* medically important difference (test for detecting)

This is the deviation from the norm or from the expected that can be possibly allowed without compromising the welfare of the subject. For example, whether the systolic blood pressure is 184 or 186 mmHg, both require same treatment. If yes, this deviation of 2 mmHg is within the clinical tolerance. However, possibly 184 and 198 mmHg are too different to have the same clinical implication. Similarly, forgetting to take a drug once a week when the regimen is three times a day can be considered within clinical tolerance; forgetting to take four to five times a week can have some implication on the rate of recovery and cannot be ignored.

The term *tolerance* is quite often used to also assess the practical utility of the statistical models that calculate certain difficult-to-measure medical parameters. If brain volume by actual measurement is 1272 cc and by using a model based on simple measurements is 1258 cc, one may say that this difference is within clinical tolerance. But if the model tells us that the volume is 1225 cc, the difference does not look like within clinical tolerance. Whereas a too wide or too small difference can be easily interpreted without error, the difficulty arises when the difference is neither too big nor too small. As of now, no objective criteria are available, and they do vary from parameter to parameter, physician to physician, patient to patient. However, there are measurements for which there is a consensus. For example, Vanhatalo et al. [1] consider 3–10 min as the tolerance range for completing four constant-work-rate knee-extension exercise bouts inside the bore of 1.5 T superconducting magnet. This might be what is generally acceptable. In addition, for laboratory measurements, principles of **quality control** as described under this topic can be used to define tolerance range.

The third use of the clinical tolerance range is in assessing the acceptable doses such as of a drug or of radiation in radiotherapy. This relates to the side effects. Minor side effects in small percentage of subjects are tolerated, but drugs with major side effects or with side effects in a large percentage of cases are not approved.

Sometimes the normal range of medical parameters is termed as *tolerance range* as done by Shin et al. [2] for HbA1c values. Normal glucose tolerance range also illustrates similar usage.

1. Vanhatalo A, Fulford J, DiMenna FJ, Jones AM. Influence of hyperoxia on muscle metabolic responses and the power-duration relationship during severe-intensity exercise in humans: A 31P magnetic resonance spectroscopy study. *Exp Physiol* 2010 Apr;95(4):528–40. http://onlinelibrary.wiley.com/doi/10.1113/expphysiol.2009.050500/pdf
2. Shin JH, Kang JI, Jung Y, Choi YM et al. Hemoglobin A1c is positively correlated with Framingham risk score in older, apparently healthy nondiabetic Korean adults. *Endocrinol Metab (Seoul)* 2013 Jun;28(2):103–9. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3811715/

## clinical trials (overview)

More for convention than semantics, medical experiments on human beings are called *trials*. The objective is to study the cause–effect relationship between a medical intervention and a health outcome in human subjects. Since the subjects are human, a large number of issues crop up ranging from stricter ethics to profound variations. Thus, trials do need extra care.

Clinical trials have a long history. Laplace wrote in 1825 that it is sufficient to "test each treatment on the same number of patients, while keeping all circumstances perfectly similar" [1]. In 1925, Fisher introduced randomization that helped eliminate biases due to uncontrolled factors and provided basis for using statistical inferential method. The first randomized clinical trial was done in 1946 on streptomycin by British Medical Research Council for the treatment of tuberculosis. The methodology of double blinding reached maturity in the 1970s [1].

Clinical trials are mostly done to investigate new modes of therapy. Research on new diagnostic procedures also falls in this category. Most clinical trials are carried out meticulously involving heavy investment. Since variation between and within subjects occurs due to a large number of factors, it is quite often a challenge to take full care of all of them. **Epistemic uncertainties** also play a significant role. It is imperative in this situation that the rules of **empiricism** are rigorously followed. This means that the trial should be conducted in controlled conditions so that the influence of extraneous factors is minimized, if not ruled out. Also, the trials should be conducted with a sufficient number of patients so that a trend, if any, can be successfully detected. Clinical trials are done in ideal conditions as much as possible to provide a realistic estimate of the **efficacy**. For estimating **effectiveness** in practical conditions, **pragmatic trials** are favored.

A human experimentation cannot be done unless sufficient reasons are present. The medical community must be sufficiently uncertain about the effect of the intervention under the trial. The regimen must have passed through rigors of preclinical phases before reaching humans (see **phases of the trials**). Preclinical phases involve studying biochemical properties and experiments on suitable animal models. It must be clearly established that the intervention is more likely to be beneficial than harmful. The motto "Do no harm" is scrupulously followed in all medical research including trials. If any unsuspecting harm is detected later on, the intervention is immediately discontinued. Patients, then clinicians, should be the overriding considerations instead of the researchers.

By their very nature, all trials are **prospective studies** where the **antecedent** is the intervention and the **efficacy** and safety are the outcomes. A follow-up is built into all the trials. Thus, many ideas discussed under observational **prospective studies** apply to the trial setup as well. All trials involve careful consideration of issues such as selection of subjects and controls, **randomization** and **matching**, and **blinding**, **masking**, and **concealment of allocation**. All these are presented separately under the concerned topic.

From the etiological point of view, clinical trials can be divided into six broad categories: therapeutic trials, diagnostic trials, screening trails, prophylactic trials, field trials, and vaccine trials. A brief description about them is given next. Clinical trials are a huge concern with several books devoted to this topic. Our description here is brief and introductory. For details, see, for example, Chow and Liu [2].

### *Therapeutic Trials—Efficacy and Side Effects*

The primary objective of a therapeutic clinical trial usually is to evaluate the safety and efficacy of a treatment regimen in individuals with different levels of severity of disease and of various backgrounds such as different age, sex, and nutritional status. Extreme care is required because therapeutic trials generally involve exogenous material that may have side effects, and may not be beneficial at all relative to the existing modes of therapy. For this reason, before a therapeutic trial is undertaken, it is necessary to be sufficiently convinced regarding intoxicity and potentiality as a beneficial regimen. Thus, the previous phases in the laboratory must have provided unequivocal results. Since the clinical trials are precarious, they are pursued in phases, particularly for a new formulation or substance as for drug development.

Efficacy is always related to a particular outcome. Terms such as recovery and discharge are vague for the outcomes. They must be specified either in terms of measurements such as glomerular filtration rate for kidney diseases, in terms of images such as x-ray for dislocated joint, or in terms of any such objective criterion. Also, the duration after which the outcome is to be assessed should be specified—within a day, within a week, etc. This applies to death also. Everybody dies, but if a death occurs 3 months after a surgery, should this be ascribed to the surgery? The follow-up period for different outcomes of interest must also be fully specified.

There are other issues as well relating to the outcomes. The actual interest may be in cardiovascular outcomes, but for expediency, change in blood pressure level can be considered as a surrogate endpoint. Large cohorts and long follow-up are expensive—surrogates tend to make them expedient. For example, microalbuminuria is a promising surrogate of renal protection in many cases. However, do not use surrogates indiscriminately. Examine first whether they are indeed valid markers for the hard endpoint you are looking for. The surrogate should accurately assess not only the benefit or the lack of it but also the harm. A correlate may not be a suitable surrogate.

Outcomes can also be assessed at prespecified interim stages. This can help in discontinuing a trial if confirmed results are available one way or the other. Desired efficacy may be proved, or unacceptable severe side effects may appear. Sample size can also be reassessed. However, interim appraisals can unblind the study. For details, see **adaptive designs** and **stopping rules**.

A regimen should not be assessed only in terms of its benefits or efficacy. Except possibly those that alter lifestyle, no intervention is without risk of **side effects** and toxicity. Thus, the benefit must be seen in relation to the possible risk. This has special relevance to potentially hazardous drugs. In some situations, **safety** is more important than efficacy. For an account of benefit–risk assessment of various regimens, see Korting and Schafer-Korting [3].

Some of the side effects, now generally termed adverse events, may be preexisting or may occur in any case in a person or even a group of persons, but some could be attributed to the regimen. How can this attribution be achieved? Methods are available that help to categorize side effects into those that are definitely due to the regimen, possibly due to the regimen, and unlikely to be due to the regimen. For this, one procedure is to withdraw the treatment and administer it again, and see if the side effect disappears and recurs.

What side effects should be considered common and what should be considered rare in a therapeutic trial? Of course, this will depend on the nature of the side effect. We have provided a guideline in the topic **adverse effects** and **adverse patient outcomes** that applies to the side effects as well. For this, it is necessary to define all side effects and record them, including time of onset, time of resolution, severity, relation to study regimen, action taken, and outcome. In addition, significant aberrations in laboratory values may also have to be defined.

Among other issues related to therapeutic trials are (i) **efficacy and effectiveness**; (ii) **equivalence and noninferiority trials**; (iii) various **designs of the trials** such as **randomized controlled trials (RCTs)**, **crossover**, ***N*-of-1**, **up-and-down**, and **sequential**; and (iv) biostatistical ethics that comprise issues such as **equipoises**. These are discussed under the respective topics.

### *Clinical Trials for Diagnostic and Prophylactic Modalities*

Diagnostic trials are for modalities that help in diagnosis rather than in therapeutics. They are almost invariably conducted in clinics. Prophylactic trials are for prophylactic measures that prevent either the disease or its complication. Prophylactic trials can be conducted in clinic as well as in the field in a community.

The intervention in a diagnostic trial is not a therapeutic agent but a procedure that can change the diagnosis and thus the course of the treatment. Thus, it has the potential to improve decision-making and patient management.

From the ethics point of view, noninvasive procedures such as measuring blood pressure (BP) and weight do not cause much anxiety except for time, cost, and inconvenience to the patient, but an invasive procedure such as endoscopy has the potential to cause harm to the health of the patient. More care is required for a trial on an invasive procedure.

As discussed under **randomized controlled trials (RCTs)**, most RCTs have a **parallel control** group, but many diagnostic trials are self-controlled. For comparison of prostatic specific antigen levels and ultrasound images for prostate cancer, both procedures would be done on the same set of patients. Agreement between the two can be evaluated, but to find which is better, a **gold standard** is needed. For example, magnetic resonance imaging can be considered gold and can be compared with arthroscopy as the reference for detection of meniscal ruptures. When a reference is available, the diagnostic efficacy in terms of **sensitivity** and **specificity**, and **predictivities** can be obtained.

When a group of suspected cases is available that meets specific inclusion and exclusion criteria, the diagnosis would be established by a gold standard in some of them. The diagnostic procedure under trial would also be used on all the suspected cases, and some would be found positive. This would yield a 2×2 table of the type shown in Table C.17. In this table, $(a + c)$ cases were found to have the disease by gold standard and $(a + b)$ by the procedure under trial. Thus, various validity parameters can be estimated such as sensitivity and specificity.

**TABLE C.17**
**Results of a Diagnostic Trial**

| Results of the Diagnostic Procedure under Trial | Actual Disease | | |
|---|---|---|---|
| | Present | Absent | Total |
| Positive | $a$ | $b$ | $a + b$ |
| Negative | $c$ | $d$ | $c + d$ |
| **Total** | $a + c$ | $b + d$ | $n$ |

Haig et al. [4] report a prospective, masked, controlled diagnostic trial to determine the sensitivity and specificity of electrodiagnostic consultation for the clinical syndrome of lumbar spinal stenosis. A total of 150 suspected cases were included. Out of these, clinical consensus on diagnosis was reached for 55 cases. This was considered the gold standard for this study. Electrodiagnostic measurements such as paraspinal EMG score and composite limb and paraspinal filtration score were evaluated for their sensitivity and specificity against this gold standard.

Prophylaxis is a procedure that promotes health or controls primordial factors that adversely affect health. A prophylactic trial is generally conducted in the field where a community is involved. This is discussed later in this section. But it can be conducted in a clinical setup also. An example is prophylactic nasal continuous positive airway pressure after major vascular surgery. Amnioinfusion for meconium stained amniotic fluid in labor at the time of childbirth is a prophylactic procedure that can be tried for specific types of births. As with all other trials, prophylactic trials are also restricted to a particular segment of the subjects. For example, for a prophylactic drug such as aspirin to reduce cardiovascular events, the participants may be vulnerable people of age 50 years or more.

The principles of clinic-based prophylactic trials are the same as for therapeutic trials. Perhaps a prophylactic procedure is somewhat insulated against harmful effects, and thus slight relaxation in ethics may allow stricter control over confounding factors. Community-based trials in the field have a different setup, whether for therapeutic modality, for prophylactic agent, or for a screening procedure. These are discussed next.

### *Field Trials for Screening, Prophylaxis, and Vaccines*

Field conditions, where a segment of the population is involved, are different from clinical conditions. Many confounders such as the severity condition of the patient can be controlled in the clinic but would be difficult to control in the field. In any case, **field trials** are done for modalities that can be used on mass scale on the general population or its segment. Thus, they can have wider health policy implications that clinical trials seldom have.

In public health, field trials are sometimes done with the health facility as the unit of experiment. The intervention could be training to the peripheral workers (such as for Pap smear) to find out whether that improves the case detection rate against a control area where no such training was given. Issues in such trials may be slightly different from those that were discussed in this section for individual-based trials.

Screening is quite in vogue for cancers. The prostate, lung, colorectal, and ovarian cancer screening trial initiated in 1992 in the United States has enrolled more than 150,000 participants [5]. Nearly half are randomly assigned to the intervention (screening), and the other half remained as control. Whether screening helps in reducing cancer mortality is yet to be seen. Many collateral benefits emerged though. For example, it was found that chest radiograph abnormalities not suspicious for lung cancer are common, and prostate volume and age are independently associated with increased prostatic specific antigen in men undergoing screening. A second example is a randomized mass screening trial for abdominal aortic aneurysm. Ten-year results show 73% reduction in mortality by this aneurysm in those screened as compared to those not screened [6].

Note how **screening trials** look like they are based on mass screening but the procedures used are hospital based. Sankaranarayanan et al. [7] report a community-based trial wherein all persons of age 35 years or older were screened for oral cancer in intervention villages

and not screened in the control villages. The screening was done three times at 3-year intervals for signs of oral cancer. There were nearly 60,000 eligible subjects in the intervention group and nearly 55,000 in the control group. The villages were **cluster-randomized** rather than individual-randomized to receive or not receive the intervention. The difficulty was, as in most field trials, low **compliance** when referred for confirmatory examination that has to be done in a hospital. In this trial, compliance was lower than 70%.

A **prophylactic trial** in the field could be for a strategy such as lifestyle changes for coronary disease or could be for vitamin intake—even drugs that are stipulated to prevent occurrence or recurrence of adverse events. Giving vitamin A supplements to infants and young children to improve their retinol level is an example of such an intervention. Although there is a fine distinction between preventive and prophylactic measures, we are including both into the prophylactic category. Such trials have tremendous value in policy formulation, in saving lives, and in improving health, but they do not receive that kind of attention.

A prophylactic trial is not necessarily conducted in the general population. The vitamin A trial cited in the preceding paragraph is for children. A trial on an educational campaign for responsible sexual behavior may target adolescents. Another trial on hematinic supplementation may target antenatal women of low socioeconomic stratum.

Consider the following example. A total of 198 refugees were selected in Kenya by multistage **cluster sampling** and were divided almost equally to receive the intervention and placebo by cluster randomization [8]. The intervention was insecticide-treated personal clothes and linen, and the placebo in this case was plain water treatment. Double blinding was done for malaria parasite smear. The odds of malaria infection in the intervention group were reduced by about 70%.

**Vaccine trials** are conducted in phases as therapeutic trials but need even more precaution. The need for extra care arises from the applicability of vaccines to a large segment of populations who are not sick but are at risk, as opposed to therapeutics that is applied only to patients and administered under close supervision. A feature of vaccines is immunogenicity, which might be an important consideration in some diseases, in addition to protective efficacy. In others, duration of protection may be important. Quality and quantity of immune responses required for protection against infection and against development of disease are scientific challenges. In the case of HIV, for example, there could be a vaccine that inhibits HIV infection, and there could be a vaccine that inhibits or retards development of disease—AIDS—in those already infected.

In view of the complexities involved in vaccine trials, an additional phase called phase IIB (see **phases of clinical trials**) is sometimes advocated. This is also called the "test of concept" phase. The aim of phase IIA could be to establish the schedule of administration for different age groups, as it would be most likely a factorial experiment with dose level as one factor and age group as the second factor. Thus, four phases are required for vaccine trials instead of the usual three. The objective of phase IIB is to evaluate whether the vaccine has any (>0%) efficacy at all. In a phase III trial for vaccines, this objective shifts generally to at least 30% efficacy. The participants in phase IIB are not necessarily representative of the target population. For phase III, a representative sample is indicated. Phase IIB also assesses the operational efficiency, whereas the objective of phase III is to produce compelling evidence of efficacy for licensure from regulatory agencies.

A phase III trial for a vaccine has to be a large-scale trial so that adequate numbers developing the disease, particularly in the control group, are available. The total number of subjects may run into thousands, and the follow-up too may go up to several years. Since phase III is an expensive trial for vaccines, phase IIB becomes a highly desirable proposition to indicate whether or not to proceed to phase III. Phase IIB, however, increases the time frame because this too can take at least a couple of years.

In addition to what have already been mentioned, other related issues separately discussed in this volume for clinical trials are **compliance**, **CONSORT statement**, **multicentric trials**, and **registration of trials**. Further details of clinical trials are discussed by Shih and Aisner [9].


1. Weisberg HI. What next for randomized clinical trials? *Significance* 2015;12(1):22–7. http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2015.00798.x/abstract
2. Chow S-C, Liu J-P. *Design and Analysis of Clinical Trials: Concepts and Methodologies*, Third Edition. Wiley, 2013.
3. Korting HC, Schafer-Korting M (Eds.). *The Benefit/Risk Ratio: A Handbook for the Rational Use of Potentially Hazardous Drugs*. CRC Press, Boca Raton, FL, 1998.
4. Haig AJ, Tong HC, Yamakawa KS et al. The sensitivity and specificity of electrodiagnostic testing for the clinical syndrome of lumbar spinal stenosis. *Spine* 2005;30:2667–76. http://journals.lww.com/spinejournal/Abstract/2005/12010/The_Sensitivity_and_Specificity_of.12.aspx
5. Oken MM, Marcus PM, Hu P et al. Baseline chest radiograph for lung cancer detection in the randomized Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial. *J Natl Cancer Inst* 2005;97:1832–9. http://jnci.oxfordjournals.org/content/97/24/1832.full.pdf+html
6. Lindholt JS, Juul S, Fasting H, Henneberg EW. Preliminary ten-year results from a randomised single centre mass screening trial for abdominal aortic aneurysm. *Eur J Vasc Endovasc Surg* 2006;32:608–14. http://www.sciencedirect.com/science/article/pii/S1078588406003364
7. Sankaranarayanan R, Mathew B, Jacob BJ, Thomas G, Somanathan T, Pisani P, Pandey M, Ramadas K, Najeeb K, Abraham E. Early findings from a community based, cluster-randomized, controlled oral cancer screening trial in Kerala, India: The Trivandrum Oral Cancer Screening Study Group. *Cancer* 2000;88:664–73. http://www.ncbi.nlm.nih.gov/pubmed/10649262
8. Kimani EW, Vulule JM, Kuria IW, Mugisha F. Use of insecticide-treated clothes for personal protection against malaria: A community trial. *Malar J* 2006;5:63. http://link.springer.com/article/10.1186%2F1475-2875-5-63#page-1
9. Shih WJ, Aisner J. *Statistical Design and Analysis of Clinical Trials: Principles and Methods*. Chapman and Hall/CRC, 2015.


## clinimetrics, *see also* scoring systems for diagnosis and for gradation of severity

Metrics is generally identified by quantitation and calculations in any science. Clinimetrics goes beyond these calculations and seeks to place the quantities and calculations in a clinical context so that decisions in the interest of the patient can be taken on the basis of measurements after assessing their implications. The focus in clinimetrics is on quality of measurements and their performance. In a way, this takes away subjectivity from decisions, which might otherwise be difficult to justify logically. You would agree that logic is the cornerstone for medical decisions. The term *clinimetrics* was introduced by Feinstein [1] in 1983.

Many mathematical formulas have unknowingly entered into medical practice back door through computer-based systems. They directly produce results without the user being aware of the back-end calculations. Examples such as the bispectral index, high-performance

liquid chromatography, and evoked potential can be cited that directly produce results, without revealing the complex calculation. These are examples of clinimetrics tools. Gill [2] has given details of clinimetrics of laser Doppler imaging in assessing depth of burns.

**Indicators** and **indexes** can be considered components of clinimetrics. But the major focus of this gradually developing science is on the methodology for developing **scoring systems** that have the requisite applicability. These components of clinimetrics are concerned with the quality of the measuring instruments and look at the process of tool development rather than the tools themselves.

**Reliability**, **validity**, and responsiveness of such measuring tools are an integral part of their quality. A measuring tool should also be sufficiently sensitive to detect clinically relevant improvements attributable to therapeutic interventions. Close collaboration among clinicians, biostatisticians, and epidemiologists is required for the development of clinimetrics as a science of consequence.


1. Feinstein AR. An additional basic science for clinical medicine: IV. The development of clinimetrics. *Ann Intern Med* 1983;99:843–8. http://www.ncbi.nlm.nih.gov/pubmed/6651026
2. Gill P. The critical evaluation of laser Doppler imaging in determining burn depth. *Int J Burns Trauma* 2013 Apr 18;3(2):72–7. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3636664/


## Clopper–Pearson bound/interval, *see also* exact confidence intervals (CIs), Wilson interval

The Clopper–Pearson interval is the most widely used "exact" confidence interval for binomial π. Whereas the CI for a large sample can be obtained by using **Gaussian** approximation, Clopper–Pearson provides almost an exact interval that is supposed to be valid for small *n* also. Agresti [1] has quoted articles that say that this interval is extremely conservative and not so exact as made out. Conservative means that the interval will actually contain higher than 95% probability when computed for 95% confidence.

Suppose the interest is in estimating the chance of uterine prolapse in women who come with complaints of micturition disturbance and vaginal discharge. If only $n = 12$ women with such complaints could be examined and three had uterine prolapse, the proportion ($p$) is 3/12 = 0.25. Another sample may give another proportion. How does one find limits within which this proportion is not unlikely to lie in all such patients? In this case, since $np = 3$ is small, the Gaussian approximation cannot be used. The confidence interval for π in case of a random sample of small *n* is obtained by using the sum of binomials and the values from the **beta distribution**, which, in turn, can be obtained from the ***F*-distribution** (see Johnson and Kotz [2] for details). This involves some approximation because an exact 95% probability cannot be obtained for **discrete variables** such as the binomial. The quantity π in this case is the actual proportion of women with uterine prolapse among the population of women with those complaints. The confidence interval (Clopper–Pearson interval) is given by

$$\left(1+\frac{n-x+1}{x[F_{\alpha/2}\text{ at }2x,\ 2(n-x+1)df]}\right)^{-1},$$

$$\left(1+\frac{n-x}{(x+1)[F_{1-\alpha/2}\text{ at }2(x+1),2(n-x)df]}\right)^{-1},$$

where $n$ is the total sample size, $x$ is the number of successes or positives, and $F$ is the value of $F$ at the subscripted probability (percentile) and at the stated degrees of freedom. In our example of uterine prolapse, $n = 12$ and $x = 3$. For 95%, CI, $F$-values are at probabilities 0.025 and 0.975 at df = (6, 20) for the lower limit and df = (8, 18) for the upper limit. These $F$-values are 0.19 and 3.02 from the $F$-table found in statistics books. Thus, the 95% CI for π is

$$\left(1+\frac{12-3+1}{3\times 0.19}\right)^{-1}\text{ to }\left(1+\frac{12-3}{4\times 3.02}\right)^{-1},\text{ or 0.054 to 0.57.}$$

For those who are comfortable with beta distribution, the interval can also be written more simply as

$$\text{Idf.Beta}(\alpha/2,\ x,\ n-x+1),\ \text{Idf.Beta}(1-\alpha/2,\ x+1,\ n-x).$$

These are the inverse distribution functions of beta distributions for cumulative probability α/2 and (1 – α/2), respectively, with 1 added to $n - x$ the first time (lower limit) and to $x$ the second time (upper limit). Statistical software packages generally follow this method in place of going through $F$-distribution. Both are equivalent, though. As you can see, the Clopper–Pearson interval could be larger than needed because of 1 added to the parameters for both sides of the limit. This interval is conservative anyway as stated earlier and becomes too conservative if the anticipated value of π is close to 0 or 1. Because of this, many people prefer **Jeffreys interval**.

Instead of going through the rigmarole of complex mathematics, presented next is a graphical method for obtaining Clopper–Pearson CI when the sampling is random. This involves some approximation but is still useful for practical applications.

The 95% CI for π corresponding to different values of the observed sample proportion $p$ can be read from Figure C.13. This is drawn for some specific values of $n$. The upper and lower limits are read off the vertical axis using the pair of curves corresponding to the sample size $n$. The sample proportion $p$ is on the horizontal axis. If your $n$ is not exactly as shown, visual interpolation can be done to get an approximate CI. The figure is large relative to the other figures in this book so that this can be actually used.

For our example of 3 women with uterine prolapse out of 12 examined with complaints of micturition disturbance and vaginal

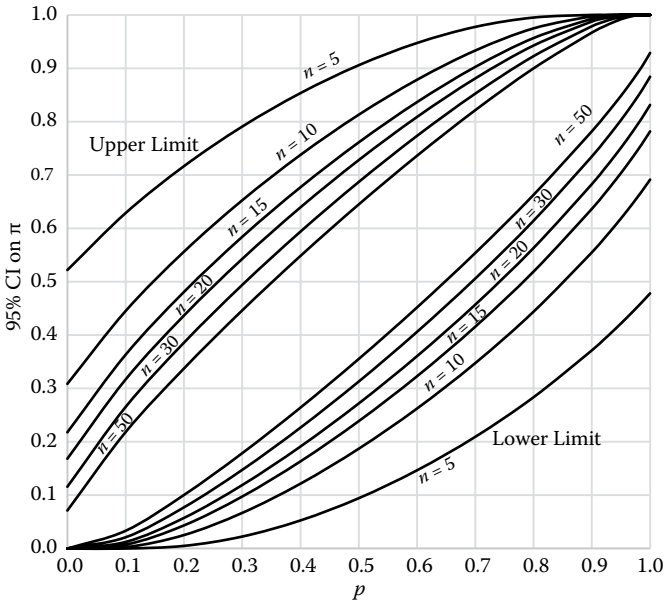



**FIGURE C.13** The 95% Clopper–Pearson confidence interval (CI) for π for different sample sizes.

discharge, $n = 12$ and Figure C.13 does not give a CI curve for this $n$. However, visual interpolation for $n = 12$ corresponding to $p = 0.25$ gives (0.06, 0.58) as the CI. Thus, the chance of uterine prolapse in women with those complaints can be anywhere between 6% and 58%. Such a wide interval may not be helpful, but that is what could be obtained on the basis of this small sample.

If $n = 50$, the 95% CI for observed $p = 0.25$ from Figure C.13 is (0.14, 0.39) approximately. This is much narrower. You may wish to calculate CI for $n = 120$, now using Gaussian approximation because of the large $n$. For this $n$, the 95% CI is (0.17, 0.33). Note how the CI narrows as $n$ increases. Two useful observations can be made on the CIs obtained for different values of $n$ in our example:

- As $n$ increases, the CI narrows, but this gain decreases with increasing $n$. The width of the interval for $n = 12$ is $0.58 - 0.06 = 0.52$, for $n = 50$ is $0.39 - 0.14 = 0.25$, and for $n = 120$ is $0.33 - 0.17 = 0.16$. Increasing $n$ from 12 to 50 reduced the width to less than half, but increasing $n$ from 50 to 120 reduced it to only two-thirds. Thus, the law of diminishing returns is applicable here also.
- Relative to $p = 0.25$, the first interval for $n = 12$ is highly asymmetric. The lower limit 0.06 is closer to $p = 0.25$ than the upper limit 0.58. This asymmetry declines as $n$ increases, and the CI for large $n$ becomes symmetric around the value of $p$. Also, symmetry increases as $p$ becomes closer to 0.5.

### *Clopper–Pearson Bound for π When the Success or the Failure Rate in the Sample Is Zero Percent*

Consider again a situation in which a surgeon performs the same operation on 10 different patients for kidney stone with complete success without a single complication. Thus, the complication rate is $p = 0$ in this sample. Can it be concluded that the complication rate would continue to be zero for all such operations in the future? Or is this just good luck for the 10 patients who happened to be operated on during that period? In statistical language, can $p = 0$ be used as an estimate of π in the calculation of SE($p$)? The answer obviously is no. In such situations, where the observed $p$ is the extreme value, the true complication rate can be estimated only by obtaining a one-sided **confidence bound** for π.

The 95% confidence bounds for extreme results for various sample sizes are displayed in Table C.18. These are again based on the exact binomial distribution and are called the Clopper–Pearson bounds.

Consider the previous example of a surgeon with no complication in 10 surgeries for kidney stone. For this surgeon, $p = 0$ and $n = 10$. The upper bound for the true complication rate, corresponding to 95% confidence, from Table C.18 is 27%. Thus, you could be 95% confident that the complication rate in the long run would not exceed 27%; that it could be as high as 27% may have set an alarm. The claim of 0% complication rate based on the experience for 10 subjects is not tenable. If no complication is observed in a series of 50 such surgeries, then the confidence bound corresponding to $n = 50$ from Table C.18 is only 6%. Note again how important the size of the sample is in determining the bound and in arriving at a focused conclusion.

An exactly similar situation arises when the observed $p$ is 1.0. In this case, for example, for $n = 15$, the lower bound for π corresponding to 95% confidence is 0.82. The limits in Table C.18 are given for specific values of $n$. For other values of $n$, such as $n = 18$, an approximate CI can be obtained by suitably interpolating between the relevant curves.

Although the discussion is for $p = 0$ and $p = 1$, it suggests extra caution in obtaining CI for π when $p$ is extremely small or extremely high. The conventional two-sided CI in this case can go beyond 1 or can start from a negative value. Obviously, such values are impossible for any proportion. In such cases, examine if one-sided bound serves the purpose instead of two-sided CIs. In many situations, bounds serve the purpose well.

There is another method to obtain CI for binomial π. For this, see **Jeffreys interval**. Jeffreys adds ½ to the value of $x$ and $n - x$ in place of adding 1. This may be more exact than Clopper–Pearson interval. Also see **Wilson interval**.

**TABLE C.18**
**95% Confidence Bounds for Extreme Results**

| If the Sample Size $n$ Is | If the Sample Percentage Is 0%, the True Percentage Could Be as High as | If the Sample Percentage Is 100%, the True Percentage Could Be as Low as |
|---|---|---|
| 1 | 95 | 5 |
| 2 | 78 | 22 |
| 3 | 63 | 37 |
| 4 | 53 | 47 |
| 5 | 45 | 55 |
| 6 | 39 | 61 |
| 7 | 35 | 65 |
| 8 | 31 | 69 |
| 9 | 28 | 72 |
| 10 | 26 | 74 |
| 15 | 18 | 82 |
| 20 | 14 | 86 |
| 25 | 11 | 89 |
| 30 | 10 | 90 |
| 35 | 8 | 92 |
| 40 | 7 | 93 |
| 45 | 6 | 94 |
| 50 | 6 | 94 |
| 55 | 5 | 95 |
| 60 | 5 | 95 |
| 65 | 5 | 95 |
| 70 | 4 | 96 |
| 75 | 4 | 96 |
| 80 | 4 | 96 |
| 85 | 3 | 97 |
| 90 | 3 | 97 |
| 95 | 3 | 97 |
| 100 | 3 | 97 |
| 150 | 2 | 98 |
| 300 | 1 | 99 |

1. Agresti A. Dealing with discreteness: Making 'exact' confidence intervals for proportions, differences of proportions, and odds ratios more exact. *Statistical Methods Med Res* 2003;12:3–21. http://www.stat.ufl.edu/~aa/articles/agresti_2003.pdf
2. Johnson NL, Kotz S. *Discrete Distributions*. Wiley, p. 59.

## cluster analysis

The problem addressed in cluster analysis is the division of subjects or units into an unspecified number of **affinity** groups. The grouping is done in such a manner that units similar to one another with

respect to a set of variables are classified to one group and the dissimilar ones to another group. Thus, natural groupings in the data are detected. For instance, this is unwittingly done in assigning grades to students in a course where the instructor looks for rational cutoff points. Sometimes only two grades A and B are considered enough; sometimes they go up to E. In another setup, the method can also create taxonomic groups for future use. An example is of cases falling into various diagnostic groups on the basis of their clinical features. A name is subsequently assigned to these groups depending on their etiology or features. Although overlapping clusters can be conceived so that one or more units belong simultaneously to two or more clusters, this section is restricted to exclusive clusters. Cluster analysis is a nonparametric procedure and does not require values to follow a Gaussian or any other pattern. Also, this procedure is primarily used for **multivariate** observations and not so much for univariate observations.

### *Measures of Similarity*

Division of subjects into a few but unknown number of affinity or natural groups requires that proximity between subjects is objectively assessed. Those with high proximity go into one group, and those with low proximity are assigned to some other group. Thus, the subjects resembling one another are put together in one group. As many groups are formed as needed for internal homogeneity and external isolation of the groups (Figure C.14). The points plotted in Figure C.14b are the same as in Figure C.14a, but now the affinity groups are shown.

Similarity between two subjects can be measured in a large number of ways. The methods are different for **qualitative** than for **quantitative** variables. In the case of **binary** qualitative variables in a $K$-variable multivariate setup, the affinity can be shown in the form of a $2\times2$ table as in Table C.19. There is a total of $K$ variables out of which matches between the $i$th and the $j$th units occur in $k_{11} + k_{22}$ variables. A popular measure of similarity in this case is

$$\text{simple matching coefficient: } s_{ij} = \frac{k_{11} + k_{22}}{K};\ i, j = 1, 2,\ldots, n.$$

Note that $s_{ij}$ is the ratio of the number of variables on which the $i$th and the $j$th units match to the total number of variables. This can be calculated for each pair of units. Similar measures are available for polytomous variables [1]. This reference also discusses strategies for mixed sets (quantitative and qualitative) of variables.

In the case of quantitative variables, the usual Pearsonian **correlation coefficient** can be used as a measure of similarity. If $y_{i1}, y_{i2},\ldots, y_{iK}$ are $K$ quantitative (multivariate) measurements on the $i$th subject and $y_{j1}, y_{j2}, \ldots, y_{jK}$ on the $j$th subject, the correlation between the two can be directly calculated. Note that this is being used here for assessing similarity between subjects, whereas the general use is to measure strength of correlation between variables. A more acceptable method in this case is to compute a **measure of dissimilarity** instead of similarity. This, between the $i$th and the $j$th subjects, can be measured by

$$\text{Euclidean distance: } d_{ij} = \sqrt{\Sigma_k (y_{ik} - y_{jk})^2};\ i, j = 1, 2,\ldots, n.$$

This is calculated after standardization (**$z$-score**) of the variables so that the scales do not affect the value. Otherwise, the variables with larger numerical values will mostly determine clustering. This distance can also be calculated for a setup with one variable ($K = 1$). In this case, this reduces to a simple difference between the values.

On the basis of such measurement of similarity or dissimilarity, the subjects are classified into various groups using one of the several possible algorithms. A very popular one is **hierarchical clustering**. With hierarchical algorithm, two units (or subjects) that are most similar (or least distant) are grouped together in the first step to form one group of two units. This group is now considered as one entity. Now the distance of this entity from other units is compared with the other distances between various pairs of units. Again, the closest are joined together. This hierarchical agglomerative process goes on in stages, reducing the number of entities by one each time. The process is continued until all units are clustered together as one big entity. Described later in this section is a method to decide when to stop the agglomerative process so that natural clusters are obtained. This process is graphically depicted by a **dendrogram** of the type shown in Figure C.15. Note that in this method, subsequent clusters completely contain previously formed clusters.

It may not be immediately clear how to compute the distance between two entities containing, say, $n_1$ and $n_2$ units, respectively.

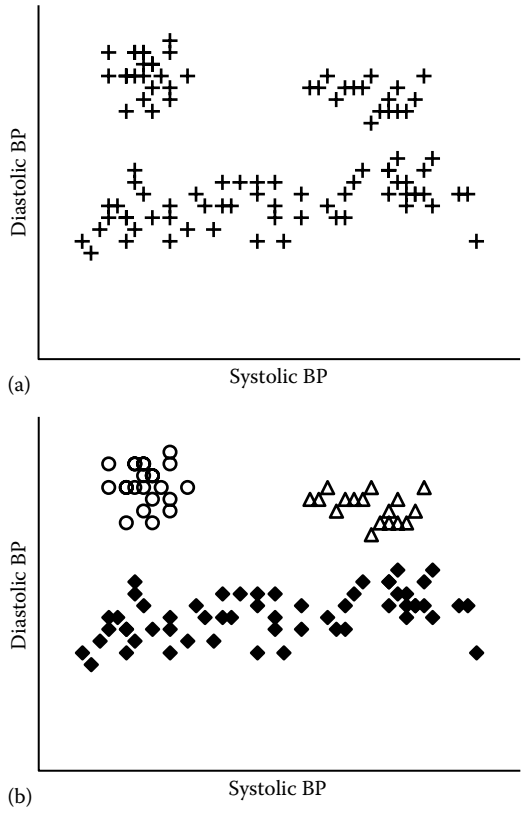



**FIGURE C.14** Scatter plot of diastolic and systolic levels of blood pressure: (a) affinity not shown and (b) one form of affinity groups.

**TABLE C.19**
**Matches and Mismatches in Variables in Two Units Measured on *K* Binary Variables**

| | ***i*th Unit** | | |
|---|---|---|---|
| ***j*th Unit** | **Yes** | **No** | **Total** |
| Yes | $k_{11}$ | $k_{12}$ | $\boldsymbol{k_{11} + k_{12}}$ |
| No | $k_{21}$ | $k_{22}$ | $\boldsymbol{k_{21} + k_{22}}$ |
| **Total** | $\boldsymbol{k_{11} + k_{21}}$ | $\boldsymbol{k_{12} + k_{22}}$ | $\boldsymbol{K}$ |

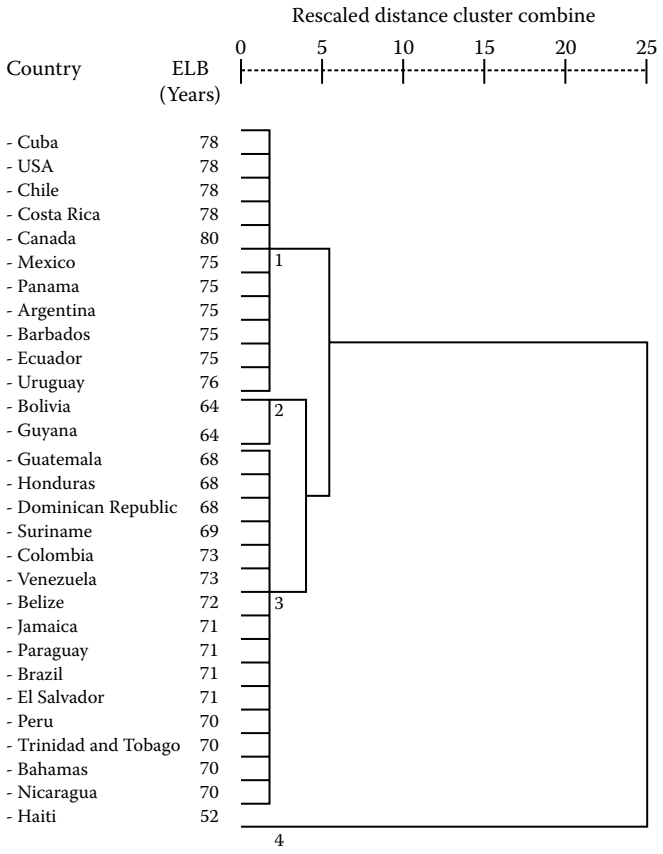



**FIGURE C.15** Dendrogram of Pan-American countries for expectation of life at birth (2004): average linkage method.

Several methods are available. One is to consider all units in an entity centered on their average. Another method is to compute the distance of each unit of one entity with each of the other entity and take the average. A third method is to base it on the nearest units. There are several others. Depending on how this distance is computed, names such as **centroid**, **average linkage**, **single linkage**, etc., are given to these methods. Different methods can give different results. Jain et al. [2] have studied relative merits and demerits of some of these methods on random data. No specific guideline can be given, but a method called **complete linkage** has been found to perform better in not discovering false cluster. This method uses the farthest distance between units belonging to different entities as the measure of distance between two entities.

Cluster analysis can also be done on just one variable. Figure C.15 shows the dendrogram obtained for Pan American countries when clustered on the basis of 2004 values of expectation of life at birth (ELB). The method followed for this clustering is the average linkage. The distance on the horizontal axis is rescaled in proportion to the actual distances between entities. The algorithm detected four clusters as shown from 1 to 4 in the figure. These are given in Table C.20. The fourth cluster has only one country, namely, Haiti, which has very low ELB. Note the gap in ELB between clusters that makes them distinct, and qualify them to be called natural clusters.

### *Deciding on the Number of Natural Clusters*

The most difficult decision in the hierarchical clustering process is regarding the number of clusters naturally present in the data. The decision is made with the help of a criterion such as pseudo-*r* or the **cubic clustering criterion** [3]. These values should be high compared with the adjacent stages of the clustering process. Another criterion could be the distance between the two units or entities that are being merged in different stages. If this shows a sudden jump, it is indicative of a very dissimilar unit joining the new entity. Thus, the stage where the entities are optimal in terms of internal homogeneity and external isolation can be identified. The entities at this stage are the required natural clusters.

**TABLE C.20**
**Clusters Discovered by Average Linkage Method for Pan-American Countries with Respect to the Expectation of Life at Birth**

| Cluster | Expectation Life at Birth (Years) |
|---|---|
| 1 | 80–75 |
| 2 | 73–68 |
| 3 | 64 |
| 4 | 52 |

The following comments regarding cluster analysis may be helpful:

- The algorithm just described is a hierarchical agglomerative algorithm. You can use a *hierarchical divisive algorithm* in which the beginning is from one big entity containing all the units, and divisions are made in subsequent stages. However, this is rarely favored because agglomeration is considered a natural clustering process.
- The other algorithm is nonhierarchical. This can be used when the number of clusters is predetermined. We do not recommend this algorithm because it does not adequately meet the objective of discovering an unspecified number of natural clusters.
- Cluster analysis methods have the annoying feature of "discovering" clusters when, in fact, none exists. A careful examination of the computer output for cluster analysis, particularly with regard to the criteria for deciding the number of clusters, should tell you whether or not natural clusters really exist.
- Since there is no target variable, the clusters so discovered may or may not have any medical relevance.
- Different clustering methods can give different clusters. One strategy to overcome this problem is to obtain clusters by several different methods and then look for consensus among them. Such consensus clusters are likely to be stable. The consensus may be difficult to identify in a multivariate setup. Indrayan and Kumar [4] have given a procedure to identify consensus clusters in the case of multivariate data.
- Cluster analysis has developed into a full subject by itself. Details of the procedures just described and of several other cluster procedures are available in a book by Everitt et al. [5].
- The clustering mentioned above is different from the clustering of subjects with disease in population units such as households. For this, papers by Mantel [6] and Fraser [7] may be helpful.

1. Romesberg MR. *Cluster Analysis for Researchers*. Lulu.com, 2004.
2. Jain NC, lndrayan A, Goel LR. Monte Carlo comparison of six hierarchical clustering methods on random data. *Pattern Recog* 1986;19:95–9. http://www.sciencedirect.com/science/article/pii/0031320386900385

3. Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 1985;50:159–79. http://link.springer.com/article/10.1007/BF02294245#page-1
4. Indrayan A, Kumar R. Statistical choropleth cartography in epidemiology. *Int J Epidemiol* 1996;25:181–9. http://ije.oxfordjournals.org/content/25/1/181.full.pdf?origin=publication_detail
5. Everitt B, Landau S, Lease M. *Cluster Analysis*. Hodder Arnold, 2001.
6. Mantel N. Re: Clustering of disease in population units: An exact test and its asymptotic version. *Am J Epidemiol* 1983;118:628–9. http://aje.oxfordjournals.org/content/118/5/628.short
7. Fraser DW. Clustering of disease in population units: An exact test and its asymptotic version. *Am J Epidemiol* 1983;118:732–9. http://aje.oxfordjournals.org/content/118/5/732

## clustering effect, *see* design effect and the rate of homogeneity

## cluster randomization, *see* block, cluster, and stratified randomization, and minimization

## clusters, *see also* cluster analysis

A cluster is a collection of units. The term is used generally for a collection of those units who have something in common. Thus, we can have a cluster of patients of a certain disease who are treated in a clinic, a cluster of households in a colony, a cluster of people living together as a family, a cluster of blood samples collected on a day for specific investigation, etc. Proximity may be physical, for a particular category of a characteristic, or for any such criteria. A distinctive statistical feature of clusters is their internal homogeneity and external isolation—each contains units similar to one another, but other clusters contain different types of units.

We can also have a cluster of variables such as height, weight, head circumference, and chest circumference, all of which are used for assessing growth of children, or a cluster of measurements for assessing kidney function, or a cluster of measurements for lung or liver function, etc. Thus, "cluster" is quite a general term, but, in this volume, we restrict to a cluster of units rather than to a cluster of variables.

Whereas clusters are many times predefined, sometimes statistical procedures are used to find internally homogeneous and externally isolated clusters. These can be understood as **natural clusters**. Generally, in this case, the number of clusters is not fixed. This depends on how many clusters will achieve the desired objective. For details, see **cluster analysis**. The other method could be to divide the subjects by a specified number of **quantiles**. This is used when the number of clusters is predetermined and internal homogeneity is a marginal issue instead of a core issue. One of the popular quantiles used for this purpose is **tertile**. This divides the subjects into three groups of equal size. For example, you can divide 1-year children into those with low, medium, and high weight using 8 and 10 kg as tertiles—which means thereby that one-third children have weight less than 10 kg, one-third between 10 and 12 kg, and one-third 12 kg or more.

The other important issue is the cluster size. This is the number of units in a cluster. If something like tertiles are used as cutoffs, the size is predetermined to be one-third of the total $n$. When natural clusters are used, however, one cluster may have very different number of units than the others, e.g., only 3 subjects in one cluster and 18 in the other depending on their **affinity**. In the case of **cluster sampling** for field surveys, the clusters are mostly villages, and these could greatly vary in size. Such clusters are chosen for managerial convenience.

## cluster sampling

The term *cluster sampling* is used when clusters of units are selected instead of individual units. **Clusters** are collections of similar units. They may be physically contiguous such as dwelling units, students in a class, or patients in a hospital ward, or any such groups.

When the size of the clusters is not large, i.e., when they generally contain a small number of subjects, then it is sometimes advisable that these units are not sampled further. All the elements in the selected clusters are then surveyed. This tends to increase the total number of subjects in the sample without a corresponding increase in the cost. When subjects in close proximity are included in the sample, the travel time and cost are saved. If a population comprises a total of $N$ clusters, then $n$ clusters out of $N$ are randomly selected. If the $i$th cluster has $M_i$ subjects, then a total of $\Sigma M_i$ elements of these $n$ clusters are investigated. This is called cluster random sampling (CRS). This sampling gives extremely good results when the units within the clusters are heterogeneous with respect to the outcome of interest. The method is used also in settings with a large number of units, but then they are divided into small clusters. A schematic presentation of cluster sampling is shown in Figure C.16, where clusters are delimited by bold lines and 3 have been selected out of 10 clusters. The units in sample are those with dots.

Besides the fact that CRS helps to increase the size of the sample and thus the **precision** (without a corresponding increase in the cost), the other advantage relative to **simple random sampling** (SRS) is that the **sampling frame** of the units is not required and the only frame required is the list of clusters. For example, it is easy to draw a list of hospitals than a list of patients of a particular disease. In this case, hospitals are the clusters of patients of that disease. In another setup, a family could be a cluster of interest. Cluster sampling is also very easy to administer. Since the survey of subjects within the cluster is quick due to close proximity, CRS is sometimes considered a **rapid assessment method**. At one time, the World Health Organization (WHO) recommended this kind of sampling

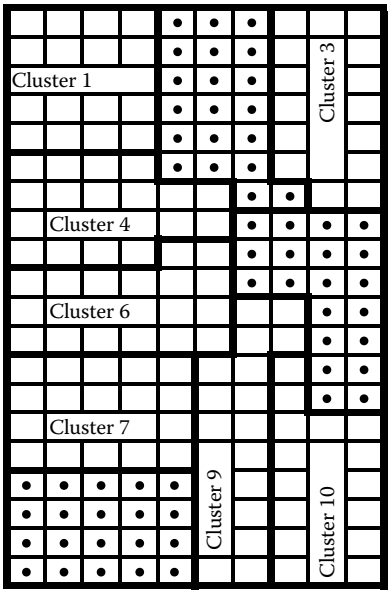



**FIGURE C.16** Schematic presentation of cluster sampling.

C

to estimate the percentage of children immunized in a community, particularly in developing countries. Their recommended strategy was 30 clusters of size 7 each—called **30 × 7 sampling**. Bennett et al. [1] have given some very useful details of this methodology. The methodology has become popular and is used in many other setups, more for convenience than for statistical propriety. The 30 × 7 methodology is appropriate when the anticipated prevalence is around 50% and the precision required is ±10% points. For other setups, a different size of cluster and a different number of clusters may be needed. Since the immunization level in most countries has reached 70% or 80%, this methodology has become outdated for immunization surveys. However, such CRS is used as a rapid assessment method for assessing the status of other aspects of health as in estimating prevalence of poor vision in old age as per the example given later in this section.

As you will see shortly, a disadvantage of CRS is that the elements within a cluster tend to be similar to one another and produce a clustering effect. This is also called the **design effect**. This effect reduces the chances of getting the full spectrum of subjects in the sample. To compensate for this, a larger sample may be required relative to SRS. However, it sometimes happens that even a large sample chosen by CRS is less expensive to investigate than a small sample chosen by SRS. Intercluster comparison is not valid in the case of CRS.

It is customary in the case of CRS that the population is divided into clusters of equal or approximately equal size, although this is not a prerequisite. The size of cluster—that it should have 20, 30, or 50 subjects—is generally determined by administrative convenience, such as the ability of a team to finish the survey in 1 day. However, the number of clusters would depend on the statistical requirement of reliability of the estimate.

The following example illustrates one application of cluster sampling. For a survey on prevalence of poor vision (visual acuity <6/36 in the better eye with corrective glasses if any) in persons of age 50 years and above (50+) in a district with half a million population, suppose 20 clusters of size 30 each are selected as per the following scheme:

a. A list of census blocks is prepared along with the population of each, and this is cumulatively added.
b. Since the sampling fraction is one cluster per (500,000/20 =) 25,000 population, one number less than or equal to 25,000 is randomly selected. Then, 25,000 is sequentially added every time in a systematic fashion, and thus a sample of 20 numbers is obtained. Twenty blocks containing the chosen 20 numbers are selected from the list made in (a). These blocks are now in the sample.
c. Home visits are made from a geographically random point in each of the selected blocks, and the first 30 persons of age 50+ residing in contiguous houses are listed and examined for visual acuity. This gives one cluster of 30 subjects from each selected block.

The scheme in this example is similar to the one recommended by WHO for surveys to assess immunization coverage in developing countries, but it is not exactly the same. Note the following features of the CRS in this example:

- The frame required is only the list of census blocks, which in any case is generally available. No listing of households or of persons of age 50+ is required.
- The selection of blocks is on the basis of the size of the population. This is inherent in step (b). Blocks with a larger population have a higher chance of being included in the sample. This is called sampling with **probability proportional to size** (PPS). Details of this method are given under that topic. The size in this case is the population in the blocks, and the subjects are the persons of age 50+. It is reasonable to expect that this age group would have nearly the same proportion in each census block. The PPS sampling makes the estimate self-weighting, and the usual sample proportion becomes a statistically valid estimate of the population proportion. Weighted calculations, as done for **stratified random sampling** (StRS), are not required.
- Starting from a geographically random point ensures random sampling but may require a map of each of the selected blocks.
- The houses to be visited are contiguous. This should make the survey substantially faster. Because of this property, it is valid to consider it as a rapid assessment method. It is up to the investigators to define contiguity. These could be houses lined up on both sides of a street or houses whose entrance is closest to the previously selected houses. This must be defined before the survey begins.
- The total number of clusters of size 30 would be large in a district that has a population of half a million. A sample of 20 clusters is relatively small. This would be generally the case in CRS.

Not all of the first 30 persons of age 50+ in a cluster selected for the survey may be available, and some may not cooperate in the examination of their vision. In practice, the **nonresponse** could be large in this kind of CRS and may have to be tackled separately.

1. Bennett S, Woods T, Liyanage WM, Smith DL. A simplified general method for cluster-sample surveys of health in developing countries. *World Health Stat Q* 1991;44:98–106. http://apps.who.int/iris/bitstream/10665/47585/1/WHSQ_1991_44%283%29_98-106_eng.pdf?ua=1

## *C*-max, *see* pharmacokinetic parameters ($C_{max}$, $T_{max}$) and pharmacokinetic studies

## Cochrane collaboration/reviews

Cochrane reviews [1] are internationally recognized systematic reviews of medical literature on specific topics relating to prevention, diagnosis, and treatment of various diseases. They follow a strict protocol using tools such as full literature search, study flow diagram [2], and **meta-analysis**, and are believed to come up with the best available evidence on specific aspects of health care. Many consider this as "research into research."

Cochrane reviews are a collaborative effort, known as Cochrane collaboration. The reviews were founded by Iain Chalmers in 1991, but he preferred to call them Cochrane collaboration [3].

These reviews are published online in The Cochrane Library and regularly updated to provide the latest information. More than 5000 reviews are available in this library as of 2015. The reviews intend to provide answers to everyday questions such as whether vitamin D supplements prevent cancer in adults, or whether zinc helps in common cold. However, the answer may be inconclusive depending upon what type of information is available in the literature. For example, Goda et al. [4] found that the percentage of inconclusive reviews in the field of pediatric gastroenterology has increased from mere

9% in the block years 1998–2002 to 24% in the years 2008–2012. They included Cochrane reviews as well as other reviews in their study. The reasons they cited for inconclusive reviews are heterogeneity of studies and small sample size. Another reason could be increased awareness among researchers about the importance of the methodology.

The Cochrane Library allows free access to the abstract of the reviews but requires a subscription for accessing the full texts. However, some governments such as in Scandinavian countries have obtained license for free viewing by their health professionals. A word of caution for those who tend to take Cochrane reviews as the gold standard: sometimes an overview of Cochrane reviews is needed to come to some definitive conclusion. For example, a large number of interventions address the problem of preterm births, and in an evaluation of 56 Cochrane reviews, Piso et al. [5] found not more than one-fourth of these interventions effective in lowering the incidence of preterm births. Remember that Cochrane reviews suffer from the same ailment that all systematic reviews do. They are based on published research, and we all know about **publication bias** that inhibits reporting and publication of the trials with negative results. Thus, the results based on published research tend to provide a biased picture in favor of the positive findings.


1. Cochrane Reviews. http://www.cochrane.org/cochrane-reviews
2. Stovold E, Beecher D, Foxlee R, Noel-Storr A. Study flow diagrams in Cochrane systematic review updates: An adapted PRISMA flow diagram. *Syst Rev* 2014 May 29;3(1):54. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4046496/
3. Champkin J. "We need the public to become better BS detectors": Sir Iain Chalmers. *Significance* July 2014;11(3):25–30. http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2014.00751.x/abstract
4. Goda Y, Sauer H, Schöndorf D, Hennes P, Gortner L, Gräber S, Meyer S. Clinical recommendations of Cochrane reviews in pediatric gastroenterology: A systematic analysis. *Pediatr Int* 2015;57(1):98–106. http://www.ncbi.nlm.nih.gov/pubmed/24978114
5. Piso B, Zechmeister-Koss I, Winkler R. Antenatal interventions to reduce preterm birth: An overview of Cochrane systematic reviews. *BMC Res Notes* 2014 Apr 23;7:265. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4021758/


## Cochran *Q* test

The Cochran *Q* test is an extension of the **McNemar test** for 2×2 tables to *T* repeated measures with **dichotomous** outcome. This tests whether or not the effect at *T* repeated measures is the same. *T* can also be the number of treatments on the same group of subjects, or different group with one-to-one **matched** subjects. This test criterion is given by

$$\text{Cochran } Q = (T-1)\frac{T\Sigma_t P_t^2 - P^2}{TP - \Sigma_i S_i^2},$$

where $T$ = number of repetitions

$P_t$ = number of positives at the $t$th ($t$ = 1, 2,…, $T$) repetition

$P$ = $\Sigma_t P_t$ = total number of positives

$S_i$ = number of positives for the $i$th ($i$ = 1, 2, …, $n$) subject in $T$ repetitions.

For large $n$, $Q$ follows a **chi-square** distribution with df = $(T-1)$. Thus, the ***P*-value** can be obtained. You may never need to use this formula since the software will give the *P*-value. The exact distribution of $Q$ is also known, which is valid for small $n$ also. A good statistical package will give you exact *P*-value as well. The test was first proposed by William Cochran in 1950 [1]. The following example illustrates the calculation for those who want to see exactly how it works.

William Cochran
(Courtesy of Fabian Bachrach Studio.)

Thirty patients of benign prostatic hyperplasia (BPH) whose disease was under control were assessed for their perceived health at monthly intervals for 4 months. The data obtained are shown in Table C.21.

**TABLE C.21**
**Self-Perceived Satisfaction with Health of BPH Patients at Monthly Intervals**

| Subject (*i*) | Time 1 | Time 2 | Time 3 | Time 4 | $S_i$ |
|---|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 1 | 3 |
| 2 | 0 | 1 | 0 | 0 | 1 |
| 3 | 1 | 0 | 1 | 1 | 3 |
| 4 | 1 | 1 | 1 | 1 | 4 |
| 5 | 0 | 0 | 0 | 1 | 1 |
| 6 | 0 | 1 | 1 | 1 | 3 |
| 7 | 1 | 0 | 0 | 0 | 1 |
| 8 | 0 | 0 | 1 | 0 | 1 |
| 9 | 0 | 1 | 0 | 0 | 1 |
| 10 | 1 | 1 | 1 | 1 | 4 |
| 11 | 0 | 1 | 1 | 1 | 3 |
| 12 | 0 | 1 | 1 | 0 | 2 |
| 13 | 0 | 0 | 1 | 1 | 2 |
| 14 | 1 | 1 | 0 | 0 | 2 |
| 15 | 0 | 1 | 0 | 1 | 2 |
| 16 | 0 | 0 | 1 | 1 | 2 |
| 17 | 0 | 0 | 0 | 1 | 1 |
| 18 | 0 | 0 | 1 | 1 | 2 |
| 19 | 1 | 1 | 0 | 1 | 3 |
| 20 | 1 | 1 | 1 | 0 | 3 |
| 21 | 0 | 0 | 0 | 1 | 1 |
| 22 | 0 | 1 | 1 | 1 | 3 |
| 23 | 0 | 0 | 1 | 1 | 2 |
| 24 | 1 | 1 | 1 | 1 | 4 |
| 25 | 0 | 1 | 1 | 1 | 3 |
| 26 | 0 | 0 | 0 | 1 | 1 |
| 27 | 0 | 1 | 1 | 0 | 2 |
| 28 | 0 | 0 | 0 | 0 | 0 |
| 29 | 1 | 1 | 0 | 1 | 3 |
| 30 | 0 | 0 | 1 | 1 | 2 |
| **Sum $P_t$** | **9** | **17** | **18** | **21** | **65** |

*Note:* 0 = not satisfied; 1 = satisfied.

A software package gives $Q = 10.862$ and exact $P$-value = 0.012. For those who want to see what is going on underneath, note for these data that $T = 4$, $P_1 = 9$, $P_2 = 17$, $P_3 = 18$, and $P_4 = 21$. Thus, $P = 9 + 17 + 18 + 21 = 65$. Also $S_1 = 3$, $S_2 = 1$, $S_3 = 3$, …, $S_{30} = 2$ as given in the last column of the table. Substituting these values gives

$$Q = (4-1)\frac{4\times(9^2+17^2+18^2+21^2)-65^2}{4\times 65-(3^2+1^2+\ldots+2^2)}$$

$$= 3\times\frac{4\times 1135-4225}{260-173} = 10.86.$$

This is the same as was just obtained by the software package. At $T - 1 = 4 - 1 = 3$ df, the critical value of chi-square is 7.815 at $\alpha = 0.05$. Since the calculated value is higher, $P < 0.05$ and the result is statistically significant. The same was obtained by the exact $P$-value. Conclude that self-perceived satisfaction in these people is different at different time points.

For $T = 2$, Cochran $Q$ reduces to McNemar $\chi^2$ as it should. If you find $Q$ statistically significant and want to find where this difference actually is, do McNemar for pairs of interest using the **Bonferroni** procedure. This procedure for multiple comparisons is separately explained. If there are a total of $K$ pairwise comparisons of interest, the Bonferroni procedure requires that you use **significance level** $\alpha/K$ for each comparison instead of $\alpha$. This keeps the total probability of **Type I error** less than $\alpha$.

1. Cochran WG. The comparison of percentages in matched samples. *Biometrika* 1950;37(3/4):256–66. http://www.jstor.org/stable/2332378

## Cochran test for linearity of trend, *see* chi-square test for trend in proportions and deviation from trend

## coding

This is the process of assigning numbers to qualities. This helps in converting long textual characteristics to a single or few digits, thereby helping in data entry and in analysis. For example, we can have code 1 for cancer of lung, 2 for cancer of cervix, 3 for cancer of breast, etc. You can see that codes are used for **nominal categories**.

It is customary to code values of **binary** variables as 0 for absent and 1 for present. For example, you can assign code 1 to people with a particular disease and 0 to those without disease. When this is done, the proportion of persons with the disease is the average of these 0's and 1's. This explains why a proportion also is a form of mean, and thus eligible to get advantage of the **central limit theorem** for a **Gaussian distribution** for large $n$. Such coding also helps in obtaining unifying **regressions** and their proper interpretation. Consider the following hypothetical regression equation for healthy adults:

[A] sysBP(mmHg) = 110 + ⅓*Age(years) – 3*Sex,

where sysBP is the systolic level of blood pressure, and sex is coded as 0 for females and 1 for males. When these codes are substituted, the equation is

for males (sex = 1): sysBP(mmHg) = 107 + ⅓*Age(years),

for females (sex = 0): sysBP(mmHg) = 110 + ⅓*Age(years).

Thus, a single equation [A] yields two equations—one for males and one for females. This kind of binary coding is also called creating **indicator variables**. In this example, sex is an indicator variable when coded as 0 and 1. This kind of coding is almost invariably used for the outcome in **logistic regression** and commonly for binary **regressors**.

The situation with multiple categories is not so simple. Code 1 for cancer of lung, 2 for cancer of cervix, 3 for cancer of breast, etc. cannot be used in equations since, in this case, code 2 is not double of code 1, and the difference between code 3 and code 2 is not code 1. They need to be recoded as described for indicator variables. The moral is that codes per se cannot be used as quantities.

Next is the coding of **ordinal categories**. Coding none, mild, moderate, serious, and critical condition as 0, 1, 2, 3, and 4, respectively, also cannot be used as quantities since two moderate cases (code = 2) are not equal to one critical case (code = 4). However, in some specific situations, these can be used as quantities, now called **scores**. For example, if you want to investigate the relationship between duration of survival in cases with different degrees of severity of any disease, such codes can be used as scores for assessing their effect on duration of survival. The underlying requirement for such use is that the difference in survival between mild and moderate cases is the same as between serious and moderate cases. If this is likely, codes for ordinal categories can be used as scores. Remember though that these codes, when used as scores, are linear since they are increasing by the same quantity (1 in this case) as we move up in severity of disease. Scores can be nonlinear also such as 0, 1, 3, 6, and 10 depending on the assessment of severity; however, codes cannot be nonlinear.

Metric categories such as age 15–19, 20–24, 25–29, etc., can also be coded as 1, 2, 3, etc. This, in fact, is a simple linear transformation—in this case, $\text{code} = \frac{\text{midpoint of the inerval} - 12}{\text{width of the interval}}$. The width of the **class interval** is actually 5 although it looks like 4. Thus, these codes can indeed be used as scores with no issues related to coding, but there are issues with using categories for a variable on **metric scale** of the type of age. For details, see **categories of data values** in this volume. There are exceptions where this does not apply. The average number of alcohol drinks per day can be categorized as 0, <1.0, 1.0–2. 9, 3.0–5. 9, and ≥6.0. Since the categories are metric, midpoints seem to be the most adequate scores for this kind of analysis. Try to resist the temptation to use equally spaced codes 0, 1, 2, 3, and 4 as scores in such cases since these categories do not follow this linear pattern.

## coefficient of correlation, *see* correlation coefficient (Pearsonian/product–moment)

## coefficient of determination, *see also* multiple correlation

Coefficient of determination of a model is a measure of its goodness of fit to the observed values. This is mostly used for regression models and can be understood in two ways:

$$\text{coefficient of determination: } \eta^2 = \frac{\text{Reg SS}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}},$$

where RegSS is the sum of squares due to regression, SST is the total sum of squares, and SSE is the sum of squares due to error (also called residual sum of squares). An explanation is as follows.

A regression model is considered a good fit if the **residuals** $e = (y - \hat{y})$ are small, where $\hat{y}$ is the predicted value of $y$. Since residuals fluctuate around zero in any case, small residuals would

necessarily yield a small sum of squares $\Sigma e^2$. This is the residual **sum of squares**, popularly called the sum of squares due to error (SSE). Its magnitude, when compared with the total sum of squares, $SST = \Sigma(y - \bar{y})^2$, provides a measure of "lack of fit" of the regression, and (SST – SSE) is called the regression sum of squares (RegSS). These have associated degrees of freedom (df's) as in the case of ANOVA. RegSS measures how much of the total sum of squares the regression has been able to account for. The larger the $\eta^2$, the better the fit. If the residual sum of squares is as small as, say, 10% of the total sum of squares, then $\eta^2 = 0.90$. The fit, then, is said to account for 90% variation in *y*. A fit with such high $\eta^2$ should be adequate in most cases, especially if *n* is large.

There are two important underlying features of the coefficient of determination. First, the dependent must be quantitative and not qualitative since the usual regression is obtained for quantitative dependent only. The independent variable could be quantitative, qualitative, or a mixture of both. Second, there must be a **model** for predicting or explaining the values.

The positive square root of the coefficient of determination, η, exists but has no particular name when the relationship is nonlinear. This square root is called the coefficient of **multiple correlation** when the relationship is *linear* and the number of independent variables is more than one. This is denoted by *R*. No meaning can be attached to the direction of correlation when the number of regressors is more than one. Thus, the sign of *R* is always considered positive. This is the correlation between *y* and its values predicted by the linear regression. Thus, the term *coefficient of determination* should be used only for nonlinear regressions. Many researchers erroneously use the term coefficient of determination for the coefficient of multiple correlation.

## coefficient of reproducibility, *see* Guttman scale

## coefficient of variation

**Variation** among values has a tendency to be higher for values that are numerically bigger. For example, variation, as measured by the **standard deviation (SD)**, would be higher for cholesterol level than for bilirubin level since the cholesterol levels are more than 100 mg/dL but bilirubin levels are between just 0.20 and 1.10 mg/dL. Thus, variation in one variable cannot be compared with variation in another variable if we use SD. To correct this "anomaly," particularly for comparison, coefficient of variation (CV) is calculated as follows:

$$\text{coefficient of variation: CV} = \frac{\text{SD}}{\text{mean}}.$$

Since the units of measurement (g/dL or mmHg or any other) cancel out in such a ratio, the CV is independent of units and makes it a valid index for comparison of variations in variables on very different scales and different units. Thus, the CV removes the contextualization of variation.

An SD of 5 mmHg for systolic BP readings is small, but an SD of even 3 g/dL for Hb level is large. If the mean systolic BP level of the subjects under study is 132 mmHg, then SD = 5 mmHg is only 3.8% of the mean. If the mean Hb level is 15 g/dL, SD = 3 is 20% of the mean. The latter SD is surely higher, relatively speaking. Thus, SD by itself cannot be used to *compare* variability in two different kinds of variables. Its ratio with the mean can be used. For example, it can be concluded on the basis of CV that the variation between healthy individuals in thermoregulation (oral, rectal, and skin temperatures) is small and that in renal functions (urinary flow, creatinine clearance, urea clearance, etc.) is high. In the former case, the SD does not generally exceed 1% of the mean value, but in the latter case, the SD could be as much as 50% of the respective averages. Cardiovascular functions (systolic/diastolic BPs, heart rate, etc.) are somewhere in between with SD around 5%–15% of the mean.

Figure C.17 shows plots of the value of serum uric acid and blood ammonia in a sample of subjects. They are plotted in a manner that the variability looks equal. It all depends on the scale chosen on the *x*-axis. In these subjects, the level of serum uric acid is between 2.5 and 9.1 mg/dL, the SD is 2.44 mg/dL, and the CV is 44%. Against this, the level of blood ammonia ranges from 80 to 110 μg/dL, and the SD is 9.78 μg/dL. The CV is only 10%. Thus, variability is much lower in blood ammonia than in serum uric acid. The graphs could be deceptive as in this example.

The following example illustrates a useful application of the CV. The R–R interval in the electrocardiograms (ECG) can be used to evaluate brainstem function. This interval is highly variable in healthy subjects. The R–R interval is measured several times in a subject in one ECG, and the mean and SD are obtained. Thus, CV can be obtained for each subject separately. Nezu et al. [1] studied 18 children with severe brain damage and 22 controls. They obtained the results shown in Table C.22.

**TABLE C.22**
**Mean Coefficient of Variation (CV) of R–R Intervals in Cases with Different Conditions**

| Condition | *n* | Mean CV of R–R Intervals (%) |
|---|---|---|
| Neurologically normal controls | 22 | 5.56 |
| Patients complicated with respiratory insufficiency (RI) | 10 | 2.19 |
| Patients with severe athetotic cerebral palsy (SA) | 8 | 11.30 |
| Patients with brain death (included in RI group also) | 4 | 1.00–1.29 |

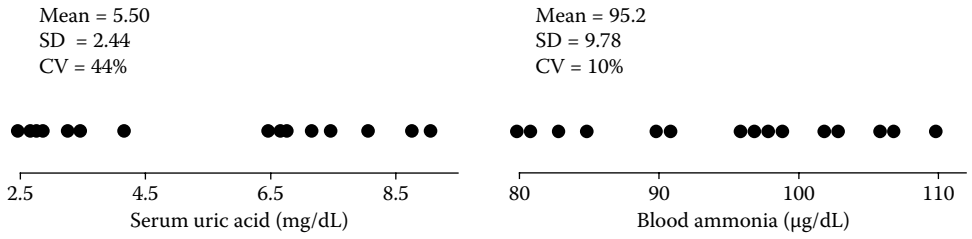



**FIGURE C.17** Less SD but higher CV of serum uric acid levels relative to blood ammonia levels.

C

The authors did not investigate the cause of the higher CV in the severe athetotic cerebral palsy (SA) group but concluded on the basis of the extremely low CV in patients with brain death that the CV of the R–R interval may be useful for quantitative evaluation of severe neurological deficit. In general, the CV of R–R intervals is also reduced in the aged and possibly also in subjects suffering from essential hypertension.

Chang et al. [2] have described another useful application of CV. They measured the nuclear diameter of cancer cells by ocular micrometry and calculated the CV. Recurrent cases and cases in which death occurred had higher CV than their respective complements. They concluded that the extent of variation as measured by the CV of the nuclear diameter of cancer cells offers a prognostic adjunct to standard clinical and histological analysis.

Although CV is unit free, it is affected by change of origin and scale since mean and SD are affected. Thus, it can give a misleading result for data on an interval scale where zero has no substantive meaning and a constant can be added or subtracted. Otherwise, also if some values of a variable are negative and others positive, the mean could be small or close to zero. In this case, the CV will undesirably inflate. This will happen if you are measuring the change from time 1 to time 2, pre to post, etc. Thus, use the CV for **ratio scale** data only where zero is defined and values are positive. Strictly speaking, from this point of view, the CV should not be used for a measurement such as BP and Hb levels since zero is not possible in living subjects. However, we see no such problem in the examples cited earlier in this section since no negative values are possible.

The CV can also be used as an index of reliability in repeated measurements. For details, see Shechtman [3].


1. Nezu A, Kimura S, Kobayashi T, Osaka H, Uehara S. Coefficient of variation of R-R intervals in severe brain damage. *Brain Dev* 1996;18:453–55. http://www.sciencedirect.com/science/article/pii/S0387760496000484
2. Chang IC, Kuo SH, How SW. Coefficient of variation of nuclear diameters as a prognostic factor in papillary thyroid carcinoma. *Anal Quant Cytol Histol* 1991;13:403–6. http://www.ncbi.nlm.nih.gov/pubmed/1807282
3. Shechtman O. The coefficient of variation as an index of measurement reliability, In: *Methods of Clinical Epidemiology* (Eds. Doi SAR, Williams GM). Springer, 2013.

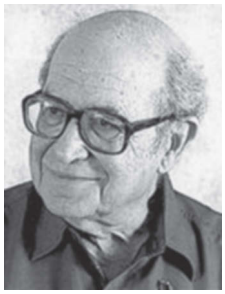

## Cohen kappa, *see also* Bangdiwala B

This measures the extent of **agreement** in qualitative assessments between two observers. Thus, this is applicable when $n$ subjects are categorized into same $K$ categories by two independent observers. This is a matched pairs setup that would lead to a $K \times K$ table. In place of two observers or two raters, you can have two methods, two sites, two laboratories, or any other such comparison.

In terms of notations,

$$\text{Cohen kappa: } \kappa = \frac{\Sigma O_{kk} - \Sigma(O_{k\bullet}O_{\bullet k}/n)}{n - \Sigma(O_{k\bullet}O_{\bullet k}/n)},$$

where $O_{kk}$ is the cell frequency in the $k$th row and the $k$th column (diagonal element) of the $K \times K$ table, $O_{k\bullet}$ is the marginal total in the $k$th row, $O_{\bullet k}$ is the marginal total in the $k$th column, and $n$ is the grand total. The first term in the numerator is the observed agreement, and the second term is the chance agreement as explained next. Thus, the numerator is the agreement in excess of chance. The denominator is its maximum possible value. Further explanation is given later in this section with the help of an example. Cohen kappa was first proposed by Jacob Cohen in 1960 [1].



Jacob Cohen

An example is comparing response evaluation criteria in solid tumors and volumetric evaluation for colorectal cancer with liver-limited metastasis using $n = 45$ computed tomography (CT) images [2]. In this case, the categories were partial response, stable disease, and progressive disease, a total of $K = 3$ categories of responses in the study. Assessing optical disc characteristics by two or more observers, results of Lyme disease serological testing by two or more laboratories and comparison of x-ray images with Doppler images are the other examples of the problem of qualitative agreement. The objective is to find the extent of agreement between two or more methods, observers, laboratories, etc. In some cases, for example, in comparison of two laboratories, agreement has the same interpretation as **reproducibility**. In the case of comparison of observers, it is termed *interrater reliability*. In all these cases, only one group of subjects is assessed twice. Cohen kappa is easy to understand when the meaning of qualitative agreement is clear.

### *The Meaning of Qualitative Agreement*

For simplicity, we will focus for the time being on the presence or absence of a characteristic assessed by two observers on the same group of subject. An example is presence or absence of a lesion in x-rays read by two observers. Suppose the observations are as given in Table C.23.

The two observers agree on a total of 29 + 11 = 40 cases in this example. This is the sum of the frequencies in the leading diagonal cells. In the other 20 cases, the observers do not agree. Apparently, the agreement = 40/60 = 66.7%. But part of this agreement is due to chance, which might happen if both are dumb observers and randomly allocate subjects to present and absent categories. This chance agreement is measured by the cell frequencies expected in the diagonal when the observer's ratings are independent of one another. These expected frequencies are obtained by multiplying the respective marginal totals and dividing by the grand total, as obtained for calculating the **chi-square**. This explains why we subtract $O_{k\bullet} \times O_{\bullet k}$ from the numerator and the denominator while calculating Cohen kappa.

**TABLE C.23**
**Presence or Absence of a Lesion Assessed by Two Radiologists in X-Rays of 60 Suspected Cases**

| | Observer 1 | | |
|---|---|---|---|
| **Observer 2** | **Present** | **Absent** | **Total** |
| Present | 29 | 7 | **36** |
| Absent | 13 | 11 | **24** |
| **Total** | **42** | **18** | **60** |

For the data in Table C.23, the chance-expected frequencies are 36 × 42/60 = 25.2 and 24 × 18/60 = 7.2 in the two diagonal cells. The total of these two is 32.4. Agreement on so many cases is expected by chance alone. Thus, agreement in excess of chance is in only 40.0 – 32.4 = 7.6 cases. The maximum possible excess is 60.0 – 32.4 = 27.6.

The measure of agreement is the ratio of the observed excess to the maximum possible excess, in this case 7.6/27.6 = 0.275 or 27.5%. Thus, the two observers in this case do not really agree much on rating of x-rays for the presence or absence of lesion. Most of their agreement is due to chance. Now consider the following example where the number of categories is three.

Detection of intrathecal immunoglobulin G (IgG) synthesis is important in patients with suspected multiple sclerosis. Isoelectric focusing is a method used for the detection of intrathecal IgG synthesis. Let this be assessed as positive, doubtful, and negative by two laboratories on 129 patients. The results are in Table C.24. In this case,

$$\kappa = \frac{(36+12+55)-(44\times 44/129+25\times 21/129+60\times 64/129)}{129-(44\times 44/129+25\times 21/129+60\times 64/129)}$$
$$= \frac{54.155}{80.155} = 0.68.$$

Generally speaking, a kappa value equal to 0.68 is adequate to conclude fair agreement, but the investigation in this example is on reproducibility between laboratories. If both the laboratories are using standardized tools and methods, the agreement should be close to 1. Thus, the reproducibility of the method of isoelectric focusing between the laboratories for assessing intrathecal synthesis cannot be considered good in this case despite a not so disappointing value of kappa.

Kappa is used mostly to assess how close the agreement is to 1 rather than how far is it from 0. The following scale is suggested.

| Kappa | Strength of Agreement |
|---|---|
| <0.3 | Poor |
| 0.3–0.5 | Fair |
| 0.5–0.7 | Moderate |
| 0.7–0.9 | Good |
| >0.9 | Excellent |

The following comments regarding Cohen kappa may be helpful:

- There are some other measures of agreement for qualitative variables. These are discussed by Agresti [3]. He also describes other agreement assessment models such as log-linear, Rasch, and latent class models. A competitor of Cohen kappa is **Bangdiwala B**. A comparison of these two is presented by Munoz and Bangdiwala [4].
- Cohen kappa is valid for nominal categories only. Ordinal or metric categories are considered nominal by this measure and the order is ignored. The other assumptions are that (i) the subjects are independent; (ii) the observers, laboratories, or methods under comparison operate independently of one another; and (iii) the rating categories are mutually exclusive and exhaustive. These conditions are easily fulfilled in most practical situations.
- Although rare, you may sometimes find reference to **weighted kappa**. Here, the cells are assigned a weight according to the degree of disagreement they exhibit. Thus, cells in the diagonal, since they are in full agreement, get zero weight. Off-diagonal cells get varying weight depending upon either the perceived importance of the involved cells or some objective criterion such as quadratic weight. All this makes kappa too complex and possibly not as useful.
- The kappa value is +1 for complete agreement and 0 if the agreement is the same as expected by chance. However, the value does not become –1 for complete disagreement. Thus, *it is a measure of the extent of agreement but not of disagreement*.
- For the formula of variance of kappa for large samples, see Fleiss et al. [5]. This variance can be used to construct a confidence interval and to test a hypothesis on the value of kappa. Standard statistical software packages generally have provision to do these calculations.
- Kappa values from different studies may not be comparable as the value also varies with prevalence, which is the number of subjects in categories relative to the total. If one study has 30% subjects in category A and another study has 50%, the value of kappa will differ even if the extent of agreement is the same. In situation of rare events, a low kappa may not necessarily reflect low level of agreement.
- Kappa does not distinguish between +/– **discordance** and the reverse –/+ discordance. Both get the same weight. Thus, kappa can be 0.9 in 100 subjects when all discordances are of the +/– type and none of the –/+ type.
- The discordance between two raters can be due to bias of one rater to classify differently or could be a genuine difference in assessment. Kappa does not distinguish these two different kinds of discordance, and both are combined.
- Cohen kappa can also be used for assessing **test–retest reliability** in qualitative data.

**TABLE C.24**
**Assessment of Intrathecal Synthesis by Two Laboratories**

| | Laboratory 1 | | | |
|---|---|---|---|---|
| **Laboratory 2** | **Positive** | **Doubtful** | **Negative** | **Total** |
| Positive | 36 | 5 | 3 | **44** |
| Doubtful | 7 | 12 | 6 | **25** |
| Negative | 1 | 4 | 55 | **60** |
| **Total** | **44** | **21** | **64** | **129** |


1. Cohen, J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20(1):37–46. https://www.researchgate.net/publication/220017506_A_coefficient_of_agreement_for_nominal_scales, last accessed April 28, 2015.
2. Fang WJ, Lam KO, Ng SC et al. Manual contouring based volumetric evaluation for colorectal cancer with liver limited metastases: A comparison with RECIST. *Asian Pac J Cancer Prev* 2013;14(7):4151–5. http://www.ncbi.nlm.nih.gov/pubmed/23991968
3. Agresti A. Modelling patterns of agreement and disagreement. *Stat Methods Med Res* 1992;1:201–18. http://smm.sagepub.com/content/1/2/201.full.pdf
4. Munoz SR, Bangdiwala SI. Interpretation of Kappa and B statistics measures of agreement. *J Appl Stat* 1997;24(1):105–12. http://www.tandfonline.com/doi/abs/10.1080/02664769723918#.UvD4ydKSxIA
5. Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa. *Psychol Bull* 1969;72:323–7. http://psycnet.apa.org/psycinfo/1970-01528-001

C

## cohort studies, *see also* prospective studies

A **cohort** is a group of subjects that share a common base and observed forward in time. In the case of a usual prospective study, the subjects can be enrolled continuously and can leave the study abruptly, whereas, in the case of cohort, enrollment, joining, or leaving in the middle is rarely admissible. In a study on use-effectiveness of oral contraceptive pills [1], the users joined the group when they started using the pills and left when they stopped using them. They could not join 2 or 3 months after starting the pill. All subjects of a cohort do not have to start from the same calendar time, but they all start from the time of occurrence of the same event.

The term *cohort* generally connotes a substantial time gap between exposure and outcome, and the observation spans all or most of this period. There could be a cohort of children born in a particular year followed up for growth pattern, or a cohort of adults residing in an area at a time followed up for diet–exercise and occurrence of coronary events. There could be a cohort of smokers and a matched cohort of nonsmokers residing in an area at a particular time followed up for 20 years for the development of chronic obstructive pulmonary disease. Many other examples can be cited. Such cohorts are called **concurrent cohorts** when exposed and unexposed groups are followed up in future. If the group identified for the study is the one that has been recently exposed to a risk factor, it is called an **inception cohort**. For example, an inception cohort of early rheumatoid arthritis can be studied for assessing predictive factors of orthopedic surgery.

It is not necessary that the follow-up chronology is in the future; it could be in the past also. Sokal et al. [2] carried out a cancer risk study in 1992 on the basis of the records of women sterilized with transcervical quinacrine hydrochloride pellets in Chile between 1977 and 1991. Traceable women were also interviewed. Despite being based on past records, it is not a retrospective study since the direction of investigation is from antecedent to outcome. A **historical cohort** or **retrospective cohort** is a group of subjects with exposure in the past, investigated for development of an outcome. Terms such as *retrospective follow-up* and *historical prospective* are also used for this kind of methods. This requires that past records are fully available. See the following examples.

Richards et al. [3] report findings of a 53-year follow-up of a 1946 birth cohort, initially consisting of 5362 children of nonmanual and agricultural workers, and a random sample of one-in-four manual workers selected from all single and legitimate births that occurred in England, Scotland, and Wales during 1 week in March 1946. (Note the rigorousness with which the specifications are stated.) The cohort was studied on 21 occasions between birth and age 53 years. Information about sociodemographic factors and medical, cognitive, and psychological function was obtained by interview and examination at each point of contact. They concluded that birthweight and postnatal growth are independently associated with level of cognition at different ages. In this case, the main outcome of interest was level of cognition, and the antecedents were birthweight and postnatal growth. The outcome was repeatedly measured over the period so that the cognition achieved at different ages could be studied. Postnatal growth may be a function of birthweight, but as far as cognition was concerned, this study found that the two act independently.

Neary et al. [4] carried out a retrospective cohort study of all nonelective general and orthopedic surgical procedures performed on a total of 1869 patients in a hospital in the United Kingdom during the calendar year 2000. Outcomes were identified from various related hospital databases, and case notes of those who died were reviewed. Note that the study was from antecedent to outcome, but the subjects belonged to a past period. The study found that increasing age, size of operation, and American Society of Anesthesiologists (ASA) grade were significantly associated with higher risk of death by 1 year. The authors concluded that a simple scoring system could be used to identify high-risk patients among those who required nonelective surgery. Such patients could be targeted for interventions for reducing the risk of death. The conclusion reached is the same as anticipated by common sense. Yet the study has value, first for documenting the evidence and second by linking it to the scoring system. The study would hold greater value had the relative risk been quantified and if its confidence interval provided.


1. Indrayan A, Bagchi SC, Verma V. Medico-social factors contributory to dropouts in a rural cohort of oral contraceptors. *J Fam Welf* 1972;18:65–75. http://www.popline.org/node/480324
2. Sokal DC, Zipper J, Guzman-Serani R, Aldrich TE. Cancer risk among women sterilized with transcervical quinacrine hydrochloride pellets, 1977 to 1991. *Fertil Steril* 1995;64:325–34. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.305.826&rep=rep1&type=pdf
3. Richards M, Hardy R, Kuh D, Wadsworth MEJ. Birthweight, postnatal growth and cognitive function in a national UK birth cohort. *Int J Epidemiol* 2002;31:342–8. http://ije.oxfordjournals.org/content/31/2/342.full.pdf+html
4. Neary WD, Foy C, Heather BP, Earnshaw JJ. Identifying high-risk patients undergoing urgent and emergency surgery. *Ann R Coll Surg Engl* 2006;88:151–6. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1964060/


## collinearity, *see* multicollinearity

## communality of a variable

In a **multivariate** setup, communality is the proportion of variance of a particular variable $x_k$ that is shared with other $x$'s. Note the social overtones of such sharing, hence the term communality. This is frequently used in the context of **factor analysis** to measure how much variables share with the extracted factors. The basic premise of factor analysis is that various variables have one or more common factors. Thus, sharing is built in into the concept of factor analysis. High communality of several $x$'s indicates that the factor analysis would be successful. The remainder variance, after subtraction of the communality, is considered unique to $x_k$ as it is the part not shared with the other variables.

In factor analysis, the calculation of communality is based on what is called **factor loadings** of the variables. This is the correlation between the variable and the factor. Loading can be negative since this is correlation. Its square is the percent variance in that variable explained by the concerned factor. If loading of the $k$th variable for the $j$th factor is $l_{jk}$, the communality of the $k$th variable is

$$\text{communality of } x_k\text{: } c_k = \Sigma_j l_{jk}^2.$$

This can be understood as the square of the **multiple correlation** coefficient ($R^2$) for the regression model predicting $x_k$ from the extracted factors. This is the proportion of variance of a variable explained by all the factors together. Factor loading is calculated for each factor and each variable, whereas communality is calculated for the variables when the factors are combined.

The results of a factor analysis of an 11-point questionnaire on family burden of a particular disease, for example, are given in Table C.25. Factor analysis extracted three factors. For variable $x_1$ (loss of earnings), communality $c_1 = (0.87)^2 + (0.07)^2 + (0.11)^2 = 0.7739$. This is high relative to, say, variable $x_9$ (dependence on

**TABLE C.25**
**Results of a Factor Analysis of an 11-Point Questionnaire on Family Burden of a Disease**

| | Loadings | | | |
|---|---|---|---|---|
| Variable | Factor 1—Economic | Factor 2—Social | Factor 3—Mental | Communality (Sum of Squares of Loadings) |
| $x_1$ Loss of earnings | 0.87 | 0.07 | 0.11 | 0.7739 |
| $x_2$ Hospital expenses | 0.75 | 0.15 | 0.23 | 0.6379 |
| $x_3$ Depression | 0.01 | 0.43 | 0.88 | 0.9594 |
| $x_4$ Disruption of life | 0.22 | 0.58 | 0.41 | 0.5529 |
| $x_5$ Desertion by friends | 0.05 | 0.92 | 0.12 | 0.8633 |
| $x_6$ No recreation | 0.02 | 0.69 | 0.38 | 0.6209 |
| $x_7$ Helplessness | 0.14 | 0.56 | 0.69 | 0.8093 |
| $x_8$ Mental agony | 0.08 | 0.24 | 0.97 | 1.0049 |
| $x_9$ Dependence on others | 0.20 | 0.24 | 0.32 | 0.2000 |
| $x_{10}$ Family secluded | 0.12 | 0.85 | 0.73 | 1.2698 |
| $x_{11}$ Loans | 0.97 | 0.23 | 0.18 | 1.0262 |

others) whose communality $c_9 = (0.20)^2 + (0.24)^2 + (0.32)^2 = 0.2000$. In this example, loss of earning is explained best by the factors, and depression is not explained well.

See **factor analysis** to further understand the application of this method. For other methods of calculating communalities, see Gorsuch [1].

1. Gorsuch RL. *Factor Analysis*. Psychology Press, 2013.

## comparative mortality ratio

Comparative mortality ratio is the ratio of two **death rates** with some common base. Common base means that both the rates should be per 1000 population, per 1000 births, per 1000 cases, etc. This is defined as

$$\text{comparative mortality ratio} = \frac{\text{death rate-1}}{\text{death rate-2}} * 100,$$

where the two death rates have some common base.

The term *comparative mortality ratio* is yet to achieve a common global meaning. One popular use of the comparative mortality ratio is in the ratio of expected number of deaths arrived by standardization to the actual deaths. If the **standardized death rate** is 5.5 per thousand population for a country where the **crude death rate** is 8.5 per thousand population, the comparative mortality ratio for this country is $100 \times 5.5/8.5 = 65\%$. For another country, this might be 138%, which would imply that the standardization has reverse effect.

The utility of the comparative mortality ratio increases when both the rates are standardized to the same standard structure. If this ratio for the United States compared to Venezuela is 95%, since both rates refer to the same standard population, this can be interpreted as saying that the death rate in the United States is 95% of that in Venezuela. Although both rates under comparison are standardized for the same age structure (thus comparable), the effect of the chosen standard is not eliminated by this comparison. If another standard age structure is chosen, the ratio can become very different.

The utility of the comparative mortality ratio increases further when several populations are compared. For example, if the US age-standardized heart disease death rate in white males is considered 100, the comparative mortality ratio for age-standardized death rate is more than 120 for many health service areas (HSAs) in the mid-southeast and less than 80 for many HSAs in the west. Thus, the comparative mortality ratio is one more method to know which segment of population has relatively higher or lower mortality. Strategies for controlling mortality can be accordingly devised.

An example of this is a study of site-specific cancer mortality in Greece. Alexopoulos et al. [1] found 3.5-fold variations in cancer mortality ratios among men across broad occupational groups. Such variation across occupations for the same cancer might stimulate thinking, and preventive actions can be taken regarding the job-related factors.

1. Alexopoulos EC, Messolora F, Tanagra D. Comparative mortality ratios of cancer among men in Greece across broad occupational groups. *Int Arch Occup Environ Health* 2011 Dec;84(8):943–9. http://www.ncbi.nlm.nih.gov/pubmed/21331610

## comparison groups

Comparison groups help to attach meaning to the findings where stand-alone may not carry much of a meaning. If you are told that the 5-year survival rate in patients of cancer of the esophagus of age 60–65 years is 32%, what do you make of this? This suddenly acquires significance when compared with, say, 85% survival in healthy subjects of corresponding age. This comparison might be inadvertent, but values acquire interpretation when put into context. Relativity such as high or low, more or less, better or worse, or even same is inherent in much of our thought process. An explicit comparison group is helpful in proper interpretation. This must be explicitly stated, particularly in a research setup, since different comparison groups can give different results.

A comparison group has special significance in clinical trials, particularly in phase III (see **phases of clinical trials** for details). Phase I is primarily to establish tolerability, and this does not require any comparison group. In phase II also, the objective might be to test the concept in the sense that the regimen has at least minimum **efficacy** for pursuing it further. In this case also, comparison group is not necessary, although this helps in confirming that the observed efficacy is not due to a **placebo** effect. In a rare case a regimen is developed for a disease with no existing treatment, the interest might

C

be in its absolute efficacy even in phase III. There are conditions for which no cure exists, but there is hardly any ailment for which no treatment is available. That treatment may have low efficacy is a different matter. Thus, the objective in almost all phase III trials is to establish that the regimen is better than the existing treatment. In this case, the comparison group is given the existing treatment. In the context of the usual clinical trials, the comparison group is called a parallel **control group**. In before–after experiments and trials, the comparison is with the preintervention values.

Ideally, the comparison group must be identical to the active group. In our example, the comparison of 32% 5-year survival in cancer cases with 87% in healthy controls is valid only when both relate to the same parent population. In a clinical trial, the control group must be similar to the active group except that the active group receives the regimen under examination and the control group receives nothing or placebo or the existing regimen. This brings in the question of **baseline equivalence** of the groups. Two popular methods to achieve this equivalence are **randomization** and **matching**. Randomization is used when the group sizes are reasonably large and ethically justified, and matching is generally used for small groups or when randomization is not feasible.

## comparison of intercepts in two or more simple linear regressions

For simplicity, we restrict ourselves to only **simple linear regression** in this section where the **intercept** is the value of the dependent variable $y$ at $x = 0$. The intercept in this setup is conventionally denoted by $\alpha$ for the population and by $a$ for the sample. It is important in some health setups such as the bilirubin level at age = 0, i.e., at the time of birth. You might be running a simple linear regression of bilirubin ($y$) on age ($x$) for male and female singleton children in a well-to-do segment of a population, and may be interested in comparing the bilirubin level at birth in the two sexes based on subsequent trend. Statistically, this is the same as comparing the intercepts, which really means testing the **null hypothesis** $H_0$: $\alpha_1 = \alpha_2$, where the subscripts are for group 1 and group 2, respectively.

In our example, the group of male children is independent of the group of female children since they do not, in any way, affect the bilirubin level of the other group. Any comparison between the independent groups is easier relative to the comparison between the dependent groups. Dependence arises when, for example, the comparison is of the intercept of the regression of bilirubin on age in male children with this intercept in female children in the same family. In this case, you can see that bilirubin and albumin levels in the same family would be correlated—thus, the comparison requires this correlation to be factored. This is a rare problem and is not discussed further. It is mentioned here so that you become aware that this can happen.

### *Comparison of Two Intercepts in Independent Groups*

Consider the intercepts of the simple linear regression of bilirubin level on age in male and female children not belonging to the same family. These two are independent groups. For simplicity, we assume that the slopes of the two regressions are the same. In this situation, the null hypothesis $H_0$: $\alpha_1 = \alpha_2$ can be tested by setting up a joint model of the type $y = \alpha + \beta z + \beta x$, where $z$ is the indicator variable for sex, taking value 0 for males and 1 for females. Thus, the intercept for males is $\alpha$ and that for females is $(\alpha + \beta)$. Statistical software can be asked to test $H_0$: $\beta = 0$ in the regression equation just mentioned, which is the same as equal intercepts.

In case you want to do it yourself or want to learn what goes behind the computer calculations, here are the two solutions. An approximate solution that might work well for large samples is to find the **standard errors** (SEs) of the respective intercept estimates, and use the fact for independent samples that estimated SE $(a_1 - a_2) = \sqrt{\text{SE}_{a_1}^2 + \text{SE}_{a_2}^2}$, where estimated $\text{SE}_a^2 = \text{MSE}\left[\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x-\bar{x})^2}\right]$ for simple linear regression. This is calculated separately for the intercepts $a_1$ and $a_2$ in the two regression equations. When the distribution of the dependent $y$ (or the residuals) is Gaussian, the test statistic as usual is

$$\text{Student } t = \frac{a_1 - a_2}{\text{estimated SE}(a_1 - a_2)}$$

with df = $(n_1 + n_2 - 4)$ for the null that the two intercepts are equal ($H_0$: $\alpha_1 = \alpha_2$). The df actually is $(n_1 - 2) + (n_2 - 2)$. For large samples, the **central limit theorem** can be invoked for non-Gaussian distributions. Reject the null in favor of the two-sided alternative at 5% level of significance when $|t|$ > table value of $t$ at the specified df. If the null is $\alpha_1 - \alpha_2 = \delta$ (that is, the intercepts differ by at least some medically important quantity $\delta$), the test statistic modifies to $t = (a_1 - a_2 - \delta)/\text{SE}(a_1 - a_2)$.

The test described in the preceding paragraph can be improved if the residuals for the two groups under comparison have nearly the same variance. This is implicit in the case of the previous test also. When variances are equal, you can improve its estimate by calculating

$$\text{pooled estimate of the residual variance: } s_{y.x}^2 = \frac{\text{SSE}_1 + \text{SSE}_2}{n_1 + n_2 - 4},$$

where SSEs are the error sum of squares of the respective regressions. This is the new MSE for this setup. When this is used, the

$$\text{pooled estimate of SE } (a_1 - a_2) = s_{y.x}^2 \left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{\bar{x}_1^2}{\text{SS}(x_1)} + \frac{\bar{x}_2^2}{\text{SS}(x_2)}\right],$$

where SSs are the sum of squares due to $x_1$ and $x_2$, respectively. All these sums of squares come from the regression analysis. Under Gaussian conditions, $t = (a_1 - a_2)/[\text{estimated SE}(a_1 - a_2)]$ has a Student $t$ distribution with df = $(n_1 + n_2 - 4)$ under the null that the two intercepts are equal ($H_0$: $\alpha_1 = \alpha_2$). Reject the null if the ***P*-value** corresponding to the calculated value of $t$ is less than the predetermined $\alpha$ **level of significance**; otherwise, do not reject the null. All this is essentially the same as described in the preceding paragraph except that we have an improved estimate of the residual variance. In place of separate MSEs of the two individual regressions, we now have one MSE, denoted by $s_{y.x}^2$, from the combined sample.

Another approach could be to find the $100(1 - \alpha)\%$ **confidence intervals** (**CIs**) for the two intercepts and see if these CIs overlap. If the CIs overlap, the difference between the intercepts is not statistically significant at $\alpha$ level of significance; otherwise, the difference is significant. For these CIs, the pooled estimate of the SE, as just stated, is used.

## comparison of one sample mean and proportion with a prespecified value

The question of comparison of one sample with a prespecified value arises when we want to assess if the sample values agree or disagree

with some known value or a hypothesized value of a **parameter**. The parameters considered in this section are either mean or proportion, which incidentally are also the most commonly used parameters. Mean is used when the data are quantitative and the proportion when they are qualitative.

### *Comparison with a Prespecified Mean*

Let the interest be in finding whether patients with chronic diarrhea have the same average hemoglobin (Hb) level as normally seen in healthy subjects in the area. Suppose the mean normal level of Hb is 14.6 g/dL. This is assumed to be known and fixed for the present example. Since chronic diarrhea can only decrease the Hb level, and not increase it, it is a one-tailed situation. In an unlikely event of sample mean being >14.6 g/dL, the evidence is immediate that Hb level does not decrease in chronic diarrhea patients, and there is no need to proceed further. Although a higher mean can occur by chance in the sample when the population mean is less than 14.6 g/dL, let us keep that aside for the moment.

Suppose further that a random sample of 10 patients with chronic diarrhea is investigated, and the average Hb level is found to be 13.8 g/dL. Thus, the sample mean is lower than normal. This would occur if the sample happens to comprise subjects with a lower level. Such subjects are not uncommon in the healthy population as well. If another sample of 10 patients is studied, the average could well be 14.8 g/dL. Can it be concluded with reasonable confidence on the basis of the previous sample that patients with chronic diarrhea indeed have a lower Hb level on average?

There is only one sample in this example, and the comparison is with the known average in the healthy subjects. It is a one-sample problem, although the comparison is of two means—one observed in the sample and the other known for the healthy population.

The **null hypothesis** in the preceding example is $H_0$: $\mu = 14.6$ g/dL. Since the possibility of a higher average Hb level in patients with chronic diarrhea is ruled out, the alternative hypothesis is one-sided. That is, $H_1$: $\mu < 14.6$ g/dL. If $H_0$ is rejected, then $H_1$ is considered true.

The first step is to choose an appropriate criterion to test the hypothesis. The value of this criterion is then calculated assuming that $H_0$ is true. Then the probability of the observed or a more extreme value is obtained. This is the ***P*-value**. If this probability is very small, $H_0$ is considered not plausible and rejected. The conclusion then reached is that $H_1$ is true. If the *P*-value is not sufficiently small, say not less than 0.05, the null hypothesis is conceded. As mentioned in **tests of hypothesis (philosophy of)**, this does not imply that $H_0$ is accepted. It is just that it cannot be rejected on the basis of that sample because **sampling fluctuation** is not adequately ruled out as a likely explanation.

Heuristically, the answer depends on the magnitude of the difference between the sample mean and the known mean of the healthy subjects. In the preceding example, this difference is 13.8 – 14.6 = –0.8 g/dL. This magnitude is assessed relative to the expected variation in means from sample to sample. This variation is measured by the **standard error (SE)** of the mean, $\sigma/\sqrt{n}$. In a rare case when $\sigma$ is known, the criterion for testing this hypothesis is

$$\text{Gaussian test: } z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

This follows a **Gaussian distribution** provided that the underlying distribution of *x*'s is also Gaussian. The value of this criterion is used to find if *P*-value is sufficiently small.

In practice, the standard deviation (SD) $\sigma$ would be rarely known and is replaced by its sample estimate *s*. This converts it to **Student *t***. Thus, the criterion for this setup is

$$\text{Student } t\text{-test (one sample): } t_{n-1} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

where $\mu_0$ is the value of the mean under $H_0$. This criterion is valid under mild conditions. One of the conditions is that the observations are independent. In the context of this example, this means that the Hb level in one subject should not have any influence on the level in other subjects. This is clearly satisfied in this example, but there are situations where independence is violated. The other condition is that the distribution of *x*'s themselves is Gaussian, which also implies that the distribution of sample mean $\bar{x}$ is Gaussian. Because of the **central limit theorem**, this condition can be relaxed if sample size is large. The higher the value of *t*, the greater the chance that the sample has not come from a population with mean $\mu_0$. In that case, the decision to reject $H_0$ would have less chance of error. When this probability of **Type I error** is low, say less than 0.05, $H_0$ can be safely rejected. The exact *P*-value for a particular value of *t* is provided by most of the standard statistical packages. As mentioned under **Student *t***, the distribution of *t* depends on the degrees of freedom, df. For *t* in the formula just given, df = $(n - 1)$. This is specified as a subscript of *t*.

Consider our example of the Hb level in chronic diarrhea patients. Suppose that in a random sample of size $n = 10$, the levels in grams per deciliter are as follows:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 11.5 | 12.2 | 14.9 | 14.0 | 15.4 | 13.8 | 15.0 | 11.2 | 16.1 | 13.9 |

These give mean $\bar{x} = 13.8$ g/dL and SD $s = 1.67$ g/dL.

The hypothesis under test is that the average Hb level in the patients with chronic diarrhea is the same 14.6 g/dL that is normal in healthy subjects. Thus, $H_0$: $\mu = 14.6$ g/dL. The alternative as already explained is $H_1$: $\mu < 14.6$ g/dL. In this case, under $H_0$,

$$t_9 = \frac{13.8 - 14.6}{1.67/\sqrt{10}} = -1.51.$$

A statistical package gives $P(t < -1.51) = 0.0827$. This is the probability of getting the sample mean this much or more extreme in favor of $H_1$. Since this $H_1$ is one-sided, the probability required is also one-tailed. In this case, *P* is more than 0.05. Thus, the chance is more than 5% that the sample has come from a population with mean 14.6 g/dL. Therefore, this $H_0$ cannot be rejected. The difference between the sample mean 13.8 g/dL and the population normal 14.6 g/dL is not statistically significant. This sample does not provide sufficient evidence to conclude that the mean Hb level in chronic diarrhea patients is less than normal.

The conclusion in this example is partly the result of the high variability in the Hb level in the sample patients. Whereas it was only 11.2 g/dL in one patient, it was 16.1 g/dL in another. Widely scattered values gave a high value of sample SD *s* and led to the expectation of high intersample variability. Consequently, it became difficult to say anything definite about the lower mean Hb in these patients.

### *Comparison with a Prespecified Proportion*

Now consider qualitative data where the interest is in proportion. Suppose that a pharmaceutical company claims that its particular

drug has at least 72% efficacy in the long run. Thus, the null is $H_0$: $\pi = 0.72$. To test this claim, you try the drug on a random sample of $n = 40$ patients and 28 respond positively. Thus, the efficacy is 70%. Now the question is, can population proportion still be at least 72%? If not, the inference would be that it is less than 72%. The null would be rejected in favor of this alternative. Thus, the alternative hypothesis is $H_1$: $\pi < 0.72$. Inference on any **binary variable** such as this is always based on **binomial distribution**, but a Gaussian approximation can be used when the sample size is large. For small $n$, see **exact tests** based on binomial distribution described for the **McNemar test**.

How much $n$ is large? For small $\pi$, we need really large $n$, and for $\pi$ around 0.5 not as large. The rule generally followed for being able to use the Gaussian approximation is $n\pi \geq 8$. This implies that $n$ should be at least 400 for $\pi = 0.02$ (2%) and should be just about 20 for $\pi = 0.4$. When $n$ is this large, the criterion to test the null $H_0$: $\pi = \pi_0$ is

$$\text{large sample test for one sample proportion: } z = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}.$$

This is referred to the Gaussian distribution to find the $P$-value. In our example,

$$z = \frac{0.70 - 0.72}{\sqrt{0.72 \times 0.28/40}} = -0.28\,.$$

Since the alternative is of the "less than" type, the $P$ is the probability in the left tail (i.e., $P(z < -0.28)$). This gives $P = 0.39$ for a Gaussian distribution. Since this is much more than the conventional level 0.05, the null cannot be rejected. Conclude that the efficacy could be (not that it is) 72%.

## comparison of one sample proportion with a prespecified value, *see* comparison of one sample mean and proportion with a prespecified value

## comparison of two or more correlation coefficients

We restrict this section to comparing **Pearsonian correlation** coefficients. This is also what is most commonly required in practice. Other kinds of correlations such as Spearman and point biserial are excluded from the present discussion. You may be aware that Pearsonian correlation coefficient is a measure of only the **linear** relation. When the relationship is nonlinear, only the linear component is considered by this correlation coefficient.

The problem of comparing two or more correlations can arise in two different setups:

1. Comparing the correlations between the same variables in two independent groups such as the correlation between kidney size and height in males compared with the same correlation in females. In this setup, males and females are independent groups since the values in one group do not alter the values in the other group.
2. Comparing different correlations in the same group such as comparing the correlation between kidney size and height with the correlation between creatinine level and weight when both the correlations are obtained for the same group of male subjects. In this setup, since the values belong to the same group of subjects, the two correlations are not strictly independent. In this example, an overlap can also occur when the two correlations share a common variable, such as kidney size appearing in both the correlations. This kind of dependence gives rise to further statistical problems.

Both these setups are discussed below.

### *Comparing Correlations in Two or More Independent Groups*

The distribution of the Pearsonian correlation coefficient ($r$) in general is not Gaussian. However, for large samples, the **central limit theorem** (CLT) may help. You may have noticed that the formula for calculating $r$ has *sum* of cross-products in the numerator. Because of this sum, the central limit theorem is applicable, and we can say that the distribution of the sample correlation coefficient follows a **Gaussian distribution** for large $n$ even when the distributions of the variables $x$ and $y$ are not Gaussian. Under these conditions, one can think of using the **Student *t*-test** for comparing the two independent correlations. For this test, an estimate of the **standard error (SE)** of $r$ is needed for calculating the SE of the difference $r_1 - r_2$. SE of $r$ is a complex expression except when the population correlation coefficient $\rho = 0$ (in this special case, estimated $\text{SE}(r) \approx \sqrt{1-r^2}/\sqrt{n-2}$ ). Also, this $t$ requires comparatively very large $n$ for CLT to be effective. Thus, for comparing two independent Pearsonian correlations, it is customary to use **Fisher *z*-transformation**. This is given as

$$\text{Fisher } z\text{-transformation: } z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right),$$

$$\text{and estimated SE}(z) = \sqrt{\frac{1}{n-3}}\,.$$

This transformation was proposed by Fisher in 1921 and was shown to follow an approximate Gaussian distribution even when $n$ is not very large [1]. This also works well for relatively small samples but requires that at least one of the variables follows a Gaussian distribution.

For comparing two independent correlations, under the null hypothesis that the two correlations in the populations are equal ($H_0$: $\rho_1 = \rho_2$), calculate

$$z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}\,.$$

Reject the null at 5% **level of significance** when the calculated value of $|z| > 1.96$. Since the estimated SE is being used, this, in fact, should be Student $t$ with $\text{df} = (n_1 + n_2 - 6)$ but is still considered Gaussian $z$ because both are equivalent for large samples.

There are cautions, though. (i) Correlations in two groups are comparable (just as almost any other sample summary) when the two groups follow the same method of measuring the variables, same range restriction, etc. (ii) Make a distinction between a correlation coefficient and a **regression coefficient** (slope). In one setup, the variable $y$ may increase on average by 5 units for each unit increase in $x$ and in the other setup by only 2 units, yet both can have equal correlations (Figure C.18b). Because of higher scatter, solid dots have lower correlation in Figure C.18a compared with hollow squares, but both have nearly the same regression coefficient.

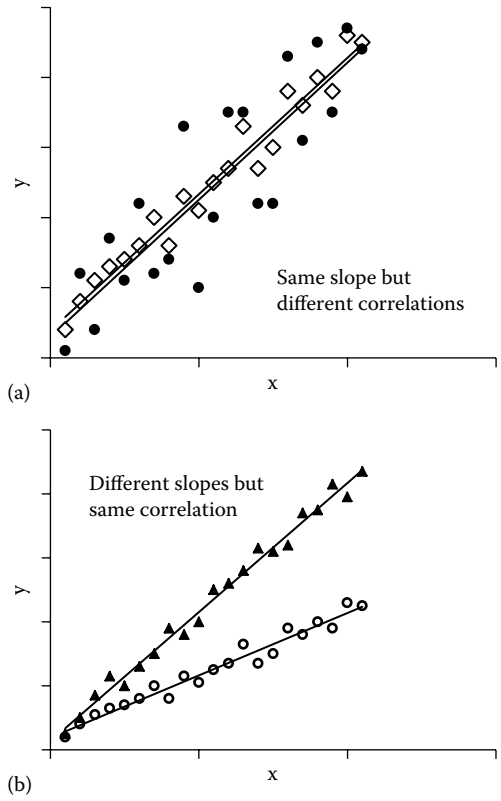



**FIGURE C.18** (a) Nearly same slope but different correlations. (b) Different slopes but nearly same correlations.

In both these figures, the intercept is the same but that also can be different.

The procedure for comparing more than two correlations is not straightforward. The details are provided by Fleiss [2] in the context of **meta-analysis** where the researcher would like to know that the correlations reported in different studies are homogeneous. A relatively simple procedure for large $n$ in each sample is as follows. Suppose you have $K$ independent samples and the correlation coefficient in the $k$th ($k = 1, 2, \ldots, K$) sample is $r_k$. The test criterion is

$$Q = \sum_{k=1}^{K} W_k (z_k - \bar{z})^2,$$

where $z_k = \frac{1}{2}\log[(1 + r_k)/(1 - r_k)]$, $W_k$ is the reciprocal of the variance of $z_k$, i.e., $W_k = (n_k - 3)$ since $1/(n_k - 3)$ is the variance of $z_k$, and $\bar{z} = (\Sigma_k W_k z_k)/(\Sigma_k W_k)$. This is the **weighted average** of the Fisher $z$-transformations of the values of the correlation coefficients. Under the null hypothesis of equality of correlations, $Q$ follows a chi-square distribution with df = $(K - 1)$. This is a fairly general procedure that can be applied for comparing any parameter in $K$ independent samples whose estimate follows a Gaussian distribution for large $n$. That is, you can have another summary in place of the correlation coefficient. Use the chi-square distribution to find the ***P*-value** corresponding to the calculated value of $Q$, and reject the null if the $P$-value is less than the prespecified level of significance.

If a statistical package is used to do these calculations, examine if the facility for such comparison is available in the software. Most packages as of now do not have direct facility to do these tests. You may have to write the commands or programming codes with the assistance of a biostatistician, or search the Internet for such commands prepared by someone else. See Weaver and Wuensch [3] for more details.

### *Comparing Two or More Related Correlations*

The problem of comparing two correlated correlations is not simple. Correlated correlations can arise in a variety of setup. First are the correlations $r_{xy}$ and $r_{xz}$ where $x$ is the common variable, and both are obtained from the same set of data. This comparison will tell whether $x$ is related more with $y$ than with $z$. Second are the correlations among nonoverlapping variables such as comparing $r_{xy}$ with $r_{uv}$. When calculated from the same data, these correlations between ostensibly unrelated variables may also be related. Comparison of correlated correlations in both these setups leads to complex test criteria. Weaver and Wuensch [3] have provided the formulas and also examples. Further details are available in Raghunathan et al. [4] and Meng et al. [5].

### *Test of Significance of One Sample Correlation Coefficient*

The $z$-transformation can also be used to test the null $\rho = \rho_0$, rejection of which would indicate that the correlation coefficient is statistically significantly different from $\rho_0$. For this, calculate $z_0 = z_0 = \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)$ and

$$z\text{-test for one-sample correlation coefficient } (H_0\colon \rho = \rho_0)\colon\ z = \frac{z - z_0}{\sqrt{\frac{1}{n-3}}},$$

provided $n$ is large and at least one of $x$ and $y$ has a Gaussian distribution. In case the objective is to find whether a particular correlation coefficient observed in the sample is significant or not (this is the same as testing the null $\rho = 0$), this changes to

$$\text{The Student } t\text{-test for one sample correlation coefficient } (H_0\colon \rho = 0)\colon\ t = r\sqrt{1 - r^2}/\sqrt{n-2},$$

with df = $(n - 2)$. This does not require a large sample but requires Gaussian distribution of the values. If the $P$-value from this is less than the level of significance, conclude that the correlation coefficient is statistically significant.


1. Fisher RA. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* 1921;1;3–32. https://digital.library.adelaide.edu.au/dspace/bitstream/2440/15169/1/14.pdf
2. Fleiss JL. The statistical basis of meta-analysis. *Statistical Methods Med Res* 1993;2:121–45. http://www.ncbi.nlm.nih.gov/pubmed/8261254
3. Weaver B, Wuensch KL. SPSS and SAS programs for comparing Pearson correlations and OLS regression coefficients. *Behavior Research Methods* 2013;45(3):880–95. http://link.springer.com/article/10.3758/s13428-012-0289-7
4. Raghunathan TE, Rosenthal R, Rubin DB. Comparing correlated but nonoverlapping correlations. *Psychol Methods* 1996;1(2):178–83. http://psycnet.apa.org/index.cfm?fa=buy.optionToBuy&id=1996-04469-006
5. Meng X, Rosenthal R, Rubin DB. Comparing correlated correlation coefficients. *Psychol Bull* 1992;111(1):172–5. http://psycnet.apa.org/psycinfo/1992-15158-001

## comparison of two or more means, *see* Student *t*-test, analysis of variance (ANOVA)

## comparison of two or more medians, *see* Wilcoxon signed-ranks test, Kruskal–Wallis test

C

## comparison of two or more odds ratios

Two kinds of comparison can arise in case of **odds ratios**. First is when the **logistic regression** is run separately for two groups such as for males and females with the same dependent and regressors, and you want to compare the odds ratio (OR) in one group with the other for any particular regressor. For example, the dependent can be logit of probability of death within 5 years with stage of cancer as one of the regressors, and this regression is run separately for males and females. In this situation, the two groups are independent and

$$\text{SE}[\ln \text{OR}_1 - \ln \text{OR}_2] = \sqrt{[\text{SE}(\ln\text{OR}_1)]^2 + \text{SE}[(\ln\text{OR}_2)]^2}.$$

The SEs on the right-hand side of the equation are **standard errors** of $\ln(\text{OR}_1)$ in the first group and of $\ln(\text{OR}_2)$ in the second group. For large samples, these SEs can be obtained by the formula given for **confidence interval for odds ratio**. Since OR is a ratio, it does not directly follow a Gaussian distribution—logarithm makes it linear and allows you to exploit the premise of the **central limit theorem**. Thus, lnOR becomes Gaussian for large samples, so does the difference in lnORs. Thus, the test criterion for testing the null $\text{H}_0$: $[\ln\text{OR}_1 - \ln\text{OR}_2] = \delta$ is

$$z = \frac{[\ln\text{OR}_1 - \ln\text{OR}_2] - \delta}{\text{SE}[\ln\text{OR}_1 - \ln\text{OR}_2]}.$$

For equality of ORs, $\text{H}_0$: $\delta = 0$ that would be rejected if the calculated value of $z$ for the available samples is less than the threshold at the specified level of significance. For 5% level of significance, this threshold for a **two-tailed test** is 1.96. Note that this test is for lnOR and not for OR itself. Also, this is valid only for large samples. In any case, the concept of OR is hardly ever applicable to small samples. This test presumes that the two groups have been studied with similar methodology and similar care, and there is no factor other than the group (sex in our example) that can affect the ORs. The procedure is extendable to more than two groups provided they are independent and each has a large sample.

The second setup is when you want to compare the OR of one regressor with the OR of another regressor in the same regression equation. If the logistic regression is

$$\text{logit}(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \varepsilon,$$

the null hypothesis now is $\text{H}_0$: $\beta_1 = \beta_2$. These are the effects of $x_1$ and $x_2$, respectively, on logit($\pi$). Thus, this null tests whether the effect of the regressor $x_1$ is the same as of $x_2$ in the presence of the other/s when both $x_1$ and $x_2$ are binary. In our example, we can divide stages of cancer into two categories such as stage I/II and stage III/IV, and age as <70 and ≥70 years, and the question to be answered is whether age has more influence on 5-year survival or the stage has more influence. Since both $x_1$ and $x_2$ are in the model, the question actually is about the effect of cancer stage when age is fixed, and the effect of age when stage is fixed. This setup is too complex and is explained by Hosmer et al. [1].

Another question is regarding the additional effect of $x_2$ when $x_1$ is present. The measurement of effect is in terms of OR in the present context and is called sequential OR. This question is not really a comparison of two ORs but is related. In our example, this is the effect of age on mortality when we know the cancer stage. Statistical significance of this can be tested by calculating the difference in $-2\ln L$ under the model with $x_1$, $x_2$ and with only $x_1$, and checking the additional contribution with chi-square at 1 df for large $n$. This procedure can be extended to find the statistical significance of combined effect for any number of regressors in the presence of (or adjusted for) other regressors.

1. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*, Third Edition. Wiley, 2013.

## comparison of two or more proportions

Under **Gaussian conditions**, the difference between proportions $p_1$ and $p_2$ in two independent groups is statistically assessed by

$$\text{Gaussian test: } z = \frac{|p_1 - p_2|}{\sqrt{p(1-p)(1/n_1 + 1/n_2)}},$$

where $n_1$ and $n_2$ are the respective sample sizes and $p = (n_1p_1 + n_2p_2)/(n_1 + n_2)$. This is the combined proportion in the two samples. This combination is applicable in this case since under the **null hypothesis**, the two proportions are the same ($\text{H}_0$: $\pi_1 = \pi_2$) and the test criterion is obtained under this null. The denominator is the estimated **standard error (SE)** of the numerator in the case of independent samples. Because of *estimated* SE in the denominator, this actually is **Student *t***, but for large sample sizes where this is applicable, Student $t$ is almost the same as Gaussian $z$ and it is customary to use Gaussian $z$. The value of $z$ is referred to a Gaussian distribution, and a two-sided ***P*-value** is obtained.

For the data in Table C.26, for comparing prevalence of anemia in parity ≤ 2 with the prevalence in parity ≥ 3,

$$\cdots z = \frac{|14/60 - 16/40|}{\sqrt{30/100(1-30/100)(1/60+1/40)}} = 1.78.$$

Corresponding to this value of $z$, two-sided $P = 2 \times 0.0375 = 0.075$ when both negative and positive sides are considered, i.e., parity ≤ 2 may have more anemia or may have less anemia than parity ≥ 3. Since the $P$-value is not less than 0.05, the null cannot be rejected.

For large $n$, this procedure should give the same $P$-value as obtained by the **chi-square** criterion. In fact, there is a theoretical relationship saying that $z^2 = \chi^2$ with 1 df. This is called the equivalence of the 1 df $\chi^2$ to $z$. An advantage of the $z$-criterion is that $\text{H}_0$ can be directly tested against the **one-tailed alternative**. Then the $P$-value is obtained for one tail only. This would be twice the two-tailed probability if the one-tailed probability is not directly available. In case of chi-square, the $P$-value obtained would be two-tailed.

**TABLE C.26**
**Anemia and Parity in a Cross-Sectional Study of 100 Women**

| | Observed in the Survey | | |
|---|---|---|---|
| **Anemia** | **Parity ≤ 2** | **Parity ≥ 3** | **Total** |
| Present | 14 | 16 | **30** |
| Absent | 46 | 24 | **70** |
| **Total** | **60** | **40** | **100** |

Most software packages require only the entry of data in a specific format and a command to compute chi-square. *P*-values with and without **Yates correction** will be automatically obtained. Many packages issue a warning in case *n* or **expected cell frequencies** are small or automatically compute the **Fisher exact test** in this case. The purpose of providing formulas here is to explain the underlying principles. Many users these days would not actually use these formulas and would not manually calculate the value of $\chi^2$ or $z$ or, for that matter, any other given in this book.

### *Comparing Proportions in More Than Two Populations*

Suppose in Table C.26, parity is measured in five groups as 0, 1, 2, 3, and 4+. If the null hypothesis is that the proportion of anemic women is the same in each parity, this can be tested by the usual chi-square. The hypothesis of equality of proportions is the same as of independence. When we say that the proportions are the same in each parity, it implies that the proportion anemic is not affected by parity. However, keep in mind that the alternative hypothesis for the usual chi-square is that the proportions are not the same—there is some difference somewhere. If the alternative is some kind of trend that says that the proportion anemic increases as the parity increases, another test called **chi-square for trend** should be used. This question will arise only for ordinal categories such as parity in our example and not for nominal categories.

## comparison of two or more regression coefficients

You may be aware that the **regression coefficient** in simple linear regression measures the slope or gradient of the linear relationship. This defines the steepness. For example, the relationship between age and blood pressure is generally steeper in females than in males. That is, rise in BP with age is faster in female adults compared with males. This is one example where the comparison of the regression coefficients can provide useful information. We cite some other examples later in this section.

Statistically, the null hypothesis under consideration is $H_0$: $\beta_1 = \beta_2$, where β's are the two regression coefficients in two regression lines—one for each group. Their sample estimates are denoted by $b_1$ and $b_2$. Note that the subscripts here are for the groups and not for the regressors. Thus, these $\beta_1$, $\beta_2$ and $b_1$, $b_2$ have different meanings from what you see in multiple regressions.

The problem of such a comparison can arise also when the regression coefficients are for different variables within the same group. An example of the latter is comparing the regression coefficient of carbohydrate intake on total cholesterol level in a group of obese females and the regression coefficient of fat intake on total cholesterol level in the same group of subjects. You can see that this comparison would be valid only when the regression coefficients are **standardized**. This comparison would tell that cholesterol level is affected more by carbohydrate intake or by fat intake in obese subjects when nothing else is considered.

Another setup could be comparing the regression coefficient of carbohydrate intake on total cholesterol level with the regression coefficient of carbohydrate level on HDL cholesterol. When standardized, this comparison will tell whether the total cholesterol is affected more by carbohydrate intake or the HDL cholesterol is affected more. One can think of other setups as well. In these latter setups, the values of the regression coefficients could be related. These setups are complex and we do not discuss them any further. One method for this kind of setups is given by Clogg et al. [1].

For simplicity, we present the method of comparison for two groups. The method is extendable to more than two groups—only that the notations become messy.

### *Comparing Two or More Regression Coefficients in Independent Groups*

A simple but approximate method for independent groups is as follows. Obtain the estimated **standard errors (SEs)** of the sample regression coefficients by using the formula $\text{SE}_b^2 = \dfrac{\text{MSE}}{\Sigma(x-\bar{x})^2}$, where MSE is the mean square error (see **mean square in ANOVA**). This is obtained separately for the two groups. Then calculate the estimated SE of the difference in sample regression coefficients as $\text{SE}(b_1 - b_2) = \sqrt{\text{SE}_{b1}^2 + \text{SE}_{b2}^2}$. Now the test criterion is

$$t = \frac{b1 - b2}{\text{estimated SE}(b1 - b2)}.$$

This follows the **Student *t*** distribution with df = $(n_1 - 2) + (n_2 - 2) = n_1 + n_2 - 4$ under the **Gaussian conditions** and the null that the two regression coefficients are equal.

The method just mentioned is approximate since it does not use combination of the samples from both the groups for estimation of the SE. These two samples can be combined if the two groups have nearly the same variance of the **residuals**. In this case, an improved estimate of the MSE is given by

$$\text{pooled estimate of the residual variance: } s_{y.x}^2 = \frac{\text{SSE}_1 + \text{SSE}_2}{n_1 + n_2 - 4},$$

where SSEs are the **error sum of squares** of the respective regressions. When this is used,

$$\text{estimated SE}^2\,(b_1 - b_2) = s_{y.x}^2 \left[\frac{1}{\text{SS}(x_1)} + \frac{1}{\text{SS}(x_2)}\right],$$

where SSs are the sum of squares of $x$'s in the two regression equations. The test criterion is $t = (b_1 - b_2)/[\text{estimated SE}(b_1 - b_2)]$ with df = $(n_1 + n_2 - 4)$ as before under the Gaussian conditions and under the null hypothesis of equality of the regression coefficients.

Another approach for this setup could be to find the confidence intervals (CIs) for each of the two regression coefficients at $100(1 - \alpha)\%$ confidence level, where α is the level of significance, and see if these CIs overlap. If they do, the difference is not statistically significant; if not, the regression coefficients are significantly different. The method can be extended to regression coefficients in more than two groups.

1. Clogg CC, Petkova E, Haritou A. Statistical methods for comparing regression coefficients between models. *Am J Sociol* 1995; 100(5)1261–93. http://www.jstor.org/discover/10.2307/2782277?uid=3738256&uid=2&uid=4&sid=21106058631561

## comparison of two or more regression lines

Quite often the objective is to compare the trends in two groups such as treatment and control groups or males and females. Most statistical software packages have still not been given intelligence to do this comparison directly. You may be required to specify the right model or use a modified procedure. Among various ways that this comparison can be done, the following illustrates one for comparing **simple linear regressions**. This procedure compares both the **intercepts** and the **slopes** at the same time.

C

Define an **indicator variable** $z$ and assign it a value 0 for group I and a value 1 for group II. The variable has no meaning except to indicate the group. For this reason, this is also called a **dummy variable**. Remind yourself that in the case of regression, $x$'s are considered fixed and they can have any appropriate value. Thus, a regressor variable of the type of $z$ is admissible. Now fit the following regression model to the data. The data should have values of $x$ and $y$ as before, as well as the values of $z$ as assigned above.

$$y = b_0 + b_1x + b_2z + b_3xz.$$

You can easily see that for group I, since $z = 0$, the model is

$$y = b_0 + b_1x,$$

and for group II, since $z = 1$, the model is

$$\begin{aligned} y &= b_0 + b_1x + b_2 + b_3x \\ &= a + bx, \end{aligned}$$

where $a = (b_0 + b_2)$, and $b = (b_1 + b_3)$.

Compare the latter two equations and note that $\beta_3 = 0$ implies same slope but different intercepts, i.e., the lines in the two groups are parallel, and $\beta_2 = 0$ implies same intercept in the two groups but different slopes; $\beta_2 = 0$ and $\beta_3 = 0$ together imply that the regression lines in the two groups are identical.

The test of these hypotheses can be done as usual for **regression coefficients**. The only additional requirement is that the variance in the two groups is the same. If the two lines are found not significantly different when $n_1$ and $n_2$ are sufficiently large to provide enough statistical **power**, the data from two groups can be pooled and a unified regression can be fitted that would have better reliability because of increased $n$. For this, pool the data and run the regression on $x$ and $z$ again.

The procedure can be easily extended to compare more than two lines. We are not going to discuss here this extension as it becomes mathematically complex. If interested, see Milliken and Johnson [1].

1. Milliken GA, Johnson DE. *Analysis of Messy Data, Volume III: Analysis of Covariance*. Chapman and Hall/CRC, 2001.

## comparison of two or more relative risks

For this section, we presume that you are familiar with the basics of relative risk (RR) and understand why we use its logarithm (lnRR) for most statistical inferences on RR. You may like to review the topic **relative risk**.

Two kinds of comparison can arise in case of RR. One is when this is calculated separately for two independent groups. For example, this may be RR of death within a period of 10 years for stage I of oral cancer versus no cancer for males and females. In this case, one RR is for males and one for females, and these RRs are compared to find whether they are different or not in the corresponding populations. Another comparison is RR of low birth weight in anemic women of age 20–24 years versus nonanemic, and similarly RR in women of age 25–29 years, to check if they are significantly different or not. In both these setups, the groups are independent since the outcome in one group does not affect the outcome in the other—the cases are different in the two groups. In this situation,

$$\mathrm{SE}[\ln \mathrm{RR}_1 - \ln \mathrm{RR}_2] = \sqrt{[\mathrm{SE}(\ln \mathrm{RR}_1)]^2 + [\mathrm{SE}(\ln \mathrm{RR}_2)]^2}.$$

The SEs of $\ln(\mathrm{RR}_1)$ in the first group and $\ln(\mathrm{RR}_2)$ in the second group on the right-hand side of the equation can be obtained by the formula given for **confidence interval for relative risk** for large samples.

RR is a ratio and cannot be expected to follow a Gaussian pattern. Logarithm linearizes this and helps in exploiting the premise of the **central limit theorem** to say that the statistical distribution of lnRR will be approximately Gaussian for large $n$. Thus, the test criterion for testing the null $\mathrm{H}_0$: $[\ln \mathrm{RR}_1 - \ln \mathrm{RR}_2] = \delta$ is

$$z = \frac{[\ln \mathrm{RR}_1 - \ln \mathrm{RR}_2] - \delta}{\mathrm{SE}[\ln \mathrm{RR}_1 - \ln \mathrm{RR}_2]}.$$

For equality of RRs, $\delta = 0$ and reject equality of RR in the two groups if the calculated value of $z$ for your samples is less than the threshold at the specified level of significance. For 5% level of significance, this threshold for a **two-sided alternative** is 1.96. Note that this test is for lnRR and not for RR itself. Also, this is valid only for large samples. The concept of RR is hardly ever applicable to small samples. This test presumes that the two groups have been studied with similar methodology and similar care, and there is no factor other than the group that can affect the RRs. The procedure is extendable to more than two groups provided they are independent and each has a large sample.

The second setup is when you want to compare the RR of one risk factor with the RR of another risk factor in the same group of subjects. In our example, besides the stage of cancer, we may be interested in the effect of age also. Age may be divided as <70 and ≥70 years. The question is whether age has more influence on 10-year risk of death or the stage has more influence. For this, you may like to refer to Hosmer et al. [1].

1. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*, Third Edition. Wiley, 2013.

## competing risks

Competing risks are the **mutually exclusive and exhaustive** factors that compete with one another to cause the same outcome. For example, cardiovascular diseases and malignancies are the competing risks for death in old age. In this case, death is the outcome of interest. Both diseases can occur together in the same person, but the **cause of death**, when medically established, has to be only one of these. A large number of risks simultaneously operate, but when only one of them can succeed in reaching to the outcome, they will be called competing. In this case, occurrence of one will prevent any other to cause the outcome.

In another setup, surgery (say, radical prostatectomy) and radiotherapy are competing strategies for treating prostate cancer. But only one of them can be applied. When death is considered the outcome of interest, 15-year follow-up and competing risks analysis shows that the nonmetastatic (less severe) patients with surgery have nearly 1.76 times chance of survival compared to those on radiotherapy, but the performance is nearly the same in metastatic patients [1].

The objective of competing risks methodology is to delineate the risk from each competing factor such that the sum of these risks is 1. The methodology is based on cause-specific hazards and provides a kind of "real-world" probabilities occurrences of an event due to competing causes compared to single cause probabilities that could be contaminated by other causes. Take the example of deaths due to lung cancer in people of age 60 years and above who smoked for at least 200 pack-years of cigarette. These deaths will not be just because of smoking but also because of ageing and other ailments afflicting these people. The conventional analysis will attribute

deaths at different ages as though they are from lung cancer alone since the subjects are suffering from this disease. The fact is that the deaths are contaminated by ageing process and comorbidities. The competing risk methodology will filter out causes and provide risk "purely" due to lung cancer.

The methodology is mathematically complex; those interested may see Andersen et al. [2]. However, the concept and applicability of competing risks may be clear from the preceding discussion. The following example may help in appreciating the application of competing risk methodology.

Kandala et al. [3] analyzed 239,000 patients in the National Joint Registry for England and Wales for 10-year revision rate of total hip replacement. This was the outcome of interest. This was studied for each of the five prosthesis types such as cemented and cementless with ceramic-on-ceramic bearing surfaces or ceramic-on-polyethylene bearing surfaces. These were considered as competing risk for revision of hip replacement. Cemented prostheses with ceramics-on-polyethylene bearing surfaces had the lowest revision rates.


1. Sooriakumaran P, Nyberg T, Akre O, Haendler L, Heus I, Olsson M, Carlsson S, Roobol MJ, Steineck G, Wiklund P. Comparative effectiveness of radical prostatectomy and radiotherapy in prostate cancer: Observational study of mortality outcomes. *BMJ.* 2014 Feb 26;348:g1502. http://www.bmj.com/highwire/filestream/688258/field_highwire_article_pdf/0/bmj.g1502
2. Andersen PK, Geskus RB, de Witte T, Putter H. Competing risks in epidemiology: Possibilities and pitfalls. *Int J Epidemiol* 2012;41:1–10. http://ije.oxfordjournals.org/content/early/2012/01/08/ije.dyr213.full.pdf+html
3. Kandala NB, Connock M, Pulikottil-Jacob R et al. Setting benchmark revision rates for total hip replacement: Analysis of registry evidence. *BMJ* 2015 Mar 9;350:h756. http://www.bmj.com/content/350/bmj.h756.long


## complete linkage method of clustering,
*see also* **cluster analysis**

See **cluster analysis** for basic details of this kind of analysis. Complete linkage is one of the several methods of hierarchical clustering. In hierarchical clustering, a measurement of dissimilarity such as **Euclidean distance** is used to classify the units into various groups using one of the several possible algorithms. Two units (or subjects) that are most similar (or least distant) are grouped together in the first step to form one group of two units. This group is now considered as one entity. Now the distance of this entity from other units is compared with the other distances between various pairs of units. Again, the closest are joined together. This hierarchical agglomerative process goes on in stages, reducing the number of entities by one each time. The process is continued until all units are clustered together as one big entity. See **hierarchical clustering** for a method to decide when to stop the agglomerative process so that natural clusters are obtained.

The primary problem in clustering is computing the distance between two entities containing, say, $I$ and $J$ units, respectively. Several methods are available. One is to consider all units in an entity centered on their average. Another is to compute the distance of the units that are farthest in the two entities. A third method is to base it on the nearest units. There are several others. Complete linkage is one such method. Different methods can give different results, but complete linkage is considered the most appropriate for many applications. This method was first proposed by Sorensen [1] in the context of similarity in plant species.

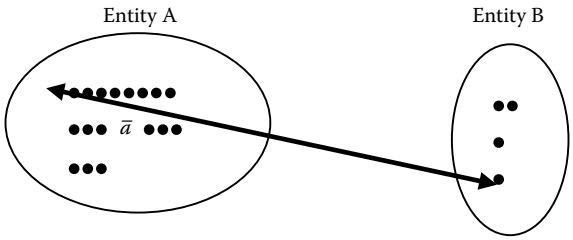



**FIGURE C.19** Complete linkage distance between entity A and entity B.

In complete linkage method of clustering, the distance between two clustered entities is measured by the distance between the farthest located units of the entities. If one entity has 7 units with values denoted by $(x_1, x_2, \ldots, x_7)$ and the other has 4 units with values denoted by $(y_1, y_2, y_3, y_4)$, the first step would be to calculate the distance between $x_1$ and $y_1$, $x_1$ and $y_2$, …, $x_2$ and $y_1$, …, $x_7$ and $y_4$. The maximum of these 28 distances will be considered as the distance between these two entities (Figure C.19). If entity A contains $I$ units $(a_1, a_2, \ldots, a_I)$ and entity B contains $J$ units $(b_1, b_2, \ldots, b_J)$, then, under the complete linkage method, the distance between these two entities is measured as $d_{AB} = \max_{ij}(d_{ij})$, where $d_{ij}$ is the distance between the $i$th unit of the first entity and the $j$th unit of the second entity. This method can be easily extended to a multivariate setup.

A large distance indicates that the entities are really very different from each other and thus should not be clustered together. If this distance is small, the entities can be considered similar, and you can merge these two entities together to form a bigger entity.

In a Monte Carlo comparison of six methods of hierarchical clustering on random data, complete linkage has been found the most adequate method not to discover false clusters [2]. Wentzensen et al. [3] used the complete linkage method for hierarchical clustering of disease groups and human papilloma virus genotypes for 2780 women who underwent colposcopy at baseline.


1. Sorensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analysis of the vegetation on Danish commons. *Biologiske Skrifter* 1948;5:1–34. http://books.google.co.in/books/about/A_Method_of_Establishing_Groups_of_Equal.html?id=rpS8GAAACAAJ
2. Jain NC, lndrayan A, Goel LR. Monte Carlo comparison of six hierarchical clustering methods on random data. *Pattern Recog* 1986;19:95–9. http://www.sciencedirect.com/science/article/pii/0031320386900385
3. Wentzensen N, Wilson LE, Wheeler CM et al. Hierarchical clustering of human papilloma virus genotype patterns in the ASCUS-LSIL triage study. *Cancer Res* 2010 Nov 1;70(21):8578–86. http://cancerres.aacrjournals.org/content/70/21/8578.long


## completely randomized designs

In the context of biostatistics, design is the prefixed structure for collecting the observations for an empirical research. Such a study is likely to have at least two series of measurements that could be for different groups of subjects or the same group repeatedly measured. The design of an experiment is called completely randomized when the allocation of subjects to the groups is entirely at random with no consideration of any other factor. For example, if you have 60 eligible subjects for a trial and you divide them randomly into three groups to receive placebo (no drug), low dose, and high dose of

drug A, this is a completely randomized design. But if you consider something like age, sex, and severity of disease for allocation, the design is no longer completely randomized. Completely randomized is also called single-factor design—the factor being the dose of the drug in this example.

C

Suppose the interest is in conducting the same trial in males and females. This classifies the subjects. You will obviously have three groups of males with none, low, and high doses of drug A, and similar three groups of females. These cannot be considered six independent groups for **randomization**, since the randomization is now restricted to males and females separately. This is not a completely randomized design with one factor at six levels. This is a two-factor $3 \times 2$ design since one factor (dose) has three levels and the other factor (sex) has two levels. This is analyzed by a two-way analysis of variance.

There is a twist, though. If the dose of drug A is considered as factor 1 at three levels and another regimen B as factor 2 at two levels (yes/no), this can still be conducted as a completely randomized design. Since there is no classification factor for subjects, the eligible subjects can be randomly allocated to these six groups. These groups will have a group, for example, with high dose of drug A and no regimen B and another group with low dose of drug A and with regimen B. Thus, the **one-way ANOVA** of the data from this design will reveal whether the outcome in these six groups is significantly different or not. But this analysis will not find the significance of the **main effects and interactions** of the factors. For this, the analysis needs to be done by **two-way ANOVA**.

An experiment with repeated measures can also follow a completely randomized design when subjects are unrestrictedly randomly allocated to the groups receiving different regimens. However, because of repeated measures, this cannot be analyzed by one-way ANOVA. For analysis of this design, see **repeated measures ANOVA**.

In contrast to the completely randomized design is the **randomized block design** where the subjects are first sorted into homogeneous groups, called blocks, and the levels of the intervention are then randomly assigned within the blocks. The blocks can also be considered as strata or a factor. We have given the example of an experiment in males and females where sex is the blocking factor that divides subjects into homogeneous groups assuming that male subjects are homogeneous within themselves and so are the females. But the two groups are expected to provide very different responses. If you already know that the response does not differ by sex, you may as well plan completely randomized design without classification of the subjects by sex.

## compliance (in clinical trials)

Compliance in clinical trials is adherence to the **protocol**. This term is particularly used for the subjects when they comply with the regimen they are allocated. Noncompliance, differential compliance, and poor compliance adversely affect the validity of the results of a clinical trial. This is one of the issues that distinguishes **efficacy** with **effectiveness**. Effectiveness of a regimen is the response in actual practical conditions, including lack of compliance.

An intervention study such as a clinical trial requires active participation and cooperation of the study subjects. After agreeing to participate, some subjects may deviate from the protocol for a variety of reasons. These include developing side effects, forgetting to take their medication, or simply withdrawing their consent after randomization. Analogously, in a trial of surgical therapy, those who were randomized to one group may choose to obtain alternative treatment on their own initiative. In addition, there may be instances in which participants cannot comply, such as when the condition of a randomized patient rapidly worsens to a point where continuation in the trial becomes contraindicated. Following the regimen partially is called partial compliance.

The problem of compliance could be particularly acute in a field trial. Sankaranarayanan et al. [1] report a community-based trial wherein all persons of age 35 years or older were screened for oral cancer in intervention villages and not screened in the control villages in India. The difficulty was, as in most field trials, low compliance when referred for confirmatory examination that has to be done in a hospital. In this trial, compliance was lower than 70%.

In a clinical trial setup, subjects in the active treatment group may drop out more because of discomfort or poor taste of the drug, even when the placebo looks like the drug and the trial is randomized. On the other hand, the subjects in the treatment group may stay if they see improvement in their condition, whereas the placebo group can become noncompliant. The compliance rate in this case is related to the efficacy of the regimen, and the comparison can be jeopardized. The difference may also be related to prognostic factors such as occurrence of nausea in the treatment group but not in the control group. If so, the comparability may be lost despite full randomization and despite random selection of subjects from the target population. Such differential compliance is statistically more challenging when it comes to the analysis of the data.

The other problem is lack of compliance with the protocol by the observers or the investigators. This problem can be severe if many observers are involved. Multiple observers are natural in a **multicentric trial**, and therefore such trials are especially vulnerable to such bias. More intensive efforts in terms of training and periodic assessment of the methods and the procedures followed by different observers may be helpful in minimizing this bias.

Poor compliance can force a trial to stop midway if the results are not going to meet the validity requirements. Take preemptive steps to minimize such losses. The problem of achieving and maintaining high compliance is an issue that needs consideration in the design and conduct of all clinical trials. **Blinding and masking** are among the strategies that can help to minimize loss in compliance. Whenever noncompliance or partial compliance occurs, this should be disclosed at the time of the reporting. Statements such as **CONSORT** require that this be fully described so that the reader can make his/her own judgment.

Some participants in a trial may become noncompliant despite all reasonable efforts. Try to obtain complete information on these subjects. Adjust the results in accordance with the outcome observed in those who are similar to the ones who dropped out.

1. Sankaranarayanan R, Mathew B, Jacob BJ, Thomas G, Somanathan T, Pisani P, Pandey M, Ramadas K, Najeeb K, Abraham E. Early findings from a community based, cluster-randomized, controlled oral cancer screening trial in Kerala, India: The Trivandrum Oral Cancer Screening Study Group. *Cancer* 2000;88:664–73. http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0142(20000201)88:3%3C664::AID-CNCR25%3E3.0.CO;2-V/full

## components of variance, *see* variance components analysis

## composite curve

Composite curve is the one that results from a mixture of two or more distinct curves. We illustrate this for a **distribution** curve that results from mixing of two different distributions.

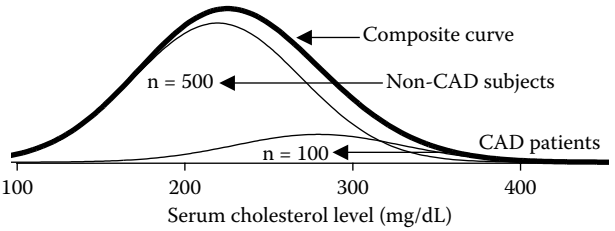



**FIGURE C.20** Mixing of two distinct groups resulting in a nice composite curve.

The pattern of distribution of the serum cholesterol level in two distinct groups as well as jointly in their mixture is shown in Figure C.20. The group of patients with coronary artery disease (CAD) has mean 280 mg/dL and standard deviation (SD) 50 mg/dL. There are 100 such subjects. The non-CAD group contains 500 subjects, and its mean is 220 mg/dL (SD = 50 mg/dL). The pattern is **Gaussian** in both groups. When these two groups are combined, the resultant again is a Gaussian curve with mean 230 mg/dL and SD 50 mg/dL as before. This is shown by the thick curve in Figure C.20. The composite curve is nice and masks the fact that this actually is made up of two distinct groups. When only the composite curve is available, it is extremely difficult to make out that two very different groups constitute the population. It is only through extraneous information that any clue to the presence of such distinct groups would be available.

## composite index

A composite index is a combination of two or more indexes. An **index** by definition itself is composite in the sense that it puts together information on several aspects into a single metric. Composite index is a further combination. The need of a composite index arises because many aspects are sometimes combined to measure health in its comprehensive form, and it is difficult to express this by one index.

An index can be defined as a combination of two or more **indicators** stated relative to a base or a standard. Indicators measure a single and specific aspect of health. No index is yet available that measures health in its entirety. Since health in its comprehensive form is also understood as state of well-being, it can encompass diverse aspects such as physical health, income, satisfaction, social relations, and cultural activities. The Office of National Statistics of UK is trying to develop a measure of national well-being. A debate is going on regarding what should this contain. For example, this may have indexes that measure how people feel in control of themselves, make choices, and have a sense of purpose and belonging. The following are two categories of some of the currently more popular composite indexes of community health.

The first one is the set of indexes of comprehensive health. These try to incorporate several aspects of health, although they still fail to capture the complete well-being. One popular such composite index at public health level is the **human development index**. This has components on education, income, and life expectancy at birth. Each of these components is measured by an index. The other popular index at community level is **physical quality of life index** that consolidates infant mortality rate, life expectancy at 1 year, and literacy. For individual level, see **quality of life index** that considers domains such as physical well-being, psychological well-being of the person, and psychological well-being divided into negative and positive feelings, self-esteem, memory, etc.

The second set of composite indexes comprises indexes of health gap at community level. The most prominent among these is the **disability adjusted life years (DALYs)** based on years of life lost due to premature mortality and the equivalent life lost due to various forms of disability. Also see multidimensional poverty index [1] and index of need for health resources [2], which also are composite indexes of different aspects of health needs.

1. UNDP. Human Development Reports. *Multidimensional Poverty Index (MPI)*. http://hdr.undp.org/en/content/multidimensional-poverty-index-mpi
2. Sekhar CC, Indrayan A, Gupta SM. Development of an index of need for health resources for Indian States using factor analysis. *Int J Epidemiol* 1991 Mar;20(1):246–50. http://ije.oxfordjournals.org/content/20/1/246.long

## computer-aided diagnosis, *see* expert systems

## concealment of allocation, *see* blinding, masking, and concealment of allocation

## concomitant variable, *see* dependent and independent variables (in regression)

## concordance and disconcordance

Concordance is similarity. In health and medicine, we may be interested in concordance between methods, between observers, between laboratories, between measurements, etc. The underlying connotation in this term is qualitative assessment than quantitative assessment. The corresponding term for quantitative "concordance" is agreement where each value is assessed against its paired value. For this, see **agreement assessment (overall)** in this volume. Whereas assessment of the difference between values of body mass index (BMI) before and after an exercise–diet regimen requires agreement assessment, comparing the degree of obesity by BMI and the category of waist/hip ratio in the same group of subjects would require concordance. If both the methods agree for the degree of obesity in most people, the methods can be said to have good concordance for assessment of obesity. The degree of this kind of concordance is statistically measured by **Cohen kappa** and **Bangdiwala B**.

The term *concordance* is also used for consistent observations, particularly in the context of *ordinal* categories. Consider two characteristics on ordinal scale such as severity of disease and obesity. One may have $R = 4$ categories and the other $C = 3$ categories. The interest is in measuring the strength of association between these two ordinal characteristics. The data can be arranged in an $R \times C$ table. The association is high if the higher category of one is more frequently seen with the higher category of the other. The association between severity of disease and obesity is high if more severe cases are obese. This is also called concordance. If less severe cases are mostly obese, this is discordance. For various measures of ordinal association using concordant and disconcordant pairs, see **association between ordinal characteristics (degree of)**.

Disconcordance is the lack of concordance—the values do not match. The most pronounced application of this is in the analysis of pair-matched data by the **McNemar test** that is based on disconcordant pairs rather than concordant pairs.

**concurrent controls,** *see* **controls**

**concurrent cohort,** *see* **cohort studies**

C

**concurrent validity,** *see* **validity (types of)**

**conditional probability,** *see* **probability**

## confidence band for regression line/curve

As happens with all sample estimates, **regression** lines and curves also fluctuate according to the sample values. Statisticians have worked out limits outside which the regressions are unlikely to lie with a certain degree of confidence. Since we have a line or a curve, these limits produce a band instead of limits, which is called confidence band. For simplicity, we explain this band for a regression line (**simple linear regression**) and illustrate also for a simple **curvilinear regression**. The concept and the methods are the same for a complex-shaped curve. Confidence band requires **standard error** (SE) of the regression coefficients. These SEs measure the variation expected from sample to sample.

### *SEs and Confidence Intervals (CIs) for the Regression*

When there is only one independent variable $x$, the regression equation is $\hat{y} = a + bx$. The estimated SEs for this simple linear regression are as follows:

$$\text{SE}(a) = \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x-\bar{x})^2}\right)}, \quad \text{and} \quad \text{SE}(b) = \sqrt{\frac{\text{MSE}}{\Sigma(x-\bar{x})^2}},$$

where MSE = SSE/($n$ – 2), and SSE is the **error sum of squares**. If α is the intercept and β is the slope parameter in the population, under Gaussian conditions,

CI for α: $a \pm t_\nu$*SE($a$)

CI for β: $b \pm t_\nu$*SE($b$),

where $t_\nu$ is the value of **Student *t*** at $\nu = (n - 2)$ df. An example in the next paragraph illustrates these CIs. Statistical software packages work out these CIs easily. Similar procedure is used for various regression coefficients in multiple linear regression. Statistical software packages do this also easily.

### *Confidence Band for Simple Linear Regression*

When we consider the CI for the regression line, which is based on both $a$ and $b$, we get a confidence band. As always, this band depends on the SE of $\hat{y}$. This is different when it is for the mean of $y$, also called the *estimated value* of $y$, and when it is for an individual value of $\hat{y}$, called the *predicted value* of $y$. You may be aware that any regression pertains to the *mean* of $y$ for given $x$'s, although this is often used for predicting individual values. Predicting individual values has a much larger SE. These SEs have complex forms for multiple regression, but they can be easily stated for a simple linear regression. Sometimes the software does not provide these SEs, and you may have to calculate these yourself. Under certain general conditions, for simple linear regression of $y$ on $x$,

$$\text{SE (estimated mean of } y_x) = \sqrt{\text{MSE}\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\Sigma(x-\bar{x})^2}\right]} \quad \text{(C.1)}$$

$$\text{SE (predicted individual value of } y_x) = \sqrt{\text{MSE}\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\Sigma(x-\bar{x})^2}\right]}, \quad \text{(C.2)}$$

where $y_x$ is the value of $y$ at given $x$. MSE is always generated by the software.

If the dependent of interest is duration of analgesia ($y$) induced by different doses of a drug ($x$), and $n$ = 15, MSE = 6.50, $\bar{x}$ = 11.2 μg, and $\Sigma(x-\bar{x})^2$ = 18.07, then for $x$ = 8, SE(estimated mean of $y$ at $x$ = 8) = $\sqrt{[6.50\{1/15 + (8-11.2)^2/18.07\}]}$ = 2.03 from Equation C.1. The SE(predicted individual value of $y$ at $x$ = 8) = $\sqrt{[6.50\{1 + 1/15 + (8-11.2)^2/18.07\}]}$ = 3.26 from Equation C.2. The SE of the predicted value is much higher. Prediction of an individual value of $y$ is always less precise than the prediction of mean of $y$. These SEs are used to generate the confidence band for the regression line. Confidence band would be different for mean than for individual values depending upon which SE is used. Since $x$'s are considered known, everything is fixed in this SE. As $x$ moves away from its mean, the SE increases because of the term $(x-\bar{x})^2$. This provides a band of the type shown in Figure C.21a.

Under Gaussian conditions, the confidence band is obtained by calculating the limits $\hat{y} \pm t_{n-2}\text{SE}(\hat{y})$ for different values of $x$ in the given range, where the value of $t$ corresponds to the required

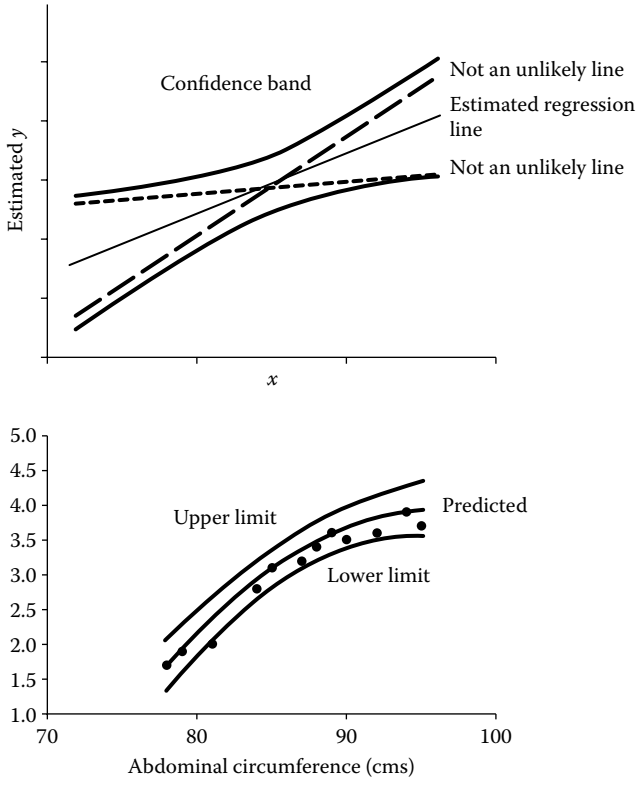



**FIGURE C.21** Confidence band (a) for a regression line and (b) for a quadratic curve.

confidence as usual. The confidence band specifies the *area* within which the regression line can lie—better phrased as outside of which the regression is extremely unlikely to lie. For example, it could be as steep as shown by the dashed line in Figure C.21a, or as flat as shown by the dotted line. Only statisticians seem to know the wide variation that can occur in the estimation of $y$ from a regression line. Many medical professionals use the line obtained from their data as the true line without realizing the variation around it. Figure C.21b has such a band for a quadratic regression. This is just to illustrate the shape of the confidence band for a curvilinear regression.

In the case of multiple linear regression, the width of the CI for estimated $y$ decreases in most situations as the number of regressors increases because of reduced SSE and MSE. However, in some situations, width increases because of loss of df. Loss of df can cause the MSE to increase, which consequently increases the width of the CI. For this reason, an adequate sample size becomes doubly important in case of multiple regression.

## confidence bounds, *see also* confidence intervals

Confidence bound is a one-sided lower or upper limit established by the sample that is likely to cover the value of the **parameter** under estimation with a prespecified degree of confidence. It contrasts with **confidence interval**, which is two-sided and provides both lower and upper limits. Confidence bound is used when it is known that the parameter value is almost surely in one direction as illustrated in the following example.

Suppose a new iron supplement is formulated that is quickly absorbed by the body and has low cost compared to the existing supplements. Previous experimentations and early phases of clinical trials have established that this formulation cannot harm, but a bigger trial is planned to find how much difference it makes in the mean hemoglobin (Hb) level of those who are anemic. Clearly, in this case, the interest is in the minimum gain in the mean Hb level below which it is unlikely. Under **Gaussian conditions** (either the underlying distribution is nearly Gaussian if the sample size is small or the sample size is large), at $100(1-\alpha)\%$ confidence,

$$\text{lower confidence bound for mean } \mu\text{: } \bar{x} - t_{\upsilon(1-\alpha)}\frac{s}{\sqrt{n}},$$

where $\bar{x}$ is the sample mean, $s$ is the sample standard deviation (SD), $n$ is the sample size, and $t_{\nu(1-\alpha)}$ is the value of **Student *t*** at $\nu$ degrees of freedom (df) for probability $(1-\alpha)$. In this case, $\nu = n-1$. Note that $s/\sqrt{n}$ is the **standard error (SE)** of $\bar{x}$. In an unlikely case of known SD $\sigma$, $s$ is replaced by $\sigma$ in this equation, and $t$ is replaced by the $z$-value from Gaussian distribution. For 95% confidence bound, $z = 1.645$.

In our example, if the average gain in Hb level in a sample of $n = 60$ subjects is 1.7 mg/dL with SD = 0.4 mg/dL, the lower bound for mean at 95% confidence level is

$$1.7 - 1.671 \times 0.4/\sqrt{60} = 1.61,$$

where 1.671 is the value of Student $t$ at 59 df corresponding to 0.95 probability. Thus, there is only 5% chance that the mean gain in the population will be less than 1.61 mg/dL. Note that this inference is for mean in the population and not for individual values. Individual gain in Hb level would vary much more.

$$\text{Upper confidence bound for mean } \mu\text{: } \bar{x} + t_{\upsilon(1-\alpha)}\frac{s}{\sqrt{n}}.$$

Similar to Gaussian conditions for being able to use Student $t$ in calculating the bounds, the requirement for calculating confidence bounds for population proportion $\pi$ is that the sample size must be large such that $np$ and $n(1-p)$ are both more than 8, where $p$ is the sample proportion. When this requirement is met,

$$\text{lower confidence bound for proportion } \pi\text{: } p - z_{(1-\alpha)}\sqrt{\frac{p(1-p)}{n}},$$

and

$$\text{upper confidence bound for proportion } \pi\text{: } p + z_{(1-\alpha)}\sqrt{\frac{p(1-p)}{n}},$$

where $z_{(1-\alpha)}$ is the value of Gaussian $z$ at $(1-\alpha)$ probability. For 95% confidence, $z_{0.95} = 1.645$, and for 90% confidence, $z_{0.90} = 1.28$. Such bounds for proportion are useful when you want to say, for example, that a minimum of 4% children in an area are affected by a particular ailment rather than saying that somewhere between 2% and 9% are affected.

Similar confidence bounds can be obtained for other parameters such as difference in means, difference in proportions, odds ratio, and relative risk. Under Gaussian conditions, these take the form of sample estimate plus (for upper bound) or minus (for lower bound) the $z$-value times the SE of that sample estimate. The method is similar to the one mentioned for **confidence interval** for that parameter. However, the case of $p = 0$ or $p = 1$ deserves special consideration. It is possible in the sample that you get all positives or all negatives. For confidence bounds in this case, see **Clopper–Pearson bounds/interval**.

## confidence intervals (the concept of)

Confidence interval (CI) for a population **parameter** is the range within which it is likely to lie with high probability. A more correct interpretation is that CI is the range outside of which the parameter value is unlikely to lie in repeated samples. This likelihood is called the **confidence level**. A confidence level of 95% is almost universally accepted in health and medicine.

Jerzy Neyman

Jerzy Neyman introduced the CI in statistical testing in 1937 [1]. (For the relation between the CI and the test of hypothesis, see **confidence interval versus test of significance**.) The CI envisages that the parameter value is fixed, but repeated samples give different CIs depending on the sample values such that it is highly likely to contain the value of the parameter. Figure C.22 aptly explains the CI and shows that different samples can give varying CIs, but the chances are small that it does not contain the value of the parameter. In this illustration, sample no. 7 does not contain the value of the parameter—all others do. CIs can be explained further as follows.

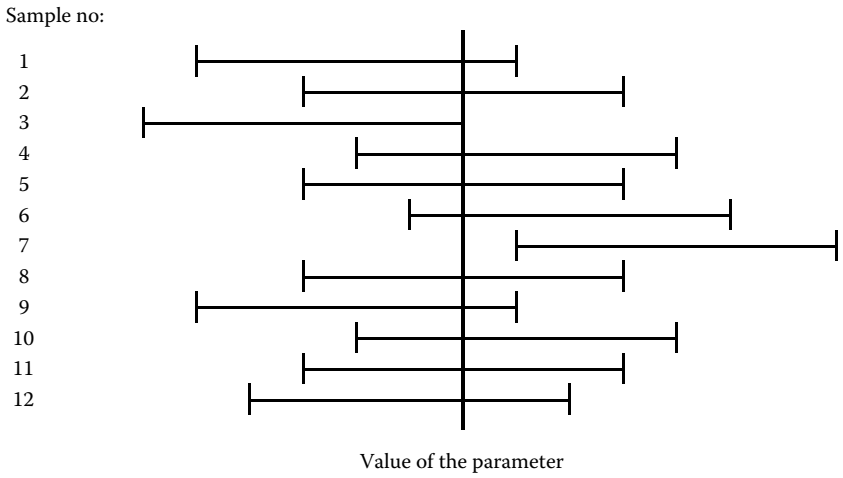



**FIGURE C.22** Varying CIs in different sample, mostly containing the value of the parameter.

Medical empiricism is all about groups rather than individuals. Suppose a research finds that the positive predictivity of a new diagnostic procedure is 70%. How confident could you be that a similar study on another group of subjects would not give predictivity less than 70%? Since sample values differ from sample to sample, it is useful to find how different the results are likely to be in different samples. However, repeated samples actually are not studied. Methods are available that would provide an interval within which the actual result is likely to lie in repeated samples. This interval can be obtained by using the data of only one sample when randomly drawn, and is called the CI. This is also called an **interval estimate**. The basic function of a CI is to provide a range that cannot be denied for the value of the parameter.

Because of profound medical uncertainties, particularly the sampling fluctuations, it is never possible to work out an interval with 100% confidence. Generally, a confidence level of 95% is used. Statistically, a confidence interval gives a range with a substantial hope that it will include the parameter of interest. The confidence level associated with the interval (say 90%, 95%, or 99%) gives the percentage of all such possible intervals in repeated samples that will actually include the true value of the parameter. Note that a CI tells us what to expect in the long run. It does not say anything about a particular sample.

The CI gives rise to the **bikini syndrome**. *What it reveals is interesting but what it conceals is vital.* Some researchers may argue that they already know about 95% and want to know about the other "behind-the-scene" 5%. Luckily for biostatistics, rarely in medical research, if ever, 95% will be known a priori. Even when 5% is masked, the revelation of the other 95% as plausible turns out to be a good piece of information.

The general strategy for calculating a 95% CI that works fairly well in most cases when $n$ is large is as follows:

1. Find the sample analog of the population parameter for which the CI is needed.
2. Estimate its **standard error (SE)**.
3. 100*(1 – α)% CI is the (sample analog ± $t_{\nu(1-\alpha/2)}$SE), where $t_{\nu(1-\alpha/2)}$ is value of Student $t$ at $\nu$ df and probability (1 – α/2).

This is what is followed in our description of CI for various parameters under separate topics in this volume. This assumes that the actual value of the population SD σ is not known as would happen in almost all practical situations. Discussion in this volume is mostly restricted to confidence level 95% because that is the most commonly used level. However, the CI can be obtained with any other confidence level. In the case of sufficiently large $n$, or when the population SD σ is known, the multiplier $t_{(1-\alpha/2)}$ reduces to the corresponding Gaussian $z$. The Gaussian multiplier for large $n$ is approximately 1.96 for 95% confidence and 2.58 for 99% confidence. That is, the 95% CI is obtained by ±1.96*SE limits and 99% by ±2.58*SE limits. The 90% CI is obtained by ±1.645*SE, and the 80% CI by ±1.28*SE. The exact value of the multiplier can be obtained from the Gaussian distribution.

Some intricacies of CI can be explained with the help of an example. The following comments largely apply to all CIs, although they are explained in the context of one example. Suppose a random sample of 100 hypertensives with mean diastolic BP 102 mmHg is given a new antihypertensive drug for 1 week as a trial. The mean level after the therapy came down to 96 mmHg. The mean decrease in these 100 subjects is $\bar{d} = (\bar{x}_1 - \bar{x}_2) = 6$ mmHg, and suppose the SD of the decrease is $s_d = 5$ mmHg. This $s_d$ is obtained as usual for the differences between the matched pairs. What is the 95% CI for the actual mean decrease in the population? After taking the difference, the problem reduces to one-sample mean. For this, the procedure described under **confidence interval (CI) for mean** is followed.

Since $n = 100$, $v = 100 - 1 = 99$. Thus, the 95% CI for the mean decrease in this example is

$$(6 - t_{99} \times 5/\sqrt{100},\ 6 + t_{99} \times 5/\sqrt{100}).$$

From a software package, $t_{99} = 1.99$ for 0.975 probability. Because of the symmetry of **Student *t***, this means that the range (–1.99, +1.99) contains 95% of values of $t$. Therefore, the 95% CI is

$$(6 - 1.99 \times 5/\sqrt{100},\ 6 + 1.99 \times 5/\sqrt{100}) = (5,7) \text{ mmHg}.$$

- Note that the CI (5, 7) mmHg in this example is for the *mean* decrease. The decrease in 95% of *individual* patients is likely to vary in the interval $(\bar{x} - 1.96s,\ \bar{x} + 1.96s)$ or in the interval (6 – 10, 6 + 10), that is, (–4, 16) mmHg, provided the underlying distribution is Gaussian. Individual variation is much higher than the variation in means. The minus sign indicates that some patients may show a rise in diastolic BP to the extent of 4 mmHg despite the drug instead of a decline. Care is required to maintain the distinction between SD and SE.

- It is important to realize that a 95% CI gives limits that are likely to contain the value of the parameter. Thus, the CI is to be interpreted as a useful quantity in the long run. If 100 such trials are done, nearly 95 of them may give a mean decrease between 5 and 7 mmHg. The other 5 trials can give either a higher or a lower decrease. As emphasized earlier, the value of the CI is more in what it does not contain. In this example, the CI suggests that the chance of a mean decrease being either less than 5 mmHg or more than 7 mmHg is very remote.
- Sometimes, an approximate multiplier 2 is used instead of exact 1.96. Such an approximation is often preferred not only for convenience but also because this slight increase tends to cover the mild departure from the Gaussian pattern as the distribution is seldom exactly Gaussian in practice. This approximation, however, can create an anomaly in some situations. In our example, $t_{99} = 1.99$, and this is less than 2. Thus, the CI based on $t$ would be smaller than the CI based on the Gaussian pattern if 2 is used. The fact is that $t$ has a larger variance, and it should always give a larger CI relative to the Gaussian CI. The exact value of $t$ does not allow any departure from the Gaussian condition that the approximate multiplier 2 based on $z$ does.
- For small $n$, when the underlying distribution is Gaussian and σ is not known, the value of $t_v$ could be very different from the Gaussian values. For example, if $n = 6$, then $v$ is 5 and $t_v$ for 95% CI is 2.571 and for 99% CI is 4.032. These are very different from Gaussian values 1.96 and 2.58, respectively. For large $n$, the $t$-value can be approximated by the Gaussian $z$-value, but there is no need to use such an approximation because exact values of $t$ are available for any specific df.
- *Role of sample size in obtaining a narrow CI.* In our example with $n = 100$, the 95% CI for a decrease in diastolic BP level is between 5 and 7 mmHg. This is fairly narrow. If $n = 10$, with no other change, the 95% CI is

$$(6 - t_9 \times 5/\sqrt{10},\ 6 + t_9 \times 5/\sqrt{10}),$$

or

$$(6 - 2.262 \times 5/\sqrt{10},\ 6 + 2.262 \times 5/\sqrt{10}),$$

or (2.4, 9.6) mmHg.

Small $n$ increases the CI in two ways: (i) by increasing the SE and (ii) through an increase in $t$-value. Thus, the CI for $n = 10$ is relatively very wide compared with $n = 100$. A focused conclusion is illusive when $n$ is small. This again underscores the statistical importance of a large sample size. Also note that the width of CI depends mostly on $n$ and SE and not on the **effect size**.

Proportion and mean (or their function such as difference) are just about the two most common parameters on which confidence intervals are drawn. There might be isolated examples in which the interest is in the CI for the median or for a decile, even σ. The basic methodology to obtain 95% CI is to get the 2.5th and 97.5th percentiles of the distribution of the corresponding "statistic" in the sample. The difficulty is that the distribution of statistics other than the mean, and hence the 2.5th and 97.5th percentiles, is not easy to obtain. Regular statistical packages are not of much help in obtaining such CIs. Some feel comfortable with bootstrap method, which can be used for any parameter and any distribution, but it looks like this method lacks the rationale that the methods just described provide. **Confidence interval (CI) for the median** is separately described, which might be a good alternative in some situations. For details of how to obtain the CI for other **quantiles**, see Conover [2]. The CI for ratios such as **odds ratio** and **relative risk** are also discussed separately in this volume.

1. Champkin J. Timeline of Statistics: Pull out. *Significance* Dec 2013; 10(6):23–6. http://www.significancemagazine.org/details/webexclusive/5774761/The-timeline-of-statistics.html
2. Conover WJ. *Practical Nonparametric Statistics*, Third Edition. Wiley, 1999.

## confidence interval (CI) for attributable risk (AR)

As explained for **attributable risk**, this actually is the difference between two proportions—one measuring risk in the group with exposure and the other measuring risk in the group without exposure, provided that the groups are independent. If the 5-year risk of death in 70-year-old patients with a disease is 22% and in patients without the disease is 15%, the attributable risk of death due to the disease is the difference 7%. Note that the groups are independent in this case. Thus, the confidence interval for attributable risk in this case is obtained as for difference in proportions in independent samples. For this, see **confidence interval (CI) for difference between means/proportions**. We do not wish to repeat that here.

Attributable risk can also be obtained in case of **matched pairs**. Table C.27 illustrates one such application where the groups with therapy and without therapy are one-to-one matched. The *risk* of interest in this case is Relieved within 1 Week.

The attributable risk in this case is also the difference between proportions. This is measured in a peculiar way. For an explanation and the CI, see **confidence interval (CI) for difference between means/proportions**.

## confidence interval (CI) for correlation coefficient

See the topic **correlation coefficient** to apprise yourself with this statistical measure. The confidence interval (CI) on this depends on its **sampling distribution** as does any other CI. The central limit theorem can be invoked to stipulate a Gaussian distribution of $r$ for large $n$, where $r$ is the correlation coefficient in the sample and $n$ is the sample size. This is attained for moderate $n$ when the correlation coefficient (ρ) in the population is close to zero, since then the distribution of $r$ is symmetric if the underlying distribution of $x$ and $y$ is bivariate Gaussian. As the value of |ρ| becomes large, the sampling distribution of $r$ becomes increasingly skewed. Thus, a larger $n$ is

**TABLE C.27**
**Trial for Therapy for Common Cold: Matched Pairs**

| | Without Therapy (Control Group) | | |
|---|---|---|---|
| **With Therapy (Experimental Group)** | **Relieved within 1 Week** | **Not Relieved within 1 Week** | **Total** |
| Relieved within 1 week | 22 | 15 | **37** |
| Not relieved within 1 week | 5 | 8 | **13** |
| **Total** | **27** | **23** | **50** |

required for it to become nearly Gaussian. However, the standard error (SE) of $r$ is complicated. In view of these issues, the following method for finding CI on ρ is used.

This method is based on Fisher $z$-transformation given by $z_r = \frac{1}{2}\ln\frac{1+r}{1-r}$. This tends to become Gaussian as $n$ increases. The beauty of this transformation is $\text{var}(z) = \frac{1}{n-3}$. The corresponding population parameter is $\zeta_\rho = \frac{1}{2}\ln\frac{1+\rho}{1-\rho}$. Thus, for large $n$,

$$95\%\text{ confidence CI on } \zeta_\rho: \left(\text{L} = z_r - 1.96\sqrt{\frac{1}{n-3}},\ \text{U} = z_r + 1.96\sqrt{\frac{1}{n-3}}\right).$$

This CI is converted back to ρ as follows:

$$95\%\text{ CI on } \rho: \left(\frac{e^{2L}-1}{e^{2L}+1},\ \frac{e^{2U}-1}{e^{2U}+1}\right).$$

We illustrate this with the help of an example. Suppose you find $r = -0.62$ between shock index and quality of life index in a random sample of $n = 84$ subjects suffering from acute coronary syndrome. Since $n$ is large, we can use our method to find the CI. This gives $z_r = \frac{1}{2}\ln\frac{1-0.62}{1+0.62} = -0.725$ and standard error $= \sqrt{\frac{1}{84-3}} = 0.111$. Thus, the 95% CI on $\zeta_\rho$ is $-0.725 \pm 1.96 \times 0.111$, or (–0.943, –0.507). When converted back to ρ, this gives

$$95\%\text{ CI on } \rho: \left(\frac{e^{-2\times0.943}-1}{e^{-2\times0.943}+1},\ \frac{e^{-2\times0.507}-1}{e^{-2\times0.507}+1}\right), \text{ or } (-0.737, -0.468).$$

There are extremely few chances (<5%) that this interval will not contain the population correlation coefficient.

## confidence interval (CI) for difference between means/proportions

The interest in many situations is in the magnitude of the difference in proportions or in means of two groups under study. The types of differences that are of special importance in medicine are between a placebo and a drug, between drug 1 and drug 2, between males and females, etc. The procedure to obtain the **confidence interval (CI)** for the difference between means is different from the procedure for the CI for difference in proportions, and these procedures are different for independent samples and for matched pairs. The following is restricted to a setup that follows **Gaussian conditions**. These conditions generally require that the underlying distribution is not far away from **Gaussian** if the sample size is small. For small samples from non-Gaussian distribution, **nonparametric methods** are used.

### *Two Independent Samples*

In the case of sample means in two independent groups, the **standard error (SE)** of the difference in means is given by

$$\text{SE}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

where $s_1^2$ and $s_2^2$ are the sample variances in group 1 and group 2, and $n_1$ and $n_2$ are the respective sample sizes. If population variances are nearly equal (**Levene test** is used to check this), you can pool the variances of the two samples and get a better estimate. Then

$$\text{SE}(\bar{x}_1 - \bar{x}_2) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where $s_p = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ is the **pooled variance**. Then

$$\text{CI for difference in two means: } (\bar{x}_1 - \bar{x}_2) \pm t_v{*}\text{SE}(\bar{x}_1 - \bar{x}_2).$$

The value of the **degrees of freedom** for this $t$ is given by $\upsilon = n_1 + n_2 - 2$ and is obtained from the Student $t$ distribution at probability $(1 - \alpha/2)$.

For unequal variances, Welch $t$ is used. The CI remains the same, but the df of $t$ reduce to some extent by using a complex formula stated under **Welch *t***.

For difference in proportions, several forms of *estimated* SE are available [1]. But the following is widely used when both samples are large and independent of one another:

$$\text{SE}\,(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}},$$

where $p_1$ and $p_2$ are the proportions observed in the samples from groups 1 and 2, respectively, and the sample sizes are $n_1$ and $n_2$. Now, under Gaussian conditions,

$$95\%\text{ CI for difference in two proportions: } (p_1 - p_2) \pm 1.96{*}\text{SE}(p_1 - p_2).$$

For illustration, consider the following example. In a randomized controlled trial, patients with peptic ulcers were put on two treatment regimens: one based entirely on drugs, and the other based on minimal drugs but supplemented by a change in lifestyle. In the first group, 12 out of 30 responded after a month, and in the second group, 28 out of 50 responded. How does one find the 95% CI for the difference in proportions in the two groups, and how does one explain its meaning?

In this example, $p_1 = 12/30 = 0.40$ and $p_2 = 28/50 = 0.56$. Since $n_1$ and $n_2$ are both sufficiently large, the general procedure as stated above can be used. The sample analog of population difference in proportions $(\pi_1 - \pi_2)$ is $(p_1 - p_2)$. We need to use its SE to get the CI.

$$\text{SE}\,(p_1 - p_2) = \sqrt{\frac{0.40\times0.60}{30} + \frac{0.56\times0.44}{50}} = 0.1137.$$

The 95% CI for $(\pi_1 - \pi_2)$ is

$$(0.40 - 0.56) - 1.96 \times 0.1137,\ (0.40 - 0.56) + 1.96 \times 0.1137$$

or (– 0.38, +0.06).

The 95% CI for the difference in the response rate is –0.38 to +0.06, but it also says there is a 5% chance that the difference will be outside these limits. The negative difference indicates that it is not unlikely in the long run that the response rate in the "change in the lifestyle" group will be even lower than that in the exclusive drug group.

Note also that $(\pi_1 - \pi_2)$ in some cases can be interpreted as **attributable risk (AR)**, and the same method is used for finding the CI for AR.

### *Matched Pairs*

Paired samples in medicine arise primarily in two situations: (i) in a **before–after study** with no parallel **controls** where the same subjects are measured twice—before and after the stimulus; and (ii) in a strict one-to-one **matching** where a parallel control group of different subjects is present, but each control subject is matched to the corresponding case for nearly all the characteristics that can affect the outcome, except of course the stimulus itself. The control group is not necessarily a **placebo**. It can be an active control that has received another stimulus under comparison. For matched controls, mild matching by one or two characteristics such as age and sex generally is not considered adequate for paired analysis—instead such control group is considered independent and analyzed by the methods previously described in this section.

The procedure for calculating the CI for mean difference in the case of matched pairs is the same as that for the mean of one sample after obtaining the differences. The mean and SD of these differences can be obtained, and for the CI, see **confidence interval (CI) for mean**.

Proportions in the case of paired samples are not so straightforward. The responses in this case would be as given in Table C.28. The counts are of subjects who remain at the same level after the stimulus and those who change.

Consider a sample of $n = 70$ patients with kidney disease. They were dialyzed by the regular machine for 1 week and 49 (70%) showed significant improvement. The following week, they were dialyzed by a new machine and 56 (80%) showed significant improvement. Since the patients are the same, these are not independent samples, and the method for difference in proportions in independent samples is not applicable. These results are stated in Table C.29.

In this example, 40 patients responded well to both the machines. This is *a* in Table C.28. Suppose that 5 did not respond well to either machine. This is *d* in Table C.28. In a paired setup such as this, the difference in proportions is measured in terms of the ratio

$$d = \frac{|b-c|}{n},$$

**TABLE C.28**
**Counts for Paired Sample Response**

| Characteristics before the Stimulus | Characteristics after the Stimulus: Present | Characteristics after the Stimulus: Absent | Total |
|---|---|---|---|
| Present | *a* | *b* | ***a + b*** |
| Absent | *c* | *d* | ***c + d*** |
| **Total** | ***a + c*** | ***b + d*** | ***n*** |

**TABLE C.29**
**Example of Counts in Paired Setup**

| Significant Improvement with Machine 1 | Significant Improvement with Machine 2: Yes | Significant Improvement with Machine 2: No | Total |
|---|---|---|---|
| Yes | 40 | 9 | **49** |
| No | 16 | 5 | **21** |
| **Total** | **56** | **14** | **70** |

with

$$\text{SE}(d) = \sqrt{\frac{(b+c)n-(b-c)^2}{n^3}},$$

where the notations are the same as in Table C.28. Thus, for large samples,

95% CI for difference in proportions in paired setup:
$d \pm 1.96*\text{SE}(d)$

For the dialysis data in Table C.29, this CI is

$$\frac{|9-16|}{70} \pm 1.96\sqrt{\frac{(9+16)70-(9-16)^2}{70^3}},$$

or

$$0.10 \pm 1.96 \times 0.0704,$$

or $\quad (-0.04, 0.24).$

The improvement by machine 2 could be up to 24%, but machine 2 can give 4% less performance than machine 1. This result is not unequivocal for these machines.

1. Brown L, Li X. Confidence intervals for two sample binomial distributions. *J Stat Planning Inf* 2005;130:359–75. http://www.sciencedirect.com/science/article/pii/S037837580400271X

## confidence interval (CI) for mean

**Confidence interval** for mean is the range outside which the population mean is unlikely to lie. Consider a new herbal drug, which is tried on a group of 50 coronary disease patients and which reduced lipoprotein(a) level by an average of 9 mg/dL in 3 months. Because of **sampling fluctuation**, another group may show a very different reduction or no reduction at all. The CI delineates the limits of the likely values on average; in fact, it specifies values beyond which the average reduction is very unlikely.

The **central limit theorem (CLT)** tells us that the distribution of the sample mean is always nearly **Gaussian** for large *n*. The underlying distribution of measurements on individual subjects is then immaterial. The distribution of the duration of survival of patients after detection of leukemia is skewed and far from Gaussian. Yet, when mean survival times are obtained in many samples, each of large size, these means still follow a nearly Gaussian pattern. The property of the Gaussian distribution can be invoked to obtain 95% CI for μ as $[\bar{x} - 1.96\text{SE}(\bar{x}), \bar{x} + 1.96\text{SE}(\bar{x})]$. This gives

95% CI for μ (σ known): $(\bar{x} - 1.96 * \sigma/\sqrt{n},\ \bar{x} + 1.96 * \sigma/\sqrt{n})$. (C.3)

However, σ is rarely known. It is then replaced by the sample SD *s*. Because of this replacement, the Gaussian distribution can no longer be used, and we need to use **Student *t*-distribution** instead. Thus

95% CI for μ (σ not known): $(\bar{x} - t_v * s/\sqrt{n},\ \bar{x} + t_v * s/\sqrt{n})$, (C.4)

where $t_v$ is the value of *t* at $v = (n - 1)$ df. For 95% confidence, this corresponds to the probability of 0.975 so that the total probability in the two tails outside $\pm t_v$ is 0.05. The CI in the preceding equation is valid only when the underlying distribution is Gaussian, especially

for small $n$. For large $n$, this can be used even when the underlying distribution is not Gaussian because the distribution of $\bar{x}$ is still approximately Gaussian for such $n$. In other words, when the underlying distribution is Gaussian, use Equation C.3 for known σ and Equation C.4 for unknown σ, irrespective of $n$ being small or large. For non-Gaussian distribution, they are valid for large $n$ only. For small $n$, generally, **confidence interval (CI) for median** is obtained.

## confidence interval (CI) for median, *see also* exact confidence intervals (CIs)

The **confidence interval** (CI) for a population median is the range outside which it is unlikely to lie. The CI for median is generally obtained where mean is not a good measure of central value. The methods for obtaining the CI for underlying **Gaussian** and for non-Gaussian distributions are different.

### *CI for Median of a Gaussian Distribution*

When the underlying distribution is Gaussian, the standard error SE(median) = 1.253*σ/√$n$ and the sampling distribution of median is Gaussian. You can see that this SE is about 25% larger than the SE for mean. This indicates that sample median is not a precise estimate and should not be used in situations where mean can be used. If your interest remains firm with median because of specific interest in the middle value, when the underlying distribution is Gaussian and $n$ is large, use

95% CI for population median: sample median ±1.96*1.253*σ/√$n$.

If you do not know σ, replace it with sample SD $s$ and replace 1.96 by the $t$-value corresponding to the appropriate df as in the case of computing **confidence interval (CI) for mean**. Realize though that you may never use median and may never need CI for median when mean is appropriate. This CI is given here only for the sake of completeness.

### *CI for Median under Non-Gaussian Conditions*

Median becomes relevant when the measurements follow a highly skewed pattern, as often seen in sick subjects. Median is not a summation-type statistic like mean is, and thus the **central limit theorem** does not generally apply, and large $n$ does not help median to attain Gaussianity. Thus, a different method is needed if the underlying distribution is far from Gaussian, particularly if $n$ is small. For example, it is known that the distribution of serum glucose level in diabetics is **right-skewed** and that of Hb level in anemics is **left-skewed**. (In both these cases, it may be truncated depending on the threshold used for defining diabetes and anemia, respectively.) When no such a priori information about the form of the distribution is available, it is difficult to judge the shape of the distribution from small samples. When the distribution is known or suspected to be far from Gaussian, the methods to be used for small $n$ are given for **exact confidence intervals (CIs)**. These methods do not depend on the shape of the underlying distribution.

## confidence interval (CI) for odds ratio (OR)

You may like to review the topic **odds ratio** (OR) if the basics of OR are not clear. We use the same notations. This section is divided into two parts—for independent samples and for matched pairs. The OR is generally calculated for case–control studies, and the two groups are independent samples when the cases comprise a different group of subjects than the control group, and there is not much matching. If the cases and controls are one-to-one matched for many antecedent factors or those who are naturally matched such as twins, the method of CI is different and is described separately in this section.

### *Confidence Interval for OR (Independent Samples)*

OR is a ratio and its natural logarithm (ln) is a linear function. The distribution of OR can be shown to be highly skewed, but lnOR has a nearly Gaussian pattern for large $n$. It has been established for large $n$ that the estimated

$$\text{SE (ln OR)} = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}},$$

where $a$, $b$, $c$, and $d$ are as follows:

- $a$ = number of subjects with presence of the antecedent factor among those with positive outcome (cases)
- $b$ = number of subjects with absence of the antecedent factor among those with positive outcome (cases)
- $c$ = number of subjects with presence of the antecedent factor among those with negative outcome (controls)
- $d$ = number of subjects with absence of the antecedent factor among those with negative outcome (controls)

For a sufficiently large sample, none of these numbers would be zero or small. If any of these is zero or small despite large $n$, add ½ to each of these in the denominators in the formula of SE given above. For large $n$,

95% CI for OR: exp[lnOR ± 1.96*SE(lnOR)],

where exp is the exponent on Naperian base $e$. Similarly, one-sided **confidence bound** can also be obtained.

For the data in Table C.30 on average extramarital contacts among HIV+ and controls, SE (ln OR) = $\sqrt{1/38+1/42+1/51+1/109} = 0.2809$. The sample size $n$ is sufficiently large in this case, and a Gaussian distribution of lnOR is expected. Thus, the 95% CI for OR is

exp[ln(1.93) ± 1.96 × 0.2809],

or ($e^{0.107}$, $e^{1.208}$), or (1.11, 3.35).

The odds ratio in this sample is ($ad$)/($bc$) = 38 × 109/(51 × 42) = 1.93, but it can well range from 1.11 to 3.35 in the target population. If only the one-sided (lower) bound with 95% confidence is required, this is exp[ln(1.93) – 1.645 × 0.2809] = 1.22. It can be stated with 95% confidence that the OR in the population is not less than 1.22.

### *Confidence Interval for OR (Matched Pairs)*

In the case of matched pairs, the odds ratio is calculated as odds ratio (matched pairs): $OR_M = B/C$, where $B$ and $C$ are the disconcordant

**TABLE C.30**
**Extramarital Contacts in HIV+ Subjects and Controls**

| | Average Extramarital Contacts/Month | | |
|---|---|---|---|
| **Group** | **One or More** | **Less Than One** | **Total** |
| HIV+ | 38 | 42 | **80** |
| Controls | 51 | 109 | **160** |

pairs. The distribution of $\ln OR_M$ is also nearly Gaussian for large $n$. This also implies that no $B$ or $C$ is small, say less than 5, where $B$ and $C$ are the disconcordant pairs of the two types. Statisticians have established for large $B$ and $C$ that

$$\text{SE}(\ln \text{OR}_M) = \sqrt{\frac{1}{B}+\frac{1}{C}}.$$

The 95% CI for log of odds ratio, as usual, is $\ln OR_M \pm 1.96$ SE($\ln OR_M$). In this case, this becomes

$$95\%\ \text{CI for } \ln \text{OR}_M:\ \ln\frac{B}{C} \pm 1.96 * \sqrt{\frac{1}{B}+\frac{1}{C}}\ .$$

Take the exponential of the limits and get

$$95\%\ \text{CI for OR in matched pairs: } \frac{B}{C}e^{\pm 1.96\sqrt{\frac{1}{B}+\frac{1}{C}}}.$$

While the CI for $\ln OR_M$ is symmetric, it is not symmetric for the OR itself. The following example illustrates the procedure to get the CI for $OR_M$ in case of matched pairs.

Consider a case–control study of births with multiple malformations. The malformations considered are cleft lip, cleft palate, anal atresia, heart defects, hypospadias, etc. They are considered multiple when at least two are present. The controls were one-to-one matched for birth order, maternal age, socioeconomic status, and the place of delivery. The objective is to find any excess of one gender over the other in such births. Suppose the data obtained are as shown in Table C.31.

In these data, the disconcordant pairs are $B = 60$ and $C = 50$. Thus,

$$\text{OR}_M = \frac{60}{50} = 1.20.$$

The odds are 1.2 times in these subjects that the malformed child is male and not female. Now,

$$\text{SE}(\ln \text{OR}_M) = \sqrt{\frac{1}{60}+\frac{1}{50}} = 0.1915.$$

Thus, the 95% CI for the odds ratio is

$$(1.2 \times e^{-1.96 \times 0.1915},\ 1.2 \times e^{1.96 \times 0.1915}),\ \text{or } (0.82,\ 1.75).$$

**TABLE C.31**
**Sex of Children with and without Multiple Malformations**

| | Matched Control | | |
|---|---|---|---|
| **Cases** | **Male** | **Female** | **Total** |
| Male | 70 | 60 | **130** |
| Female | 50 | 40 | **90** |
| **Total** | **120** | **100** | **220** |

## confidence interval (CI) for predicted *y* in simple linear regression

Predicted $y$ ($\hat{y}$) refers to an individual opposed to estimated $y$ that pertains to the mean of the group of subjects with specified $x$'s. The standard error (SE) of $\hat{y}$ is different than it is for the mean of $y$. Prediction of individual values has a much larger SE. These SEs have complex forms for multiple regression, but they can be easily stated for a simple linear regression. Sometimes software packages do not provide these SEs, and you may have to calculate these yourself. Under certain general conditions, for simple linear regression of $y$ on $x$,

$$\text{SE (estimated mean of } y_x) = \sqrt{\text{MSE}\left[\frac{1}{n}+\frac{(x-\bar{x})^2}{\Sigma(x-\bar{x})^2}\right]},$$

and

$$\text{SE (predicted individual value of } y_x) = \sqrt{\text{MSE}\left[1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\Sigma(x-\bar{x})^2}\right]},$$

where $y_x$ is the value of $y$ at given $x$. MSE is always generated by the software. If the dependent of interest is duration of analgesia ($y$) induced by different doses of a drug ($x$), and $n = 15$, MSE = 6.50, $\bar{x} = 11.2$ μg, and $\Sigma(x-\bar{x})^2 = 18.07$, then for $x = 8$, SE(estimated mean of $y$ at $x = 8$) = $\sqrt{[6.50\{1/15+(8-11.2)^2/18.07\}]} = 2.03$. The SE(predicted individual value of $y$ at $x = 8$) = $\sqrt{[6.50\{1+1/15+(8-11.2)^2/18.07\}]} = 3.26$. The SE of the predicted value is much higher. Prediction of the individual value of $y$ is always less precise than the prediction of the mean of $y$.

These SEs can be used to generate the CI for an estimated value of $y$ and for a predicted value of $y$. Under Gaussian conditions, these are

95% CI for estimated mean of $y$: $\hat{y}_x \pm 1.96$*SE(estimated mean of $y_x$)

95% CI for predicted $y$: $\hat{y}_x \pm 1.96$*SE(predicted value of $y_x$)

In our example, suppose $\hat{y}$ from the simple linear regression at $x = 8$ is 10 h. Then

95% CI for estimated mean duration of analgesia at dose = 8 μg: $10 \pm 1.96 \times 2.03$

or

(6.02, 13.98) h

and

95% CI for predicted duration of analgesia for one patient at dose = 8 μg: $10 \pm 1.96 \times 3.26$

or (3.61, 16.39) h.

For duration of analgesia, the mean would be useful for comparing the efficacy of one drug with the other. In this example, the prediction interval for one patient may provide more useful information. For one patient, the interval from 3.61 to 16.39 h is too wide for any worthwhile conclusion. This highlights the limitation of prediction by regression—a limitation that is many times ignored.

C

## confidence interval (CI) for proportion, *see also* exact confidence intervals (CIs)

Confidence interval for population proportion $\pi$ is the range of values outside of which it is unlikely to lie. The procedure of obtaining the CI for proportion is easy for large $n$ but is relatively difficult for small $n$. We describe both in this section.

### *CI for Proportion $\pi$: Large* n

When $n$ is sufficiently large and $p$ not too small such that $np > 8$ and $n(1-p) > 8$, the sample proportion $p$ has an approximately Gaussian distribution. For this situation, a property of the Gaussian distribution is invoked to calculate the CI for $\pi$:

$$95\%\ \text{CI for } \pi \text{ (large } n\text{)}: p \pm 1.96*\text{SE}(p)$$

or

$$\left(p - 1.96\sqrt{\frac{p(1-p)}{n}},\ p + 1.96\sqrt{\frac{p(1-p)}{n}}\right),$$

where SE($p$) stands for standard error of $p$. In the long run, if repeated samples were taken, this CI would include the proportion $\pi$ in the target population. A property of the Gaussian distribution says that a distance of 1.96*SE on either side of $\pi$ would contain $p$ in 95% of samples. This is transformed to the statement that 1.96*SE on either side of $p$ is not an unlikely range for the value of the parameter $\pi$. Note the reverse direction of this statement.

The limits in 95% CI are sometimes called ±2SE limits as approximation. The lower limit is 2SE less than $p$, and the upper limit is 2SE more than $p$. It is customary to use the estimate of SE in place of SE itself because the actual SE would not be known. This can be safely done when $n$ is large, but not for small $n$. The multiplier 2 is an approximate value of exact 1.96. For the confidence levels other than 95%, this multiplier will change as per the Gaussian values.

Consider the following example. The management of cases of bronchiolitis in infants may become easier if somehow the course of the disease can be predicted on the basis of the condition at the time of hospital admission. One simple criterion for this could be the respiration rate (RR). Consider an investigation in which cases with RR ≥ 68 per minute are observed during their stay in the hospital. Suppose in a random block of 80 consecutive cases of bronchiolitis coming to a hospital with RR ≥ 68, a total of 51 (64%) are ultimately observed to have had a severe form of the disease, i.e., they either had a prolonged stay in the hospital, developed some complication, required endotracheal intubation or mechanical ventilation, or died. This 64 is the percentage observed in the present sample. Another sample from the same hospital may give a different percentage. What could be the percentage of cases with a severe form of the disease in the entire population of patients admitted to the hospital with a diagnosis of bronchiolitis and RR ≥ 68 per minute?

The best **point estimate** according to this sample is 64%. However, this estimate is likely to differ from the actual percentage in the whole population or when another sample is taken. Since $n$ = 80 and $p$ = 0.64, $np$ and $n(1-p)$ are large enough for an approximately Gaussian pattern. The ±1.96*SE limits in this example are

$$p - 1.96\ \text{SE}(p) = 0.64 - 1.96\sqrt{0.64 \times 0.36/80} = 0.53$$

and

$$p + 1.96\ \text{SE}(p) = 0.64 + 1.96\sqrt{0.64 \times 0.36/80} = 0.75.$$

Thus, the percentage with a poor prognosis can be anywhere between 53 and 75 in cases of bronchiolitis with RR ≥ 68. In other words, there is a chance between 53% and 75% that a case of bronchiolitis with RR ≥ 68 per minute at the time of hospitalization will require special handling.

Suppose that 6% of those with RR ≥ 68 per minute in the preceding example fail to survive. The 95% CI for the proportion dying is

$$(0.06 - 1.96\sqrt{0.06 \times 0.94/80}, 0.06 + 1.96\sqrt{0.06 \times 0.94/80}),$$

$$\text{or} \quad (0.01, 0.11).$$

Thus, the actual fatality rate could be anywhere between 1% and 11%. This is a wide interval relative to the case fatality of 6% observed in the sample in the sense that the case fatality could, in fact, be nearly double of what was actually observed. Compare it with the (53%, 75%) interval obtained earlier for cases with poor prognosis. This interval is narrow relative to the 64% rate observed in the sample. In general, the CI is narrow relative to $p$ when $p$ is around 0.5, say between 0.3 and 0.7, and it is wide relative to $p$ when $p$ is either very low or very high.

Other points to remember are as follows:

- Where the study group size is small and the proportion of interest is also small, use exact methods based on **binomial distribution** as mentioned next in this section.
- CI as just stated is valid only when the observed sample proportion $p$ is neither zero nor one. Confidence bounds for such extreme values are given by the **Clopper–Pearson bounds**. Also, if $p$ is too small or too large, the CI as just stated can yield a lower limit less than 0 or an upper limit greater than 1. This clearly is not possible for $\pi$. In such cases, it is customary to keep 1 as the upper limit and 0 as the lower limit, although this amounts to an approximation. Examine whether lower or upper **confidence bounds** as discussed separately are more appropriate in such cases.
- Strictly speaking, a 95% CI implies the probability is 0.95 that such a *random* interval contains the value of the parameter. The value of the parameter is fixed. However, in practical applications, 95% CI is interpreted as though this interval contains the parameter with probability 0.95. This is a loose statement but helps to grasp the essential feature of a CI. This book adopts an easy but not so accurate course and interprets 95% CI on a more practical basis as the interval that contains the parameter with probability 0.95.

**Sensitivity**, **specificity**, positive **predictivity**, and negative predictivity are all proportions, and their CIs are obtained as is usually done for proportions. On the contrary, parameters such as **relative risk (RR)** and **odds ratio (OR)** are ratios and they require a **logarithmic** transformation as described separately.

### *CI for Proportion $\pi$: Small* n

Suppose the interest is in estimating the chance of uterine prolapse in women who come with complaints of micturition disturbance and vaginal discharge. If only $n$ = 12 women with such complaints could be examined and 3 had uterine prolapse, the proportion is 3/12 = 0.25. Another sample may give another proportion. The quantity $\pi$ in this case is the actual proportion of women with uterine prolapse among the population of women with those complaints. How does

one find limits within which this proportion is not unlikely to lie in all such patients? In this case, since $np = 3$ is small, the Gaussian approximation cannot be used. The **exact** confidence interval for $\pi$ in case of small $n$ is obtained by using a **binomial distribution**. Instead of going through the rigmarole of complex mathematics, it can be obtained graphically by using the figure given for **Clopper–Pearson bounds/interval**. This involves some approximation but is still useful for practical applications.

## confidence interval (CI) for regression coefficient and intercept in simple linear regression

The confidence interval for **regression coefficient** is the range outside of which the actual value of this coefficient is unlikely to lie. As all sample estimates, the estimates of regression coefficients are also subject to **sampling fluctuation**. They also have a **standard error (SE)**, which provides a measure of their precision. This is used to obtain the CI under **Gaussian conditions**.

Most of the standard statistical software packages readily provide various SEs corresponding to the data entered for regression. Each regression coefficient $b_0$, $b_1$, $b_2$, …, $b_K$ will have its own SE, and the regression-estimated $\hat{y}$ also has an SE. Under the usual conditions, particularly when $n$ is large, these estimates follow a Gaussian pattern. Thus, a CI can be constructed. A standard statistical software package will readily provide these CIs. The expression for SEs in case of multiple regression is too clumsy. For those interested, the estimated SEs for simple linear regression $y = a + bx$ are as follows:

$$\text{SE}(a) = \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x-\bar{x})^2}\right)} \quad \text{and} \quad \text{SE}(b) = \sqrt{\frac{\text{MSE}}{\Sigma(x-\bar{x})^2}},$$

where MSE = SSE/($n$ – 2) and SSE is the **error sum of squares**. If $\alpha$ is the intercept and $\beta$ is the slope parameter in the population, under Gaussian conditions,

$$\text{CI for } \alpha\text{: } a \pm t_\nu * \text{SE}(a)$$

$$\text{CI for } \beta\text{: } b \pm t_\nu * \text{SE}(b),$$

where $t_\nu$ is the value from Student $t$ distribution at $\nu$ df corresponding to the desired level of confidence. In this case, $\nu = n - 2$.

For further details, see Chatterjee and Hadi [1].

1. Chatterjee S, Hadi AS. *Regression Analysis by Examples*, Fifth Edition. Wiley, 2012.

## confidence interval (CI) for relative risk (RR)

It would not be wrong to surmise that all users of statistical methods utilize software to perform the calculations. Nearly all standard statistical software packages do have the provision to calculate the RR and the CI for the RR corresponding to the prefixed confidence level. Nevertheless, the following details may be helpful in understanding the underlying procedure.

Suppose there are two populations in which the probabilities that an individual shows an outcome of interest are $\pi_1$ and $\pi_2$, respectively. Suppose also that a random sample of size $n_1$ from the first population has $a$ subjects showing the outcome (and a proportion $p_1 = a/n_1$), while the corresponding values for an independent sample from the second population are $n_2$, $b$, and $p_2 = b/n_2$, respectively. Then the estimated RR is the simple ratio of these proportions. Thus, lnRR (natural logarithm of RR) is a linear combination of the frequencies. The **central limit theorem** works not just for mean but for any linear combination. Because of this, lnRR has a Gaussian distribution for large $n$. It is thus easy to use lnRR for inference when $n$ is large. It is known for large samples that

$$\text{SE (lnRR)} = \sqrt{\frac{1}{a} - \frac{1}{n_1} + \frac{1}{b} - \frac{1}{n_2}}.$$

Therefore,

$$95\%\ \text{CI for RR: exp[lnRR} \pm 1.96 * \text{SE(lnRR)]},$$

where exp is the exponent to the Naperian base $e$ and is the inverse of the logarithm. As you can see, the CI is first obtained for lnRR, and then the exponent is taken to convert it to RR. The CI for lnRR will be symmetric, but this interval will not be symmetric to RR when exp is taken. Note, however, that the CI 0.125–0.5 when estimated RR is 0.25 has the same implication as CI 2–8 when estimated RR is 4. The width looks larger in the latter case, but both are one-half to two times of the estimated RR. The width of CIs for ratios is tricky. Ratios make better sense when examined on a log scale.

Martinez et al. [1] reported a prospective study on wheezing lower respiratory tract illness (LRI) during the first year of life of 500 boys and an almost equal number of girls. They were enrolled in Tucson, Arizona. One of the objectives was to explore maternal age as a risk factor for wheezing LRI. The data for the boys are given in Table C.32.

The outcome is the LRI. The estimated relative risk of wheezing LRI in boys with their mothers' age < 26 years compared to those with their mothers' age ≥ 26 years is

$$\text{RR} = \frac{48/165}{65/335} = 1.50.$$

Thus, boys born to younger women were 1.5 times likely to get LRI during their first year of life as compared to boys born to older women. Also, lnRR = ln(1.50) = 0.4055, and

$$\text{SE(lnRR)} = \sqrt{\frac{1}{48} - \frac{1}{165} + \frac{1}{65} - \frac{1}{335}}$$

$$= 0.1648.$$

from the formula given earlier.

Therefore, 95% CI for RR is

$$\exp(0.4055 \pm 1.9\,6 \times 0.1648)$$

$$\text{or} \quad (e^{0.082}, e^{0.729}) \quad \text{or} \quad (1.09, 2.07).$$

**TABLE C.32**
**Maternal Age and Lower Respiratory Tract Illness (LRI) in Infant Boys**

| Maternal Age (Years) | LRI | | Total |
|---|---|---|---|
| | Yes | No | |
| <26 | 48 | 117 | **165** |
| ≥26 | 65 | 270 | **335** |

It can be stated with 95% confidence that this interval contains the true RR in the population provided that the sample subjects can be considered random representatives.

1. Martinez FD, Wright AL, Holberg CJ, Morgan WJ, Taussig LM. Maternal age as a risk factor for wheezing lower respiratory illnesses in the first year of life. *Am J Epidemiol* 1992 Nov 15;136(10):1258–68. http://www.ncbi.nlm.nih.gov/pubmed/1476148

C

## confidence interval (CI) for variance

Though rarely used, CI for the population variance can be easily obtained, particularly when the sample is from a Gaussian distribution. It is well known for Gaussian distribution that the criterion $\frac{(n-1)s^2}{\sigma^2}$ follows a chi-square distribution with df = ($n$ – 1), where $n$ is the sample size, and $s^2$ and $\sigma^2$ are the usual notations for the sample and population variance, respectively. Thus, $P\left(\chi_i^2 \le \frac{(n-1)s^2}{\sigma^2} \le \chi_u^2\right) = 1-\alpha$, where $\chi_i^2$ is obtained from the chi-square distribution with df = ($n$ – 1) with probability α/2 in the left tail and $\chi_u^2$ has α/2 probability in the right tail. Statistical software packages would easily get these values; otherwise, one should take help of a chi-square table in statistics books. For 95% CI, these will have 2.5% probability in either tail. This statement can be converted to provide the required CI for $\sigma^2$. That is,

$$\text{95\% CI for variance: } \left(\frac{(n-1)s^2}{\chi_u^2}, \frac{(n-1)s^2}{\chi_i^2}\right).$$

If a sample of $n$ = 18 subjects gives $s^2$ = 30.00 for fasting blood glucose level, the 95% CI for variance is $\left(\frac{17\times 30}{30.19}, \frac{17\times 30}{7.56}\right)$ or (16.89, 67.46). The value 30.19 is the chi-square values at 17 df corresponding to 2.5% probability in the right tail, and 7.56 is the chi-square value with 2.5% probability in the left tail. You can see that this CI is not symmetric to the sample variance of 30.19 mg/dL$^2$.

Although theoretically not sound, the square root of these limits is considered 95% CI for the population standard deviation (SD) σ. In our example, these limits are (4.1, 8.2) mg/dL. The sample SD is $\sqrt{30.19} = 5.49$.

## confidence interval (CI) versus test of significance

There is a considerable debate about the actual utility of the concept of **statistical significance**. Two aspects in particular are discussed in this connection. The first pertains to the relationship between the **confidence interval (CI)** and **tests of significance**, and which of these is better for inference. This is the subject matter of this section. The second concerns **medical significance versus statistical significance**. Sometimes these two significances may be in conflict. A way to resolve this is suggested under that topic.

It may not be immediately evident that CI and the statistical tests are intimately related. In fact, they have a one-to-one correspondence. Consider the example where 8 births in a random sample of 60 are observed to be premature. The 95% CI for the proportion of premature births, as per the procedure given under **confidence interval for mean/proportion**, is

$$\frac{8}{60} - 1.96\sqrt{\frac{(8/60)(52/60)}{60}}, \frac{8}{60} + 1.96\sqrt{\frac{(8/60)(52/60)}{60}}$$

$$= (0.047,\ 0.219).$$

That is, there is a 5% chance that the actual percentage of premature births is less than 4.7 or more than 21.9. Any percentage between 4.7 and 21.9 is not unlikely. In other words, any percentage between 4.7 and 21.9 is plausible, whereas any percentage outside these limits is not. Thus, this sample does not provide sufficient evidence to reject the **null hypothesis** at **significance level** α = 0.05 if the hypothesized value of π is between 0.047 and 0.219, and it will reject the hypothesis if the hypothesized value is either less than 0.047 or more than 0.219. This is how the two procedures are equivalent.

A test of hypothesis, which is also called a test of significance, can be alternatively performed as follows. Calculate the CI corresponding to the level of confidence (1 – α = 0.95 or any other) you desire. Check whether the value of the parameter under null hypothesis ($H_0$) falls within the CI. If yes, do not reject $H_0$; otherwise, reject it. The level of significance for this test is α—the complement of the confidence level. If the test is to be carried out at α = 0.02, then obtain a 98% CI and check whether this contains or does not contain the hypothesized value of the parameter. Conversely, set the value of the test criterion equal to its critical values (upper and lower) at the desired level, and solve for the values of the parameter. That will give the required 100(1 – α)% CI. For a two-sample situation, if CIs for individual group means do not overlap, the sample means are significantly different. However, the converse is not true. It can be demonstrated that the CIs overlap, yet the two means are significantly different. An example quoted frequently is means = 17 and 9 based on independent large samples with **standard error (SE)** for each = 2.5. The 95% CIs are (12.1–21.9) and (4.1–13.9), respectively. These overlap, but the two-sample **Student** $t = (17 - 9)/\sqrt{[(2.5)^2 + (2.5)^2]} = 2.26$, providing evidence of statistical significance ($P$ < 0.05) of the difference.

The first point of the debate arises from the equivalence of CI with the corresponding test of significance. If both are equivalent, which one is preferable? Many consider CI a better procedure than the test because CI gives the spectrum of unacceptable values that are outside the interval. Marginal values, both inside and outside CI, can be identified. It is thus possible to exercise more caution when the hypothesized value is on the margin. A CI provides a range of probable values of a population parameter rather than a dichotomy as significant or not significant the way a statistical test does. A test of significance considers only the null value, whereas medical implications of the range of plausible values may be an important element to reach a valid conclusion.

There is another view that supports CIs. They use inductive logic. From the observed data, you come up with a range that is not unlikely. Thus, CIs can generate new hypotheses and provide new learning. They do not depend on the null hypothesis, which is assumed true for calculating $P$-values. In that sense, statistical tests use deductive logic of inference. Deductive logic is easy, direct, objective, and definitive, but it does not expand our knowledge. Inductive knowledge provided by CIs expands the horizon and is not limited to preconceived values, although it tends to be tentative.

On the other hand, the statistical tests have particularly valid applications in situations in which the present knowledge or a claim is to be refuted. If a decision is to be taken one way or another on the basis of a statistical result alone, perhaps the test of hypothesis provides clear evidence. The statistical tests also have a special place in the comparison

of two or more groups where the only objective is to find out whether they are different and the exact effect is not of immediate concern. Some feel lost without a ***P*-value**, which comes only for the statistical test. In a **regression** setup, the test of $H_0$ helps to decide whether to keep a variable or not as a suitable predictor. For further details of the debate about CI versus test of hypothesis, see Gardner and Altman [1].

The scientific community tends to accept new (data-based) findings when their statistical significance is adequately demonstrated. A $P < 0.05$ is generally accepted. Thus, $P = 0.06$ receives the same fate as $P = 0.30$. These two are clearly different and should have different implications. Note that the *P*-value is the probability of **Type I error**, i.e., rejecting $H_0$ when it is true. A 6% chance of error is higher than the threshold 5%, yet it is fairly low. A 30% chance of error is very high. Similarly, a *P*-value 0.03 is different from a *P*-value of 0.001. Using a cut point of 0.05 as is sometimes done for testing of hypothesis masks this difference. Thus, another point of debate is whether *P*-values, as they are, should be stated or only a cutoff such as 0.05 should be stated. The former is a more exact way of describing the situation, but the latter is simple to understand. Opinion is getting around to stating the exact *P*-values. However, as mentioned earlier, a cutoff is required in any case to make a decision one way or another without forgetting that *P*-values can provide graded evidence instead of simply the binary yes or no.

1. Gardner MJ, Altman DG. Confidence intervals rather than P-values: Estimation rather than hypothesis testing. *BMJ (Clin Res Ed)* 1986;292:746–50. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1339793/

## confidence level, *see* confidence intervals (the concept of)

## confounders

A confounder or a confounding factor is an **antecedent** characteristic that can be described as a possible explanation of the outcome in addition to the antecedent under investigation. Thus, it is an extraneous factor that plays spoilsport. It can also be understood as the one that is related to the antecedent as well as to the outcome but is not in the causal chain under consideration. It mixes the association of disinterest with the ones of interest and blurs the causal pathway. Necessary but not sufficient criteria for a factor being a confounder are (i) it causes the outcome, (ii) it is associated with exposure, and (iii) it is not a descendent of exposure or outcome. One simple way to identify a confounder is to imagine that if there is no exposure under study, the association between the suspected factor and the outcome will remain or will vanish. If the association still remains, the factor is a confounder.

In a study on smoking and hypertension, one confounder is obesity (Figure C.23). Smokers tend to be obese, and hypertension is also related to obesity. In other words, obesity can also be at least a partial explanation for hypertension in addition to smoking, and it is not in the causal chain under study since smoking does not cause obesity. The effect of smoking and obesity on hypertension cannot be disaggregated unless obese and nonobese subjects are separately studied. The second confounding factor in this example is age. As age increases, the lifelong burden of smoking increases for smokers, and the chance of developing hypertension also increases because of age-related arterial changes. Thus, again, age as a possible explanation should be ruled out to find net effect of smoking on hypertension.

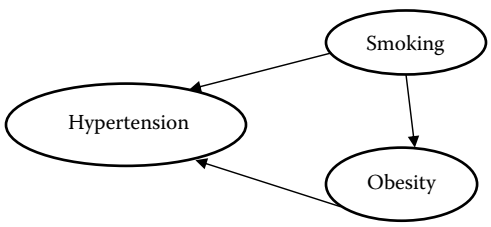



**FIGURE C.23** Obesity as a confounding factor for smoking and hypertension.

Confounders should be identified before data are collected for any research. Previous research, clinical insight, and clarity about the disease process can help in this identification. Age and sex are potential confounders in almost all medical setups. One easy method of identifying confounders is to draw a list of all possible factors that might influence the outcome of interest. The list may be based on your own knowledge, wisdom of seniors, or review of literature. Out of this list, choose the ones that would be studied as risk factors or antecedents for their role in the outcome. Among whatever remains in the list, confounders for that outcome are those that are related to any of the antecedents. This is different from a covariate. For relationship between smoking and hypertension, cholesterol level is a **covariate** but not a confounder assuming that cholesterol level has no relationship with smoking. On the other hand, sex is a confounder as it affects smoking as well as hypertension.

Note, however, that the confounders so identified would be restricted to those that are known. The knowledge could be incomplete, and the list may not be comprehensive. If so, epistemic uncertainties would remain in the results. Nothing can be done to remove this lacuna except to expand the horizon and look for factors that are not in the conventional domain.

The effect of confounders on the results can be eliminated either by developing a proper design or by performing a suitable statistical analysis. Any one of the following specific steps will help in controlling the effect of confounders:

1. Stratify the subjects by the level of the confounder. In our example on smoking and hypertension, stratify the subjects by levels of obesity. This will tell you the relationship between smoking and hypertension in obese subjects and in nonobese subjects separately.
2. Limit the study to one particular group of confounder such as only obese subjects in our example. At least for this group, you will have clear results.
3. Conduct multivariable analysis such as ordinary and logistic regression. In our example, when obesity is also entered as a regressor along with smoking, the effect of smoking will be automatically adjusted for any effect of obesity. In a regression setup, you can see that confounding is closely related to **multicollinearity**. If some of the regressors are highly correlated with one another, the results of regression are less reliable. Thus, confounding among regressors should not too pronounced.
4. Note also that these strategies can be adopted only when information on the confounding factor is available. In fact, the first two strategies require that this information should be available even before the data are collected at the planning stage. The third strategy can be adopted when the data on the confounding factor are collected along with the other data. When stratification is done, it should be sufficiently fine so that the strata are homogeneous. For

C

example, age into ≤45 and 45+ may not be fully effective. Also, in case of stratification, the result would be available separately for each stratum while the objective could be to get a combined result. Methods such as **Mantel–Haenszel** can be used to get a pooled estimate of the effect that gives larger weight to the bigger stratum. This does not work if **interactions** are present. In that case, do not pool the strata but present results separately for each stratum. In case the regression strategy is adopted, substantial difference between crude and adjusted effect implies that the factor is indeed a confounder. In this analysis, the adjusted effect will be the valid estimate as it removes the effect of the confounders.

## conjoint analysis

Conjoint analysis is done to find how preferences of the users are affected by the combination (when considered jointly) of the characteristics of the product. For hospital treatment, some prefer a big hospital because they are likely to have highly specialized doctors and latest gadgets, whereas others are happy with small hospitals where the care is more personalized; some would like to be treated by super-specialists, whereas for others, cost is a big constraint and do not mind being treated by general physicians; some would like shared rooms for social reasons and others like single room; some would like domiciliary care and others institutional care; and so on. In most cases, it is the combination of these various characteristics that determines whether a patient would use a particular facility or not. How does one find which factor is more important for a group of patients than the other factors? The answer is through conjoint analysis. The method is rooted in conjoint measurements proposed by Luce and Tukey [1] in 1964.

For conjoint analysis, the data collected from the users are in the form of their ranking for various combinations of the characteristics of the type shown in Table C.33. In this table, we have considered three **binary** characteristics of the treatment facility so that there are a total of $2 \times 2 \times 2 = 8$ combinations. Three patients were asked to rank these combinations from 1 to 8, from least preferred to most preferred. Patient 1 has the lowest preference for big hospital, treatment by a generalist, and a single room in the hospital. Patient 2 has the lowest preference for big hospital, treatment by a generalist, and a shared room. Similarly for other patients. On the basis of these data, can we find whether hospital size, the type of doctor, or type of room is more important for these patients?

The initial analysis is done separately for each patient. The first step obviously is to find the average rank of big hospital and the average rank of small hospital by patient 1, then by patient 2, and then by patient 3. For example, the average rank of small hospital by patient 2 is $(8 + 3 + 4 + 6)/4 = 5.25$, and the average rank of big hospital by this patient is $15/4 = 3.75$. For this patient, small hospital is more important than big hospital. If you repeat the exercise for patient 3, you will find that this patient also has preference for small hospital. We can find the average of these averages over the patients and find how different the ranking of the big hospital is from the ranking of the small hospital. The same can also be done for the type of doctor and the type of room. However, this simple procedure is not favored because this does not consider the effect of the combination. Our objective is to find which of these characteristics is a more important determinant when considered together with others, or to list them in order of their importance. The procedure described below is a simplistic version of an intricate method and is illustrated in Table C.34 for the data in Table C.33.

For each subject separately proceed as follows:

(i) Find the average rank for each level of each characteristic. Find its deviation from the mean. Call this $d_{ij}$ for subject $i$ and combination $j$ ($i = 1, 2,\ldots, n$; $j = 1, 2,\ldots, J$). In our example, $n = 3$ patients and $J = 8$ combinations. The sum of $d_{ij}$ will be zero for each characteristic.

(ii) Note that, in our example, the number of combinations is 8 but the total number of levels of all the three characteristics is $2 + 2 + 2 = 6$. Calculate the standardizing value as $b$ = (total number of levels of the characteristics)/$\Sigma d_{ij}^2$. This value is 0.6957 in Table C.34 for patient 1.

(iii) Calculate standardized deviation $b^*d_{ij}$ and obtain $u_{ij} = \sqrt{|b^* d_{ij}|}$. This is called the part-worth utility and measures the utility of the level of the characteristic. The utility is obtained while ignoring the sign for square root, but the sign is put back. This is one of the many ways that part-worth utility is calculated.

(iv) Find the range of utility for each characteristic, $r$ = (max $u$ – min $u$) for that characteristic.

(v) Importance of the characteristic for this subject, $I = r/(\Sigma r)$. In Table C.34, for the size of hospital, this is $100 \times 1.1796/4.3728 = 26.98\%$.

Similar calculations are done for each subject. Their mean, standard deviation, etc., can be computed as usual for drawing inferences. For example, the mean of the importance of hospital size, doctor, and room will tell us the relative importance of these characteristics in determining the preference of the patients. There are

**TABLE C.33**
**Rank for Combination of Three Characteristics of the Treatment Facility by Three Patients**

| Combination No. | Hospital (Big/Small) | Specialist/ Generalist | Room (Shared/Single) | Rank Given to This Combination by | | |
|---|---|---|---|---|---|---|
| | | | | Patient 1 | Patient 2 | Patient 3 |
| 1 | Big | Generalist | Shared | 4 | 1 | 4 |
| 2 | Big | Generalist | Single | 1 | 7 | 7 |
| 3 | Big | Specialist | Shared | 6 | 2 | 1 |
| 4 | Big | Specialist | Single | 5 | 5 | 2 |
| 5 | Small | Generalist | Shared | 2 | 8 | 6 |
| 6 | Small | Generalist | Single | 3 | 3 | 8 |
| 7 | Small | Specialist | Shared | 7 | 4 | 3 |
| 8 | Small | Specialist | Single | 8 | 6 | 5 |

**TABLE C.34**
**Conjoint Analysis for Patient 1 in Table C.33**

| | Ranks | Ave. Rank | Deviation from Mean ($d_{ij}$) | Squared Deviation $\left(d_{ij}^2\right)$ | Standardized Deviation ($b^*d_{ij}$) | Part-Worth Utility ($u_{1j}$) | Range of Utilities ($r$) | Importance (%) |
|---|---|---|---|---|---|---|---|---|
| Hospital | | | | | | | 1.1796 | 26.98 |
| Big | 4, 1, 6, 5 | 4.00 | –0.50 | 0.25 | –0.3479 | –0.5898 | | |
| Small | 2, 3, 7, 8 | 5.00 | 0.50 | 0.25 | 0.3479 | 0.5898 | | |
| Doctor | | | | | | | 2.3592 | 53.95 |
| Generalist | 4, 1, 2, 3 | 2.50 | –2.00 | 4.00 | –1.3914 | –1.1796 | | |
| Specialist | 6, 5, 7, 8 | 6.50 | 2.00 | 4.00 | 1.3914 | 1.1796 | | |
| Room | | | | | | | 0.8340 | 19.07 |
| Shared | 4, 6, 2, 7 | 4.75 | 0.25 | 0.0625 | 0.1739 | –0.4170 | | |
| Single | 1, 5, 3, 8 | 4.25 | –0.25 | 0.0625 | –0.1739 | –0.4170 | | |
| **Sum** | | | **0** | **8.625** | | | **4.3728** | **100** |
| | | | **$b$ = 6/8.625 =** | **0.6957** | | | | |



several examples of useful applications of conjoint analysis to medical problems. Wilson et al. [2] used this to delineate patient preference attributes of multiple sclerosis disease-modifying therapies, and Espanol et al. [3] proposed improved immunoglobulin therapy for patients with primary immunodeficiency by conjoint analysis of attributes such as quality of life and views on treatment.


1. Luce RD, Tukey JW. Simultaneous conjoint measurement: A new types of fundamental measurement. *J Math Psychol* 1964;1:1–27. http://www.sciencedirect.com/science/article/pii/002224966490015X
2. Wilson LS, Loucks A, Gipson G, Zhong L, Bui C, Miller E, Owen M. Patient preference attributes of multiple sclerosis disease-modifying therapies: Development and results of a rating-based conjoint analysis. *Int J MS Care* 2015;Mar–Apr;17(2):74–82. http://ijmsc.org/doi/pdf/10.7224/1537-2073.2013-053
3. Espanol T, Prevot J, Drabwell J, Sondhi S, Olding L. Improving current immunoglobulin therapy for patients with primary immunodeficiency: Quality of life and views on treatment. *Patient preference and adherence*. 2014;8:621–9. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4014377/


## consecutive sampling

Consecutive sampling is the process of taking a block of eligible subjects occurring together in some sense in a specified population as the sample. These may be those who consecutively report in an outpatient department of a medical facility, say, every Tuesday in a year; those living in consecutive dwellings in a specified area; those admitted in a hospital from, say, 1st of April to 30th of September of a year; etc. There is no bar in defining the eligibility; for example, you may say that all consecutive patients with complaints of pain in the abdomen of age at least 40 years would be selected. Inclusion and exclusion criteria will apply.

Questions are sometimes raised about the validity of this kind of sampling because consecutive subjects can be of similar nature. Whereas those living in consecutive houses may really not represent the cross-section of the population, generally those that consecutively come to a hospital are adequate representative of hospital patients since their order of arrival has little to do with the type or severity of cases if there is no epidemic. Thus, these can be considered as a random sample. Sansone et al. [1] studied borderline personality disorder in consecutive patients of age at least 18 years undergoing cardiac stress testing in a community hospital from June 6, 2010 to September 3, 2010. Hoffman et al. [2] have even taken a consecutive sample of reports of randomized trials on nonpharmacological interventions published in 2009 in six leading general medical journals for studying the adequacy of description of the interventions. Thus, this kind of sampling is widely accepted.


1. Sansone RA, Dittoe N, Hahn HS, Wiederman MW. The prevalence of borderline personality disorder in a consecutive sample of cardiac stress test patients. *Prim Care Companion CNS Disord* 2011;13(3). pii: PCC.10l01087. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3184559/
2. Hoffmann TC, Erueti C, Glasziou PP. Poor description of nonpharmacological interventions: Analysis of consecutive sample of randomised trials. *BMJ* 2013 Sep 10;347:f3755. http://www.bmj.com/content/347/bmj.f3755.pdf%2Bhtml


## consistency, *see also* internal consistency

As in common parlance, consistency is getting the conforming results across a variety of conditions, but it can have different meaning depending on the context. A high degree of agreement among observers, laboratories, and methods measuring the same characteristic is called consistency. Similar is intraobserver and intralaboratory consistency when the measurements of the same characteristic from time to time have a large degree of agreement. Variations would occur, but those must be minor, possibly negligible. A group of long-term heavy smokers having no or little incidence of pulmonary problems is an inconsistent result with our existing knowledge. A perfectly healthy person getting aberrant results on liver parameters is an inconsistent finding. A patient reporting acute pain one time and no pain another time in the same interaction with a medical care facility is an inconsistent response. Such inconsistencies are not impossible but need to be explained so that there is no suspicion about such findings. Inconsistency is different from incorrect reporting, such as a person with injury from a sharp object saying that it occurred due to falling on a flat floor just to avoid legal hassles.

The most common form of statistical consistency is the **reliability** of the results across different samples from the same population, which means they do not vary much. Sample summaries such as mean and proportion from sample to sample would be called consistent when the differences across samples are minor. The greatest assurance for this is a large sample and random selection. When the sample size is large and it is truly random, mean or proportion would

not differ too much from sample to sample. In other words, large random samples do provide reliable results. The measure of such consistency is the inverse of the **standard error (SE)** of the sample summaries. The larger the SE, the lower the consistency.

Consistency of quantitative measurements across observers, time, methods, etc., is measured by **intraclass correlation** and **limits of disagreement**. For details, see these topics.

Internal consistency of an instrument such as a survey questionnaire is considered an important property for its validity. This is explained under the topic **internal consistency**. Methods such as **item analysis** are used for this purpose and are evaluated by measures such as **split-half consistency**, **Cronbach alpha**, and **Kuder–Richardson coefficient**. These are described in detail separately.

In epidemiological studies, incidence, prevalence, remissions, case fatality, and cause-specific deaths are related with one another. This kind of consistency becomes important when information on different aspects of a disease is obtained from disparate sources. This is explained under the topic **epidemiological consistency**.

## CONSORT statement

The CONsolidated Standards Of Reporting of Trials (CONSORT) statement prescribes the format in which trial results should be reported. The objective is to provide complete details to the reader about the plan and execution of the trial so that the reader himself/herself can make judgment about the validity of the results. The statement is periodically revised to reflect new realizations about such reporting. This was developed in 1996 and revised in 2001 and 2010.

The basic premise of CONSORT is that it is important to properly report a clinical trial to the fraternity after so much of resources are spent on its planning and execution. A good trial such as a randomized controlled trial must also be reported in a format that could be appreciated by the readers. The features of CONSORT statement are the same as generally known, namely, the report of the trial should indicate why the study was undertaken, and it should include scientific background and explanation of rationale, structured review of all pertinent literature not leaving out the opposite view, selection of subjects and sample size, allocation of subjects and blinding, baseline data, transparency regarding analytic methods including those for missing data, noncompliance, etc. [1]. Though widely known, many researchers were not as comprehensive in their reporting—thus the need to develop such a standard format.

To help minimize confusion and promote clarity in reporting the methods and results, the CONSORT statement comprises a 25-item checklist and a flow diagram (Figure C.24). This list ensures clear and transparent reporting of key elements of clinical trials. However, there are some deficiencies. As of 2015, the flow diagram does not include information on blinding, although it contains information on randomization. Moreover, there is no exclusive mention of the number of cases actually followed up, although this can be deduced. CONSORT guidelines do not require much information on statistical errors. The statement outlines the minimum requirement, and one can always supplement to make reporting more complete and accurate.

1. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomized trial. *BMJ* 2010;340:c332. http://www.bmj.com/content/340/bmj.c332

## constructs (statistical), *see* factor analysis

## construct validity, *see* validity (types of)

## consumption units

In the context of food, consumption units rescale the calorie requirements of individuals relative to an adult male with average physical activity. The calorie requirement of an adult male who performs heavy physical activity is greater compared to that of a woman or a

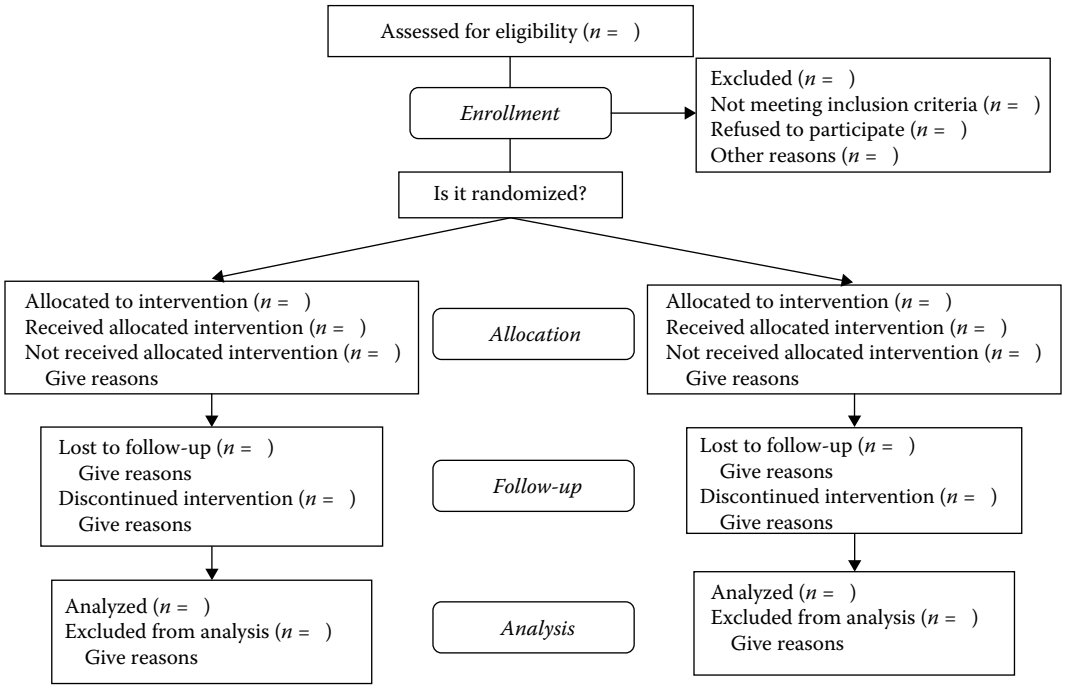



**FIGURE C.24** Flow chart of CONSORT statement.

child. Consumption units (CUs) for others are generally considered as follows: female adults who perform average physical activity, 0.9; lactating and prenatal women, 1.1; infants, 0.0; children 1–3 years, 0.4; children 3–5 years, 0.5; children 5–7 years, 0.6; children 7–9 years, 0.7; children 9–12 years, 0.8; and children 12+ years, 1.0. Note that the children units are for males; for female children, multiply these by 0.9.

These units can be used to find the total calorie requirement of a family or a group that has people of different age and sex. If a family has one male adult who performs average physical activity, one female adult not pregnant or lactating, one boy of age 4 years, and one girl of age 8 years, the total CU of the family is $1.0 + 0.9 + 0.5 + 0.7 \times 0.9 = 3.03$. If the average requirement of an adult male is considered to be 3000 calories, the total calorie requirement of the family is $3000 \times 3.03 = 9090$ calories. You can evaluate through diet survey whether the family is nearly consuming these calories or not. Such an assessment can be made for individuals also. This may be helpful in the assessment of food needs and health status. For the population of a country, the age–sex structure can be used to find the calorie requirement for the nation and can be assessed against availability of food. Conversely, food availability converted to calories per CU of one group of people can be compared with that of another group if the people are getting similar calories or one group outscores the other.

In an economic sense, such as requirement of income for a family comprising individuals of various age and sex, the kind of rescaling done for food may not be appropriate. For economic consumption, the following is recommended: 1.0 CU for the first adult in the household; 0.5 CU for other persons aged 14 years or older; and 0.3 CU for children under 14 years [1].

1. National Institute of Statistics and Economic Studies. *Consumption Units.* http://www.insee.fr/en/methodes/default.asp?page=definitions/unite-consommation.htm, last accessed April 25, 2015.

## content validity, *see* validity (types of)

## contingency coefficient, *see* association between polytomous characteristics (degree of)

## contingency tables

When a group of subjects is cross-classified into **mutually exclusive and exhaustive** categories, you obtain a contingency table. Categories are called mutually exclusive when only one of them is applicable to one subject. While measuring body mass index (BMI), categories such as –14, 15–24, 25–34, 35+ kg/m$^2$ are mutually exclusive because a person's BMI can be in only one of these categories. These are exhaustive too because no BMI can be beyond these categories. For example, a table giving the number of children with severe, moderate, and mild forms of a disease and no disease, attended to in a clinic, is a contingency table. Note that categories are necessary for representation of **frequencies** in a contingency table form. For a variable like heart rate (per minute), the categories could be (60–64), (65–69), (70–74), etc. For family size, each number by itself could be a category.

When the number of characteristics is two or more, a cross-classification can be done. A contingency table is called a one-way, two-way, or *K*-way table depending upon the number of characteristics on which the subjects are cross-classified. Table C.35 is a three-way table in which 1000 subjects coming to a cataract clinic are divided by age, gender, and the visual acuity (VA) in the worse eye. Totals collapse one or more characteristics and yield two-way tables, and further collapsing yields one-way tables. The last column of Table C.35 is a one-way classification of subjects by age, and the bottom row totals provide the two-way classification by VA and gender. When totals and columns P, which are the sum of M and F, are excluded, this table has 30 cells. The number of subjects in a cell is called the **cell frequency**. Since VA has three categories, age has five categories, and gender has two categories, the **order of the table** is $3 \times 5 \times 2$, which makes a total of 30 possible categories. Cells obtained by totals are not counted.

In Table C.35, gender is on the **nominal scale** while age and VA are **metric** but are categorized into groups. Such specifics of scale are considered when analyzing a contingency table. A table continues to be a contingency table when percentages are mentioned in place of the cell frequencies as long as the total $n$ is stated somewhere. When $n$ is known, the percentages can be readily converted to the respective frequencies.

A popular form of a two-way table is a $2 \times 2$ table that has two rows and two columns. These generally represent antecedent present/absent and outcome positive/negative. This kind of contingency table is commonly used for calculating **odds ratio** and **relative risk**. This can be extended to an $R \times C$ table that has $R$ rows and $C$ columns. The last three columns of Table C.35 form a two-way $5 \times 2$ table since this has $R = 5$ rows and $C = 2$ columns (excluding the totals). An $R \times C$ table is called a **square table** when $R = C$.

The primary interest in contingency tables is the association between two or more characteristics. This is explored almost

**TABLE C.35**
**Distribution of 1000 Subjects Coming to a Cataract Clinic by Age, Gender, and Visual Acuity (VA) in the Worse Eye**

| Age Group (years) | VA ≥ 6/60 | | | 6/60 < VA ≤ 1/60 | | | VA < 1/60 | | | Total | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | M | F | P | M | F | P | M | F | P | M | F | P |
| –49 | 11 | 8 | 19 | 37 | 32 | 69 | 12 | 10 | 22 | **60** | **50** | **110** |
| 50–59 | 18 | 21 | 39 | 69 | 73 | 142 | 13 | 16 | 29 | **100** | **110** | **210** |
| 60–69 | 25 | 21 | 46 | 183 | 142 | 325 | 42 | 47 | 89 | **250** | **210** | **460** |
| 70–79 | 10 | 11 | 21 | 54 | 44 | 98 | 26 | 25 | 51 | **90** | **80** | **170** |
| 80+ | 3 | 4 | 7 | 9 | 14 | 23 | 8 | 12 | 20 | **20** | **30** | **50** |
| **Total** | **67** | **65** | **132** | **352** | **305** | **657** | **101** | **110** | **211** | **520** | **480** | **1000** |

*Note:* F: female; M: male; P: person.

invariably by chi-square, but the actual calculations depend on the order of the table (see **chi-square—overall** for an overview of a large number of methods based on chi-square). For three-way tables, **log-linear models** are the method of choice, although these can also be analyzed by chi-square. **Three-way tables** are described separately in this volume.

### *Some Intricate Contingency Tables*

It is not uncommon in medical data that some cells in a contingency table have no frequency or zero frequency. It is important for the purpose of analysis to distinguish between observed zeroes and structural zeroes. An observed zero is one where some frequency could have occurred but happens to be zero in the sample. This is not much of a problem and is treated just like any other small frequency. A structural zero occurs when it is just not possible to have any subject in the cell. Both together are called **empty cells**.

Consider a study on multiple births—twins, triplets, and quadruplets. The births are classified by gender. The distribution may be as displayed in Table C.36. There are some observed zeroes in this table. In addition, note the × sign in some cells where no frequency is possible. Such tables are called **incomplete tables**. Special methods are required to analyze such tables. For a summary of such methods, see Kateri [1]. The present book excludes such incomplete tables, but observed zeroes are admissible for the methods discussed in this book.

### *Problems in Preparing a Contingency Table on Metric Data*

A problem frequently encountered in preparing a contingency table on all continuous and most metric variables is in deciding the number and width of intervals. When age in years is divided into (0–4), (5–14), (15–49), (50–69), and (70+) categories, the number of intervals is five and they are all unequal. Reporting of systolic BP (in mmHg) is mostly done in equal groups (120–129), (130–139), (140–149), etc.; and for diastolic BP in groups (70–74), (75–79), (80–84), etc. The choice mostly depends on commonsense evaluation of the utility of such groups in conveying the basic features of the data. Generally, the number of such groups should be between four and eight. The more groups the better for describing the variability, but this should also not take away the advantage of compactness of a contingency table. In the case of Table C.35, VA can be divided into (6/6–6/9), (6/9–6/18), (6/18–6/60), (6/60–3/60), (3/60–1/60), and (1/60–PL+) and (PL–), where PL is for perception of light. That certainly would give a better view of the subjects but would also make the table clumsier and less intelligible. Also, for cataract management, such detailed categorization may not be necessary.

All contingency tables give the distribution of subjects over various values of a measurement. Thus, these describe a frequency **distribution** although in an attenuated form. In this distribution, the values of a continuous variable would be grouped such as for age and VA in Table C.35. For a discrete variable, the values may appear as such, for example, 0, 1, 2, 3, 4, and 5+ for parity. The last category in this case is also a group. Ordinal and nominal groups may appear as such without any numeric assigned. A one-way contingency table describes a univariate distribution, a two-way table a bivariate distribution, and a three-way table a trivariate distribution. Table C.35 contains a trivariate distribution of cases coming to a cataract clinic by age, gender, and VA.

In contrast to contingency tables, in multiple response tables, the categories are not mutually exclusive. Details are presented under the topic **multiple response** tables.

1. Kateri M. *Contingency Table Analysis: Methods and Implementation Using R*. Birkhauser, 2014.

**TABLE C.36**
**Distribution of Multiple Births (Twins, Triplets, and Quadruplets) by Gender**

| Number of Female Children | Number of Male Children 0 | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|---|
| 0 | × | × | 17 | 5 | 0 | **22** |
| 1 | × | 25 | 6 | 2 | × | **33** |
| 2 | 12 | 3 | 0 | × | × | **15** |
| 3 | 4 | 0 | × | × | × | **4** |
| 4 | 1 | × | × | × | × | **1** |
| **Total** | **17** | **28** | **23** | **7** | **0** | **75** |

## continuity correction

Continuity correction is required when the distribution of a **discrete variable** is approximated by a distribution of a **continuous variable**. This helps in obtaining the approximate probability relating to a discrete variable since, in most cases, the probability relating to a discrete variable is relatively very difficult to obtain. Described below are three different situations where continuity correction helps in obtaining the probabilities.

The **Gaussian distribution** is meant for continuous variables. For a really continuous variable, $P(z > 2.33) = P(z \geq 2.33)$, that is, it does not matter whether or not the equality sign is used. Consider the following. A variable such as heart rate (HR) per minute is measured as discrete such as 70 or 71, but not as 70.4 per minute. Strictly speaking, when HR is measured, it is not necessarily exactly 70 per minute. While counting beats, it is possible that they are 70 in 59.7 s, and still 0.3 s remains. In other words, if these are counted for 10 min, the number may reach 704. Thus, a rate of 70.4 per minute is not impossible. In that sense, it is not wrong to say that the rate 70 really means that it is between 69.5 and 70.5. This is the correction for continuity in this case.

When this is acknowledged, HR between 65 and 70 (both inclusive) is actually HR between 64.5 and 70.5. Thus, to be exact, the probability that HR is between 65 and 70 (both inclusive) is actually HR between 64.5 and 70.5. If the mean in healthy subjects is 72 per minute and the standard deviation (SD) is 3 per minute, the probability that HR is between 65 and 70 is

$$P(64.5 \leq \text{HR} < 70.5) = P(\text{HR} < 70.5) - P(\text{HR} < 64.5)$$

$$= P\left(\frac{\text{HR} - 72}{3} < \frac{70.5 - 72}{3}\right) - P\left(\frac{\text{HR} - 72}{3} < \frac{64.5 - 72}{3}\right) = 0.30,$$

if the distribution of HR is Gaussian. Now, with the correction for continuity, nearly 30% of subjects in this healthy population are expected to have an HR between 65 and 70. This answer is more accurate than the 24% you would get without the continuity correction. Note how this correction can provide more accurate probability.

The other popular continuity corrections are Yates correction and the correction in the McNemar test. Both of these arise because they require the probability of Type I error by using the chi-square distribution, which is the distribution of a continuous variable, whereas the calculations are based on discrete values (frequencies) in the contingency tables. These corrections are described under the topics **Yates correction for continuity** and **McNemar test**, respectively.

## continuous variables (distribution of), *see also* discrete variables (distribution of)

A variable is continuous when it can take innumerable values in a specified range. The most glaring example in health and medicine is age. For an age between, say, 23 and 24 years, it could be 23.4 years or 23.9874 years depending upon how accurately you want to measure it. When called for, age can be in terms of years, months, days, hours, and minutes. However, this kind of accuracy is not needed since medical decisions do not change whether the age is 23 years 7 months or 23 years 8 months. In the case of neonates, possibly days are important, and in the case of a child less than 3 days, may be even hours are important. Similarly, hemoglobin level, cholesterol level, creatinine level, and duration of survival are continuous variables. On the contrary, family size, parity, and the number of patients coming to a clinic are not continuous—they are **discrete**. These cannot have a value such as 4.7.

All continuous variables are measured to only a necessary degree of accuracy. Age, in case of adults, is measured in terms of completed years; blood pressure (BP) to the counted mmHg; and hemoglobin (Hb) level to one decimal place of g/dL. This does not make them discrete despite, for example, the fact that measured diastolic BP values between, say, 86 and 88 (both inclusive) can only be 86, 87, and 88 and not innumerable. As mentioned under **variables**, many continuous variables are measured as continuous but are described in categories. If you have recorded the blood sugar level of 800 people, how do you report them? A frequency table with appropriate categories is the only way out. Such a categorical presentation of a continuous variable leads graphically to a **histogram** (Figure C.25a). In this figure, each category has Hb level width of 1 g/dL, and the number of categories is 8. When the width of categories is small, say just 0.1 g/dL in this example, the number of categories is large, and the number of subjects in the population is huge, the shape of the histogram approximates a curve (Figure C.25b). The smooth shape of this curve is called the distribution of the variable—in this case the hemoglobin level. It is called distribution as it tells how various values are distributed—which values are common, whether they are symmetric to a middle value, etc.

The shape of a distribution of a variable is best seen in a graph such as in Figure C.25b, but its mathematical properties are best studied when this is represented by an equation. Graph, however, is adequate to see the basic features of the distribution. The shape tells that the distribution is **left-skewed** (as in Figure C.25b) or **right-skewed** or symmetric, and also tells where the peak is and where it tapers off. The peak corresponds to the most common level, called the **mode**. Several other shapes such as **bimodal distribution**, **bathtub distribution**, and **uniform distribution** can also be identified by the figure alone. Among common distributions of continuous variables used for statistical inference are **Gaussian**, **chi-square**, **Student *t***, and ***F***. Mathematical formulation of these distributions helps in accurately working out the theoretical mean, variance, and such other parameters, and also the probabilities of any value bigger than or less than a specified value. Such probabilities are required primarily to find ***P*-values** for tests such as chi-square and Student *t*. Also, the properties of distributions also help in working out **confidence intervals** for various **parameters** of the distribution.

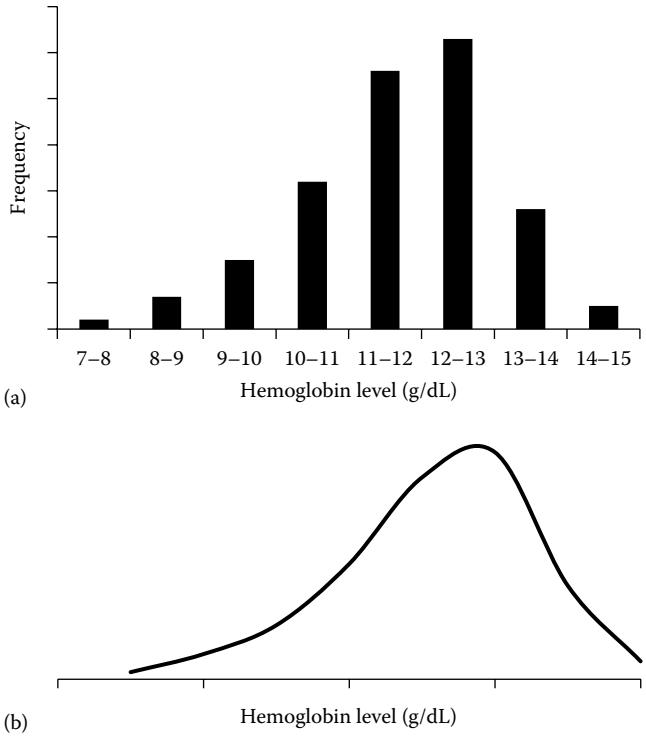



**FIGURE C.25** Distribution of hemoglobin level: (a) histogram and (b) curve.

## contour-like diagram, *see* Lexis diagram

## contrasts (statistical)

Contrasts are the custom-made hypotheses constructed to test, for example, that the effect in control group in **one-way design** with three levels of a factor is less than the average of the effect of the other two levels. In terms of notations, this hypothesis is $\mu_1 < (\mu_2 + \mu_3)/2$, where the subscript 1 is for the control group. In null form, the hypothesis we seek to *reject* in this case is $H_0$: $\mu_1 \geq (\mu_2 + \mu_3)/2$. This **null hypothesis** can also be written as $H_0$: $\mu_1 - (\mu_2 + \mu_3)/2 \geq 0$. The contrast in this case is $\mu_1 - (\mu_2 + \mu_3)/2$. Conventionally, the contrast is tested as being equal to zero or not. Since there is no square, cubic, logarithmic, etc., term in this contrast, this is called *linear contrast*. We are generally interested in such contrasts only, and the following discussion is restricted to linear contrasts.

In the case of $K$ levels of a factor in one-way design, the general form of a contrast is $C = a_1\mu_1 + a_2\mu_2 + \ldots + a_K\mu_K$. In our contrast just mentioned, $K = 3$ and $a_1 = 1$, $a_2 = -½$, and $a_3 = -½$. The contrast can also be of the type $\mu_2 - \mu_3 = 0$. Note that the conventional null hypothesis in one-way **analysis of variance (ANOVA)** with three levels of a factor is $\mu_1 = \mu_2 = \mu_3$, but the contrast $\mu_2 - \mu_3$ is only for groups 2 and 3. In this contrast, $a_1 = 0$, $a_2 = 1$, and $a_3 = -1$. These coefficients that we are denoting by $a_k$'s ($k = 1, 2, \ldots, K$) are important for further analysis as many software packages require that these coefficients be specified. Also note that the sum of these coefficients is zero. Some consider this as a defining requirement for a function to be a contrast.

C

Under the usual conditions, the estimate of the contrast is obtained by replacing the population means μ's by their sample estimates $\bar{x}$'s. Thus, the estimate of the contrast $\mu_1 - (\mu_2 + \mu_3)/2$ is $\bar{x}_1 - (\bar{x}_2 + \bar{x}_3)/2$, and that of the contrast $\mu_2 - \mu_3$ is $\bar{x}_2 - \bar{x}_3$. In general, the estimate of the contrast $C = a_1\mu_1 + a_2\mu_2 + \ldots + a_K\mu_K$ is $\hat{C} = a_1\bar{x}_1 + a_2\bar{x}_2 + \ldots + a_K\bar{x}_K$. For finding the **confidence interval (CI)** of the contrast, we need its standard error (SE). This is given by

$$\text{standard error of the contrast: } \text{SE}(\hat{C}) = \sigma\sqrt{\Sigma\frac{a_k^2}{n_k}},$$

where σ is the variance of the residuals in the ANOVA, and $n_k$ is the sample size in the $k$th group. This is the usual SE we get for linear combination of sample means of independent groups. Independence of groups is an important requirement for the ANOVA. The CI for the contrast $C$ under Gaussian conditions is

$$100(1-\alpha)\%\text{ CI for contrast } C\text{: } \hat{C} \pm t_{1-\alpha/2} * s\sqrt{\Sigma\frac{a_k^2}{n_k}},$$

where $s$ is the estimate of σ and is the same as the √(MSE) in ANOVA, and $t$ is the Student $t$ at df = $(n_1 + n_2 + \ldots + n_K - K)$. When any $n_K$ is large, the value of $t$ is the same as the corresponding value of Gaussian $z$. If this CI contains 0, $H_0$ cannot be rejected.

Study of the contrasts raises several intricate issues at advanced level. (i) For two-way or higher-way designs, contrasts can involve **main effects** of the factors as well as the **interactions**. You may be aware how interpretation of the main effects becomes complicated when interactions are present. Contrasts become even more complicated. Just be careful in this setup. (ii) Even for one-way design, you can construct several contrasts. Their simultaneous consideration requires that they are independent, which is called orthogonality. Two contrasts are orthogonal when the sum of their cross-products is zero. The number of possible orthogonal contrasts is the one less than the number of groups. If you have $K = 3$ groups in one-way design, you can have only two orthogonal contrasts. (iii) Simultaneous consideration of two or more contrasts raises the issue of **experiment-wise** error rate (**Type I error**). For this, the procedure such as that of **Bonferroni** can be used. (iv) Contrasts can be preplanned or post hoc. While preplanned contrasts do not cause any problem, post hoc contrasts (those done after looking at the results) are more likely to be false positive. Interpretation of all post-hoc comparisons should be done as indicative rather than definitive. They can give rise to important hypotheses for planning subsequent studies.

## control charts, *see also* cusum chart

Control charts for any measurement typically plot the limits within which most of these values are expected to lie when they meet a desired goal. Observed values are assessed against these limits. Such charts are considered essential statistical tools for quality control of a product. In health and medicine, they are commonly used for quality control of the laboratories.

Laboratories differ in their methods, chemicals, skills of the staff, etc., and thus, results for aliquots of the same specimen may differ from laboratory to laboratory. Part of this variability can be eliminated by standardization across laboratories. Differences also occur within the laboratory from time to time. If such differences are substantial, it shakes the clinician's confidence in the values reported. Quality control helps to keep a check and maintain a high level of performance. This can be done with the help of a control chart.

For quality control in medical laboratories, a specimen with known composition is analyzed at least once everyday. Such a specimen is called a *control specimen*. This is preserved under standard conditions for repeated analysis. The daily readings for this control specimen are plotted on a graph called the *control chart*. This is prepared by adopting the following steps:

Step 1. At the initial stage, carefully analyze the control specimen at least 20 times under standard conditions. This will delineate the random error (that occurs due to **chance**) and should be small in this setup. They can occur due to the factors that vary in the operation of the method such as timing, temperature, humidity, staining, and actual measurement. However, this may be restricted to the factors operating in a short period of time. It fails to consider long-term variation such as in different seasons, even between night and day. For better control, the random errors should be studied separately "within day" and "between days." Also the random error must be estimated for the same solution (aqueous, serum, etc.) as proposed for final use.

Step 2. Calculate the mean and standard deviation (SD) of these 20 readings. Because the same specimen is being repeatedly analyzed, the distribution pattern would be **Gaussian** and SD would be small.

Step 3. The tolerance range is (mean – 2SD, mean + 2SD). The justification for the tolerance range is that this leaves out nearly 2.5% of the highest values and 2.5% of the lowest values of a Gaussian distribution. Thus, a useful property of this range is that, on average, only 1 in 20 samples is expected to fall outside the range, if no systematic error is committed.

Step 4. Draw the lines corresponding to mean, (mean – 2SD), and (mean + 2SD) on a chart as in Figure C.26. Your control chart is ready.

Step 5. Plot the readings for the control specimen on this chart every morning before using the laboratory for other incoming specimens. If a reading on any day is outside the tolerance range, consider this as an indication that the laboratory requires scrutiny. Take corrective measures as revealed by this scrutiny before starting the analysis of other specimens.

Consider the following data on serum glucose level on repeated analysis of a control specimen with a known level of 80 mg/dL.

| Repetition number | 1 | 2 | 3 | … | … | 20 |
|---|---|---|---|---|---|---|
| Serum glucose level (mg/dL) | 78 | 80 | 84 | … | … | 82 |

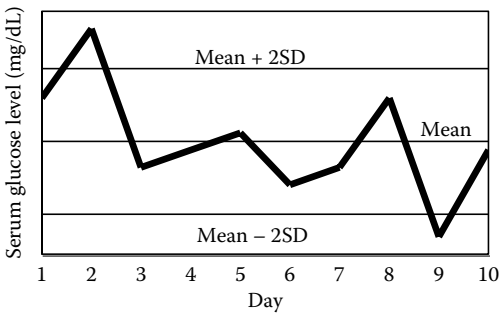



**FIGURE C.26** Control chart and the observed values.

The data for intermediary days are not shown. Suppose mean = 80.5 mg/dL and SD = 1.1 mg/dL. The lines corresponding to mean, (mean – 2SD), and (mean + 2SD) are shown in Figure C.26. This completes the first four steps just outlined. Step 5 is taking readings of the control specimen on each day. Suppose these readings for 10 days are as shown in the above table. These values are plotted on the control chart (Figure C.26). Any reading above or below the respective tolerance limits is considered a potential error and is investigated further for reasons that produced that kind of unusual value. In most such cases, an assignable reason can be identified and corrective steps can be taken before actual specimens from patients are analyzed. In this example, the value on day 2 is unacceptably high, and the value on day 9 is unacceptably low. When corrective steps are regularly taken to minimize the occurrence of such outliers, the laboratory's performance is considerably improved. Future help is provided by, what is called, a **cusum chart**.

Walter Shewhart invented control charts in 1924 to aid industrial production and management [1]. Scientists subsequently found useful applications in several fields including health and medicine.

1. Champkin J. Timeline of statistics: Pull out. *Significance* Dec 2013; 10(6):23–6. http://www.statslife.org.uk/history-of-stats-science/1190-the-timeline-of-statistics

## controls

Control is the mechanism to keep a system under check. In medical research, controls serve the purpose of a comparison against which the performance of the test group is evaluated. In this setup, controls provide the baseline, and test regimen delineates the gain or loss over this baseline. Selection of an appropriate control group is critical for the success of most analytical studies. They should ideally have same characteristics as the cases except for the antecedents under investigation as risk factors.

The most prominent use of controls is in the clinical trials, and we will come to this in a while. First consider observational studies such as **case–control** and **cohort studies**. Details are given later but realize for now that in case–control studies, one group has the disease, called the case group, and the other is without that disease, called the control group. Relevant risk factors are investigated among **antecedents** in both groups, and the factors more commonly present in one relative to the other are identified as possibly associated with the disease, either as aggravating or protective factors. In cohort studies, sometimes, but not always, a group with exposure and a group without exposure are followed up for the development of the disease. In this setup, the group without exposure is called the *control group*. Note how same nomenclature can be adopted for entirely different groups in different setups—without disease in case–control studies and without exposure in cohort studies.

Notwithstanding the strong argument made above for some kind of controls, there might be situations where they are not needed. If a treatment is being tried for a rapidly fatal disease such as tuberculous meningitis, where is the need of a control group? Saving of some cases is enough evidence of efficacy. Drugs with dramatic effects such as penicillin did not need a control. The utility of Pap smear was established without recourse to a controlled trial. However, such instances are becoming increasingly rare, and need for controls is widely accepted.

There is a debate whether surgical trials need a group with sham surgery as **placebo**. Perhaps evidence is not enough that sham surgery has the same psychological benefits as a placebo in a drug trial. Nevertheless, a sham surgery group can be adopted for a setup where it is not too expensive and is harmless to the participants.

### *Types of Controls*

A control is usually perceived to be the one that does not possess the factor of whom the effect is to be studied. Thus, this is a subject without the disease or without the exposure. While that is true in all setups, to be clear, for example in a case–control setup, the group that does not have the disease under study but has another disease can also be a good control. This depends on what you plan to study. In a study on antecedents of cardiovascular disease versus those of malignancy, any one of these groups can serve as control. No group is without disease in the setup. This is called an active **control**. This is more clearly defined in clinical trials where active control is not the placebo group but is a group receiving the existing treatment. Active controls help to assess the efficacy and side effects of the test regimen in comparison with those of the existing therapy.

In a **before–after study**, the preintervention values serve as control. Thus, there is self-control whatever its limitations may be. In contrast to this, **parallel controls** are a separate group of subjects that get an active or inert (placebo) treatment. Experiments in the laboratory and clinical trials where before–after study is not feasible or does not serve the purpose will have a parallel control group. This increases the requirement of subjects but many times is an essential "evil" for reaching a valid result.

Parallel control can be concurrent or historical. **Historical controls** comprise the group that was previously assessed. This kind of control group could be particularly considered when existing treatment is the control. This group may be derived from previous clinical trials or records such as registries or databases. This reduces the requirement of the subjects as well as the cost. Ethical issues regarding recruiting and exposing subjects to the control regimen are avoided. Historical controls may be appropriate for a disease that has a relatively stable natural history, and understanding of prognostic aspects has not changed. They must come from the same milieu from which the cases are coming. This must be verified by comparing the background characteristics. Generally, it is extremely difficult to locate a historical control group that matches with your current case group.

Despite demonstrable equivalence, the results are rarely accepted as definitive when based on historical controls. The flaw is that some factors may have changed over time. Known changes can be accounted for in the interpretation of results, but there might be some obscure changes that could affect the results. Diagnostic techniques and evaluation procedures may have improved over time. Lack of **randomization** also compromises the credibility of results in this setup. Historical controls may not have been monitored with the same keenness as concurrent controls would. If all this is fairly assured, examine the possibility of multiple controls. Multiple controls in this setup may help increase the confidence if the results replicate in each group of controls. In any other setup also, multiple controls help to increase the statistical **power** to detect a specified effect when present.

In contrast to historical control, **concurrent control** is a group that is studied at the same time as the study group. This helps to keep similarity in time such as both the groups available in 2015 since historical controls studied, say, 10 years ago may have been unwittingly overexposed or underexposed due to medical and other advances during the interregnum. Because of several other problems just enumerated, most clinical trials will have a concurrent control group.

An external group can also be used for comparison in some situations. An adequate number of nondiabetics may not be available in a

C

diabetes clinic for assessing the development of coronary incidents. *External controls* can be included in such a situation; however, they should come from the same milieu and should preferably be matched for all the factors except the exposure. In a rare situation when an appropriate external group is also not available, comparison can be done with the outcome rates in the general population. For example, incidence of birth defects in babies born to women of age 45 years or older can be compared with that in the births to women of childbearing age in the general population. The actual control group in this setup should be births to women of age less than 45 years, but a separate incidence of birth defects in them may not be easily available. The incidence in births to women of age less than 45 years may not be much different from that in all women of childbearing age since births after that age are rare. However, in many situations, the rate in the general population is not comparable with the rate in the unexposed group, and a great degree of precaution is required in using such a general group as control.

If the control subjects are in their homes, it is difficult to know if they have received some other therapy that can affect their status as controls. In the prostate cancer detection project reported by Concato et al. [1], the control subjects are those who are under routine care. But some of these may be screened outside the study and treated. Thus, their survival rate would not be sufficiently "pure" to be compared with the survival of those who were screened by the test procedures. In a field situation, *contamination* in a control group can occur if the control group is in close proximity with the unblinded test group and learns from the experience of the latter. The neighboring area may not be the test area of the research, but some other program may be going on there that has spillover effect on the control area.

### Controls in Observational Studies

The basic premise in the selection of controls in a case–control study is that the controls must come from the same group of people from whom cases have developed. For example, they could be patients from the same hospital but suffering from another disease, or peer workers, neighbors, relatives, and friends if that does not mask the relationship. Imagine a group that could have given rise to the cases and select the controls from the same group. In any case, the choice should be such that all **confounders** are ruled out as much as possible. For this, matched controls are advocated. Matched controls are those who are deliberately chosen to have the same characteristics as the cases with the exceptions as mentioned for **matching** in this volume. It has been observed, though, that matching confounders in case–control studies does not necessarily remove its confounding effect. You still have to **adjust** this in the analysis to remove its confounding effect. However, matching in case–control studies does increase the efficiency of the analysis in terms of rejecting the **null hypothesis** when false. A comprehensive discussion on selections of controls in case–control studies is available in a series of articles by Wacholder et al. [2–4].

In a prospective study, quite often, the control group comes from within the cohort, since some subjects are naturally exposed and some are not. For valid comparison, the exposed and unexposed groups must be similar at baseline, particularly with regard to the factors that can influence the outcome. If the objective is to study the effect of recently acquired central obesity on the electrocardiogram changes over time, factors such as age, gender, personality traits, stress conditions, and smoking need to be matched between the study group (with central obesity) and the control group (without central obesity). If complete matching is not possible, as would generally happen in practice, statistical methods are used to do the required adjustment at the time of analysis. Such an adjustment can become incomprehensible if done for a large number of factors.

It is sometimes impossible to find a group that is completely nonexposed. An example is exposure to dichlorodiphenyltrichloroethane (DDT). Even people in remote locations, such as Canada's Baffin Island, harbor traces of DDT. In such cases, the comparison effectively would be between the less exposed and the more exposed.

### Control Group in a Clinical Trial

As in other experiments, clinical trials can have one or more treatment regimens, but a parallel control group is almost invariably required except in crossover and before–after setup. Trials with parallel control group are also called *reference-controlled* trials. Real controls are those that are similar at baseline but follow the natural course of disease without any intervention. In practice, however, the reference control group is either treated with an existing regimen or administered a placebo. Randomization is a great strategy that could insulate against imbalance between the case and control groups.

The placebo is important because some subjects tend to behave or respond differently when no treatment is given compared with when a sham treatment is given. This may happen due to activation of mu-opioid reception in the brain by the *expectation* of relief [5]. The placebo should look exactly like the therapeutic agent under trial, perhaps with same taste, and should be given in a parallel dose. See the topic **placebo** for more details, particularly for situations where placebo can be justified.

The control group in a clinical trial should undergo the same medicinal rituals, such as dietary regulations, as the treatment group. This is more easily said than done. There are procedures for which a placebo group is nearly impossible. Examples are renal dialysis and fitting of an artificial limb. If a parallel group is a control, then appearance, schedule of administration, discomfort, etc., may cause differential compliance. Such a trial cannot be **double blind** as discussed separately. Finding a strategy that minimizes bias in such cases can be a challenging task. Whatever bias creeps in will have to be tackled at the time of analysis of data, and this too may not be easy.

In **crossover trials**, the availability of natural **controls** is automatic because the same experimental unit is used twice. This tends to reduce the impact of experimental error. The benefit of randomization is also available. Fewer subjects are required in this design relative to a parallel design with a separate control group. It is not necessary to recruit all subjects at a particular point of time. Period has a relative meaning. Some patients may have completed both periods before others are even recruited. But a big problem with crossover trials is the carryover effect and the limitation that the disease under study must bounce back to its original severity after the washout period.

1. Concato J, Peduzzi P, Kamina A, Horwitz RI. A nested case-control study of the effectiveness of screening for prostate cancer: Research design. *J Clin Epidemiol* 2001;54:558–64. http://www.sciencedirect.com/science/article/pii/S0895435600003553
2. Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case-control studies: I. Principles. *Am J Epidemiol* 1992;135:1019–28. http://www.tc.umn.edu/~alonso/Wacholder_AJE_1992_1.pdf
3. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies: II. Types of controls. *Am J Epidemiol* 1992;135:1029–41. http://www.tc.umn.edu/~alonso/Wacholder_AJE_1992_2.pdf

4. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies: III. Design options. *Am J Epidemiol* 1992;135:1042–50. http://www.tc.umn.edu/~alonso/Wacholder_AJE_1992_3.pdf
5. Zubieta JK, Bueller JA, Jackson LR et al. Placebo effects mediated by endogenous opioid activity on μ-opioid receptors. *J Neurosci* 2005; 25:7754–62. http://www.fisica.uh.cu/rationalis/etica-placebo/quitar-dolor/Placebo-Zubieta.pdf

## convenience sample, *see* sampling techniques

## Cook distance

This is one of the three methods to check whether any value is an outlier or not in the sense of unusually affecting the results, particularly for ANOVA. The first is called *leverage*, which is the proportion of the total sum of squares (TSS) contributed by the value under suspicion. For this, calculate TSS with and without that value and find how much is the contribution. If each observation is contributing equally, this would be TSS/*n*, and a substantial difference from this would indicate unusual contribution, indicating that this is an outlier. The second is called *Studentized residual* and is obtained as $\frac{r}{\sqrt{\text{MSE}}}$, where $r$ is the residual for the value under consideration, and MSE is the **mean square due to error**. If this exceeds, say, 5, conclude with confidence that this is an outlier. The third is

$$\text{Cook distance: } D_i = \frac{\Sigma_j(\hat{y}_j - \hat{y}_{j(-i)})^2}{K * \text{MSE}},$$

where $\hat{y}_j$ is the predicted value of $y_j$ ($j = 1, 2, \ldots, n$), $\hat{y}_{j(-i)}$ is the predicted value when the suspected $i$th observation is excluded, and $K$ is the number of parameters in the model. This measures the scaled change in the predicted value and in a way combines the first two. Cook distance is commonly used to assess if an observation is an outlier. Most statistical software packages have provision to calculate Cook distance and to test its statistical significance with the help of *F*-test. When significant, the $i$th value is most likely an outlier.

## COREQ (reporting of qualitative research)

An abbreviation for consolidated criteria for reporting qualitative research, this provides a checklist for conducting and reporting such research. A necessity for this arises because many vital aspects of health defy hard measurement and need special care. Among these are opinions, satisfaction, perspective, and contextual circumstances. Many medical decisions, both by the patients and the clinicians, are based on such soft aspects rather than hard facts such as risk factor, efficiency, and side effects. **Qualitative research** focuses on such hard-to-measure aspects. This research tries to find the why and how of decisions instead of what and when.

Most qualitative research is descriptive and exploratory rather than analytical. Techniques such as focus group discussion and consensus among opinion makers are used in this kind of research. Many consider this as anecdotal and not sufficiently scientific. Whatever its worth, such suspicion underscores a definite need to develop guidelines on how qualitative research should be conducted. Poses and Isen [1] cited several articles suggesting different methodological standards for this kind of research for clinical and health decisions, which further underscores the need for such guidelines.

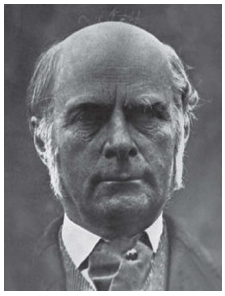
COREQ is a 32-item checklist for reporting of qualitative research and is a big help also in conducting this kind of research as it includes several vital aspects of interviews and focus groups. The details have been provided by Tong et al. [2]. Items included in the checklist comprise settings, respondent validation of findings, description of the deviation of themes, etc., in addition to the usual sampling method and the method of data collection. The authors have divided these into three domains, namely, research team and reflexivity, study design, and data analytics and reporting. Following these guidelines can make research more complete, transparent, and usable.

Alnahedh et al. [3] used the COREQ checklist to report their findings on an exploratory qualitative study to gauge understanding and assess opinions about evidence-based predictions in a sample of Australian and Saudi Arabian optometrists. Hole et al. [4] used the COREQ tool to appraise the quality of papers included in their review of the patients' experiences of rehabilitation after stroke influences their outcome.

1. Poses RM, Isen AM. Qualitative research in medicine and health care: Questions and controversy. *J Gen Intern Med* 1998;13(1):32–8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1496891/
2. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *Int J Qual Health Care* 2007;19(6):349–57. http://intqhc.oxfordjournals.org/content/19/6/349
3. Alnahedh T, Suttle CM, Alabdelmoneam M, Jalbert I. (2015). Optometrists show rudimentary understanding of evidence-based practice but are ready to embrace it: Can barriers be overcome? *Clin Exp Optometr* 2015;98(3):263–72. http://openaccess.city.ac.uk/6847/1/AlNahedh%20et%20al%202014.pdf
4. Hole E, Stubbs B, Roskell C, Soundy A. The patient's experience of the psychosocial process that influences identity following stroke rehabilitation: A metaethnography. *The Scientific World J* 2014; Article ID 349151. http://www.hindawi.com/journals/tswj/2014/349151/

## correlation (the concept of), *see also* correlation coefficient (Pearsonian/product–moment)

Correlation between two variables is the tendency of one moving with the other. Note that the correlation indicates only the tendency and does not imply any definite relationship. Francis Galton introduced the concept of correlation in 1888 [1].



Francis Galton

Conventionally, the term *correlation* is used for quantitative measurements. The corresponding term for qualitative measurements is **association**. Correlation is called positive when both variables tend to move in the same direction and negative when they tend to move in reverse direction. Height and weight have positive correlation, whereas age and lung functions in adults have negative correlation. Possibly lung functions have no correlation with kidney functions.

As explained for a **correlation coefficient**, a correlation is generally considered linear, but it is possible for two variables to be

C

strongly correlated without being linearly related. Relation of weight with age in children is not linear (see **growth charts**), but the two are strongly related. Relation between age and forced vital capacity (see **linear regression**) over the age from 0 to 80 years is also strong but not linear (it is parabolic). For more examples, see **regression (types of)**. The correlation coefficient is an inadequate measure of the strength of relationship.

There is a tendency in public discourse to interpret a correlation as **cause–effect relationship**. This can be true in some cases, but mostly this is far from truth. A strong correlation between heights of siblings in a family exists not because one is the cause of the other but because both are affected by parental height. Similarly, correlation between visual acuity and vital capacity in subjects of age 50 years and above is not of the cause–effect type but arises because both are products of the same degeneration process. No one expects vision to improve if vital capacity is improved by some therapy. Maternal mortality rate declined in Mexico between 1960 and 2015, and the proportional mortality from coronary diseases increased. These two have a strong negative correlation, but they do not have a cause and effect type of relationship. Counterfactuals provide useful armory to refute causality. These are examples of what is called the **spurious correlation**—the correlation between two variables that, in fact, have no direct relationship—it arises due to a third intervening factor that might be related to both.

Two variables are said to have **nonsense correlation** when the correlation defies logic. Would you imagine that persons with depressed moods can have an elevated risk of lung cancer? But this can occur because of a third intervening factor, namely, smoking [2]. Depressiveness seems to modify the effect of smoking on lung cancer either by a biologic mechanism or by affecting smoking behavior. This example illustrates how a seemingly nonsense correlation can sometimes lead to a plausible hypothesis. Another example is the parental age gap influencing the sex ratio of the firstborn children [3].

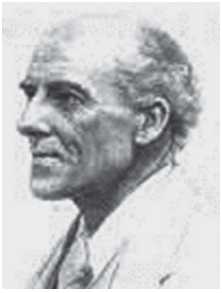
A correlation in a sample of subjects does not necessarily imply that a correlation really exists. Your sample might exhibit some correlation purely because of sampling fluctuations. This is called *incidental correlation*. Before dismissing it as incidental, make sure that no hidden factor is working to cause this correlation.

The degree of correlation in quantitative data is measured by either a product–moment correlation coefficient or a **rank correlation coefficient**. The latter is also called **Spearman correlation**. Both are described separately in this volume. The product–moment is also called the Pearsonian correlation coefficient. This coefficient is so pervasive in statistics parlance that it is simply called *correlation coefficient*. This has many positive features but also has some negative features. Artifacts in the data such as a wide range of values and outliers can also give rise to statistically significant correlation with no medical significance. See **correlation coefficient** for details. This coefficient has been devised separately to measure canonical, intraclass, intraobserver, interobserver, auto, multiple, and partial correlations. Additionally, there are also **point-biserial correlation** and **tetrachoric correlation**, which are described separately in this volume. Also see **attenuated correlation**.


1. Champkin J. Timeline of statistics: Pull out. *Significance* Dec 2013; 10(6):23–6. http://www.statslife.org.uk/history-of-stats-science/1190-the-timeline-of-statistics
2. Knekt P, Raitasalo R, Heliovaara M, Lehtinen V, Pukkala E, Teppo L, Maatela J, Aromaa A. Elevated lung cancer risk among persons with depressed mood. *Am J Epidemiol* 1996;144:1096–103. http://aje.oxfordjournals.org/content/144/12/1096.full.pdf?origin=publication_detail
3. Manning JT, Anderton RH, Shutt M. Parental age gap skews child sex ratio. *Nature* 1997;389:344. http://www.nature.com/nature/journal/v389/n6649/full/389344a0.html


## correlation coefficient (Pearsonian/ product–moment)

The correlation coefficient is the most widely used and abused measure of the degree of correlation between two variables. Note that its objective is to measure the strength of correlation. Among its several forms, the one called product–moment is the most common. This is also called Pearsonian after Karl Pearson, the person who proposed this coefficient somewhere around 1893 [1]. This section is restricted to the product–moment correlation coefficient. Others are described separately as mentioned at the end of this section.



Karl Pearson

The product–moment correlation coefficient is obtained between paired quantitative values. The units of study may be persons such as correlation between age and systolic level of blood pressure measured for $n = 80$ persons, blood samples such as between prostatic specific antigen values obtained by two methods on $n = 45$ blood samples, areas such as between infant mortality rate and life expectancy in a group of $n = 193$ countries, or any such entity. For two quantitative measurements $x$ and $y$ on $n$ units, this is calculated for a sample as

product–moment correlation coefficient:

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}} = \frac{s_{xy}}{s_x s_y},$$

where $\bar{x}$ and $\bar{y}$ are the respective means; $s_x$ and $s_y$ are the standard deviations (SDs) of $x$ and $y$, respectively; and $s_{xy}$ is the **covariance**. If you want to calculate this for a population, replace $s$ by $\sigma$, but this opportunity seldom arises in practice. In that case, this is denoted by $\rho$. $s_{xy}$ is the product–moment from which this correlation derives its name. You do not need to calculate it yourself unless you want to understand the underlying procedure—any statistical software will easily calculate this for you. The formula is just for your information and for you to realize that this correlation coefficient is covariance divided by the product of the SDs.

The crucial quantity in the product–moment correlation coefficient is the covariance in its numerator. This measures how much $x$ varies linearly when $y$ varies. The magnitude of covariance also depends on the magnitude of $(x-\bar{x})$ and $(y-\bar{y})$, i.e., on the individual variation in $x$ and $y$. The variation is measured by $\sqrt{\Sigma(x-\bar{x})^2/n}$ and $\sqrt{\Sigma(y-\bar{y})^2/n}$. These are the respective SDs. To make the covariance independent of such variation, it is divided by the product of SDs. This finally gives the formula of the product–moment correlation coefficient. The correlation coefficient thus obtained also becomes dimensionless and very easy to interpret. This correlation coefficient is not affected if you add any constant to

all values of $x$ or all values of $y$ (or both), and is also not affected if you multiply all values of $x$ by any positive constant or all values of $y$ (or both). In other words, the correlation coefficient between $ax + b$ and $cy + d$ is the same as that between $x$ and $y$, where $a$, $b$, $c$, and $d$ are any positive constants. Any of them can be zero. Also note that it is symmetric in the sense that $x$ and $y$ can be exchanged without affecting the correlation coefficient.

Now note the following for a correlation coefficient:

- The correlation coefficient is a pure number without any unit and ranges from –1 to +1. A value close to zero indicates that the two variables are linearly uncorrelated. That is, a change in the value of one is not accompanied by any linear change in the other. The scatter can then be random along a horizontal line with no slope. An example is cholesterol level and Hb level, which are generally uncorrelated. A value close to +1 indicates a strong positive relationship. Body temperature and heart rate have a strong positive correlation. A value close to –1 indicates a strong negative correlation. Homocysteine and transthyretin have a strong negative correlation among vegetarians. A perfect $r = \pm 1$ would mean that the scatter of $y$ with $x$ forms an exact straight line (see Figure C.27a,b). Such extreme correlation would seldom occur in health or medicine.
- Generally speaking, a correlation coefficient greater than 0.9 in absolute value can be considered strong, between 0.9 and 0.6 moderate, between 0.6 and 0.4 weak, and between 0.4 and 0 almost nonexistent. An illustration of scatter plots in weak and strong correlation is shown in Figure C.27. These are just rules of thumb for many applications in health and medicine, but in situations where no correlation was suspected, even a correlation of 0.2 can have great medical relevance as explained slightly later.
- An interpretation of $r$ is that $r^2$ measures the extent of variation accountable by linear relation between $x$ and $y$. Note, for example, that $r = 0.3$ means $r^2 = 0.09$. That is, for this $r$, only 9% of the total variation in $y$ is attributable to the linear variation in $x$. This explains why a correlation coefficient 0.3 is considered to indicate weak correlation.
- The magnitude of the correlation coefficient becomes unusually high when only extreme values (very high and very low) of $x$ and $y$ are observed, and becomes low when only a small range of values are observed. Thus, it is not desirable to deliberately restrict the observations of $x$ and $y$ to a narrow range or observe only extreme values. The distribution of observed values of at least one of them should be fairly spread out. Sometimes just one high value of $x$ and $y$ can yield a high value of the correlation coefficient.
- As stated earlier, the product–moment correlation $r$ measures only the *linear* component of the relationship, and this line must have some slope (small or big, negative or positive). If all age groups are combined for the relation between age and forced vital capacity (FVC), you will most likely get a scatter nearly evenly spread in all four quadrants (see **covariance**). Now the negative products $(x - \bar{x})(y - \bar{y})$ will mostly cancel out with positive products. Thus, the covariance will be nearly equal to zero. When all ages between 0 and 80 years are considered together, these data will give a small value of the product–moment correlation coefficient. This should not be construed to mean that there is no relationship between age and FVC. The relationship, in fact, is strong and can be expressed by a nice parabolic curve, although not by a straight line. The product–moment coefficient of correlation does not measure nonlinear relationships. Thus, a low value of the correlation coefficient needs to be interpreted with caution. Two variables can be heavily interdependent but may not appear correlated by this measure. This may sound like a paradox but is clear when the concept of correlation is correctly understood as concerned only with a linear form of relationship. The right criterion for measurement of the degree of relationship is the **coefficient of multiple correlation** $R^2$ when a curvilinear form is considered and the **coefficient of determination** $\eta^2$ when a nonlinear form of the parameter is considered.
- Sometimes weak correlations too can be medically important. One such example is the correlation between prostate volume and urethral resistance parameters in patients with prostatism [2]. This correlation was found to be less than 0.4 and indicated that urethral resistance is determined mostly by factors other than the volume of the prostate, or may be that they are not linearly related. This is a significant finding. Another example of weak correlation is that between age and FVC just mentioned. The *linear* relationship, as measured by the correlation coefficient, is weak, but they actually have a high degree of relationship of a curvilinear form. A third example is the correlation of systolic BP of persons with the level seen in their children. This correlation is weak in many populations but is definite. Such a correlation indicates that parents do have an influence on the level of systolic BP of their offspring, although the influence is minor.

See the explanations given for the concept of **correlation** to understand that it can be spurious, incidental, or even nonsense. This does not indicate a cause–effect relationship, neither does it measures

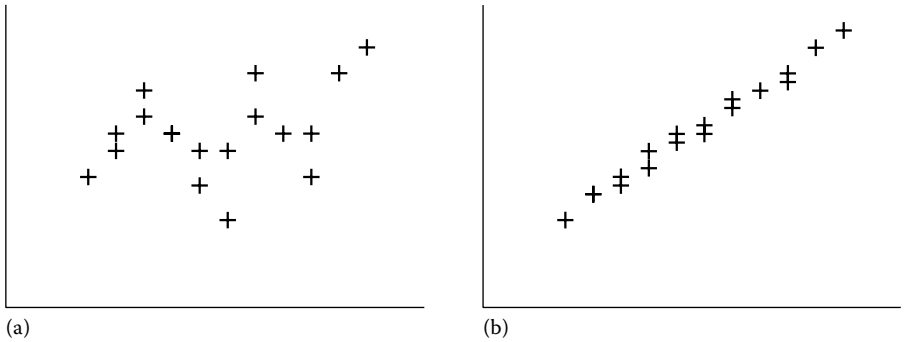



**FIGURE C.27** Scatter plot in the case of (a) weak and (b) strong correlations.

**agreement**. Modifications of the product–moment correlation coefficient are available that are used to find **intraclass correlation**, **autocorrelation**, **canonical correlation**, **multiple correlation**, and **partial correlation**. **Confidence interval for correlation coefficient** is also discussed separately, and for its test of significance, see **comparison of two or more correlation coefficients**.

Other correlation coefficients are **rank correlation**, **point-biserial correlation**, and **tetrachoric correlation**. These are not exactly product–moment correlations, but their modifications are separately described under the respective topics.


1. *Encyclopedia.com*. Karl Pearson: Complete Dictionary of Scientific Biography 2008. Quotes Section 4b of "Regression, Heredity, and Panmixia," 1896. http://www.encyclopedia.com/topic/Karl_Pearson
2. Bosch JL, Kranse R, van Mastrigt R, Schröder FH. Reasons for the weak correlation between prostate volume and urethral resistance parameters in patients with prostatism. *J Urol* 1995 Mar;153(3 Pt 1):689–93. http://www.ncbi.nlm.nih.gov/pubmed/7532234


## correlation matrix

Correlation matrix is a square arrangement of **correlation coefficients** between pairs of a large number of variables. This is used in a **multivariate** setup where the characteristic under study has several variables. For example, liver function is a characteristic that may be assessed by bilirubin level (Bil), albumin level (Alb), alanine transaminase (ALT), aspartate aminotransferase (AST), and total protein (TP). These five variables together measure liver function in this example. Suppose these are measured for $n = 40$ subjects. Now you can calculate the correlation between Bil and Alb, between Bil and ALT, etc. These correlation coefficients can be arranged as follows:

| | Bil | Alb | ALT | AST | TP |
|---|---|---|---|---|---|
| Bil | 1 | 0.67 | 0.35 | 0.47 | 0.18 |
| Alb | 0.67 | 1 | 0.59 | 0.41 | 0.32 |
| ALT | 0.35 | 0.59 | 1 | 0.83 | 0.24 |
| AST | 0.47 | 0.41 | 0.83 | 1 | 0.21 |
| TP | 0.18 | 0.32 | 0.24 | 0.21 | 1 |

This is the correlation matrix of liver function variables. This matrix has order 5, which is the number of the variables measuring liver function. Note that all the values in the diagonal are always 1 as the correlation of a variable with itself is always 1. Also note that the correlation coefficient between ALT and Alb is the same as between Alb and AST—thus, this is always a symmetric matrix. For this reason, it is not necessary to write it as a full matrix. Only the upper half triangle or lower half triangle is enough.

The correlation matrix provides an overall view of the correlations among variables of a multidimensional characteristic in a compact form. This can help in deciding which variables to be pursued with vigor or what action to take next. In statistics methods, correlation matrix is helpful for procedures such as running the regression, in estimation of the parameters, and in tests of significance in situations where several variables have to be considered together.

The correlation coefficients in a correlation matrix are mostly obtained by product–moment method, but they can also be, for example, rank correlations.

## correlogram, *see* autocorrelation

## correspondence analysis

This analysis is done to find the categories or combination of categories of one **qualitative** variable that have significant relation with specific categories of another qualitative variable in the same **contingency table**. Thus, this method is similar to the **principal component analysis** for **quantitative** variables and is used more as an exploratory tool than for conclusions. The contingency table should be large either in terms of many rows and many columns in a two-way classification or in terms of cross-classification of many qualitative variables. Simply stated, correspondence analysis uses the cellwise decomposed values of the overall value of **chi-square** in a large contingency table to locate which cells are contributing more to the value of chi-square—thus more associated than others. The full method is mathematically intricate, but we explain its essential features in a simplified manner in this section so that you understand where it is or can be used, and for which purpose. The method was developed by Benzécri in 1973 [1] and described fully by Greenacre [2].

Consider the distribution of 1000 trauma cases by the major part of the injury and the severity at the time of reporting to a trauma center (Table C.37). The table gives the observed numbers in the left panel and the contribution of each cell to the value of chi-square in the right panel. These are calculated as usual by $(O - E)^2/E$, where $O$ and $E$ are the observed and expected frequencies, respectively, under

**TABLE C.37**
**(a) Distribution of Trauma Cases by Major Part Affected and Severity of Injury; (b) Contribution of Cells to the Value of Chi-Square**

**(a) Observed Frequencies**

| Major Part Affected | Severity at the Time of Reporting | | | | | |
|---|---|---|---|---|---|---|
| | Mild | Moderate | Serious | Critical | Fatal | Total |
| Head | 30 | 87 | 68 | 123 | 49 | **357** |
| Neck | 3 | 11 | 12 | 20 | 7 | **53** |
| Chest | 18 | 42 | 59 | 41 | 78 | **238** |
| Abdomen | 22 | 20 | 18 | 35 | 47 | **142** |
| Arm | 43 | 44 | 9 | 9 | 2 | **107** |
| Leg | 28 | 32 | 19 | 21 | 3 | **103** |
| **Total** | **144** | **236** | **185** | **249** | **186** | **1000** |

**(b) Contribution of Each Cell to Chi-Square**

| Major Part Affected | Severity at the Time of Reporting | | | | | |
|---|---|---|---|---|---|---|
| | Mild | Moderate | Serious | Critical | Fatal | Sum |
| Head | 8.9 | 0.1 | 0.1 | 13.1 | 4.6 | **26.7** |
| Neck | 2.8 | 0.2 | 0.5 | 3.5 | 0.8 | **7.8** |
| Chest | 7.7 | 3.6 | 5.1 | 5.6 | 25.7 | **47.7** |
| Abdomen | 0.1 | 5.4 | 2.6 | 0.0 | 16.0 | **24.2** |
| Arm | 49.4 | 13.9 | 5.9 | 11.7 | 16.1 | **97.0** |
| Leg | 11.7 | 2.4 | 0.0 | 0.8 | 13.6 | **28.6** |
| **Sum** | **80.7** | **25.6** | **14.1** | **34.7** | **76.9** | **232.1 = $\chi^2$** |

the null that there is no relationship between the major part affected and the severity of the injury.

This table has 6 rows and 5 columns (excluding the totals)—and thus has 30 cells. The overall value of chi-square = 232.1, which is highly significant ($P < 0.001$) for $5 \times 4 = 20$ df. This helps to conclude that the severity of injury almost surely depends on the part affected. As a first step in correspondence analysis, we examine the contribution of each row to the value of chi-square given in part (b) of the table. Arm injuries with a sum of 97.0 have large contribution to chi-square followed by chest injuries with a contribution of 47.7. These two injuries do not follow the general pattern of severity. In the second step, note for within arm injuries that the contribution of mild injuries is relatively large. If you want to know whether it is unusually high frequency or unusually low frequency contributing to chi-square, two options are available. The first one is that you go back to the observed frequency table and note that the cell frequency for mild injury in arm is high. This tells that mild severity is much more common in these data when arm is the major part affected. The second option is to calculate $(O - E)/\sqrt{E}$ in place of $(O - E)^2/E$, which will have negative and positive sign, and will help to know that the frequency is unusually low or unusually high in any cell. These results can be shown in a *mosaic plot* (Figure C.28).

This initial correspondence analysis also tells that the first row (Head) in the second and third columns (moderate and serious) in Table C.37 is not contributing much to the overall value of chi-square. One can subjectively say on the basis of the total that the contribution of the second, fourth, and sixth rows and of the second and third columns also is not high. Thus, these rows and columns can be deleted from this table without substantially affecting the relationship. This reduces the dimension of the table and possibly makes it more intelligible. This dimensionality reduction is similar to what we try to achieve by principal component analysis for quantitative data. For details, see Greenacre [2].

The steps mentioned so far are essentials of correspondence analysis, but this analysis does not end there. It goes on to compute quantities called singular value and inertia, and column and row coordinates, which finally lead to a correspondence analysis plot that displays the association between different categories in a contingency table. The method can be extended to multiway tables; in this case, this is called *multiple correspondence analysis*.


1. Benzécri J-P. *L'Analyse des Donnée: L'Analyse des Correspondances*. Dunod, 1973.
2. Greenacre M. *Correspondence Analysis in Practice*, Second Edition. Chapman & Hall/CRC, 2007.


## covariance

Covariance between a variable $x$ and a variable $y$, when both quantitatively measured for the same group of $n$ subjects, is the average of the product of their deviations from the respective means. In terms of notations,

$$\text{sample covariance between } x \text{ and } y\text{: } s_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n-1}.$$

This is the average of the products of $(x-\bar{x})$ and $(y-\bar{y})$, which are the deviations from the respective means. You can see that when $x$ is replaced by $y$, the numerator becomes $(y-\bar{y})^2$ and this becomes the variance of $y$. This explains the prefix *co-* in covariance—it is for how variation in one variable $x$ is related to the variation in $y$ when both belong to the same subjects. As in the case of the sample **variance**, the divisor $(n - 1)$ is used for samples instead of $n$. For an illustration of the calculations, see Table C.38, where the covariance between creatinine and blood urea measured for 5 persons is +0.2. In this case, the plus or minus sign is important. Covariance measures the extent and direction of the *linear* relationship between two variables, but it is incomplete because it is very much affected by the variations in $x$ and $y$ and by the units of measurement. For a complete measure, it is divided by the respective standard deviations and that yields, what is called, the product–moment **correlation coefficient**.

An explanation for covariance is as follows. Consider the scatter diagrams in Figure C.29 between age ($x$) and forced vital capacity (FVC) ($y$). Figure C.29a is for the age group 10–24 years, Figure C.29b for 25–39 years, and Figure C.29c for 40–79 years. Figure C.29d gives the joint scatter of all the age groups combined. Quadrants are drawn by lines corresponding to the mean of $x$ and the mean of $y$ in each case. In Figure C.29a, FVC is increasing as age increases from 10 to 24 years. Both $x$ and $y$ are moving in the same direction. This is an indication of positive correlation between $x$ and $y$. Most points are in the first and third quadrants. In the first quadrant, both $(x-\bar{x})$ and $(y-\bar{y})$ are positive, so their product is also positive. In the third quadrant, both $(x-\bar{x})$ and $(y-\bar{y})$ are negative, so their product again is positive. Thus, the *covariance* would be a positive quantity, signifying a positive relationship.

In Figure C.29b, FVC is neither increasing nor decreasing when age is between 25 and 39 years. The points are almost equally scattered in the four quadrants. In the second quadrant, $(x-\bar{x})$ is negative but $(y-\bar{y})$ is positive. In the fourth quadrant, $(x-\bar{x})$ is positive

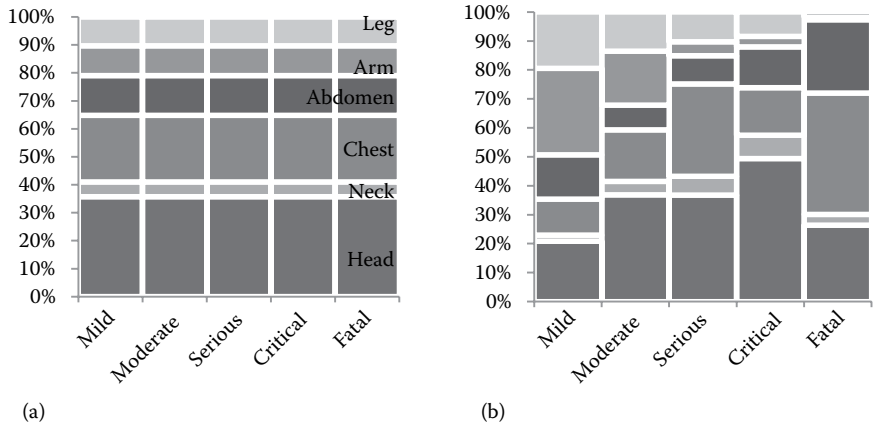



**FIGURE C.28** Mosaic plot of (a) the expected under the null of no association and (b) the observed frequencies in Table C.37.

C

**TABLE C.38**
**Illustration of Calculation of Covariance**

| | | | | | | |
|---|---|---|---|---|---|---|
| Creatinine in blood in mg/dL ($x$) | 0.7 | 1.4 | 0.9 | 1.1 | 0.9 | Mean ($\bar{x}$) = 1.0 |
| Blood urea nitrogen in mg/dL ($y$) | 8 | 12 | 16 | 17 | 17 | Mean ($\bar{y}$) = 14 |
| $(x-\bar{x})$ | –0.3 | +0.4 | –0.1 | +0.1 | –0.1 | |
| $(y-\bar{y})$ | –6 | –2 | +2 | +3 | +3 | |
| $(x-\bar{x})(y-\bar{y})$ | +1.8 | –0.8 | –0.2 | +0.3 | –0.3 | $\frac{\Sigma(x-\bar{x})(y-\bar{y})}{n-1} = +0.8/(5-1) = +0.2$ |

*Note:* $(n-1)$ used in place of $n$ for sample.

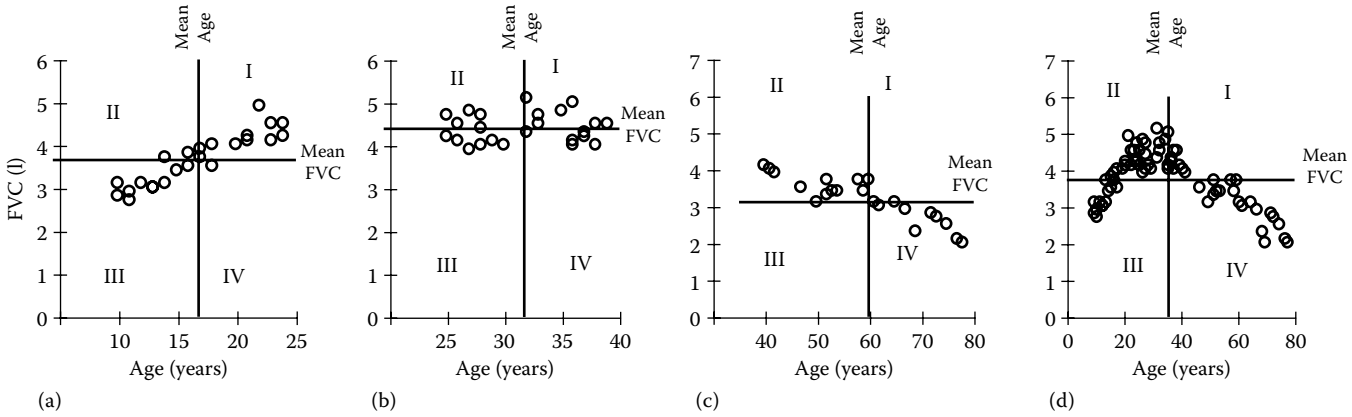



**FIGURE C.29** Scatter of forced vital capacity (FVC) by age. (a) For the age group 10–24 years; (b) for the age group 25–39 years; (c) for the age group 40–79 years; and (d) combined for all the age groups.

but $(y-\bar{y})$ is negative. In both cases, the product $(x-\bar{x})(y-\bar{y})$ is negative. The sum of negative products in the second and fourth quadrants would mostly cancel out with the sum of positive products in the first and third quadrants, thus giving a final sum nearly equal to zero. Thus, the covariance between $x$ and $y$ in Figure C.28b is nearly zero. In Figure C.29c, FVC decreases as age increases from 40 to 79 years. This inverse relationship yields a negative correlation. Most points are in either quadrant II or quadrant IV. They give negative products. The covariance in this case is negative. The illustration in all three cases indicates that covariance does measure the direction of the relationship. When all these are combined (Figure C.29d), the sum total is close to zero.

Beside its profound use in correlation coefficients, covariance is used in other setups as well. For example, it is used in analysis of covariance where the effect of a quantitative covariate on an outcome is studied.

In a multivariate setup, covariances between pairs of the variables can be expressed in a matrix, called the **covariance matrix**.

## covariance matrix, *see* dispersion matrix

## covariates

Covariate is a variable that is not of any direct interest for the study but affects the outcome of our interest. If our interest is in studying the relationship between use of tobacco and education in adults, you may like to have age as a covariate since it affects use of tobacco. Age is not the one of your interest but is included to get a clearer picture of the relation between use of tobacco and education. If the interest is in studying the effect of a hematinic on hemoglobin (Hb) level, you may like to have preexisting Hb level as a covariate since the effect of hematinic depends on the preexisting level. Preexisting level of Hb is not of any interest but is required to find the net effect of the hematinic.

Although it is customary to use the term covariate for a quantitative variable, it can be used for a qualitative variable also. Quantitative can be **categorical** divided into two or more categories, and qualitative can be **binary** or **polytomous**; it can be **ordinal** or **nominal**. There is no restriction. However, the method of analysis changes depending on such scale of the covariate.

A large number of statistical methods are available for adjusting the effect of one or more covariates on the results. First is the analysis of covariance (ANCOVA). This is used for adjusting for a quantitative covariate when the **dependent variable** is also quantitative. The **independent variables** of interest generally are qualitative. Details are under **analysis of covariance (ANCOVA)**. The second is ordinary least squares **regression** where also the dependent is quantitative. In this case, the independents also are mostly quantitative. Both these methods generally require a **Gaussian distribution** of the residuals. The third is **logistic regression** where the dependent is qualitative, mostly binary, and the independent could be qualitative and quantitative. The fourth is **Cox regression** where the dependent is the **hazard rate** and is generally applied to durations such as survival duration.

All these methods can be used to study the effect of several covariates together but require that the covariates are not much related with one another. Such multicollinearity adversely affects both reliability and validity of the results. In all these regression methods, quantitative covariate can be entered as such, but if it is

nominal, then recoding in terms of **indicator variables** would be needed. Many statistical software packages would automatically do this once you specify that the covariate is nominal. Some software packages may erroneously call nominal as categorical, forgetting that categorical can be quantitative (metric) also. If the covariate is quantitative categorical such as age categorized into 15–24, 25–34, 35–44, etc., examine if it would be wise to enter this as 1, 2, 3, etc., or as midpoints 20, 30, 40, etc. If so, enter these values. For ordinal covariates also, examine if some kind of **scores**, including linear scores 1, 2, 3, etc., are legitimate, and enter those scores. Else, use indicator variables. However, realize that indicator variables consider categories as nominal and ignore the order.

You may also like to make a distinction between a covariate and a **confounder**. For the relationship between smoking and hypertension, cholesterol level is a covariate but not a confounder since cholesterol level has no relationship with smoking. But sex is a confounder as sex affects smoking as well as hypertension. Statistical methods stated in the preceding paragraph do not distinguish between a covariate and a confounder. However, this distinction helps in proper interpretation of results.

## Cox–Mantel test, *see* log-rank test (Mantel–Cox test)

## Cox model, *see* Cox regression

## Cox proportional hazards models, *see* proportional hazards; *see also* Cox regression

## Cox regression

Cox regression is a form of regression model where the dependent or the outcome is the hazard of developing the outcome. This is especially applicable to durations such as survival duration but can be used for other setups also where outcome depends on time. To understand this, be clear about the concept of **hazard**. Cox regression is a voluminous topic. We explain its essentials in this section. The method was developed by David Cox in 1972 [1].

David Cox

The hazard of developing an outcome depends not just on time but also on several other factors. In a clinical setup, the hazard of a serious side effect may depend on the characteristics of the person such as age, gender, and nutritional status, as well as on the type of regimen, type of domiciliary care, alertness, competence of the attending physician, etc. Since many factors are involved, one of which is time, it is sometimes helpful to obtain the hazard as a function of various regressors. One such model is

$$\text{Cox regression: } h(t) = h_0(t)e^{b_1x_1+b_2x_2+\cdots+b_Kx_K},$$

where $h(t)$ is the estimated hazard at time $t$ for given values of the regressors, now better understood as covariates ($x_1$, $x_2$, …, $x_K$), and $b_1$, $b_2$, …, $b_K$ are the estimates of the corresponding **regression coefficients**. $h_0(t)$ is the baseline hazard at time $t$ when all $x$'s are zero. These coefficients measure the hazard ratio on a logarithmic scale relative to the baseline, as is odds ratios do in a **logistic model**, and are treated nearly the same way. For example,

$$\begin{aligned} b_k &= \ln(\text{hazard with } x_k/\text{baseline hazard}) \\ &\quad - \ln(\text{hazard without } x_k/\text{baseline hazard}) \\ &= \ln(\text{hazard with } x_k/\text{hazard without } x_k) \end{aligned}$$

A positive value of any $b_k$ indicates that higher values of $x_k$ increase the hazard of the outcome or indicate worse prognosis. A negative value indicates that the corresponding covariate reduces hazard. The first component on the right-hand side of the above equation, namely, the $h_0(t)$, is the estimate of the time-dependent baseline hazard that is present in any case even when all $x$'s are zero. This is considered the same for all subjects. In some cases, this can be understood as the hazard of death at time $t$ for a healthy subject (no risk factor) because of the time factor that in any case operates on all life forms. The hazard ratio $h/h_0$ can be interpreted as the relative death rate when the hazard of death is under study. The covariates $x_1$, $x_2$, …, $x_K$ could be continuous such as age, polytomous such as type of treatment, or dichotomous, such as gender, and of course would vary from person to person. The difference [$\ln h(t) - \ln h_0(t)$] is the combined effect of all the covariates.

Because of the presence of time-dependent $h_0(t)$ in the Cox regression, the usual method of maximum likelihood cannot be used to compute the $b$'s. Special methods are needed. For details, see Kleinbaum and Klein [2]. The following comments contain some useful information about Cox regression that may help to understand it better:

1. The term *hazard* is generic and not restricted to death. It can be used for any other event of interest such as appearance or reappearance of symptoms or even for a favorable event such as discharge from the hospital, cessation of smoking, or resumption of daily activities. The coefficients are interpreted accordingly.
2. The Cox regression also assumes that the covariates affect the hazard in a multiplicative manner. This means that when two factors are simultaneously present, the hazard multiplies instead of increasing additively. Multiplicability amounts to additivity in logarithm terms. This condition, too, is generally satisfied in many situations.
3. In clinical studies, the variables $x_1$, $x_2$, …, $x_K$ may contain not only the personal characteristics of the patient such as age, gender, and nutritional status but also the treatment indicators such as dosage of drug, type of treatment, and kind of care provided. These may change over time. For example, drug dosage may be heavy in the beginning and moderate later on. Such flexibility is not available in the usual quantitative regression. However, when any $x$ is time-dependent in a Cox regression, the estimation of the $b$'s becomes complex. While using statistical software yourself, or trying to understand the results of someone else's, ensure that the right package for time-dependent covariates has been used when such covariates are present. If a time-invariant model is used for time-dependent covariates, the results can be misleading.

C

4. Consider a simple situation where there is only one $x$ in the Cox regression such that $x = 0$ for standard treatment and $x = 1$ for experimental treatment. The Cox model sometimes assumes that the difference between the logarithms of hazards in the two treatment groups is constant over the follow-up period. This has been seen to be true in many applications. This makes it a *proportional* hazards model. See **proportional hazards** for details.

The following example illustrates one application of the Cox model. Ahmad and Bath [3] obtained data from a nationally representative sample of 1042 community-dwelling people in the United Kingdom of age 65 years or more. Their survival time since 1985 was recorded with censoring in 2000. Data pertain to 460 independent variables on cognitive impairment, physical health, physical activity, psychological well-being, etc. Six Cox models were run that were asked to select 1, 2, 4, 8, 12, and 16 most important variables, respectively, to predict survival time. Besides age, the analysis found handgrip strength as an important marker of frailty in predicting early death. Pain in joints causing difficulty in carrying bags and self-rated activity compared to peers were important predictors of long-term mortality.

Not many researchers test goodness of fit of Cox model. For test of this hypothesis, the deviance $-2\ln L$ is used as in the logistic regression. The test for the overall model is provided by the difference in deviance of the full model and the deviance for just $h_0(t)$, which measures the effect of time only. This will tell whether all covariates together are of any help or not in explaining the hazard. To assess the utility of any particular covariate, calculate $-2\ln L$ with and without that covariate and refer the difference to chi-square with 1 df. Statistics similar to square of the multiple correlation coefficient ($R^2$) in quantitative regression can also be used for this purpose. For testing significance of individual covariates, the **Wald test** is preferred as it gives slightly better results. In this case, this is just $[b/\mathrm{SE}(b)]^2$ and referred to **chi-square** at 1 df.

As an illustration of how the Cox model works, consider the hazard of complications in case of peritonitis with different APACHE scores. If the coefficient $b$ for APACHE is 0.0487, what does it imply? This means that if APACHE score increases by 1, the hazard ratio (relative to APACHE score = 0) of complication is exp(0.0487) = 1.050. If APACHE score increases by 8, the hazard ratio becomes exp(8 × 0.0487) = 1.476, i.e., about 1 ½ times. In place of continuous APACHE, consider a binary variable such as gender. If $b$ for gender is 0.675 and the coding is 1 for men and 0 for women (so that women is the reference category), the hazard ratio of men to women is exp(0.675) = 1.96. Thus, men have nearly twice as much hazard of developing complication in peritonitis as that of women. For these two risk factors, the model would be

$$h(t) = h_0(t)\ [\exp(0.0487)*\mathrm{APACHE}]*[\exp(0.675)*\mathrm{gender}] \text{ for all } t.$$

As in the case of logistic, any unusually large coefficient in Cox model or large SE should be regarded as wrong. It can arise either due to strong multicollinearity or due to an extremely small number of subjects in a particular subgroup. Such instances highlight the importance of scrutinizing the data and exploring their suitability for a particular analysis. As in the regression models, the covariates in Cox model also should not have high multicollinearity.

## Cox Model in Survival Analysis

The duration of survival after organ transplantation may depend on the age of the patient, coexisting diseases, and the drug regimen followed. Because of the highly **skewed distribution** of duration of survival and **censored observations**, the conventional regression approach is not applicable. Two approaches are available to study the effect of covariates on survival duration.

One approach is the parametric models such as the **exponential**, **Weibull**, and **log-normal**. These can be extended to include covariates. For example, the exponential cumulative hazard model would look like

$$\text{Exponential model with covariates: } F(t) = 1 - e^{-(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K)t},$$

where $x_1$, $x_2$, …, $x_K$ are the covariates. Similarly, other models can also be parameterized to include the covariates. These models work well when the cumulative hazard function indeed follows the modeled pattern. A priori, this is difficult to assess, although the **goodness-of-fit** can be tested once the data are available.

The second approach is the Cox model. This has found wide acceptability. The Cox model is semiparametric since it is nonparametric for time and parametric for covariates. We do not have to worry about the specific shape of the distribution of survival time under certain conditions. The following is an introductory discussion. For details, see Lee and Wang [4].

The Cox proportional hazards model lets you get away with not specifying the underlying survival function. This can be helpful as the underlying hazard function may not be of our interest. For example, many epidemiology studies want to know, "Does exposure decrease the time until event X?" The interest is in the difference in patients who have exposure and who do not have exposure. In that case, the underlying hazard does not really matter, and the risk of misspecifying it is worse than the consequences of not knowing it. Knowing exact hazard is not important—only the ratio of the situation when the factor is present to the situation when it is absent is required.

The Cox model for survival provides better estimates of the hazard at any point of time than the **Kaplan–Meier method**. The log-rank test associated with Kaplan–Meier is for the difference in the survival pattern but does not provide the magnitude of difference. Cox model delineates the magnitude through hazard ratio. Hazard ratio associated with a covariate is given by the exponent of its coefficient in the model, and provides the estimate of the independent effect of that covariate after adjusting the effect of other covariates present in the model. The estimate would come from an appropriate software package.

Time invariance of covariates is quite a strict condition in the context of survival. Most baseline values of all covariates can be incorporated as covariates because they remain the same. Generally, lifestyle and physiological variables change over time, while background characteristics such as gender, blood group, ethnicity, and urban/rural residence remain stable. If you are not sure, check that interaction between the covariate and time has statistically not significant coefficient. If significant, abandon the usual Cox proportion hazards model in favor of its extension for **time-dependent covariates**. See Thomas and Reyes [5] for details.

For Cox survival model, all polytomous covariates are converted to a series of binary classes by using indicator variables. Most statistical software packages would do this once you specify that a particular covariate is polytomous.

1. Cox DR. Regression models and life tables: *J Royal Stat Soc* B1972;187–220. http://www.jstor.org/2985181
2. Kleinbaum DG, Klein M. *Survival Analysis: A Self-Learning Text (Statistics for Biology and Health)*, Second Edition. Springer, 2005.
3. Ahmad R, Bath PA. Identification of risk factors for 15-year mortality among community-dwelling older people using Cox regression and a genetic algorithm. *J Gerontol A Biol Sci Med Sci* 2005;60:1052–8. http://biomedgerontology.oxfordjournals.org/content/60/8/1052.full

4. Lee ET, Wang JW. *Statistical Methods for Survival Data Analysis*, Third Edition. Wiley-Interscience, 2003.
5. Thomas L, Reyes EM. Tutorial: Survival estimation for Cox regression models with time-varying coefficients using SAS and R. *J Stat Software* October 2014;61:Code Snippet1. http://www.jstatsoft.org/article/view/v061c01/v61c01.pdf

## Cramer V, *see* association between polytomous characteristics (degree of)

## criterion validity, *see* validity (types of)

## critical region/value, *see* acceptance and rejection regions

## Cronbach alpha

Cronbach alpha is a measure of internal consistency of a multiple-item instrument such as a survey questionnaire or a recording schedule, or any other multiple-item test. This was first proposed by Lee Cronbach in 1951 [1].

Lee Cronbach

Internal consistency is said to be present when the items provide information that does not contradict one another—not because the respondent gave wrong information but because the items are not properly constructed. For details, see **internal consistency**. One method to check this is the **split-half** coefficient, and the second method for assessing internal consistency is

$$\text{Cronbach alpha} = \frac{M\bar{r}}{1+(M-1)\bar{r}},$$

where $M$ is the number of items in the test, and $\bar{r}$ is the average of $M(M-1)/2$ correlations between each pair of $M$ items. For this purpose, the product–moment **correlation coefficient** is used. Thus, this alpha is applicable only when the responses to various items are scores on a metric scale. For this, the items may be on a **Likert scale** that provides such scores or of any other type such as the graded response on activities of daily living (ADL) (e.g., ability to dress in infirm people graded from 0 for complete inability to 5 for complete independence). A basic assumption is that the variances of the scores on individual items are equal. If this is not the case, the measure stated in the preceding equation needs modification as indicated by Carmines and Zeller [2].

The formula shows that the Cronbach alpha depends on the number of items in the test. When $\bar{r}$ is 0.3, alpha = 0.68 for $M = 5$, and alpha = 0.87 for $M = 15$ for the same $\bar{r}$. Thus, the Cronbach reliability of a test can be increased simply by increasing the number of items, provided that the average correlation does not deteriorate. The reason for such behavior is not too hard to find. Cronbach alpha assumes that the items included in the test are a random sample from a universe of many possible items. The higher the number of items, the better the representation.

The following comments give some more information regarding Cronbach alpha:

- Cronbach alpha is based on all items of the test and thus is a better indicator of reliability than the split-half coefficient. The latter is based on only half the items. Also, split-half is based on the sum total of the scores on the items in the half-test, whereas Cronbach alpha considers each item separately and uses the average correlation.
- Cronbach alpha measures how well the items are focused on a single idea or an entity or a construct. All items are supposed to be focused on the same entity such as disability in the ADL example. But it does not tell you how well the entity is covered. A high value of this alpha is possible even when the breadth of the construct is covered only partially.
- Cronbach alpha measures the internal consistency component of **reliability** of an instrument as a whole and not of individual items. However, correlations of individual items with one another can be examined to find whether or not any particular item is adequately related to the others.
- Cronbach alpha can also be viewed as the correlation between the test and all other possible tests containing the same number of items on the same entity or construct. In the ADL example, the questions can be on being able to sit on a chair instead of walking around or can be on being able to climb stairs. If the test is reliable, the alpha will not materially change when an item is replaced by another equivalent item.
- All measures of reliability, including Cronbach alpha, range from 0 to 1. Thus, these measures are easy to interpret. A value less than 0.4 indicates poor reliability, between 0.4 and 0.6 a toss-up, more than 0.6 is generally considered acceptable, and a value more than 0.8 excellent. Use your judgment for values on the borderlines of these cutoffs.
- If the response to items is binary (yes/no, true/false, or present/absent type), then Cronbach alpha reduces to what is called the **Kuder–Richardson coefficient**. For further details, see Carmines and Zeller [2].
- All measures of internal consistency, including Cronbach alpha, are affected by the length of test, homogeneity of test items, heterogeneity of the subjects in the sample to whom the test was administered, and objectivity of the test items.
- In case needed, Spearman **rank correlation** can be used in Cronbach alpha in place of product–moment correlation coefficient.

The following example illustrates how Cronbach alpha is computed. A questionnaire consisting of four items on knowledge of acquired immunodeficiency syndrome (AIDS) was administered to 80 students of grade VII. These are as follows:

Item 1: Knowledge about mode of spread of human immunodeficiency virus (HIV) infection
Item 2: Knowledge about its prevention
Item 3: Knowledge about social handling of AIDS patients
Item 4: Knowledge about symptoms of AIDS

**TABLE C.39**
**Correlation between Items of Knowledge about AIDS**

| | Item | | | |
|---|---|---|---|---|
| **Item** | **1** | **2** | **3** | **4** |
| 1 | | 0.87 | 0.62 | 0.45 |
| 2 | | | 0.75 | 0.33 |
| 3 | | | | 0.36 |

Each item is graded on a four-point scale from 0 for no or completely wrong knowledge to 3 for perfect knowledge. Suppose the correlations are as shown in Table C.39. These give $\bar{r} = 0.563$. Thus,

$$\text{Cronbach's alpha} = \frac{4 \times 0.563}{1 + 3 \times 0.563} = 0.84.$$

The internal consistency, thus, is "excellent," despite the fact that the fourth item has poor correlation with the other items. Low correlation of this item with others is an indication that knowledge of symptoms is probably not part of the same entity that is being measured by the other three items.


1. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951;16(3):297–334. http://link.springer.com/article/10.1007%2FBF02310555#page-1
2. Carmines EG, Zeller RA. *Reliability and Validity Assessment*. Sage, 1979.


## crossover designs/trials, *see also* crossover trials (analysis of)

A design is called crossover when the same groups of subjects are administered two or more regimens one after the other in one sequence in one group and in another sequence in the other group. In a simple 2 × 2 situation, this means that one group of subjects receives AB sequence (regimen A followed by regimen B) and the other group receives BA sequence. Since the subjects are the same in each group, this design helps to control between subjects variation that would occur if the groups of subjects are different and thus can provide substantially more believable results when the stringent conditions under which it is applicable are fulfilled. Crossover design has pronounced applications to clinical as well as laboratory experiments. Thus, the following discussion is mixed for laboratory and clinic setups.

Medicine is an incomplete science as it is today; many drugs work for a limited period. They have to be administered again, else the disease returns with the same intensity. Some drugs do have some **carryover effect**, but in others, there is practically none. When all traces of their effect vanish, say 2 or 3 days after the last ingestion, the same subject can be used again to try the other drug, provided the condition of the subject reverses to what it was. This is called the steady state. In the case of humans, epilepsy, migraine, enlarged prostate, and end-stage renal failure are examples of such diseases. In some patients, hypertension and diabetes require daily medication; otherwise, the disease returns to the baseline. Yet, a high-cholesterol diet consumed by older subjects for a long time when they were young because of lack of awareness might still have a carryover effect that persists despite a change to a low-cholesterol diet. **Cross-sectional studies** of the type envisaged by crossover design fail to take care of such confounders.

In the case of animals, the experiment could be for the effect of cocaine versus saline on behavior in terms of distance traveled in 30 min following the injection separately in long-sleep and short-sleep mice. In this experiment, some mice may receive cocaine on day 1 and saline on day 2, and others saline on day 1 and cocaine on day 2. The process of trying two regimens on the same subject after providing a time gap (called the **washout period**) for complete disappearance of the carryover effect of the first regimen is called crossover (Figure C.30).

You can see that crossover is more suitable for quantitative outcomes rather than binary (yes/no) response. Nonetheless, it can be used for such binary response as well. Suitability of crossover also depends on the regimen. Quickly reversible regimens are obviously more suitable than long-acting regimens such as steroids. Crossovers are difficult to implement for multidose trials.

An important prerequisite of a successful crossover experiment is the absence of carryover effect. The carryover effect need not be just physical; it can also be psychological if that affects pharmacological outcome. Also, a drug may have caused damage to the liver that would affect metabolization of future drugs for a long time. If a patient receives regimen A followed by regimen B, the effect seen after regimen B would be the direct effect of regimen B plus the residual effect of regimen A received earlier. This produces bias. The problem confounds further if the carryover effect of regimen A is different from the carryover effect of regimen B.

As just mentioned, the time gap required for carryover effect to vanish is called the washout period. This generally should be at least four times the half-life of the regimen. **Half-life** is the period at which half of the peak concentration of drug is available in the body and the other half is eliminated or metabolized by the body after the time when the peak is reached. Nothing should occur during the washout period that can affect the outcome. Absence of the carryover effect can be statistically tested in a crossover design. If carryover effect is significant, think of an alternative such as **Balaam design** where AA and BB sequences are also included in addition to AB and BA sequence.

In a crossover experiment, the subjects are divided into two equal groups by **randomization**—one group receives drug A followed by drug B, and the other receives drug B followed by drug A. Allocation can be done by tossing a coin or by any other random method. It is desirable to continue observation till the end of the washout period of the second drug. This provides confidence regarding the drug exiting the system. However, this sequence itself

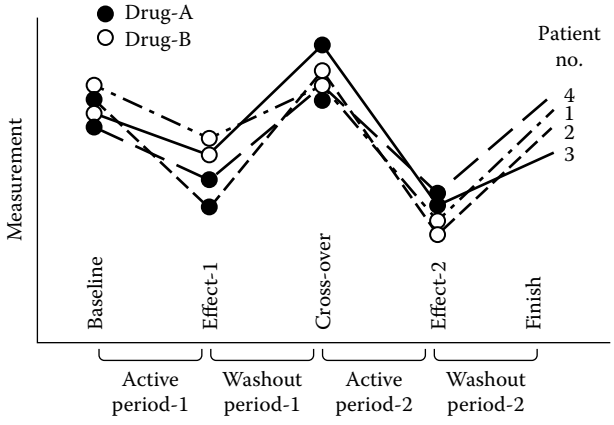



**FIGURE C.30** Representation of a crossover design.

can cause differential effect. This can particularly happen when the intervening period is favorable to one regimen than the other. For example, when lisinopril is given first and then losartan after a 2-week washout gap to nondiabetic hypertensive patients for insulin sensitivity, the first drug may still be in the system. This may not happen when losartan is given first. Crossover design is not appropriate for such regimens that have significant sequence effects.

The advantage of crossover is the availability of natural controls because the same experimental unit is used twice. This tends to reduce the impact of experimental error. The benefit of randomization is also available. Fewer subjects are required in this design relative to a **parallel** design with a separate control group. It is not necessary to recruit all subjects at a particular point of time. Period has a relative meaning. Some patients may have completed both periods before others are even recruited.

There are many cautions for using a crossover design. Besides the basic requirement of no carryover effect, crossover is recommended when the effect of the intervention can be assessed quickly so that the patient can be crossed over without a large time gap to deny opportunity for confounders to appear. If baseline condition can affect the outcome, all the subjects in the trial should be homogenous. Crossover is not a suitable design for cyclic diseases such as asthma and arthritis because they have natural remission periods. As for any design, there should not be any concomitant medication or comorbidity that can affect the outcome. Note that in this design, the subjects must stay in the experiment twice as long as otherwise needed in an experiment with parallel control group. This can cause higher dropouts. Crossover should not be used when substantial dropout is anticipated because then the statistical analysis becomes complex and the reliability steeply declines. However, periods in the case of crossover are not necessarily calendar periods. It is possible that one subject receives first-period treatment in the month of June and the second period in the month of July, and the second subject is recruited in August for the first-period treatment. Also, the entire trial should not be modified midway in any manner, such as for dosage, period, and sequence, because that could complicate the analysis.

Crossover is relatively easy when the treatment can be rapidly started and stopped. However, crossover strategy with random allocation of the order of therapy and adequate washout period as discussed earlier is essential for valid results from such self-control designs. A crossover design cannot be adopted if cure or death is the outcome. In addition, regulatory agencies seldom approve crossover because of the possibility of carryover effect even if demonstrated to be absent. Also, side effects and tolerability are difficult to study in a crossover setup since the administration is for a short period. Thus, crossover is not recommended for phase III trials. But this design can be easily used in phase II, particularly for identifying the right dosage. The conventional crossover is two regimens–two periods–two sequences strategy. But it can be extended for multiple periods and multiple regimens. For further details, see Senn [1].

Here are some examples that illustrate situations where a crossover strategy can be used. In rats, the ability of drugs to lower brain stimulation reward (BSR) thresholds often correlates well with high abuse reliability. Gill et al. [2] implanted bipolar electrodes into the medial forebrain bundle of male mice for assessing BSR thresholds after intraperitoneal alternative saline and cocaine in a crossover design. For an example involving humans, consider a representative group of 90 patients with essential hypertension. To study the effect of different drugs, they can be given atenolol, propranolol, and acebutalol for 6 days each after a gap of 3 days in between as a washout period for the drug to exit the system and for the disease to reappear at the same level. In this example, one patient is used three times. The order in which a patient receives the drugs is randomized. If the conditions are favorable, you can have four or five regimens in a trial.

Particularly in a crossover and in **repeated measures studies** without parallel control in general, the results are obtained by comparing the average within subject responses at different time points. In the usual experiments with a separate control group, the results are obtained by comparison of averages across groups. Crossover, just as **before–after experiments**, has smaller variability because the subjects are the same, and such experiments require fewer subjects to detect the same effect than the design with parallel control group.

For small cell frequencies in a crossover design, the statistical significance of the difference between treatments A and B can be investigated by using the Fisher exact test on the discordant pairs. The concordant pairs are ignored. This is discussed under **crossover trials (analysis of)**.

1. Senn S. *Crossover Trials in Clinical Research*, Second Edition. Wiley, 2002.
2. Gill BM, Knapp CM, Kornetsky C. The effects of cocaine on the rate independent brain stimulation reward threshold in the mouse. *Pharmacol Biochem Behav* 2004;79:165–70. http://www.sciencedirect.com/science/article/pii/S0091305704002436

## crossover trials (analysis of)

As mentioned for **crossover designs/trials**, crossover could be a very efficient strategy for trials on regimen that provide temporary relief. In this design, one group receives regimen A and then B (AB sequence), and the other group receives regimen B and then A (BA sequence). The two groups contain different sets of individuals; thus, the groups are independent. The primary objective of a crossover trial is to test the significance of the difference in effects of the regimens. Primary difficulty in analyzing data from crossover designs arises from duplicity of factors. If there are three patients, two periods, and two treatments, a crossover could yield only six observations instead of $3 \times 2 \times 2 = 12$ in a conventional design. Two factors determine the third. If patient 1 gets the AB sequence, the observation for this patient in period 2 must be under treatment B. This section describes the method of analysis of data from crossover trials—first when the outcome is quantitative and second when it is binary.

### *Analysis of Data from Crossover Trials with Quantitative Response (Gaussian Conditions)*

Consider a trial on $n = 16$ chronic obstructive pulmonary disease (COPD) patients who were randomly divided into two equal groups of size 8. The first group received treatment A and then treatment B, and the second group received treatment B and then treatment A. Abbreviate these treatments as trA and trB. An adequate washout period was provided before switching the treatment so that there was no carryover effect. The response variable is forced expiratory volume in one second ($FEV_1$). The data obtained are in Table C.40.

*Test for Group Effect*: In the case of crossover trials, the groups identify the sequence, and the group effect is the same as the sequence effect. If the sequence does not affect the values, the mean difference between trA and trB should be the same in the AB group as in the BA group. A significant effect means that trA has different effect when in period 1 than when in period 2. Thus, this

C

**TABLE C.40**
**$FEV_1$ (L/s) in COPD Patients in a Crossover Trial**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Group I—AB Sequence** | | | | | | | | | |
| Subject no. | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Period 1 | trA | 1.28 | 1.26 | 1.60 | 1.45 | 1.32 | 1.20 | 1.18 | 1.31 |
| Period 2 | trB | 1.25 | 1.27 | 1.47 | 1.38 | 1.31 | 1.18 | 1.20 | 1.27 |
| **Group II—BA Sequence** | | | | | | | | | |
| Subject no. | | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| Period 1 | trB | 1.27 | 1.49 | 1.05 | 1.38 | 1.43 | 1.31 | 1.25 | 1.20 |
| Period 2 | trA | 1.30 | 1.57 | 1.17 | 1.36 | 1.49 | 1.38 | 1.45 | 1.20 |

is also called treatment–period interaction. Note for crossover that group effect = sequence effect = treatment*period interaction. In this example,

| **trA – trB Values** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Group I (AB) | +0.03 | –0.01 | +0.13 | +0.07 | +0.01 | +0.02 | –0.02 | +0.04 |
| Group II (BA) | +0.03 | +0.08 | +0.12 | –0.02 | +0.06 | +0.07 | +0.20 | +0.00 |

Since the patients in Group I are different from the patients in Group II, equality of means in these groups can be tested by the two-sample **Student *t*-test** under **Gaussian conditions**. In this case, for these differences, the means, variances, and pooled variance are

$\bar{x}_1 = 0.03375$ $\bar{x}_2 = 0.06750$

$s_1^2 = 0.0023125$ $s_2^2 = 0.0048786$ $s_p^2 = 0.0035955.$

Thus, for two independent samples,

$$t_{14} = \frac{0.03375 - 0.06750}{\sqrt{0.0035955(1/8 + 1/8)}} = -1.126.$$

This is not statistically significant ($P > 0.05$). Thus, the evidence is not enough for the presence of sequence effect. The practical implication is that there is no sequence effect, although this can be ascertained only when the sample is large and the power is high. If a sequence effect is present, the reasons should be ascertained and the trial done again after eliminating those causes. In most practical situations where a crossover trial is used, the sequence of administering the drugs does not make much difference. The real possibility is that of a carryover effect that can also make a dent in the sequence effect.

*Test for Carryover Effect*: If a positive carryover effect were present in any regimen, the performance of that regimen in period 2 would be better than its performance in period 1. Thus, the presence of a carryover effect can be assessed by comparing the performance of each regimen in the two periods. In the preceding example, this is obtained by comparing trA values in period 1 with trA values in period 2, and similarly for trB. It is possible that only one of the regimens has a long-term effect so that carryover is present for that regimen and the other has no such effect. Two two-sample *t*-tests would decide whether one or both have a carryover effect. In this example, these are as follows:

**trA**

| | | | |
|---|---|---|---|
| Period 1 | Mean = 1.325 | SD = 0.1386 | Pooled variance |
| Period 2 | Mean = 1.365 | SD = 0.1387 | 0.019224 |

$$t_{14} = \frac{1.365 - 1.325}{\sqrt{0.019224(1/8 + 1/8)}} = 0.577.$$

**trB**

| | | | |
|---|---|---|---|
| Period 1 | Mean = 1.298 | SD = 0.1391 | Pooled variance |
| Period 2 | Mean = 1.291 | SD = 0.0952 | 0.014206 |

$$t_{14} = \frac{1.291 - 1.298}{\sqrt{0.014206(1/8 + 1/8)}} = -0.117.$$

*P*-values associated with these values of *t* show that the carryover effect is not statistically significant for any of the treatments in this example.

*Test for Treatment Effect*: The two tests just mentioned are preliminaries. The primary purpose of the trial, of course, is to find whether one treatment is better than the other. This can be done only when the sequence (or the group) effect is not significant.

(a) *If there is no carryover effect*, the procedure is as follows. Consider two groups together as one because the sequence is not important and there is no carryover effect. Calculate differences trA–trB and use the paired *t*-test on the joint sample. In this example, these differences are

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| +0.03 | –0.01 | +0.13 | +0.07 | +0.01 | +0.02 | –0.02 | +0.04 |
| +0.03 | +0.08 | +0.12 | –0.02 | +0.06 | +0.07 | +0.20 | 0.00 |

These give mean difference = 0.0506 and $s_d = 0.06049$. Thus,

$$t_{15} = \frac{0.0506}{0.06049/\sqrt{16}} = 3.346.$$

This gives $P < 0.01$. Thus, statistically the treatment difference is highly significant.

(b) *If the carryover effect is present*, crossover is not a good strategy. You may then increase the washout period and ensure that no carryover effect is present. If the data from a crossover trial are already available and the carryover effect is found to be significant, then analyze as usual by Student *t* after ignoring the second period. Thus, half of the data (and efforts) will become redundant and the advantages of the crossover design lost. Remember the following for crossover designs:

- It is easy to say that a washout period will eliminate a carryover effect. In fact, it can rarely be dismissed on a priori grounds. A psychological effect may persist even in case of blinding of the subjects. Thus, a crossover design should be used only after there is fair assurance that a carryover effect is practically absent.

- The test for a carryover effect is based on variation between subjects and thus has less power. A small but real effect may not be detected unless a big trial with a large number of subjects is done. Then the advantage of the economy of subjects in a crossover design may be lost.
- Details are given under **multiple comparisons** but carrying out so many statistical tests on the same set of data increases the total chance of Type I error. To keep this under control to, say, less than 5%, you should carry out the three *t*-tests at the 2% level each.
- The preceding is an easy method based on the Student *t*-test. This is not so elegant. Quantitative data from crossover trials can be analyzed more meticulously by using the analysis of variance method. For this, see Everitt [1]. Neither methodology is fully standardized.

### *Analysis of Data of Crossover Trials with Binary Response*

The concern now is with crossover experiments in which the response or the outcome is binary. This is a yes/no, present/absent, or relieved/not relieved type of response. In such cases, the number of subjects with different responses in a crossover trial can be listed as in Table C.41. In this table, for example, $b_2$ is the number of subjects who were not relieved by drug B but were relieved by drug A when the sequence of administration was B followed by A. Other notations also have similar meaning.

The analysis of binary data from crossover trials is not fully standardized. One of the methods is as follows. For finding statistical significance in this case, the concordant pairs in the first and last columns are ignored. Since the response with drug A is the same as with drug B in these two types of pairs, they cannot help in deciding which drug is better. The decision is based on the discordant pairs in the second and third columns. The frequencies in these two middle columns are analyzed as a usual 2×2 table. For large $n$, the chi-square criterion is computed and the inference drawn as per the procedure for chi-square. The relevant numbers for this setup are given in Table C.42. This table will indicate whether the response with one treatment is different from the response with the other treatment. If $(a_2 + b_3)$ is large relative to $(a_3 + b_2)$, then drug B is more effective. If $(a_3 + b_2)$ is larger, then drug A is more effective. The prerequisite for validity of this test, however, is that there is no effect of different sequence of drug administration and no carryover effect.

It is helpful to investigate whether sequencing (A precedes B, and B precedes A) alters the response. This can be done by making another 2×2 table (Table C.43) that gives counts of relieved subjects in the two groups. Again, only discordant pairs are counted. Chi-square is computed for large samples and an inference drawn as usual. If sequence effect in Table C.43 is significant, crossover is not a suitable strategy. Redo the trial with some other design. If interaction is to be statistically tested, prepare a table, such as Table C.44, and calculate chi-square as usual.

These three procedures are stated in reverse order. In practice, do ③, then ②, and then ①. If the presence of interaction is detected by ③, it is not worthwhile to do ② or ①. In this case, find out why the interaction is occurring and do the trial again after taking steps to remove the likelihood of interaction. The following example illustrates the calculations.

A new drug A for relief from urinary problems in subjects with enlarged prostates was compared with an existing drug B. Each was given for 1 month to 50 subjects in sequence BA and to another 50 subjects in sequence AB. There was a washout period in between as needed in a crossover design. The results are shown in Table C.45.

For the purpose of comparison, the first and last columns are ignored. The frequencies in the discordant cells are small, and so the

**TABLE C.41**
**Format of Binary Response in a Crossover Trial**

| | Response[a] | | | | |
|---|---|---|---|---|---|
| **Group (Sequence)** | **(0, 0)** | **(0, 1)** | **(1, 0)** | **(1, 1)** | **Total** |
| Group I—sequence AB | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $\boldsymbol{n_1}$ |
| Group II—sequence BA | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $\boldsymbol{n_2}$ |

[a] 0 = no relief; 1 = relief.

**TABLE C.42**
**Relevant Numbers for Analyzing a Crossover Design**

| | | Response | |
|---|---|---|---|
| | | **(0, 1)** | **(1, 0)** |
| ① | Group I—sequence AB | $a_2$ | $a_3$ |
| | Group II—sequence BA | $b_2$ | $b_3$ |

**TABLE C.43**
**Numbers for Testing Sequencing Effect in a Crossover Trial**

| | | **Relieved with A** | **Relieved with B** |
|---|---|---|---|
| ② | Group I—sequence AB | $a_3$ | $a_2$ |
| | Group II—sequence BA | $b_2$ | $b_3$ |

**TABLE C.44**
**Numbers for Testing Interaction in a Crossover Trial**

| | | Response | |
|---|---|---|---|
| | | **(0, 0)** | **(1, 1)** |
| ③ | Group I—sequence AB | $a_1$ | $a_4$ |
| | Group II—sequence BA | $b_1$ | $b_4$ |

**TABLE C.45**
**Crossover Trial with Small Frequencies in Some Cells**

| | Response[a] | | | | |
|---|---|---|---|---|---|
| | **(0, 0)** | **(0, 1)** | **(1, 0)** | **(1, 1)** | **Total** |
| Group I—sequence AB | 2 | 1 | 7 | 40 | **50** |
| Group II—sequence BA | 7 | 5 | 2 | 36 | **50** |

[a] 0 = no relief; 1 = relief.

Fisher exact test would be used. The given configuration and more extremes are as follows:

| | | | | | |
|---|---|---|---|---|---|
| 1 | 7 | 8 | 0 | 8 | 8 |
| 5 | 2 | 7 | 6 | 1 | 7 |
| 6 | 9 | 15 | 6 | 9 | 15 |

No other extreme configuration is possible because marginal totals have to remain the same, and one cell frequency is already zero in the second configuration. Therefore, from **multinomial distribution**,

$$P = \frac{8!7!6!9!}{15!1!7!5!2!} + \frac{8!7!6!9!}{15!0!8!6!1!}$$

$$= 0.0336 + 0.0014$$

$$= 0.035.$$

This is less than 0.05. Since it is a one-tailed probability, the conclusion too would be **one-sided**. In this case, there are seven subjects in sequence AB that had relief from drug A but not from drug B and five subjects with relief from A but not from B in sequence BA. Of a total of 15 discordant pairs, 12 favor drug A and only 3 favor drug B. A small *P*-value shows that this is statistically significant. Thus, the conclusion is that the new drug A is significantly better than the existing drug B. For such one-sided conclusions, one-sided α is used.

1. Everitt BS. *Statistical Methods for Medical Investigation*, Second Edition. Edwin Arnold, 1994: pp. 77–91.

## cross-product ratio, *see* odds ratio

## cross-sectional studies

**Observational studies** where both antecedent and outcome are elicited together are called cross-sectional studies. For example, one format for studying the effect of anemia on gestation period in pregnancy is that a group of deliveries are chosen irrespective of maternal anemia and gestational age, and both are elicited. This is a cross-sectional study since both antecedent and outcome are observed at the same time. Both are responses in this setup, and none is a factor. The presence or absence of either antecedent or outcome is not a consideration at the time of selection of subjects in this kind of design, but the objective still is to investigate the **association** as required for any **analytical study**. Although this kind of format is regularly used for descriptive studies, such as for estimating the prevalence of a health outcome, the term *cross-sectional study* is appropriate for analytical studies where the objective is to investigate antecedent–outcome relationship. They can be used to generate hypothesis regarding etiology. Cross-sectional descriptive studies are better understood as surveys. Cross-sectional studies are not surveys.

A cross-sectional study is categorized as analytical because it also intends to investigate associations and differences. However, the validity of conclusions would depend on the proportional representation of various levels of responses. Thus, it is important that a cross-sectional study is done on a randomly selected sample of adequate size, obtained from the target population without recourse to any consideration of presence or absence of antecedent or outcome. If this is compromised, the conclusions based on a cross-sectional study could be biased.

Cross-sectional studies are done where the distinction between antecedent and outcome is blurred. Consider cleft lip and thalassemia in children; neither is a known cause of the other, yet dependence of one on the other can be investigated for generating a hypothesis. Such studies are analytical in this sense and do not remain purely descriptive. For studying the association between blood group and gender, if male and female subjects are selected and tested for their blood group, it could be difficult to categorize it either as a prospective or as a retrospective study, though it might merit to be called a case–control study if one gender is regarded as case and the other as control. When the antecedent and the outcome are not identifiable, the study may be called descriptive rather than analytical—thus not really cross-sectional. It would only indicate the blood group profile in the two genders without aspersions on the exposure–outcome type of relationship. Such a study would continue to be of a descriptive type until one is identified as an antecedent or at least as a suspected antecedent for the purpose of the investigation. A study on evaluation of concordance between two or more methods is also cross-sectional.

Cross-sectional studies are more appropriate for assessing the relationship between fairly stable entities, such as gender and hypertension, which do not change during the course of the study. Note that a cross-sectional study provides a one-time snapshot of the status of the characteristics, and not a long-term perspective.

In a cross-sectional study, **confounding factors** can also be investigated at the same time. For example, in a study on determinants of the increase in serum cholesterol with age in adults, a group of subjects could be elicited for sex, diet, body mass index, etc., in addition to age and serum cholesterol level. All measurements would be considered valid for the date of investigation, although assessment of diet in this case could be based on the food intake during the previous 3 days. The investigation might have the objective of determining the role of age in increasing the cholesterol level, but the design is such that the role of sex, obesity, and diet can also be investigated with the same validity. Statistically, any characteristics in a cross-sectional study can be considered dependent on the rest; the only restriction is the plausibility or justifiability of the relationship obtained.

The first step for sampling for a cross-sectional study is, as usual, to identify the target population for which the results would generalize. Then decide which sampling method would be appropriate. For age- and sex-related disease such as hypertension and diabetes, stratification by age and sex might be useful. For a community-based study in a large population, a **multistage sampling** involving selection of counties and households can be adopted, or a **cluster sampling** might be more convenient. In a clinical setup where subjects come in queue, **systematic sampling** could be adopted.

Consider the following example. A study was conducted on 137 extremely obese subjects (mean BMI = 46.9 kg/m$^2$), and their diabetes status and obstructive sleep apnea (OSA) were elicited [1]. Thus, this is a cross-sectional study. Among subjects with normal glucose tolerance, 33% had OSA. This was 67% in prediabetic subjects and 78% in type 2 diabetes patients. Thus, the association between OSA and diabetes status in extremely obese subjects was clear. This continued to be so after age, sex, BMI, etc., were adjusted. This cross-sectional study excluded the role of various confounders, yet the conclusion rightly is of association and not cause–effect.

### *Merits and Demerits of Cross-Sectional Studies*

The following paragraphs describe the demerits of cross-sectional designs before describing the merits because, for such studies, demerits should be considered first when investigating antecedent–outcome relationship.

Cross-sectional studies give a snapshot view and cannot measure risk. They may turn out to be a poor choice in situations in which either the antecedent or the outcome or both are rare. If the outcome of interest is testicular cancer or the antecedent under investigation is exposure to synthetic estrogen, then a cross-sectional study is not appropriate. An extremely inadequate number of subjects with the characteristics of interest may render the entire exercise futile.

There are other demerits too. The analysis can certainly be extrapolated to evaluate the net relationship between any two characteristics, keeping the others constant, but it should be clear that a cross-sectional design investigates the presence and not the appearance of the condition. Transient cases or rapidly fatal cases may inevitably remain underrepresented in this kind of design. The causes that determine the appearance are confounded with those influencing the duration of the disease, and it may be difficult to draw a clear inference about either set of causes. Dropouts or migrants tend to be excluded in a cross-sectional study.

The other serious difficulty in a cross-sectional study is that it cannot be ensured that the antecedent has actually preceded the outcome. This might have important implications for a causal inference. A firm conclusion on cause–effect can rarely be drawn from cross-sectional studies, and this is a major limitation for such studies to be truly analytical. The concept of a control group is not relevant to cross-sectional studies. The biases seen in other observational studies are largely applicable to cross-sectional studies as well. However, information bias or memory lapse could be practically absent in this setup.

Caution is required in interpreting the results of a cross-sectional study. Such a study might reveal, for example, that the prevalence of hypercholesterolemia increases with age, but the fact could be that it is not age-induced but is due to the changes in the diet pattern of younger subjects resulting from increased awareness of the harmful effects of a high-cholesterol diet. Such awareness was less common 30 years ago and practically absent 50 years ago. A high-cholesterol diet consumed by older subjects for a long time when they were young because of lack of awareness might still have a carryover effect that persists despite a change to a low-cholesterol diet. Cross-sectional studies fail to take care of such confounders.

It follows from the preceding discussion that the cross-sectional design is particularly well suited for acute conditions with a short latent period or for chronic diseases that are stable and nonfatal. As already stated, this design can be recommended for situations in which the distinction between antecedent and outcome is blurred. In disease–anxiety syndrome, disease can cause anxiety and anxiety can cause disease—thus either could be an antecedent. Also, a cross-sectional study is generally considered a rapid and inexpensive way to provide clues for further and more valid investigations.

For analysis of data from a cross-sectional study, any one of the characteristics that can plausibly be considered as the outcome can be considered dependent on the others. **Logistic regression** or the usual **multiple regression** can be used to find the joint or net effect of each of the independent factors in the model. In the case of cross-sectional studies, the assessment is generally made in terms of **odds ratios** for qualitative dependent. Sometimes the **prevalence rate** ratio is used.

1. Fredheim JM, Rollheim J, Omland T, Hofsø D, Røislien J, Vegsgaard K, Hjelmesæth J. Type 2 diabetes and pre-diabetes are associated with obstructive sleep apnea in extremely obese subjects: A cross-sectional study. *Cardiovasc Diabetol* 2011;10:84. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3206416/

## *C*-statistic, *see also* receiving operating characteristic (ROC) curve

*C*-statistic is the area under a **receiving operating characteristic (ROC) curve**. This curve plots the **sensitivity** of a quantitative medical test against (1 – **specificity**). *C*-statistic measures the performance of the test in terms of combination of sensitivity and specificity. Thus, an understanding of these concepts is necessary to understand *C*-statistic. In brief, sensitivity of a test is the true positive rate that measures its ability to show positive result among those who are known to have disease, and (1 – specificity) is the false positive rate that measures the dubious ability of a test to show positive result among those who are known not to have that disease. The maximum value of *C* is 1.0.

A test such as serum omentin-1 can be considered as an inflammatory activity marker in Crohn disease [1]. This can be measured for the established cases as well as the healthy controls. Thus, sensitivity and specificity can be obtained for different levels of serum omentin-1. When these are plotted, the **area under the ROC curve** was $C = 0.87$. This was higher than $C = 0.76$ for *C*-reactive protein (CRP), indicating that serum omentin-1 has better performance than CRP as an inflammatory marker in Crohn disease. As in medical literature, we use the term *C*-statistics and AUC interchangeably in the context of ROC curves. This is also called *C*-index.

### *Area under the ROC Curve*

The primary utility of ROC curve lies in the area under the curve (AUC), denoted by statistic *C*. The larger the AUC, the better the overall performance of the medical test to correctly identify diseased and nondiseased subjects. The closer the ROC curve to the left and top border (see Figure C.31), the larger the AUC and the more valid the test in terms of sensitivity and specificity. In this figure, test A is better than test B, and test B is better than test C. If the test is lousy, for every true positive, as the level (e.g., of serum omentin-1) increases, you are likely to encounter a false positive. The ROC curve tends to flatten in this case, and nearly a diagonal line is obtained.

Because of inherent variations and uncertainties in all biological phenomena, no test can be perfect. It is considered excellent if *C* is 0.90 or more and good if *C* is between 0.80 and 0.89. An area

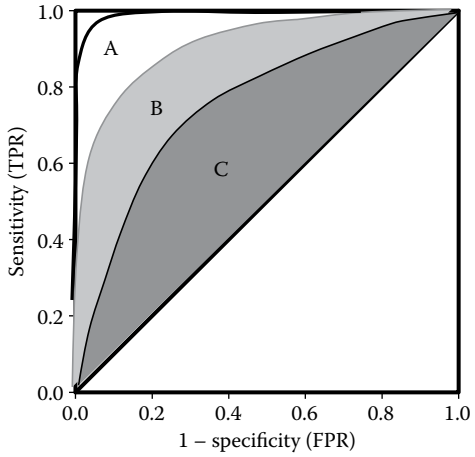



**FIGURE C.31** Three ROC curves with different areas under the curve.

of 0.50 corresponds to the diagonal and indicates that the test is absolutely not helpful. In a rare case, if this area is less than 0.50, conclude that the test is misleading. In this case, reverse the definition of positive and negative, i.e., in place of higher values signifying the presence of disease, examine if lower values will correctly pick up the disease.

As is true for sensitivity and specificity, validity of ROC and AUC depends on the "gold" being really so. If the gold is suspect, a high AUC does not necessarily mean a good test. Also, for ROC to be valid, the test must not be affected by what gold is. All other constraints for sensitivity–specificity also apply. AUC measures the performance with regard to combination of sensitivity and specificity and fails to balance the differential risks of misdiagnosis and missed diagnosis effectively. Note that these risks cannot come from the present data and should come from experience of the clinicians.

While statistically comparing two tests, the decision regarding which test is better depends on AUC. Thus, this area should be obtained by using an appropriate method. Statistical software packages provide nonparametric and parametric methods for obtaining the area under the ROC curve. The user has to make a choice. The following details may help.

**Nonparametric methods** are distribution-free and the resulting area under the ROC curve is called empirical. First such method uses trapezoidal rule. If sensitivity and specificity are denoted by $s_n$ and $s_p$, respectively, the trapezoidal rule calculates the area by joining the points $(s_n, 1 - s_p)$ at each interval value of the continuous test and draws a straight line joining the $x$-axis. This forms several trapezoids, one corresponding to each interval, and their area can be easily calculated and summed. This method is illustrated under the topic **area under the curve**, although that is area under the (drug) concentration curve. Another nonparametric method uses Mann–Whitney statistics, also known as **Wilcoxon rank-sum** statistic. Both these nonparametric methods of estimating AUC have been found equivalent [2].

Parametric methods are used when the statistical distribution of test values in diseased and nondiseased is known. **Binormal distribution** is commonly used for this purpose. This is applicable when test values in both diseased and nondiseased follow a normal distribution. In this case, the relevant parameters can be easily estimated by the means and variances of test values in diseased and nondiseased subjects. For details of how to obtain the area, see Zhou et al. [3].

The choice of method to calculate AUC essentially depends upon availability of statistical software. The binormal method produces a smooth ROC curve, and further statistics can be easily calculated but gives biased results when data are degenerate and when the distribution is bimodal. When software for both parametric and nonparametric methods is available, conclusion should be based on the method that yields greater precision of the estimate of AUC.

### *Issues in C-Statistic*

Equal AUCs of two tests represent similar overall performance of tests, but this does not necessarily mean that both the curves are identical. They may cross each other. Figure C.32a has hypothetical ROC curves of two medical tests A and B applied on the same subjects to assess the same disease. Tests A and B have nearly equal area but cross each other. Test A performs better than test B where high sensitivity is required, and test B performs better when high specificity is needed.

In cases where ROC curves for two tests cross each other and in some other situations, the interest may be restricted to specific values of sensitivity or specificity. You may be interested in a test with high sensitivity as for a disease with grave prognosis (cancer). Then the interest will be in test A and that too for specificity ≤ 0.65 or (1 – specificity) > 0.35. In that case, the area of interest is 3+4 as shown in Figure C.32b. This is called partial area under the curve. Some software packages calculate this also, and if you want, for easy interpretability, you can standardize it to 1 by considering the total area = 1 of rectangle from (1 – specificity) = 0.35.

Variance of AUC can also be obtained by using parametric and nonparametric methods. The formulas are complex. Software will give you this easily. This variance can be used to obtain the **confidence interval** assuming Gaussian pattern.

Formulas of sample size for testing hypothesis on comparing AUC with a prespecified value and for comparison of two on the same subjects or different subjects are available. See Zhou et al. [3] for details.

We have provided the details of *C*-statistic for binary groups, namely, with disease and without disease. Sometimes, the groups are **polytomous** such as with none, mild, moderate, and serious disease. In this situation, *C*-statistic can be used to see if a medical test is able to discriminate mild from moderate cases and moderate from serious cases. One alternative is to calculate *C*-statistics for mild versus moderate and another for moderate versus serious cases using

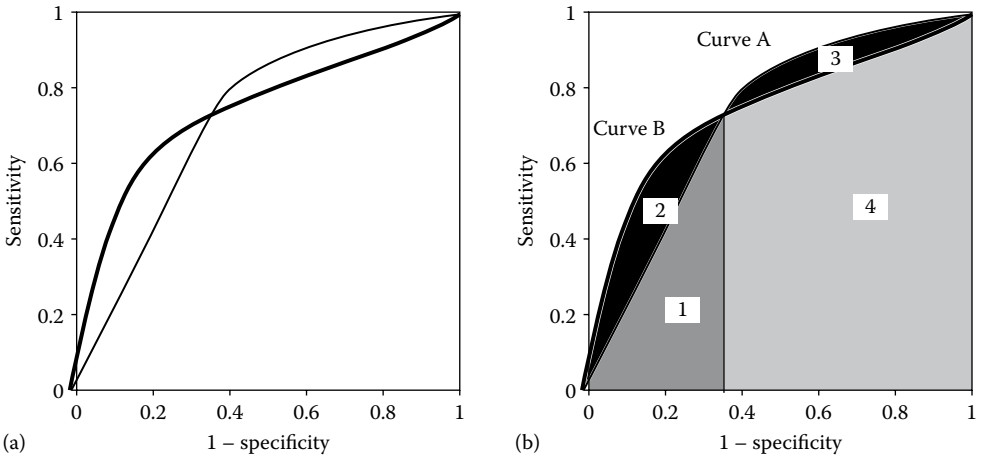



**FIGURE C.32** (a) Two ROC curves crossing each other but with nearly the same area. (b) Illustration of partial area under the ROC curve.

the binary procedure as stated in this section. But a *polytomous discrimination index* is also available based on *C*-statistic [4]. An example of such application is the study on discriminating between primary invasive and metastatic ovarian tumors, and between borderline and metastatic tumors [5].


1. Lu Y, Zhou L, Liu L, Feng Y, Lu L, Ren X, Dong X, Sang W. Serum omentin-1 as a disease activity marker for Crohn's disease. *Dis Markers* 2014;2014:162517. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3934650/pdf/DM2014-162517.pdf
2. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143:29–36. http://www.med.mcgill.ca/epidemiology/hanley/software/hanley_mcneil_radiology_82.pdf
3. Zhou Xh, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine*. Wiley, 2002.
4. Van Calster B, Van Belle V, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg EW. Extending the C-statistic to normal polytomous outcome: The Polytomous Discrimination Index. *Stat Med* 2012;31:2610–26. http://www.ncbi.nlm.nih.gov/pubmed/22733650
5. Van Calster B, Valentin L, Van Holsbeke C, Testa AC, Bourne T, Van Huffel S, Timmerman D. Polytomous diagnosis of ovarian tumors as benign, borderline, primary invasive or metastatic: Development and validation of standard and kernel-based risk prediction models. *BMC Med Res Methodol* 2010 Oct 20;10:96. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2988009/pdf/1471-2288-10-96.pdf


## cubic clustering criterion

Cubic clustering criterion (CCC) is used to decide when to stop a hierarchical clustering process so that the clusters obtained are externally isolated and internally homogeneous. For these terms, see **hierarchical clustering**. Hierarchical clustering is the process of merging most similar to less similar entities (group of units) into clusters in stages and sequentially obtaining from $n$ clusters containing one unit each to one cluster containing all the units. However, this process requires identifying a stage when it is considered that the clusters are optimal grouping of units in the sense that the similar ones are clustered together and the dissimilar ones are in the other clusters. CCC is one of the measures calculated at each stage of clustering process and is used to get a feeling of where such natural clusters have been formed. Thus, this helps to determine the number of natural clusters. CCC was developed by SAS Institute [1].

The expression and the method of computing CCC are complex and are explained by Sarle [1]. This is based on the **multiple correlation** coefficient $R^2$. A larger value of CCC indicates better clustering. Thus, the stage where CCC is the highest can be considered best for obtaining optimal clusters. However, this criterion fails if the units are correlated, and in that case, other criteria should be used such as semipartial $R^2$, pseudo-$F$, and pseudo-$t^2$ statistic. Gerlinger et al. [2] have used all of these including the CCC to find the number of clusters of women on sex hormones by studying their menstrual diaries. They found four clusters with desirable patterns and two with undesirable patterns—a total of six clusters.


1. Sarle WS. *Cluster Cubic Criterion*: SAS Technical Report A-108. SAS Institute, 1983. https://support.sas.com/documentation/onlinedoc/v82/techreport_a108.pdf
2. Gerlinger C, Wessel J, Kallischnigg G, Endrikat J. Pattern recognition in menstrual bleeding diaries by statistical cluster analysis. *BMC Women's Health* 2009 Jul 16;9:21. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2717079/


## cubic splines

This method uses cubic polynomials to extract a smooth pattern out of a scatter plot. The method is generally used in a setup where the usual methods of linear, quadratic, exponential, logarithmic, and such other regression functions fail to adequately represent the full pattern. Perhaps the most common use of cubic splines in health and medicine is in obtaining smooth growth curves of children such as for various percentiles of height, weight, and head circumference versus age.

A spline is an extension, say a wiggly shaped pipe, used to join another wiggly shaped pipe such that they can function together. Each pipe may have a different shape. For our purpose, these pipes are curves; for example, a cubic spline will draw a cubic curve, whereas a linear spline will draw a line. In statistics, a spline is a numeric function that defines polynomial for different pieces of the data range with a high degree of smoothness at connecting points. When the scatter does not have a uniform pattern over the entire range, it is split into smaller parts within which the scatter points can be adequately represented by a line or a curve (Figure C.33). In this figure, the entire scatter is divided into five parts. This is said to have four *knots* demarcating the parts—in this case at $x = 22$, 24, 29, and 31. Lines show linear splines, and the serpentine curve shows cubic splines. Note how cubic splines have extracted a smooth shape out of the scatter and how they are smoothly connected with the next at each knot. Cubic splines can take a lot of different shapes—Figure C.33 shows just one of them. The number of knots depends on the natural bents in the scatter. Generally, three knots for small samples and five to six knots for large samples are considered adequate. Location and the number of the knots can be left for the data to decide.

To understand how a smoothing problem can be challenged, consider centile curves that are often used to delineate norms of medical

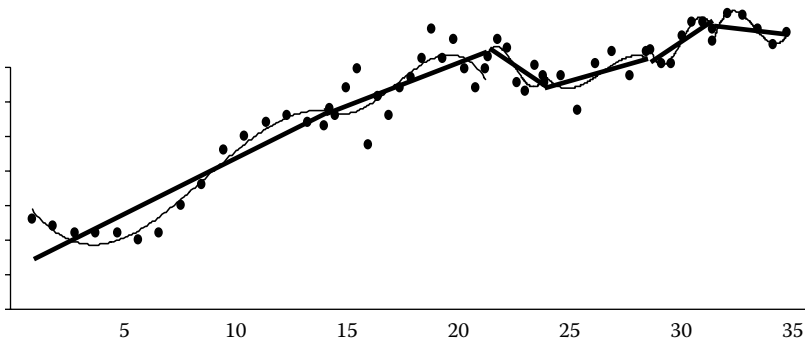



**FIGURE C.33** Linear and cubic splines with five knots.

C

parameters in a population. For these, estimation is done for many points in time. For example, for weight curve, one would obtain, say, 97.5th percentile for age 1 year, age 2 years, age 3 years, etc., to be able to obtain the centile curve. It is fair to expect that these percentiles at different ages would follow some kind of smooth trend. But, unfortunately, that would not be so in the data, mostly due to sampling fluctuations. These percentiles would, in all probability, follow an irregular pattern as are the **BCPE method**–based hypothetical estimates in Figure C.34. These are 95th percentile of aortic cross-sectional areas at the ascending aorta at different ages, similar to those obtained by Voges et al. [1].

In Figure C.34, observed values are shown by the solid line, linear trend by the dashed line, polynomial of degree 4 by the dotted line, and polynomial of degree 2 by the spaced dashes. The difficulty is to ascertain that flattening at age 15 years in this figure is real (does the aorta area really increase slowly between age 10 and 15 years?) or that it is just because of sampling fluctuation—another sample may not give this trend. If this flattening is genuine and we ignore this in our trend, important information regarding a slowdown just before age 15 years is lost. For delineating norms, no genuine information can be sacrificed. Also, a similar slowdown is noted after age 25 years. None of the four trends in Figure C.34 seems adequate to provide a real picture. Biological knowledge suggests that the aorta area increases rapidly till age 20 or 25 years and then the increase slows down, particularly in those who have a relatively large area for their age (say, those on the 95th percentile). Thus, a flattening between 25 and 30 years seems real but not between age 10 and 15 years. This is resolved by smoothing as it can provide the contextual shifts where needed.

It would be unrealistic that one age gives a high aorta area and the next a relatively low area unless there is a biological explanation. Extracting a realistic trend from erratic values without sacrificing useful information is a great statistical challenge. Eyeball trend can be fitted, but that lacks scientific basis and could vary from person to person. Thus, we take the help of methods such as cubic splines. This is done after suitably dividing the range of $x$-values (age, in our example) and fitting separate splines for each division.

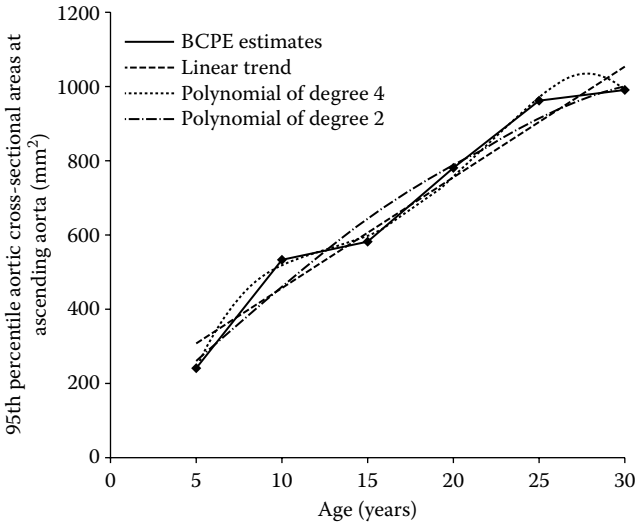



**FIGURE C.34** Erratic trend in plotted points: 95th percentile aortic cross-sectional areas at ascending aorta and the fitted polynomials. (From Indrayan A, *Ind Ped* 2014;51:37–43. http://www.indianpediatrics.net/jan2014/jan-37-43.htm, last accessed May 4, 2015.)

The calculations are intricate, and software help will be required for using cubic splines as a smoothing tool. For example, *GAMLSS* [3] software for growth curves uses cubic splines for smoothing the curves. But cubic splines seem to have not made much inroad to the popular statistical software yet. There is functionality SRS1 that adds cubic spline to MS Excel spreadsheet.

1. Voges I, Jerosch-Herold M, Hedderich J, Pardun E, Hart C, Gabbert DD, Hansen JH, Petko C, Kramer H-H, Rickers C. Normal values of aortic dimensions, distensibility, and pulse wave velocity in children and young adults: A cross-sectional study. *J Cardiovasc Magn Reson* 2012;14:77. http://jcmr-online.com/content/pdf/1532-429X-14-77.pdf, last accessed May 4, 2015.
2. Indrayan A. Demystifying LMS and BCPE methods of centile estimation for growth and other health parameters. *Ind Ped* 2014;51:37–43. http://www.indianpediatrics.net/jan2014/jan-37-43.htm, last accessed May 4, 2015.
3. GAMLSS. *Generalized Additive. Model for Location, Scale and Shape*. http://www.gamlss.org

## cumulative frequency, *see* ogive

## curvilinear regression, *see also* regression (types of)

This is a **regression** of $y$ on $x$ (or set of $x$'s) that graphically yields a curve instead of the usual straight line. A curvilinear regression is obtained when quadratic (square), cubic, logarithmic, or exponential type of terms is added among the regressors. Another term for this is **polynomial regression**.

For simplicity, we explain this with one dependent and one independent variable. Linear regression in this case is $y = a + bx$. It becomes curvilinear when

$$\text{curvilinear regression of power } K\text{:}$$
$$y = a_0 + a_1x + a_2x^2 + a_3x^3 + \ldots + a_Kx^K.$$

Depending upon the maximum degree of $x$, this is called polynomial regression of degree $K$. When $K = 2$, this is called quadratic or **parabolic curve** because of the shape it gets; for $K = 3$, it is called cubic; for $K = 4$, it is called quartic; etc. Each power of $x$ is considered as a new variable in this regression, although this may raise the question of **multicollinearity**. Figure C.35 shows a linear regression by a line, a quadratic regression by a thick curve, and a cubic regression by a dotted curve. A linear regression is a straight line with no turn, a quadratic will have one turn, a cubic will have

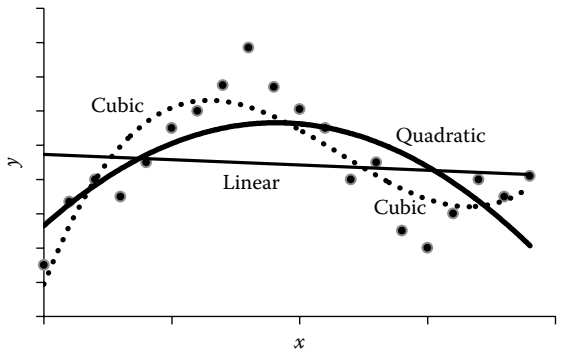



**FIGURE C.35** Linear and two curvilinear regressions with one regressor ($x$).

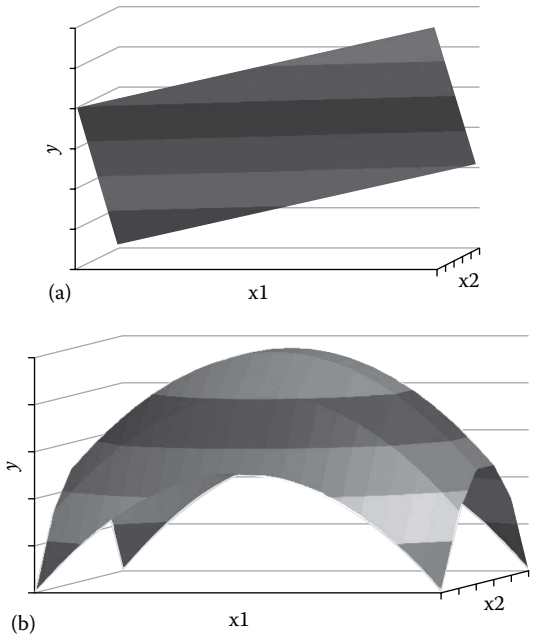



**FIGURE C.36** Regressions with two regressors: (a) linear and (b) curvilinear.

two turns, etc. A linear regression is appropriate when $y$ is increasing at a constant rate for each unit increase in $x$ such as between body temperature and pulse rate; quadratic is appropriate when $y$ shows an increase for an initial value of x and then levels off of or decreases (or vice versa); and cubic is appropriate when $y$ shows an increase then decrease, and then increase again. Regression of total glomerular filtration rate plasma on creatinine level is quadratic [1], and that of chronological age on dental age is cubic [2]. Polynomials can be fitted sequentially—first try linear then examine quadratic, and then go to cubic, etc. However, the regression coefficients will also change: the regression coefficient of $x$ in a quadratic equation will not be the same as in a linear equation.

Let us extend this concept to two regressors. A linear regression with two regressors will graphically give a plane (Figure C.36a). When quadratic terms are introduced, the plane bends as in Figure C.36b. Because of the complexity, this kind of regression is rarely used in health and medicine.

Exponential and logarithmic regressions also come under the ambit of curvilinear regression. For their shape, see **exponential curve**. This kind of relationship is frequently seen for bacterial growth over time (exponential growth) and for drug concentration in the body (exponential decay). Graham [3] has explored exponential regression of incidence of complete moles per 1000 viable conceptions on follicle stimulating hormone and luteinizing hormone in women of England and Wales. Kim et al. [4] used exponential regression for estimating the mean rate of annual endothelial cell loss for patients with no ocular surgery, those with pharmacoemulsification only, and those undergoing keratoplasty with cataract surgery.

1. Indrayan A. *Medical Biostatistics*, Third Edition. Chapman & Hall/CRC Press, 2012: p. 626.
2. Kiran ChS, Reddy RS, Ramesh T, Madhavi NS, Ramya K. Radiographic evaluation of dental age using Demirjian's eight-teeth method and its comparison with Indian formulas in South Indian population. *Forensic Dent Sci* 2015 Jan–Apr;7(1):44–8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4330618/
3. Graham IH. Are the pituitary gonadotrophins determinants of complete molar pregnancy? An investigation using the method of least squares. *JRSM Short Rep* 2013 Nov 21;4(12):2042533313505514. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3899734/
4. Kim YW, Kim MK, Wee WR. Long-term evaluation of endothelial cell changes in Fuchs corneal dystrophy: The influence of phacoemulsification and penetrating keratoplasty. *Korean J Ophthalmol* 2013 Dec;27(6):409–15. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3849303/



## cusum chart

A cusum chart is a tool of quality control similar to a **control chart**, but it depicts the cumulative sum of the differences from the standard. For quality control in a medical laboratory, a control specimen with a known standard value is analyzed each day, and the difference from this known standard is plotted. A control chart is for detecting the outliers, which are considered beyond tolerance, and cusum chart is for plotting trend. Sometimes this plot reveals an increasing or decreasing trend, indicating a need to revisit the functioning of the laboratory. Some subtleties of the daily change in the values are revealed much better by a cusum (cumulative sum) chart as in Figure C.37 for the data in Table C.46. This table gives differences of the values in the control specimen from the known standard each day and the cumulative differences.

Even though no trend is discernible in this plot and no value is possibly outside control limits, all values are on the positive side of zero. This shows that there is some tendency to overestimate the value. Thus, it is worthwhile to examine the laboratory procedures and functioning for possibility of a positive bias. This drift is more easily detected by a cusum chart than by the conventional control chart.

Wolfe et al. [1] used a cusum chart to monitor the rate of reexploration for excessive bleeding after cardiac surgery in adults. They observed that the cusum chart has become an important part of the quality feedback of clinical care outcomes. Sibanda and Sibanda [2] have illustrated its use in clinical monitoring of clinical outcomes using routinely collected data. Thus, cusum chart has applications outside laboratory also.

1. Wolfe R, Bolsin S, Colson M, Stow P. Monitoring the rate of reexploration for excessive bleeding after cardiac surgery in adults. *Qual Saf Health Care* 2007 Jun;16(3):192–6. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2464986/
2. Sibanda T, Sibanda N. The CUSUM chart method as a tool for continuous monitoring of clinical outcomes using routinely collected data. *BMC Med Res Methodol* 2007 Nov 3;7:46. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2204022/

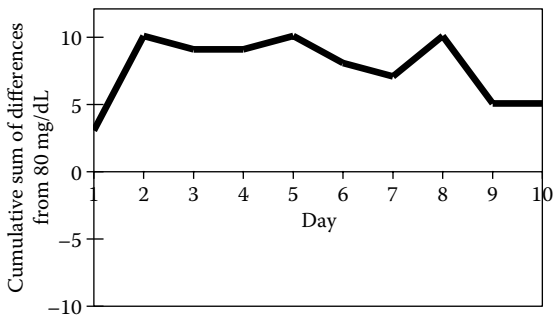



**FIGURE C.37** Cusum chart for the data in Table C.46.

**TABLE C.46**
**Serum Glucose Readings in Control Specimen**

| | Day | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| Serum glucose level (mg/dL) | 83 | 87 | 79 | 80 | 81 | 78 | 79 | 83 | 75 | 80 |
| Difference from the standard (80 mg/dL) | +3 | +7 | –1 | 0 | +1 | –2 | –1 | +3 | –5 | 0 |
| Cumulative sum of difference from 80 | +3 | +10 | +9 | +9 | +10 | +8 | +7 | +10 | +5 | +5 |

## cyclic model/trend

Cyclic trend is periodic high and low values of a variable at regular intervals. Basically this represents repetition. A model that represents a cyclic trend is called a cyclic model. Sometimes this is called an *oscillating model*.

Intraday circadian rhythm is well known for sleep, blood pressure, and body temperature. They follow a 24-h cycle. Menstrual bleeding and everything associated with it follow a 4-week cycle in most women. Incidence of seasonal diseases such as dengue, viral fever, and heat stroke follows a 1-year cycle. Asthma and inflammatory polyarthritis are cyclic as they have natural remission periods, although the period may differ from person to person. The highest point attained during each cycle is called *peak* and the lowest is called *trough*. The gap between one peak to the other peak is the periodicity of the cycle. This may remain constant or may increase or decrease with time. The distance of peak (or trough) from the middle is called the amplitude provided that the trend is symmetrically up and down (it goes up as much as it comes down). Cyclic trend may also be accompanied by increasing or decreasing or any other **secular trend**. In Figure C.38a, the secular trend is increasing but also leveling off with time. Many experts distinguish between seasonal and cyclic trends—a peak within each year is seasonal, and a peak every few years is cyclic. For example, a country may have a 6-year cycle for influenza.

Although regular cycle is typical of a cyclic trend, this may not be so in some cases. Figure C.38b shows body temperatures recorded in a person who was diagnosed with a case of nodular sclerosing Hodgkin's lymphoma similar to the one reported by Good and DiNubile [1]. Note the peaks and troughs.

For fitting a cyclic model to any data, first attempts are made by using the sine function, which is known to oscillate. In its simplest form, this takes the equation

$$y = a^*\sin(bx - c) + d,$$

where $y$ is the variable under study, $x$ is the time, constant $a$ defines amplitude, $b$ depends on periodicity, $c$ depends on the horizontal shift, and $d$ depends on the vertical shift. In a more complex situation, all these may depend on the value of $x$, i.e., they vary with time. For details, see Hyndman and Athanasopoulos [2]. The other method is **autoregressive moving average (ARMA)** model used in **time series**. Elementary information on this method is available in this volume under these terms.

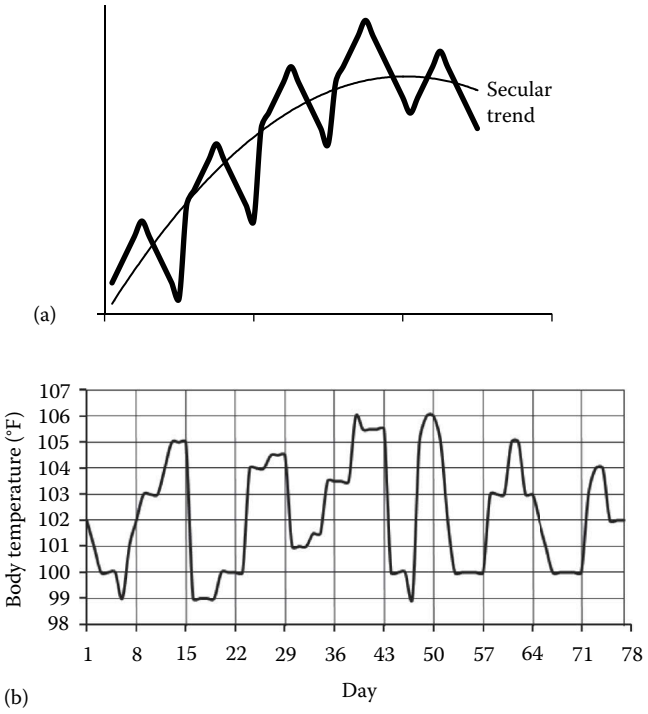



**FIGURE C.38** (a) Cyclic trend accompanied by increasing secular trend. (b) Fever recorded for 78 days in a case of Hodgkin lymphoma.

1. Good GR, DiNubile MJ. Cyclic fever in Hodgkin's disease (Pel–Ebstein fever). *N Engl J Med* 1995;332:436. http://www.nejm.org/doi/full/10.1056/NEJM199502163320705
2. Hyndman RJ, Athanasopoulos G. *Forecasting: Principles and Practice*. OTexts, 2013.

# D



## DALYs, *see* disability-adjusted life years (DALYs)

## data analytics, *see also* data mining

Data analytics is developing actionable insights by analyzing a set of data for preconceived notions using appropriate statistical methods, and its communication to the target audience. Thus, data analysis is just one component of data analytics. More important is developing insights—this could mean looking at the analysis critically and locating the right signals that could help in making decisions. Communication is mostly in terms of data visualization and tabulation that summarizes the data in a meaningful manner.

In other words, data analytics is the investigation of meaningful patterns in the data about which you have suspicion, mostly using tabulation and graphics, and communicating them. Data analytics merely tries to find the suspected statistical pattern in the data, for which statistical methods such as regressions can be used to discover the patterns. Data analytics would confirm or deny a suspected relationship, or would place a quantitative value on the extent to which this relationship exists and in what types of cases. This is more about associations and correlations than causation. Thus, the results of such analysis may not be supported by well-designed clinical trials. A natural prerequisite for data analytics is that the data are recorded in a structured format that can be systematically analyzed. Analytics relies on optimal combination of computer programming and data analysis methods that can quantify the anticipated outcome. After some trials of this nature, it is possible to develop computer programs that would drive fully automated decisions on specific aspects of a data set.

An associated term is **predictive analytics**, which is concerned about future projections, resembling statistical **regressions**. For example, when sufficient past data are available in records, a profile of cases with successful stent implants can be predicted with considerable precision. This can rule out certain types of cases, which can lead to savings in cost and inconvenience. The success of this exercise depends on correct records and their systematic organization. These aspects should be considered at the time of developing the database structure as much as possible, and corrective steps taken when something is found amiss while undertaking the analytics. In case a method such as regression is used, all the requirements for its application must be validated. All predictions assume that the past trends will continue at least for some time in the future, if no suitable adjustments are made for the anticipated changes in the future.

Statistical methods referred to in the context of data analytics are not probability-based inferential methods from a sample to the target population but just analysis without recourse to **estimation** or **testing of hypotheses**. Confidence intervals for the predicted values and ***P*-values** relating to a **null hypothesis** are excluded unless some random sampling is done. Analytics would be mostly tabulation and graphics, but statistical methods such as regression are not excluded. Realize that the inference methods from sample to population are the core of statistical science that are not used in data analytics. Although data analytics, in its pure form, is intensive data analysis that does not lead to any probabilistic conclusions, many would like to consider the data on hand as some kind of sample and include statistical methods too within the fold of data analytics. That could make the two subjects indistinguishable.

There is a thin line between data analytics and **data mining**. Analytics requires that you have some preconceived ideas, howsoever vague, about the kinds of relationships that can possibly arise, and the analysis is geared to investigate those relationships. Data mining, on the other hand, is like groping in the dark in the hope of laying a hand on gold at a place that you suspect will have treasure. It is trying to discover the patterns about which possibly nothing is known beforehand. Data mining allows you to search through enormous quantities of data without having any idea what you are specifically looking for. When a huge data set is available, a relationship may be accidentally found through data mining between the variables $x_j$ and $x_k$ in a specific type of cases that was not visualized earlier. It discovers relationships simply by performing extensive analysis in a networked manner.

Data analytics is a newly emerging science in medicine but has a lot of promise. For example, according to a study in 2013, only a small number of oncology providers in the United States (7%) possess the analytic tools and capabilities, although there is an expectation that a majority of oncology providers (60%) will have developed such capabilities within the next 2 years [1]. Google surprised the world in 2009 when they were able to track the spread of influenza in the United States faster than the Centers for Disease Control (CDC) just by finding a correlation between online searches and whether they had flu symptoms [2]. Watch out for more such surprises as data analytics gets more respectability and is used more often. Online data are piling up at an exponential speed and waiting to be exploited.


1. Barkley R, Greenapple R, Whang J. Actionable data analytics in oncology: Are we there yet? *J Oncol Pract* 2014 Mar;10(2):93–6. http://jop.ascopubs.org/content/10/2/93.long
2. Harford T. Big data: Are we making a big mistake? *Significance* 2014;11(5);14–9. http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2014.00778.x/pdf


## database systems

It is necessary first to describe what is meant by a database. A database is an organized collection of data, structured in such a manner that it may be accessed easily by a wide variety of application programs. Medical care facilities are increasingly creating large clinical databases on the patients they serve. For example, the Centers for Medicare and Medicaid Services [1] collect information on all Medicare patients' hospitalizations, surgical procedures, and office visits. These databases are primarily created to facilitate health care delivery but can be used for research through **data mining** and structured investigations. The Genome Database is another popular example of a medical database. Smaller databases are regularly generated by empirical studies that are currently the core of medical research publications.

The general format of a medical database is all the information for any one subject arranged in different mutually exclusive columns (called fields) in one row (called record), and one dedicated row for each subject. If your sample has 220 subjects, the database will have 220 rows. All the information in each column will be in a predefined format. For example, if you have a column for sex, the entries could

**TABLE D.1**
**Example of a Brief Database of Breast Cancer Patients**

| Sl. No: | ID | Present Age (Years) | Age at Detection (Years) | Nipple Discharge | Extent of Ulceration | Months of Itching | Family History |
|---|---|---|---|---|---|---|---|
| 1 | | 43 | 42 | 1 | 0 | 0 | 1 |
| 2 | | 51 | 46 | 0 | 2 | 40 | 0 |
| 3 | | 39 | 39 | 0 | 1 | 1 | 2 |
| Notes | | Age at last birthday | In completed years | 0 for no, 1 for yes | 0 for no, 1 for small, 2 for large, 3 for extensive | 0 for no itching, months if yes | 0 None<br>1 Mother<br>2 Sister<br>3 Other |



be Male or Female, or M or F, or 0 or 1, but you cannot have a mixture—M for one person and Male for another person. Table D.1 is a simple example that illustrates the essentials of a database. Note how no/yes type of information is noted in the same column.

Now for database *systems*. More fully known as database management systems, these are tools for efficiently creating and managing large databases that can persist over a long period of time [2]. These are computer-based systems, organized for systematic management of large structured collections of information that can be used for storage, modification, and retrieval of data. Querying is a special feature that lends respectability to a database. Thus, systems include some elements of **data analytics** that databases by themselves do not. Access to a database is usually restricted so that only authorized people can enter, delete, or modify the data, and query the data. However, when a database is complete, as for tobacco surveys in different countries, data can be placed for public access [3]. Anybody can download the data and use them the way they want, but the database at the access site remains undisturbed.

Database systems let you organize any type of information, including patients, prescriptions, laboratories, images, and outcomes. Sometimes, a separate database is prepared for special cases, such as for deaths, and a unique code is provided to link this to the master database. The information stored in a database can be in the form of numerical data, text, photos, PDF files, sound files, illustrations—pretty much any information can be stored on a computer. However, in this book on biostatistics, we have restricted ourselves to numerical and textual data. Textual data will be for qualitative variables. There must be linkage between different records belonging to the same subject, concerned hospital, etc., so that they can be analyzed with the proper perspective.

There are a large number of examples of useful research from hospital databases. A recent example is identifying cases of physeal fractures of the distal radius that could not be classified according to the Salter–Harris system but were fully classified by a modified nomenclature [4]. But hospital databases generally do not represent the general population, and one has to be careful in extrapolating the results.

1. Centers for Medicare & Medicaid Services. *CMS covers 100 million people...* http://cms.hhs.gov/
2. Garcia-Molina H, Ullman J, Widom J. The worlds of database systems. Chapter 1 of *Database Systems: The Complete Book*, Second Edition. http://infolab.stanford.edu/~ullman/fcdb/ch1.pdf
3. CDC. *Smoking and Tobacco Use*. http://www.cdc.gov/Tobacco/global/gtss/index.htm
4. Sferopoulos NK. Classification of distal radius physeal fractures not included in the Salter–Harris system. *Open Orthop J* 2014 Jul 11;8:219–24. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4133925/

## data collection, *see also* data quality

Data collection is the process of gathering and measuring information on variables of interest in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes [1]. Data collection can also be done for administrative purposes, but our concern is with research endeavors, particularly in health and medicine.

Systematic data collection requires methods that could achieve the objective and tools that are adequate for the purpose. These are described here.

### *Method of Data Collection*

Medical data are mostly obtained by observation, by interview, by examination, or by investigations. Data on some behavioral variables are obtained by observation only. This may be important in the case of psychosomatic disorders. The primary methods in a clinic are interview and examination, but realize that an interview may or may not reveal the full truth, and the data obtained by examination are most reliable in the sense that they can be largely believed to be correct. This is even truer for laboratory and imaging investigations, which form the core of diagnostic tools these days. Some of these methods are very expensive to adopt or just not available in a particular setup, and the clinician may have to resort to interview and physical examination, or to a test of lower quality. This can compromise the credibility of data.

For example, a patient with liver disease can be identified either by asking the individual whether the disease has already been diagnosed, or by looking at health records, or clinical examination, or carrying out some biochemical tests. Visual acuity in a subject can be assessed just by noting the power of the spectacles worn, if any, or by a proper visual examination by an optometrist. Each method of eliciting information has merits and demerits in terms of **validity** and **reliability** on one hand, and cost and time on the other. These factors have to be assessed in the context of the problem at hand and the resources available. Factual data, based on examination, are often given more weight than data obtained by the other methods. Note, however, that knowledge, opinion, and complaints as revealed by interview and observation, particularly for symptoms, have an important place in the practice of medicine.

Comprehensive documentation of the collection process before, during, and after medical maneuvers is essential to preserve data integrity. In a research setup, the documentation needs are identified at the time of writing the protocol and adhered to during the conduct of the research.

### *Tools of Data Collection*

The tools of data collection include various instruments used to collect data. These can be **questionnaires, schedules, and proformas**, besides existing **medical records**. All these can be newly developed or can be modifications of existing ones. Each has its own merits and demerits, as discussed under these respective topics in this volume. We are not discussing them all over again here, but we leave a note that perhaps more important than the structure and content of these tools are the instructions to complete the forms and faithfully following these instructions.

1. North Illinois University. Responsible Conduct in Data Management: Data Collection. http://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dctopic.html

## data dredging, *see also* data snooping

The term *data dredging* is used for two unethical practices. First is playing with the values you have actually observed. Because of availability of software packages, it is now easy to reanalyze data after deleting some inconvenient observations. Valid reasons for deleting observations include the presence of known outliers, but this can be misused by excluding some data that do not fit the hypothesis of the investigator. It is extremely difficult to get any evidence of this happening in a finished report. The integrity of the workers is not and cannot be suspected unless evidence to the contrary is available. Thus, data dredging can go unnoticed. Also, there is hardly any penalty if the data dredging was discovered.

Second, data dredging also refers to the practice of examining comparisons within a data set not specifically planned prior to the start of the study. Also called *data fishing*, data dredging is a form of data mining practice in which large volumes of data are explored for any possible relationships between data. These relationships are not those that were planned. The traditional scientific method, in contrast, begins with a hypothesis, and the data are examined for whether they support or contradict the hypothesis. Data dredging is sometimes described as seeking more information from a data set than it actually contains. This is also called *P-hacking* since hidden significant ***P*-values** are filtered after intensive search and analysis of the data. You may be aware that at a 5% **level of significance**, 1 in 20 *P*-values is likely to be significant by chance alone.

Howsoever unethical it may sound, the second type of data dredging can sometimes lead to unexpected findings. Thus, it has rewards along with the perils. To rule out that this finding is coincidental, a full-fledged study can be planned focused on that aspect so that a more reliable conclusion based on accepted scientific principles can be reached. Or, other similar data sets can be examined if they also give you the same conclusion.

While not much can be done to avoid the first type of data dredging except to call on the integrity of researchers, the second type can be avoided by investing more time and intellect at the time of planning the study. However, this kind of data dredging, when properly acknowledged, may be acceptable to generate a hypothesis for further investigations.

## data entry

This is the process of typing data values recorded in a questionnaire or schedule onto a **database**. A database will generally have a standard format—thus, data entries are made in a uniform format and in a structure that is amenable to computer analysis. This looks like a simple process but can jeopardize the analysis and findings if not done with sufficient care. The difficulty arises on two counts: first, many data forms, when filled manually, may have errors or illegible writing, and second, the data entry operator may be careless and not enter correct data as recorded in the form.

Numeric entries, dates, and text entries should be properly identified, and there must be uniformity across entries. This would require, for example, that "Male" should not be entered as "male" or "M" in other places. Also, as much as possible, there should not be any blanks. They should be clearly demarcated as Nil, No, None, Not available, Not applicable, Not answered, etc. In the case of nested studies, the level of the informant needs to be properly tracked. For example, if you have collected data from patients and doctors and hospital-level information, such as availability and utilization of staff, all these cannot be entered in the same worksheet. A separate worksheet may be required for each type of respondent, but the linkage codes may have to be provided so that the correct analysis can be done.

Substantial improvement in data entry can be ensured by planning. All steps needed for correct entries should be listed, and the data entry operators should be trained. Range checks and other consistency checks are performed to enhance the quality of entries. *Range check* means that the value or the response for a question must be within a plausible range. For example, parity can almost never be 17, and if such an entry occurs due to carelessness or otherwise, it will be immediately detected when the database flags such out-of-range entries. Large-scale medical investigations generally follow a system of dual entry of data and their cross-validation. The errors are resolved by going back to the form, and in some cases, this may even require recontacting the respondent for clarification.

Gadgets are now available that will record your response electronically and at the same time will also check for consistency. This automatically transfers data to a database format. Thus, some of the problems of data entry are eliminated. But some remain. In fact, such e-forms can provide unnecessary assurance of quality, and the investigator can become complacent in the hope that inconsistent responses will always be noticed. It is rare that all kinds of inconsistencies can be anticipated, least of all those incorporated in the program. Thus, such forms have to be devised with substantially more care and after a thorough discussion on entries that can be wrongly entered, and how to correct them in what manner. This can partly address the problem of wrong reporting too by providing checks and balances, flagging the responses requiring further probing in case any inconsistency or suspicion arises.

Use of SMS, Twitter, and a web page for surveys is becoming increasingly common. Even if their selective use and the resulting bias are ignored, such browsing has its own set of problems. Internet or mobile phone connections may terminate midsession, causing disruptions. Thus, there must be provisions for saving the information as entered. Then there are data security issues that can threaten data integrity and force you to devise steps to ensure that the responses are not used without authorization.

## data management

Data management can be considered as the mother of all data-related activities. In its broad sense, it includes planning for the collection of data; preparation of tools for data collection; and the actual process of collection of data, data collation and cleaning, and data analysis. This also includes the development and execution of policies and procedures around the data. The ultimate purpose is to have a system that can meet the information needs of an organization or of a project. In the context of a research endeavor, **data management** is the process of controlling the information generated during the project. Any research will require some level of data management,

and funding agencies are increasingly requiring scholars to plan and execute good data management practices [1].

The core of data management is data collation. This primarily starts after the data become available and comprises activities such as data scrutiny for incompleteness and misreporting, data entry such that all the data are properly transferred to the worksheet, cleaning for possible errors in entry, and arranging data in a manner such that they can be easily compiled and tabulated. Thus, the structure of the **database** in which the data are entered is important. Any data transformation, such as weight and height into body mass index, should be clearly specified and should be built-in. Complete data must have a backup that can be used in case a crash occurs. All the rules must be laid out and implemented with regard to when the entered data can be changed, who can access them, and what possible analysis is targeted. Policies for data security and data sharing should also be devised and publicized among the stakeholders. A policy for the preservation of the database should also be framed.

Data management is an essential part of the empirical research process. This provides a backbone for the research so that it can stand up to any review and be able to answer any question that might come up at a later stage. This helps the investigators in becoming accountable for their research. Data management is easy for small-scale research but can be challenging for large-scale endeavors, particularly in a **multicentric study** or when several workers are involved. All may have to follow a uniform format that is designed beforehand at a central place, although extra entries can be planned to meet the special constraints or needs of different centers.

1. Penn State University Libraries. *Publishing and Curation Services.* http://www.libraries.psu.edu/psul/pubcur/what_is_dm.html

## data mining

Data mining is discovering the patterns among different variables in a huge **database** about which possibly nothing is known beforehand. This is like a treasure hunt in a suspected forest. Data mining allows you to search through enormous quantities of data without having any idea of what you are looking for specifically. This uses the statistical method of **pattern recognition** to discover any buried relationships. When a huge database is available, a relationship may be accidentally found through data mining between the variables $x_j$ and $x_k$ in specific types of cases that were not visualized earlier. A popular perception is that it discovers relationships simply through brute-force analysis and neural network learning techniques. Thus, this is also sometimes called *knowledge discovery*, a name that has given it much respectability. The term that denigrates it is *data fishing.*

Big data are both large $n$ and large number of variables in a repository. In fact, they are enormously large compared with, say, a clinical trial setup. Contrary to the common belief, such a large $n$ does not reduce the bias but instead tends to magnify it if present. That, however, does not deter conclusions for the kind of subjects we have in our database. There is no attempt to extrapolate these results to any larger population except possibly those who are likely to become part of the same database in near future.

On a practical plane, data mining is exploring a database from several different perspectives, and analyzing its different dimensions with different angles, summarizing the analysis in several different ways, and critically looking at the results to identify some useful relationships. The art is in catching patterns that are not otherwise visible, nor even suspected. Once such unanticipated patterns are identified, they are tested for validity and further collated to convert into information and, subsequently, into knowledge. Remember, though, that data mining is not searching for credible evidence for something that was suspected. The only suspicion to begin with is that some useful relationships are hiding, without knowing the nature of these relationships.

The following quote sums up the latest about data mining very well. "Data mining and knowledge discovery techniques have greatly progressed in the last decade. They are now able to handle larger and larger data sets, process heterogeneous information, integrate complex metadata, and extract and visualize new knowledge. Often these advances were driven by new challenges arising from real-world domains, with biology and biotechnology a prime source of diverse and hard (e.g., high volume, high throughput, high variety, and high noise) data analytics problems" [1].

There is a thin line between **data analytics** and data mining. Analytics requires that you have some preconceived ideas, howsoever vague, about the kinds of relationships that can possibly arise, and the analysis is geared to investigate those relationships. Since a relationship was already suspected, data analytics would confirm or deny it, or would place a quantitative value on the extent to which this relationship exists and in what types of cases. In the case of data mining, no such hint on a preconceived relationship is available. A large number of variables are investigated for a possible relationship in the hope that accidentally, some useful pattern would be discovered. Both data analytics and data mining lack a probabilistic base and do not intend to generalize to the "population" from the sample—thus, hard-core statisticians, who are trained for probabilistic conclusions, are reluctant to join the fun. Statistical methods used in data mining are mostly restricted to cross-tabulations, and graphics although methods such as regression are also used to find relationships.

The big data deluge is occurring these days due to accumulation of data with online entries. In our context, medical records in hospitals are the warehouses of such data. They are very likely to contain useful information on the patterns and relationships between at least some of the large number of variables such a database generally has. They may conceal fascinating stories that need to be extracted and told. But discovering these patterns is a difficult process. A statistical tool called *mutual information* [2] provides a way to identify patterns without much reliance on the prior assumptions. This tool quantifies the relationship between two variables of any type. Data visualization is an important part of this process.

Data mining today is being increasingly used by large hospitals with a consumer focus. They integrate clinical data and laboratory and radiological investigation data with transactional data, and try to identify groups that have a similar usage pattern. That could possibly determine which services are amenable to promotion. This could also lead to substantive medical research, such as which signs and symptoms happen together more often in cases of a specific type, or which groups of patients happen to respond to particular therapy. For example, Raschi et al. [3] mined a database of more than a million reports to assess the association of drug-induced liver injury with antimycotics in a postmarketing setup. However, doubts are raised about the conclusions reached from data mining and other studies based on the existing data. Among nearly 50 health claims arrived at by examining existing data (including the data mined) for possible associations, not a single one could stand up when a follow-up controlled experiment was done to confirm that finding [4]. This could be partly due to the absence of **random allocation**, which tends to equalize not just the known but also the unknown factors in a clinical trial setup, and due to the fact that the existing data may not be representative of the general class of subjects. Thus, uncertainty lingers with regard to the conclusions reached by data mining.

1. Bacardit J, Widera P, Lazzarini N, Krasnogor N. Hard data analytics problems make for better data analysis algorithms: Bioinformatics as an example. *Big Data* September 2014. http://online.liebertpub.com/doi/pdfplus/10.1089/big.2014.0023
2. Kinney JB. Mutual information: A universal measure of statistical dependence. *Biomedical Computation Review* June 18, 2014. http://biomedicalcomputationreview.org/content/mutual-information-universal-measure-statistical-dependence, last accessed April 13, 2015.
3. Raschi E, Poluzzi E, Koci A, Caraceni P, Ponti FD. Assessing liver injury associated with antimycotics: Concise literature review and clues from data mining of the FAERS database. *World J Hepatol* 2014 Aug 27;6(8):601–12. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4163743/
4. Young SS, Karr A. Deming data and observational studies: A process out of control and needing fixing, *Significance* 2011;8(3):116–20. http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2011.00506.x/pdf

## data (nature of)

Data are the result of measurements of the object(s) of interest. Raw data are those that are yet to be processed; once analyzed, the results can become the raw data for the next step, if appropriate.

Blood glucose level and urea clearance are measured in terms of quantities; blood group is a quality recorded as O, A, B, or AB. Age can be measured in terms of years, months, days, and hours but is often categorized into years as 0–4, 5–14, 15–49, etc., particularly for reporting. Disease severity and extent of malnutrition are quantities but are generally measured as none, mild, moderate, or serious. Thus, the nature of data varies from measurement to measurement. The statistical analysis methods depend on the nature of the data you want to analyze.

A scale is a tool on which measurements are made. It can be quantitatively calibrated in the usual sense, or it can be qualitative. Not all measurements are necessarily in terms of quantities. Cancer sites are a quality but are still a measurement, statistically speaking. They are known only by their name: they are nominal in nature. They could be dichotomous with just two categories or polytomous with multiple categories. Such measurements are analyzed by using proportions instead of means. Methods such as odds ratio and chi-square test are prominent in this setup. Compare nominal characteristics with those that can be exactly measured in terms of a quantity (such as heart rate) on what is known as a metric scale. Sometimes, the metric scale is divided further into interval and ratio scales. In an interval scale, there is no absolute zero: body temperature, for example. In a ratio scale, a zero point can be meaningfully designated. For example, survival of 6 years is twice as much as survival at 3 years. In an interval scale, a statistically important quantity is the difference, whereas in a ratio scale, it is mostly the ratio. Both of these can be apportioned into intervals to give rise to **categorical data**. Another type of scale is the ordinal scale, made up of categorical variables in which the different levels of the variables are ordered. All these **scales of measurement** are discussed in detail under this topic.

The levels of measurements represent a successive increase or decrease in the characteristic conveyed by the variable. For example, severity of pain has the following levels: no pain, minimal pain, moderate pain, and severe pain. Note that the difference between any two consecutive levels does not necessarily represent increments of the same magnitude in this case. One should analyze ordinal data with special methods (e.g., **chi-squared test for trend**) and not with methods used for nominal variables.

Data on multiple or single measurements on subjects that are absolutely unrelated with one another are called **univariate** and are analyzed accordingly one at a time. Age, sex, and body mass index in adults are such measurements. If the measurements depend on one another, they are called **multivariate**. Liver functions such as bilirubin, albumin, and globulin are related and thus are multivariate. Repeated measurements of, say, creatinine levels in kidney patients over a period of time also are factually multivariate. These should be analyzed together by multivariate methods.

There are other aspects of the nature of data that are relevant to statistical methods. The data can be opinion based or factual. Clinical signs assessed by a qualified physician are factual, whereas symptoms told by the patient are opinions. The answer to "How are you feeling today?" is opinion, but measurement of blood pressure, heart rate, etc. is factual. One would naturally like to depend more on factual data than opinion-based data as opinion can quickly change and is subjective. Similarly, there are **hard data and soft data**. These also have to be handled differently, as mentioned under that topic.

## data quality

Data quality is a multifactorial entity comprising several aspects that together determine the fitness of data for the intended purpose. Most important of these is the **validity**—does it correctly reflect what it is supposed to? If the data are not valid, they cannot lead to right conclusions. Inaccurate or incomplete but valid data are considered better than accurate or complete but invalid data. It is better to have an approximate answer to the right question than to have a correct answer to a wrong question! Other aspects are completeness, **reliability**, freedom from errors, consistency, being up-to-date, and accessibility. Last is **accuracy**. We are using the term *accuracy* to mean being correctly measured to several digits and not as a replacement for validity, which many researchers seem to believe. Some may want to include adequacy of the data also in terms of sample size among data quality parameters, but let us keep that separate from this topic.

The importance of the quality of data can hardly be overemphasized. If the quality is poor, one cannot hope to come to valid decisions. This is like **garbage-in garbage-out** syndrome, where no amount of meticulous analysis helps. Some aspects of data quality such as **missing values** and **outliers** can be addressed to an extent by statistical methods before the data are analyzed, but there is no way to restore quality. Thus, it is important that steps are taken from the beginning so that high-quality data are available. This may require correct assessment of what data are to be obtained to achieve the stated objectives, proper identification of the subjects, adequate design of the form of recording so that the right information is elicited and recorded, detailed instructions and training for interview and examination of the subjects, use of standardized laboratory and radiological investigations, continuous monitoring and periodic appraisal, and a system of scrutiny. Another term for data quality is *data integrity.*

Obtaining correct data from subjects could be a challenge in some medical research situations. Sometimes, people deliberately hide information, and many times, they forget. Information on extramarital sex in the context of sexually transmitted diseases can be erroneous because of the stigma attached to such issues in some societies, and information on the source of injury can be misreported to avoid legal hassles. In the case of retrospective investigation into antecedents that occurred far away in time, it may be difficult to get the correct information because of memory lapses. Scrutiny of data also does not help much in these situations as the correct status is not

known. However, data scrutiny is an important step to detect inconsistencies, such as a person with male sex recorded as pregnant or a woman of age 65 years recorded as pregnant. These are examples of errors (as per the present knowledge) that may have occurred either at the time of recording in the case sheet or at the time of entry in the database. Scrutiny can also help in retrieving some of the missing data.

Most et al. [1] describe *quality assurance* and **quality control** as two approaches that can preserve data integrity and ensure the scientific validity of study results. Assurance includes activities we undertake before the collection of data to prevent corruption of the data, and control comprises activities we undertake during the process of collection of data and thereafter. Most of the steps we have listed in a preceding paragraph pertain to assurance. Besides identification of the appropriate variables and training, constant vigilance may be required to ensure that no violation of the protocol occurs. You may also have to identify steps to be taken in case aberrations occur. Scrutiny and double data entry are the steps for quality control. Feedback from this exercise may go to the data collection stage so that such errors do not recur.

1. Most MM, Craddick S, Crawford S, Redican S, Rhodes D, Rukenbrod F, Laws R; Dash-Sodium Collaborative Research Group. Dietary quality assurance processes of the DASH-Sodium controlled diet study. *J AmerDietetic Assoc* 2003;103(10):1339–46. http://www.ncbi.nlm.nih.gov/pubmed/14520254

## Data Safety and Monitoring Board

Also known as the Data Monitoring Committee, the Food and Drug Administration (FDA) of the United States defines the Data Safety and Monitoring Board (DSMB) as a group of individuals with pertinent expertise that reviews on a regular basis accumulating data from one or more ongoing clinical trials and who are interested in the trial but are strictly neutral to the results. The committee advises the sponsor regarding the continuing safety of trial subjects and those yet to be recruited to the trial, as well as the continuing validity and scientific merit of the trial [1]. The committee is supposed to be objective in its assessment, and report to the sponsors rather than the investigators.

A DSMB is primarily set up for a clinical trial because the intervention could be harmful and data quality is of prime concern, particularly for large-scale trials. All clinical trials should be designed to accommodate ethical concerns. This is true in at least two ways. Firstly, the researcher should assess the risks and benefits of treatment for the individual participants on an ongoing basis. Secondly, the researcher must make arrangements to constantly monitor the trial and determine if the evidence is strong enough to require a change in clinical practice, either with those participants already in the trial or new patients.

In many trials, such awareness regarding safety and efficacy comes about through the services provided by a DSMB. As mentioned, this committee oversees the ongoing trial with regard to treatment efficacy and safety. The board regularly meets at predetermined intervals, such as once in 6 months, and reviews the progress. They are provided unblinded data so that an **interim analysis** can be done. They can review external evidence as well if needed to support their decision. If there is strong evidence that the treatment is efficacious before the trial is due to stop, the DSMB will recommend early termination of the trial. In some cases, the trial may still be recommended to continue to monitor long-term side effects despite convincing evidence of efficacy. Interim analysis may also suggest increasing or decreasing the sample size, altering the method of allocation, changing the criteria for recruiting the subjects, or any other modification that, is in the interest of science in their opinion (see **adaptive designs for clinical trials**). Also, if serious unanticipated toxicities or side effects are discovered, the trial must be halted. Mortality and irreversible morbidity receive special consideration. The trial can also be stopped if sufficient evidence emerges that the intervention is not going to serve the intended purpose, called *futility*. There might be other logistic reasons for terminating the trial midway.

DSMB members should be completely independent of the study investigators, with no conflict of interest, either scientific or financial. Their neutrality is intended to control the adverse impact of sharing the premature information that could be sensitive to the integrity of the trial. They are typically drawn from experts in the fields of clinical, statistical, epidemiological, laboratory, data management, and ethical concerns. They will have a scientific interest in the trial, to the extent that they are looking to see a valid, well-performed trial, but will not have any interest in the success or failure of the regimen under trial. Only then can they be assumed to be neutral. However, due to a lack of subject-matter expertise of the DSMB members, they may be not as meticulous. Investigators learn quite a bit while running a trial but cannot share this practical knowledge with the DSMB. Thus, their contribution may lack scientific merit. But the advantages far outweigh these minor hiccups.

A DSMB should have standard operating procedures as decided amongst themselves before a formal meeting on substantive issues is convened. They must have the facility to record all the deliberations so that a proper review can be undertaken when needed. Conclusions of the deliberations are communicated to the sponsors in such a manner that the integrity of the trial is not compromised.

1. FDA Regulatory Information. *The Establishment and Operation of Clinical Trial Data Monitoring Committees for Clinical Trial Sponsors: Guidance for Clinical Trial Sponsors—Establishment and Operation of Clinical Trial Data Monitoring Committees.* http://www.fda.gov/RegulatoryInformation/Guidances/ucm127069.htm

## data snooping, *see also* data dredging, data mining

Data snooping is excessive unplanned analysis of data with the objective of discovering some kind of relationship or interesting finding. Planned analysis, howsoever extensive, is not snooping. This is only unplanned analysis, which is generally carried out after seeing the data. This may look like **data dredging**, and indeed, you may find overlapping usage. Dredging has an element of dishonesty, even cheating, that snooping does not have. It also can be confused with **data mining**, but data mining applies to huge data sets and rarely has probabilistic implications. Data snooping may involve repeated statistical tests of hypotheses and many confidence intervals that involve extrapolation to a target population with probabilities attached to each of them.

When considering **multiple comparisons** in **analysis of variance (ANOVA)**, there are a variety of possible tools, but many choose the **Bonferroni procedure** or the **Tukey test**. These procedures are valid only for *preplanned* comparisons. Sometimes, the idea of testing statistical significance between two or more groups arises after seeing the observed data, and this is what we are calling data snooping. Although multiple comparisons are often described as post hoc tests because they are generally done only after the $F$-test between the groups reveals significance, the tests indicated by data are also post hoc, but they are of a different type. Tests done after looking at the data distort the chances of statistical error, particularly the Type I error. Thus, they are considered indicative and

not definitive, and help generate a hypothesis for subsequent testing by another study.

## death certification, *see also* cause of death

A death certificate is a legal document that confirms that the person named has died. His/her age and sex and place of death are also mentioned. Other personal details can also be mentioned so that the person is uniquely identified. These can be parents' names, social security number, or any such identification. The certificate is helpful in claiming inheritance, insurance, etc. However, over the years, **cause of death** has become an integral part of death certificates, and this has made them an important source of information on how people are dying, at what age-sex, and where. This has great significance for developing a strategy to control deaths by certain causes and of specified age and sex. For example, you can find how many deaths of what age and sex are occurring due to fatal accidents on the road, how many due to suicides and homicides in which areas, and how many infants are dying due to asphyxia. However, the requirement of specifying the cause varies from area to area. For example, in the United States, natural causes can be assigned, although those are not medical causes. When properly recorded, such information can trigger a series of steps for the control of the dominant preventable causes.

A death is certified by a qualified person who can (i) correctly assess that the death has indeed occurred and (ii) properly identify the cause of death. Generally, only medically qualified personnel have this capability, although many who are medically qualified also may not be fully capable of correctly identifying the cause of death. The international recommendation is to record direct cause, antecedent cause, and other significant condition contributing to the death [1]. For details, see the topic **cause of death**. Although the death and its cause are certified by a qualified professional, the certificate is issued by the local registrar who is officially assigned by the government to do this job. This is mostly issued by the municipal agencies. The certificate would ordinarily contain the details of the informant; the demographic details of the deceased; the date, time, and place of death; the cause of death; and the signature and title of the registrar.

In the United Kingdom, at present, there is no legal definition of death, although guidelines do exist for the diagnosis of death in more complex situations. There has been recent guidance on the diagnosis and confirmation of death from the Academy of Medical Royal Colleges. The guidance is mainly concerned with confirmation of death in the hospital and in circumstances where the diagnosis of death may be more difficult (patients on ventilators, for example) [2].

1. WHO. *Medical Certification of Cause of Death*, Fourth Edition. World Health Organization, 1979. http://whqlibdoc.who.int/publications/9241560622.pdf
2. Patient. *Death (Recognition and Certification).* http://www.patient.co.uk/doctor/Death-(Recognition-and-Certification).htm

## death rates, *see also* mortality rates

Death rate is the number of deaths per unit of population per unit of time. The unit of time is mostly years so that most death rates are obtained per year. However, the unit of population changes depending upon the interest. Computationally, death rate is a ratio with the number of deaths in the numerator and the concerned population in the denominator. The unit of time makes it a rate that signifies the speed of occurrence. For example, the number of deaths during a year divided by the estimated size of the population midway through that year is the most common form of death rate. Frequently, this ratio is multiplied by a convenient base (such as 1000) to avoid small decimal fractions: then this becomes the annual death rate per 1000 population.

The terms *death* and *mortality* are interchangeable, but some professionals maintain a thin distinction between death rate and mortality rate. Death rate has the concerned population (exposed + unexposed) in the denominator, whereas mortality generally has only the exposed population. For example, **infant mortality rate** is the number of deaths of infants out of those born live. There are several types of death rate. More commonly used among them are as follows.

### Crude Death Rate

If the death rate is calculated for a population covering all ages and sexes, it is called a crude death rate (CDR). This is the number of deaths in an area in a year per 1000 population counted at midyear, i.e.,

$$\text{crude death rate} = \frac{\text{number of deaths in 1 year}}{\text{midyear population}} * 1000.$$

The CDR ranged from a low of 1.4 per 1000 population in Qatar to a more than 10-fold high of 16.2 in Congo in 2012 [1]. Thus, this rate varies widely from country to country.

A problem with the CDR is that it disregards the age structure of the population. For this reason, it is called *crude*. If people in an area are predominantly old, a high CDR is not as bad as in an area where the population is predominantly young. Thus, a CDR of 8 per 1000 population in Sweden should not be construed to mean that the health status is nearly the same as in India, where the CDR is nearly the same. In India, only 8% of the population are of age 60 years or more, whereas in Sweden, it is more than 20%. The death rate among old people is bound to be high, and therefore, the CDR naturally becomes high. A valid comparison is obtained when the rate is recomputed by assuming the same age structure in the two countries. This is one form of standardization. The other brings the age-wise death pattern to a common base. The details are under the topic **standardized death rates**. Both types of standardizations require age-specific death rates (ASDRs).

### Age-Specific Death Rate

When the numerator and the denominator in the equation mentioned earlier are restricted to a particular age group, we get the specific death rate for that age group. For example, the ASDR for age group 65–74 years is the deaths that occur in this age group per 1000 persons of age 65–74 years. Such a rate provides an adequate comparison of the health status in two areas or at two different times for that particular age group. In 2006, the ASDR in the age group 5–14 years was 1.0 in Peru but only 0.08 in Sweden per 1000 population of that age. The rate in Peru was more than 10 times that in Sweden in this relatively healthy age group. This legitimately highlights the qualitative difference in deaths in these two countries.

### Proportional Deaths

It is useful for health authorities to know the extent of deaths occurring because of various causes. The number one killer in the United States is vehicular accidents, whereas in Indian rural areas, it is respiratory diseases. Similar statements can also be made for the age

groups contributing to deaths. Such proportional deaths can be measured in several different ways. The most common are as follows:

$$\text{Proportional deaths due to cause A} = \frac{\text{deaths due to cause A}}{\text{total deaths}} * 100$$

$$\text{Proportional deaths in age group C} = \frac{\text{deaths in the age-group C}}{\text{total deaths}} * 100.$$

For example, proportional deaths due to cancer = 100*(total deaths due to cancer)/(total deaths by all causes). For age group 60+ years, proportional deaths = 100*(deaths in the age group 60 years and above)/(total deaths in all age groups). Multiplication by 100 puts it in terms of percentage. Both of these can be computed for any cause or any age group, and the denominator in both is total deaths. Although this is calculated for any particular time period, such as the year 2015, proportional deaths are not per year: this is not a rate. The proportional deaths for any cause are different from the cause-specific death rate.

### *Cause-Specific Death Rate*

A cause-specific death rate is the number of deaths from a specified cause for a population during a specified time period per thousand or per million population. The numerator is typically restricted to deaths of people residing in a specific geographic area to which the denominator belongs. For example, if there were 137 homicide deaths in New Mexico during the calendar year 2015 and estimated midyear population was 2,010,787 in that year, then cause-specific death rate due to homicides in New Mexico in 2015 is (137/2,010,787)*1,000,000 = 68 per million population.

$$\text{Cause-specific death rate (due to cause A)} = \frac{\text{deaths due to cause A in a year}}{\text{midyear population}} * 1{,}000{,}000.$$

This is basically the same as the CDR but is restricted to a particular cause of death. The sum of specific death rates for all causes would be the same as the CDR. Cause-specific death rates may be adjusted for the age and sex composition, or other characteristics of the population. When that is done, for instance, in the case of age adjustment, it is called an age-adjusted rate. If the rates in two areas are to be compared, then a **standardized** cause-specific death rate is computed.

1. The World Bank. *Death Rate, Crude (per 1,000 people)*. http://data.worldbank.org/indicator/SP.DYN.CDRT.IN

## death spectrum

Death spectrum is the eyeball picture of the causes of deaths that tells what percentage of deaths is occurring due to various causes in an area.

The health care profession is making every effort to prevent mortality from any cause. Few realize that the probability of death is 1—it can only be postponed and not denied. In the end, there will be some cause of death. In fact, various causes tend to compete with one another because the total probability is 1. Depending upon the biological, environmental, and demographic factors, the causes of death only change hands. If I do not die of tuberculosis, I may die of cancer, or in an accident. If one does not die in infancy, he/she may die at the age of 100 years. Perhaps the duration of survival is more important than the cause of death.

Over quite a while, the spectrum of causes of death has undergone a dramatic transition. Because of various health-promoting steps, infant deaths have substantially decreased, and correspondingly, deaths due to chronic ailments such as cancers, diabetes, and coronary artery disease have increased. This transition is the direct result of better health and increased longevity.

Since death is a certainty, this raises the question of whether some causes of death are more desirable than others. Medical science seems to have completely ignored this issue. The thrust all around is to control all the causes. This simply is not possible. It is time to debate which causes should in fact be promoted for death in *old age* and which should be controlled. Indrayan [1] has emphasized this aspect and opined that more people prefer sudden death in old age instead of a protracted slow death, which necessarily will be painful. There is no condition yet that would bring slow death but would still not be painful. A disease such as Alzheimer disease may not cause physical pain but does cause an intense psychological trauma. In Indrayan's opinion, myocardial infarction (MI) could be the most desirable cause of death in old age since it causes sudden death in many cases. But sudden death has negative features also. The person does not get time to meet near and dear ones, to pass on messages, to settle accounts, etc. Since the concern here is with death in old age, one can counterargue that the person should do all this at the time of reaching old age, say at 80 years or earlier, and not when death becomes inevitable.

If the contention that MI is the most desirable cause of death in old age is accepted, all research around the world will have to be reoriented. The risk factors for such deaths in old age have to be identified not for controlling death but for nurturing these factors during early life so that the chances of death by this desirable cause increase and, correspondingly, for other painful causes such as cancer decrease. Perhaps a choice should be available to individuals to die suddenly or slowly. The choice will be exercised not at the time of death but during one's lifetime by controlling risk factors of one type and promoting the other types that increase the chance of death by the promoted factors.

All of this is for deaths in old age only. Deaths at a young age by any cause, including MI, have to be averted as much as possible. Thus, there is a need to differentiate between risk factors of death in old age from the preferred cause and risk factors of death in young age—the former to be nurtured and the latter to be controlled. This kind of orientation is currently missing from medical research.

1. Indrayan A. Can I choose the cause of my death? *BMJ* 2001;322:1003. http://www.bmj.com/content/322/7292/1003.1

## deciles, *see* quantiles

## decision analysis/tree

Decision analysis is the formal process of making decisions after critically analyzing all available evidence for and against those decisions. It may involve several stages that act as mediators to reach a decision. These stages together form what is called a decision tree. We always make decisions in life, but the decision analysis formalizes the process so that nothing important is missed and we have more confidence in the decision we make. In health and medicine, decision analysis is an important component of **evidence-based medicine**.

In many situations, the cost of a false-positive conclusion is more than that of a false-negative one, and in some situations, it is vice versa. In the case of diagnosis, misdiagnosis of severe schizophrenia, requiring admission to a psychiatric ward, can cause severe strain on the patient, his/her family, and the medical care system. A false-positive diagnosis is more costly than a false-negative diagnosis in this case. The false-negative diagnosis would be caught in subsequent examinations if the problem persists. On the other hand, a missed diagnosis of leukemia is much more expensive in terms of loss of years of life than a false-positive diagnosis, which can possibly be rectified later on. Decision analysis in this case will consider these kinds of aspects. The chance of error cannot be eliminated, altogether but efforts can be made to keep both types of errors to a minimum. This is done by using a sufficiently valid test or by a combination of tests where feasible. The fact, however, is that errors do occur. The question is what type of error is more affordable considering the monetary cost, pain, and the risks involved. An approach can be evolved for each patient separately to minimize such costs.

Two important components of decision analysis are probabilities of various outcomes as available in the literature or record, and value judgment regarding the consequences of the action to be taken at different stages. The probabilities are assessed in terms of prevalence, incidence, risk, sensitivity, specificity, predictivity, etc. They must have an effective interface with clinical acumen so that they are examined in the context of the actual condition of a patient. Judgments regarding advising a test or not, treating or not treating, treating by medication or by surgery, discharging from the hospital or not discharging, etc. are subjective assessments based on the experience and knowledge of the physician. The final outcome depends on a judicious mix of these probabilities and judgments. A decision tree helps to visualize various possibilities and to act accordingly (Figure D.1). The value of a decision tree is substantially enhanced when utility is assigned to each possible outcome. This utility can be either to the patient, such as 0 for death and 1 for full recovery, or to society. Assigning these utilities to various judgments may be challenging in some situations. Once this is done, a decision tree maps all the pertinent courses of action and their consequences. For details, see Hunink et al. [1].

Medical decision trees generally assume the following process of patient management.

Patient → test → positive and negative predictivity → diagnosis → management strategy based on risks and benefits → efficiency of the services → outcome

Out of these, strategy based on risks and benefits is the key for evidence-based decisions. Risks and expected benefits can be assessed as follows. The most favorable situation is that there is no disease and it is correctly excluded. The patient is spared the unnecessary pain of undergoing the therapy and psychologically feels relieved, and there is no further cost (of treatment). The second most satisfying situation is that the presence of disease is correctly diagnosed, it is properly treated, and recovery is full. In between, there are several possibilities. The spectrum of possibilities is given in Table D.2.

The options provided in Table D.2 assume that the decision to treat or not to treat is guided solely by the test result—start treatment if the test is positive and no treatment if the test is negative. Similarly, no treatment is done when the diagnosis is ruled out by the test. However, test predictivity is never completely assured, and the test can mislead. Diagnosis may be missed, and a misdiagnosis can occur. If a clinician can start treatment despite a negative test and not start treatment despite a positive test, the possibilities are many more. Figure D.1 shows all such possibilities. The probabilities in this figure are positive **predictivity** of 85% and negative predictivity of 90%. The prevalence of disease among those with the reported complaints is assumed to be 70%. An oval indicates

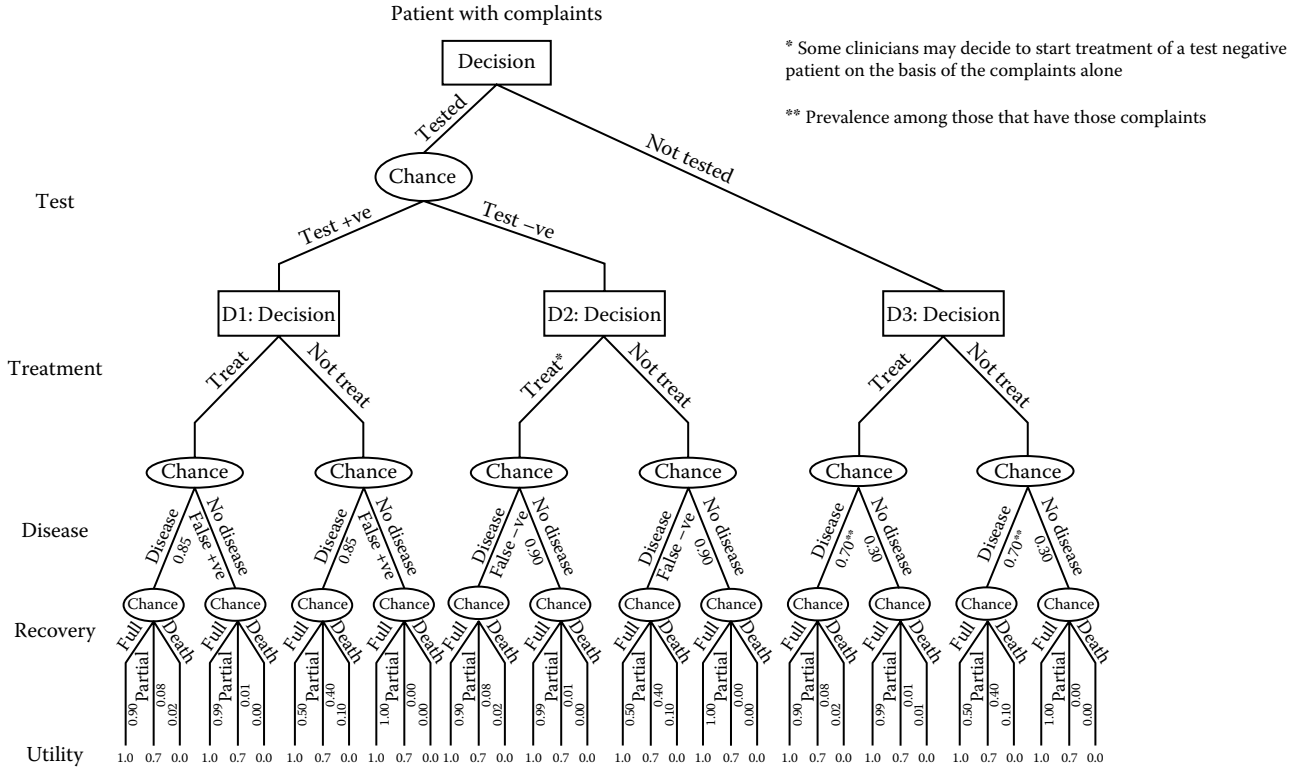



**FIGURE D.1** Example of a decision tree.

D

**TABLE D.2**
**Cost Involved in Various Situations of Disease, Diagnosis, and Treatment**

| Situation | Test | Test Outcome/ Clinical Diagnosis | Actual Status of Disease | Treatment Done | Recovery | Cost |
|---|---|---|---|---|---|---|
| 1. | Done | Test positive (Correctly diagnosed) | Present | Yes | Full | Test + treatment ① |
| | | | | | Partial | ① + disability |
| | | | | | Nil (death) | ① + loss of life |
| 2. | Done | Test positive (Misdiagnosis) | Absent | Yes | Full | Test + treatment ① |
| | | | | | Partial[a] | ① + disability |
| | | | | | Nil[a] (death) | ① + loss of life |
| 3. | Done | Test negative (Diagnosis missed) | Present | No | Full | Test ② |
| | | | | | Partial | ② + disability |
| | | | | | Nil (death) | ② + loss of life |
| 4. | Done | Test negative (Correctly excluded) | Absent | No | Full | Test ② |
| 5. | Not done | Disease diagnosed (Correctly diagnosed) | Present | Yes | Full | Treatment ③ |
| | | | | | Partial | ③ + disability |
| | | | | | Nil (death) | ③ + loss of life |
| 6. | Not done | Disease diagnosed (Misdiagnosed) | Absent | Yes | Full | Treatment ③ |
| | | | | | Partial[a] | ③ + disability |
| | | | | | Nil[a] (death) | ③ + loss of life |
| 7. | Not done | Disease ruled out (Diagnosis missed) | Present | No | Full | Nil |
| | | | | | Partial | Disability |
| | | | | | Nil (death) | Loss of life |
| 8. | Not done | Disease ruled out (Correctly excluded) | Absent | No | Full | Nil |

*Note:* Unlikely scenarios of administering the treatment when the test is negative or when the diagnosis is ruled out, or of treatment not done despite positive test or despite diagnosis of the disease, are excluded.

[a] Can occur due to side effect of treatment.

a **chance node**, where the outcome depends on probability, and a rectangle indicates a judgment node (more formally, **decision node**), where the clinician has to make a call.

A probability is assigned to each grade of recovery in different situations on the basis of the available evidence. If the evidence is not adequate, subjective probabilities based on experience are used. In Figure D.1, only three grades of recovery are shown for illustration—full, partial, and nil, where the last means death. For example, in the case of disease being present and treated, the probability of full recovery is 0.90, of partial recovery, 0.08, and of death, 0.02 in this example. When the disease is present and not treated, due to missed diagnosis or otherwise, the probability of full recovery is 0.50, of partial recovery, 0.40, and of death, 0.10.

The last row of the figure shows the utility assigned to various outcomes. This obviously is 1 for full recovery and 0 for death. For an intermediary outcome such as recovery with disability, an assessment can be made considering its lifelong implications. In this example, partial recovery is assigned a utility of 0.7. Also, 1 minus the utility can be interpreted as the cost. But a utility of 1.0 indicates that the cost of treatment is not factored in.

Depending on predictivities, the cost involved, the probabilities of various grades of recovery, and the utility assigned to various outcomes, it is possible to work out the expected benefit of different decisions. For this, the process of folding bottom-up is followed. The following is in terms of multiplication of probabilities and utilities, and their addition. Cost is not adequately factored in.

For example, the expected benefit of treatment when a test is positive is as follows.

a. When the disease is indeed present,

$$1.0 \times 0.90 + 0.7 \times 0.08 + 0.0 \times 0.02 = 0.956.$$

b. When the disease is actually not present (test is false positive),

$$1.0 \times 0.99 + 0.7 \times 0.01 + 0.0 \times 0.00 = 0.997.$$

Note how the probabilities and utilities are multiplied and added to compute the expected benefit. Since the just-calculated probabilities are $P(a) = 0.956$ and $P(b) = 0.997$, in this example, the expected benefit of treatment when the test is positive

$$= 0.956 \times 0.85 + 0.997 \times 0.15 = 0.962.$$

Similarly, the expected benefit of no treatment when the test is positive:

$$\begin{aligned} &= (1.0 \times 0.50 + 0.7 \times 0.40 + 0.0 \times 0.10) \times 0.85 \\ &\quad + (1.0 \times 1.0 + 0.7 \times 0.0 + 0.0 \times 0.00) \times 0.15 \\ &= 0.78 \times 0.85 + 1.0 \times 0.15 = 0.813. \end{aligned}$$

Clearly, in this example, when the test is positive, the expected benefit of treatment is much more than of no treatment. This takes care of decision node D1 in Figure D.1.

Now consider the expected benefit in the situation when test is negative.

c. When the disease happens to be present (the test is false negative),

$$1.0 \times 0.90 + 0.7 \times 0.08 + 0.0 \times 0.02 = 0.956.$$

d. When the disease is indeed not present,

$$1.0 \times 0.99 + 0.7 \times 0.01 + 0.0 \times 0.00 = 0.997.$$

The expected benefit of treatment when the test is negative

$$= 0.956 \times 0.10 + 0.997 \times 0.90 = 0.993.$$

Similarly, the expected benefit of no treatment when test is negative

$$\begin{aligned} &= (1.0 \times 0.50 + 0.7 \times 0.40 + 0.0 \times 0.10) \times 0.10 \\ &\quad + (1.0 \times 1.00 + 0.7 \times 0.00 + 0.0 \times 0.00) \times 0.90 \\ &= 0.78 \times 0.10 + 1.0 \times 0.90 = 0.978 \end{aligned}$$

Note that even when the test is negative, expected benefit of treatment is more than of no treatment in this example. This is based on the positive and negative predictivities, as already specified, and utilities and probabilities of various grades of recovery, as in Figure D.1. When these values change, the expected benefit also changes, and the decision to treat or not to treat would also change accordingly.

In case no test is done because of exigencies of situation or otherwise, the expected benefit of treatment

$$\begin{aligned} &= (0.0 \times 0.90 + 0.7 \times 0.08 + 0.0 \times 0.02) \times 0.70 \\ &\quad + (1.0 \times 0.79 + 0.7 \times 0.01 + 0.0 \times 0.01) \times 0.30 \\ &= 0.956 \times 0.70 + 0.997 \times 0.30 = 0.968 \end{aligned}$$

and the expected benefit of no treatment

$$\begin{aligned} &= (1.0 \times 0.50 + 0.7 \times 0.40 + 0.0 \times 0.10) \times 0.70 \\ &\quad + (1.0 \times 1.00 + 0.7 \times 0.0 \times 0.00) \times 0.30 \\ &= 0.78 \times 0.70 + 1.0 \times 0.30 = 0.846 \end{aligned}$$

Thus, when the prevalence of disease among patients with those complaints is 70% and all other values as in this example, the expected benefit from treatment is more than no treatment when the test is not done.

All these results are as you would intuitively expect. If the utility of partial recovery were only 0.2 and not 0.7, or if the prevalence of disease in this group were only 10%, the results would change. You may want to do this as an exercise.

The example illustrates the kinds of complexities involved if somebody really wants to make decisions on the basis of a decision tree such as in Figure D.1. The calculations apparently look complex but can be implemented easily with the help of a computer-based small spreadsheet. By changing values of various utilities and probabilities, the spectrum of expected benefits can be calculated, which can help decide what action to take in the best interest of the patient.

This discussion is focused on one particular application of decision trees, namely, in diagnosis and treatment. However, there are several other applications. Bayati et al. [2] developed a decision algorithm to guide decisions about postdischarge interventions in cases of heart failure and concluded that this could reduce the cost of rehospitalization by 18.2%. Su et al. [3] used this approach for an algorithm to diagnose gastric ulcer using mass spectral data. They did not consider treatment options. Lunt et al. [4] evaluated a decision tree format for classification of rheumatoid arthritis. These two approaches do not consider the utility or the cost, as illustrated in our example, and are similar to the **expert systems** described under that topic.

When resources permit, examine whether a tree diagram can help minimize the role of chance in decisions and in objective assessment of the outcome for various options that can be exercised in patient management.


1. Hunink M, Glasziou P, Siegel J et al. *Medical Decision Making in Health and Medicine: Integrating Evidence and Values*. Cambridge University Press, 2002.
2. Bayati M, Braverman M, Gillam M, Mack KM, Ruiz G, Smith MS, Horvitz E. Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PLoS One* 2014 Oct 8;9(10):e109264. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4190088/
3. Su Y, Shen J, Qian H, Ma H, Ji J, Ma H, Ma L et al. Diagnosis of gastric ulcer using decision tree classification of mass spectral data. *Cancer Sci* 2007;98:37–43. http://onlinelibrary.wiley.com/doi/10.1111/j.1349-7006.2006.00339.x/full
4. Lunt M, Symmons DP, Silman AJ. An evaluation of the decision tree format of the American College of Rheumatology 1987 classification criteria for rheumatoid arthritis: Performance over five years in a primary care-based prospective study. *Arthritis Rheum* 2005; 52:2277–83. http://onlinelibrary.wiley.com/doi/10.1002/art.21203/pdf


## degrees of freedom (df's) (the concept of)

We explain the concept of degrees of freedom (df's) with the help of an example. In Table D.3, four categories of blood group are listed, namely O, A, B, and AB. However, the frequency in only three of them can be freely chosen; the fourth is automatically determined by the total. If the frequencies chosen for O, A, and AB are 70, 20, and 10, respectively, then the frequency in group B has to be 50 because the total is 150. If the frequencies chosen for O, A, and B are 60, 30, and 20, respectively, then the frequency in the group AB has to be 40. Thus, there is freedom to choose only three out of four cells. This is the df. For $K$ cells in a one-way **contingency table**, when the sample values have no restriction other than that the total is fixed, the df = $K - 1$. In a $2 \times 2$ contingency table, when the row and

**TABLE D.3**
**One-Way Contingency Table**

| | Blood Group | | | | |
|---|---|---|---|---|---|
| | **O** | **A** | **B** | **AB** | **Total** |
| Observed frequency | 57 | 36 | 51 | 6 | **150** |

total columns are fixed, only one of the four cells can be chosen at will—all others will be automatically fixed. In general, in an $R \times C$ contingency table ($R$ is the number of rows and $C$ is the number of columns), the df = $(R - 1)(C - 1)$. The method of **chi-square** uses such df's all the time.

The preceding explanation is for frequencies. The concept can be easily extended to values. Suppose you measure the diastolic blood pressure (BP) level of a person three times and you are told that the average is 133 mmHg. If the measurements at two of the three times are 134 and 131, the measurement the third time has to be 134 mmHg for the average to be 133. You have freedom to choose only two of these three numbers; the third is automatically fixed because of the given mean. Thus, when you have $n$ measurements and the mean is fixed, the df's are $(n - 1)$. When additionally, the value of the standard deviation is also fixed, it can be shown that the df's are $(n - 2)$. This kind of df is used by the **Student *t*-test**.

The ***F*-test**, commonly used in analysis of variance, has a pair of df's—one belonging to the numerator and the other to the denominator. The purpose, as you can see, is to find the number of independent values. The distribution of chi-square, $t$, and $F$ depends on this number, the df. This is similar to BP distribution depending on age—different age groups have different BP distributions.

## Delphi method

The Delphi method is a procedure by which experts are brought to a consensus in stages by gradually eliminating the isolated differential opinion. The consensus and the isolated differential opinion are shared among the experts without identifying the names of the experts. They are asked to revise their opinion in view of the consensus, finally reaching a conclusion, which is generally agreeable. The consensus arrived at may or may not be shared by the experts who did not participate in the exercise or whose opinion was eliminated after being found not in line with those of the majority.

For example, the Delphi technique was used by Henson [1] to explore differences in expert opinion and to provide more reliable estimates of the incidence of food-borne *Salmonella* in the United Kingdom. The first step in this Delphi study was to run a workshop in which seven experts on *Salmonella* infection examined the issues to be covered in the Delphi survey. These experts wrote the precise wording to be used in the Delphi study questions and identified 62 experts to be part of the study; 42 of them agreed to participate. Five Delphi rounds were then conducted by means of questionnaires over 7 months, with three of them exploring the experts' judgments. The first question was "What would you estimate to be the total number of persons ill due to infection with nontyphoid *Salmonella* in the United Kingdom from all sources (food and nonfood) over the course of 1 year?" Second, "What proportion of experts thought the infection was food-borne?" Third, "Could the proportion of cases be related to the type of food?" For each question, experts were asked to describe how they produced their estimates, and any difficulties they encountered in doing so. The results of the first and second rounds were fed back to the participants, inviting them to revise their estimates of incidence. Importantly, the process narrowed the range of estimates for the incidence of infection as experts reflected on the median and range of responses to the incidence questions. Henson concluded that this Delphi study "provides a good summary measure of expert opinion in an area which is characterised by great uncertainty."

Zhao et al. [2] used the Delphi method to assess the validity of a medical test on prehospital stroke symptom coping in China, and Juanola Roura et al. [3] derived recommendations for early referral of patients with spondyloarthritis in Spain using the Delphi method.

1. Henson S. Estimating the incidence of food-borne *Salmonella* and the effectiveness of alternative control measures using the Delphi method. *Int J Food Microbiol* 1997;35(3):195–204. http://www.sciencedirect.com/science/article/pii/S0168160596012354
2. Zhao Q, Yang L, Zhang X, Zhu X, Zuo Q, Wu Y, Yang L, Gao W, Li M, Cheng S. Development and validation of the pre-hospital stroke symptoms coping test. *PLoS One* 2014 Oct 17;9(10):e110022. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4201517/
3. Juanola Roura X, Collantes Estévez E, León Vázquez F, Torres Villamor A, García Yébenes MJ, Queiro Silva R, Gratacós Masmitja J et al. Recommendations for the detection, study and referral of inflammatory low-back pain in Primary Care. *Reumatol Clin (Eng)* 2015;11(2):90–8. http://www.sciencedirect.com/science/article/pii/S2173574314001518

## Demographic and Health Surveys

Demography is the study of the human populations by statistical methods. The Demographic and Health Survey (DHS) program provides decision makers and program managers with the information necessary to plan, monitor, and evaluate population, health, and nutrition programs [1]. Under the DHS program, large-scale surveys are undertaken of nationally representative samples of households in several countries around the world, mostly developing countries that lack vital data, on aspects such as fertility, education, nutrition, family relations, behavior, and child mortality. A report on the findings is quickly brought out so that relevance is not lost. This program is supported by the United States Agency for International Development (USAID).

The survey process is guided through procedures and manuals developed within the DHS program. To be sure that data reflect the scenarios that they intend to describe, and that data are comparable across countries, a number of steps are undertaken, such as an almost-uniform questionnaire, common training, and the same definitions. The program also collects geographic information of the surveyed countries. Using geographic information systems, DHS data can be linked with health data, health facility locations, and local infrastructure such as roads, rivers, and environmental conditions. Regarding health issues within a country, analysis of DHS data gives a more in-depth understanding than available elsewhere for developing countries. Synthesis of information across DHS surveys can be essential to informing policy and programs.

Many countries repeat the DHS, typically every 5 years. Some countries have carried out many surveys; among them, Bangladesh did this survey in the years 1993–1994, 1996–1997, 1999–2000, 2004, 2007, 2011, and 2014. This country is practically repeating the DHS every 3 years. Since each survey is based on a nationally representative sample and since the estimates are derived using a proper weighting method for differential representation of population by age and sex, the estimates are fairly comparable across years, and a trend indeed can be obtained. Since the survey methodology is nearly the same in each country, the data are comparable across countries also.

1. USAID. *DHS Program; What We Do.* http://dhsprogram.com/What-We-Do/

## demographic cycle

Also known as *demographic transition*, this cycle depicts how a general population moves from an initial phase of high fertility/high mortality to a low fertility/low mortality phase over a period of, say, 100 years as development takes place. Figure D.2 shows this cycle.

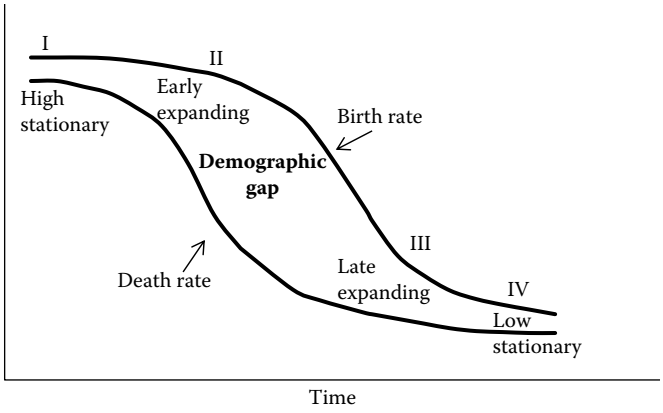



**FIGURE D.2** Demographic cycle.

The initial phase of the cycle is called high stationary as both fertility and mortality are high and the population is stable. As people become aware about health, the death rate quickly starts to come down, but the birth rate remains high. The gap increases, and the population starts expanding—this is called the early expanding phase. This is phase II of the demographic cycle. As the death rate settles down to a low level, the difference between the birth rate and the death rate—called the *demographic gap*—widens before narrowing down in phase III of the cycle. In this phase, the birth rate also relents, and the gap shrinks, although it still remains substantial. In the last phase (phase IV), called the low stationary phase, both the birth and death rates become low, and the population stabilizes.

Note a few more points about the demographic cycle. First, the death rate is easier to control than the birth rate. People easily accept advice on nutrition and treatment of diseases than on fertility control. This is paradoxical in view of the general perception that deaths are not in our hands but births are. The experience is just the reverse—we can easily control deaths but not so easily births. Second, some countries have taken 100 years in reaching from phase I to phase IV in a natural course, but countries like China have provided an unusual example of completing the cycle in just about 30 years. Some countries are still in the midst of this cycle. Third, some countries have shown a trend of reaching negative growth in phase IV, with the death rate exceeding the birth rate. This is expected in countries where fertility is low and a higher number of deaths has become inevitable due to a sizeable old-age population. According to the World Bank estimates for the years 2010–2014, Albania, Latvia, and Japan are in this club.

The demographic cycle depends much on the health of the people. Healthier people die late in life and, generally, reproduce less. Health infrastructure and health awareness play a crucial role in determining the phase of the cycle and its duration.

## demographic indicators, *see also* demographic cycle, population pyramid

Demographic indicators are those that measure the state of the population. Main among them are

- **Fertility indicators** and **mortality indicators**, including **expectation of life** and marital status
- Population growth and levels of migration
- Level of urbanization and population density
- Age-sex distribution, **dependency ratio**, and education levels
- Amenities available and their utilization (housing, water/sanitation, schools, health centers, doctors/nurses, communication, etc.)

In addition, some people include occupation, income, and expenditure pattern also under demographics. Many of these indicators are discussed as separate topics in this book.

Whereas fertility and mortality are directly related to health and medicine, other indicators also have a bearing on health. For example, age-sex distribution determines health needs—a predominantly geriatric population has different needs than a predominantly pediatric population. Urbanization affects not just the availability of amenities but also awareness and keenness to utilize those amenities. Density of population affects housing and pollution. And so on.

To get a sense of what information demographic indicators can provide, note the following from a United Nations (UN) report on ageing. "Globally, the number of older persons (aged 60 years or over) is expected to more than double, from 841 million people in 2013 to more than 2 billion in 2050. Older persons are projected to exceed the number of children for the first time in 2047. Presently, about two thirds of the world's older persons live in developing countries. Because the older population in less developed regions is growing faster than in the more developed regions, the projections show that older persons will be increasingly concentrated in the less developed regions of the world. By 2050, nearly 8 in 10 of the world's older population will live in the less developed regions" [1].

The working-age (15–64 years) population contributes most to development. Over the next few decades, infant mortality rates will be on the decline, and fertility rates will fall. Thus, with children more likely to survive into productive adulthood and fewer children being produced, the share of the working-age population will increase. The dependency ratio will decrease. This will build what is called the *demographic dividend*. When properly harnessed, this can handsomely contribute to the economic growth of nations. Those who are of working age will, on average, be more productive also. These people are more likely to save (while dependents will not), resulting in more productive investment. This has already occurred in many East Asian countries, fondly called "the East Asia miracle." The demographic dividend is also known as the *demographic bonus*.

### *Stable and Stationary Population*

A population becomes stable when its fertility and mortality rates remain unchanged for a substantial period, say, at least 5 years. This results in an unvarying age distribution: the population grows at a constant rate. The *stable population* theory provides a widely useful framework connecting a fixed set of rates to the resulting population dynamics. This allows us to trace causes and consequences of population change, to develop methods for estimating rates, and to make predictions regarding the future population. A *stationary population* is a special case of a stable population. If fertility and mortality rates remain equal over a long period of time, such a population is said to be stationary. The technical measure of a stationary population is that the **net reproduction rate** (NRR) is 1. Crudely, NRR is the number of daughters born to an average woman in her lifetime after discounting for mortality rates of the women at different ages. Under the stationary state, the population replaces itself, and there is no growth. In practice, this will happen after several years of zero population growth.

On the other end of the spectrum are populations with negative growth. In these populations, deaths exceed births, mostly because

of ageing of the population and low fertility. Albania and Latvia are current examples, and Japan has also started to show this trend [2]. Negative population growth can also occur due to large-scale deaths by widespread epidemics such as the scourge of AIDS that afflicted some sub-Saharan African countries early this century. The adult population declined, but total population did not decline in these countries due to high fertility in this region.

1. UN. *World Population Ageing 2013.* http://www.un.org/en/development/desa/population/publications/pdf/ageing/WorldPopulationAgeing2013.pdf
2. World Bank. *Data: Population growth (Annual%).* http://data.worldbank.org/indicator/SP.POP.GROW

## demography

This is the science of population—its structure, pattern, trend, and everything else that concerns the population but not individuals. Thus, counts of people done by census or otherwise, population growth, migration, fertility and mortality, family size, age-sex structure, possibly occupation and education patterns in different segments of population, housing, etc. come under the ambit of demography. We have separately discussed these various components of demography. In particular, see **death rates**, **fertility indicators**, **demographic cycle**, **demographic indicators, mortality rates**, **population pyramid,** and education indicators.

## dendrogram, *see also* cluster analysis

A dendrogram is an agglomerative diagram representing a hierarchy of categories based on degree of similarity or number of shared features among various characteristics. In statistics, this is a graphical representation of hierarchical groups, which are usually generated through a mathematical process such as **cluster analysis**. The purpose of a dendrogram is to display the relationships among distinct units by grouping them sequentially from smaller to bigger clusters.

The method to draw a dendrogram is as follows:

- Write down all of the units and their values that you want to depict in a dendrogram.
- Choose the clustering method. An overview of these methods is under the topic **cluster analysis**, and each of the hierarchical agglomerative methods is discussed in detail under respective topics. This puts like units into small groups, similar small groups into larger groups, etc., in stages until all units are in a single cluster.
- Ask the computer package to provide you the complete dendrogram depicting the agglomerative process. This will look like Figure D.3, either in this form or in transposed form. Also see the topic **cluster analysis** for another dendrogram. Minami et al. [1] analyzed basic amino acids and their derivatives in water using a turn-on fluorescent sensor array. The dendrogram obtained is in Figure D.3. Height in this diagram measures the distance between units or groups of units. As you can see from this dendrogram, lysine and histidinol were closest to each other (look at the height of the bar and joining in the first stage) followed by histidine and lisinopril (slightly bigger height of the bar and joining in the second stage). Thus, this dendrogram represents the hierarchical clustering process of the amino acids. The authors concluded that the sensor array can discriminate amino acids from the corresponding amines that are produced by the action of amino acid decarboxylases.

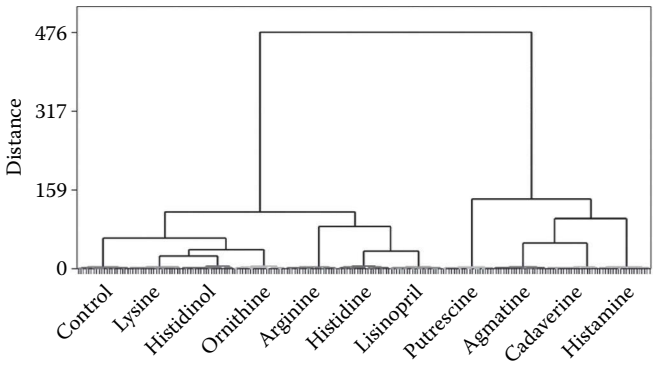



**FIGURE D.3** Dendrogram of the basic amino acids when analyzed by fluorescent sensor array. (From Minami T, Esipenko NA, Zhang B, Isaacs L, Anzenbacher P Jr. *Chem Commun* 2014;50:61–3. http://pubs.rsc.org/en/content/articlelanding/2013/cc/c3cc47416j/unauth#!divAbstract. With permission.)

1. Minami T, Esipenko NA, Zhang B, Isaacs L, Anzenbacher P Jr. "Turn-on" fluorescent sensor array for basic amino acids in water. *Chem Commun* 2014;50:61–3. http://pubs.rsc.org/en/content/articlelanding/2013/cc/c3cc47416j/unauth#!divAbstract

## dependence and independence (statistical), *see also* dependent and independent variables (in regression)

In statistics, events are said to be independent if the occurrence of one event does not affect the probability of occurrence of the other event. On the other hand, if the occurrence of one event does affect the occurrence of another event, the events are said to be dependent.

In the toss of a coin, getting heads in one toss is independent of what you get in the other tosses. If the first toss is tails, this has absolutely no effect on the second toss; the probability of getting heads or tails remains ½ regardless of the outcome of the previous tosses. Similarly, the sex of an unborn child does not depend on the sex of the past or future children. Events that do not affect each other are said to be independent.

With independent events, you can calculate the probability of both events occurring in succession simply by multiplying the probabilities of each event. Returning to the sex of children, the probability of getting a boy (B) followed by a girl (G) is

$$P(\text{BG}) = P(\text{B})*P(\text{G}) = 0.5 \times 0.5 = 0.25,$$

where $P(\text{BG})$ is the joint probability of a boy and a girl. This is the same as that of two boys or of two girls. Statistically, this is written as $P(\text{B}\cap\text{G})$ and read as B intersection G. This intersection in effect means B and G occurring together.

Statistically independent events are sometimes confused with mutually exclusive events. The fact is that if two events are independent, then they cannot be mutually exclusive, and vice versa. That is to say that it is impossible for mutually exclusive events to occur together. A subject in a study cannot be both male and female at the same time, nor can the subject be aged 40 and 50 years at the

same time. These are mutually exclusive events and not independent events. However, one person can be male and the other female. These are independent but not mutually exclusive. On a more practical plane, hypertension and diabetes are neither independent nor mutually exclusive since the presence of one increases the chance of the other and they can occur together at the same time in one person.

The idea of independence can be readily extended to more than two events. For example, the events A, B, and C are independent if and only if $P(\text{ABC}) = P(\text{A})*P(\text{B})*P(\text{C})$. This can also be extended from events to variables. A variable $x$ is considered statistically independent of another variable $y$ when $P(xy) = P(x)*P(y)$ for *all* admissible values of $x$ and $y$. If the age distribution of males in a population is the same as that of females, age and sex would be independent variables in the sense, for example, that $P$(male of age 40–45 years) = $P$(male)*$P$(age 40–45 years). This would not be so if, for example, males are mostly young and females mostly old. Intuitively, two variables are independent of each other if any specific value of one does not alter the chance of any value of the other.

There is an interesting dependence feature of the variables jointly distributed as **multivariate Gaussian**. In this setup, each variable is dependent on the others in a linear fashion. If all their product–moment correlation coefficients are 0, the variables would be independent. However, if two variables each have univariate Gaussian distribution and their correlation is 0, they are not necessarily independent. Their relationship can be nonlinear.

The terms *independent* and *dependent variable* are also used with a different meaning in the context of regression. For this, see the topic **dependent and independent variables (in regression)**.

## dependency ratio

Dependency ratio is generally defined as

$$\text{dependency ratio} = \frac{\text{population of age} < 15 \text{ years} + \text{population of age} \geq 65 \text{ years}}{\text{population of age } 15-64 \text{ years}} * 100.$$

As per this definition, the populations of children (age < 15 years) and the elderly (age ≥ 65 years) are considered to be dependent on the working-age, generally healthy population (age 15–64 years). The dependence is not necessarily only economic but might be social and psychological. The age groups at the two ends can be chosen differently according to the local context.

To be exact, the aforementioned formula gives the *total* dependency ratio. This can be divided into child dependency and old-age dependency ratios. The total dependency ratio in 2013 was 100 in Angola and only 30 in Bahrain [1]. It is generally believed that the higher this ratio, the less is the social health. This may be economically true, but interpretation of a high ratio needs to include the wisdom of the elderly as well as the contribution of children's presence to the well-being of the family.

Use of the dependency ratio is critical in terms of obvious economic consequences. But what is not generally appreciated by the population at large (although the experts realize it) is the demographic bonus (see **demographic indicators**) brought about by the transition as a population moves from the high fertility/high child mortality stage to the low fertility/low child mortality stage. A phase comes in this transition when the percentage of the adult population swells and the dependency ratio substantially declines. When all of the adult population is put to productive work, economic progress can be fast because there are few people to support, reflecting tremendous improvement in overall health. This is what turned around some countries and is now known as the East Asia miracle.

1. The World Bank. *Age Dependency Ratio (% of Working-age Population).* http://data.worldbank.org/indicator/SP.POP.DPND?order=wbapi_data_value_2013+wbapi_data_value+wbapi_data_value-last&sort=asc



## dependent and independent variables (in regression)

The terms *dependent* and *independent variable* have tremendous significance in a regression setup. The general form of regression is given by the following equation:

$$\hat{y} = f(x_1, x_2, \ldots, x_K).$$

This tries to express a variable $y$ in terms of variables $x_1, x_2, \ldots, x_K$. The variable $y$, whose estimated value appears on the left side of this equation, is called the dependent, the **outcome**, the response, or the target variable. The variables $x_1, x_2, \ldots, x_K$, which appear on the right side, are called independent or **explanatory variables** or input variables, sometimes **covariates**. These are also sometimes called determinants or **predictors** of $y$. Depending on the context, these can be **concomitant** or **confounding variables**. In general, $x_1, x_2, \ldots, x_K$ are the **regressors**. These sometimes define the intervention or specify what can be possibly manipulated. Note that there are at least seven different names for the same set of variables.

In a survey, the independent variables are the underlying variables, and the dependent variable is the target variable that is to be ultimately studied. In a clinical setup, the dependent variables may be prognostic or outcome variables. In an experiment or a trial, the independent variables define the initial state, and the dependent variable measures the subsequent or the final state. The *independents are considered fixed and known in a regression setup*, although they may actually not be fixed. Only $y$ is considered stochastic. The regression given by the aforementioned equation is interpreted as the value of $y$ for given values of $x$'s.

Independent variables are the antecedents in terms of time, and the dependent is the outcome. There may be situations where it is difficult to distinguish between independent and dependent variables. For example, between gender and blood group, which is the dependent variable? In general, *independents are those that come first in time* and possibly can be manipulated to alter the outcome. If the postulated dependence does not appeal to your conscience, try reversing the order and see if it makes more sense.

A variable can be either qualitative or quantitative. There might be a qualitative variable that occurs in one of two possible states, often coded in statistical software as 0 or 1. This is a **binary variable**, also known as dichotomous or yes/no data. Frequently found binary variables in medical investigations are of the type dead/alive, depressed/not depressed, treated/not treated, etc. Such binary dependents often require specialized techniques such as **logistic regression** for their analysis. When dealing with two categories, definitions should clearly split the subjects into two categories. For example, there might be a third sex to male/female dichotomy, and you should be able to decide the category of such subjects if dichotomy is to be retained, or whether they are to be excluded. Sometimes, the distinction is not so clear: for example, married/single. Where to put divorced people, unless the categories are currently married or currently single?

In case the dependent variable is quantitative, ordinary quantitative **regression** is used to find the relationship between the dependent variable and the independent set of variables. If the dependent is qualitative but polytomous (many categories), there are two possibilities. One is that this is **ordinal** and the other is **nominal**. In both situations, logistic regression can be extended to cover these possibilities when they are dependent. Note that the nature of the independent variables—qualitative or quantitative, binary or polytomous, or mixed—does not affect the broad method since the independents are considered fixed in a regression setup whether it is logistic or ordinary quantitative regression or any other type, but the method is different for a qualitative dependent than for quantitative dependent. A special case of regression is **analysis of variance (ANOVA)**, where the dependent is quantitative and all the independents are qualitative. If the independents are a mixture of qualitative and quantitative variables and the dependent is quantitative, **analysis of covariance (ANCOVA)** is the first method that comes to mind. Both of these are extended to multivariate quantitative dependent by **multivariate analysis of variance (MANOVA)** and multivariate analysis of covariance (MANCOVA).

In some situations, the logarithm of expected frequencies in a **contingency table** can be expressed in terms of additive factors relating to the variables under study. The name **log-linear** is used for these models. These are useful only when no variable is considered dependent on the other or others and can be used only for contingency tables. No distinction is made between outcome and antecedent in this setup. Thus, this model is best suited for data obtained from cross-sectional studies. The dependent variable in these models is the number of subjects in a cell of the contingency table. The objective is to find whether the categories of different variables, individually or jointly, are especially contributing to determining the cell frequency.

## descriptive analysis/statistics, *see also* exploratory data analysis

This term is used for the summary of all the background or current information of the subjects of a study. These would be mostly in terms of counts and percentages; means, medians, standard deviations (SDs), and ranges; and graphs and diagrams. Even if you have a small-scale study with just 60 cases, it is not helpful to the reader to see the entire database—instead, a summary in terms of descriptive statistics is helpful.

Consider a study on 60 patients with liver cirrhosis where their age and sex, nutritional intake, enzyme levels, family history, etc. are noted. All these can be summarized in the form of tables and graphs that may contain information on age distribution (number of subjects in different age groups), nutritional intake into suitable categories, the mean and SD of enzyme levels for various age-sex groups, histograms and bar diagrams, etc. Descriptive statistics provide an overview of the kind of subjects you have studied. These statistics do not include any analysis such as correlations, regressions, confidence intervals, or tests of hypothesis. This analysis is subsequently done in light of the findings revealed by descriptive statistics.

Besides communicating the basics of the subjects, descriptive statistics are also used to explore the data. For example, descriptive statistics will tell you whether you can use the usual parametric statistical procedures, such as the Student *t*-test and ANOVA, or you need to use nonparametric methods. This would depend on the nature of the statistical **distribution** of the values you have measured. The choice of descriptive statistics also depends on the nature of the distribution. If the distribution of, say, enzyme levels is highly skewed in your subjects, median and interquartile range may be more appropriate descriptors than mean and SD.

## descriptive studies

Studies that seek to assess the current status of a condition in a population of people are called descriptive studies. These are also named **prevalence studies**. These studies are mostly cross-sectional as they do not envisage any follow-up nor go into the past. The focus is only what is currently present at the time of contact. Consider the power of the following results from a 2010 National Health Interview Survey in the United States, which was a descriptive study:

- Most (82%) US children aged 17 years and under had excellent to good health.
- Fourteen percent of children had ever been diagnosed with asthma.
- Eight percent of children aged 3–17 years had a learning disability, and 8% had attention deficit hyperactivity disorder.

The importance of descriptive studies is not fully realized, but it is through such studies that the magnitude of problems is assessed. They also help to assess what is normally seen in a population. Unfortunately, even body temperature among healthy subjects is not known with precision for many populations. Thus, there is a considerable scope for carrying out descriptive studies. Such studies can provide baseline data to launch programs, such as breast cancer control or rehabilitation of elderly people, and can measure progress made.

A descriptive study can generate hypotheses regarding the etiology of a disease when the disease is found to be more common in one group than another. In some situations, a descriptive study can be designed to test a hypothesis regarding the status of a parameter, such as whether at least 30% of patients with abdominal tuberculosis come with the complaints of abdominal pain, vomiting, and constipation of long duration, or whether the prevalence of non-insulin-dependent diabetes is 10% among married females of age 50 years or more whose husbands are diabetic.

A **case study**, which generally describes features of a new disease entity, is also descriptive. It is anecdotal in nature. A story can have a definite impact, and it is easily understood. But be wary of *anecdotal evidence*. It can be interesting and can lead to a plausible hypothesis, but anecdotes are sometimes built around an exciting event and used by the media to sensationalize reporting. You must verify anecdotes even for forwarding a hypothesis. A series of such cases form **case series**. They summarize common features of the cases or may highlight the variation. This, too, can lead to a hypothesis. Initial case series of HIV positives in the United States, almost exclusively among homosexual men, led to the suspicion that sexual behavior could be a cause of HIV.

**Surveys**, too, are descriptive studies although this term is generally used for community-based investigations. When repeatedly undertaken, they can reveal time trends. Complete enumeration, such as a population census, is also descriptive. A descriptive study generally has only one group since no comparison group is needed for this kind of study. Its design is mainly in terms of a sampling plan so that a representative sample of subjects is available. Those who are aware of statistical errors realize that no Type I or Type II error arises in descriptive studies unless we test a hypothesis, such as the presence of correlation between two characteristics.

## design effect and the rate of homogeneity

Design effect is the factor by which the **variance** (Var) of the estimate is affected due to sampling other than the **simple random sampling** (SRS) of the same size. In other words,

$$\text{design effect: deff} = \frac{\text{Var(Other sampling)}}{\text{Var(SRS)}}.$$

For clarity, note that the variance of the estimate is the square of its **standard error (SE)**. When sampling other than SRS (such as stratified, cluster, or multistage) is used, the variance of the estimate of, say, the mean or proportion generally is more than what you get from SRS. Design effect is the ratio by which this is inflated. The design effect would be mostly more than 1, although sampling designs other than SRS can be constructed to give lesser variance.

Design effect is mostly talked about in the context of **cluster random sampling** (CRS). In this method of sampling, groups of units are selected in place of individual units. This can reduce the cost of travel since in this method, you will have a cluster of units in one place. These groups almost invariably will have some trait in common. They may reside close to one another, may belong to one particular class of students, may be those patients who come to consult a particular physician in a clinic, may be admitted in specific wards in a hospital, etc. Because of this affinity, the units within a cluster are likely to be more similar than the units belonging to different clusters. It is this feature that makes CRS less efficient than SRS. If you take SRS of size 100 units and CRS also of size 100—in this case, say belonging to five clusters—the CRS would not represent the same cross-section of units that an SRS would. It may sound ironical, but the CRS sample will have a larger variance.

If the data obtained by CRS are analyzed by using the regular methods that do not recognize design effect, due to clustering of units, and treat positively correlated units as independent, you may reach wrong conclusions. Bland [1] has found this quite common in analyzing clinical trials where cluster randomization was used for selection or allocation of subjects but was ignored at the time of analysis.

Many consider a design effect of 2.0 as standard for CRS, but it actually depends on the rate of homogeneity in units within a cluster and the cluster size. The rate of homogeneity measures the extent of **affinity** among the units belonging to the same cluster. The statistical measure for this for quantitative data is the **intraclass correlation coefficient** (ICC), denoted by $\rho_I$. We are avoiding mathematics in this book, but it can be shown for CRS that

$$\text{design effect: deff} = 1 + (m - 1)\rho_I,$$

where $m$ is the cluster size. This reduces to 1 when $m = 1$ because then CRS is the same as the SRS. As the cluster size increases, the deff also increases. Deff varies from variable to variable in the same study depending upon whether that the variable has more affinity in within-cluster units or less affinity.

From this formula, you can see that clustering may have a large design effect if the ICC is large or if the cluster size is large. Any one of these conditions needs to be met for large deff. For example, if the ICC is 0.001 (a very small correlation) and the cluster size is 500, deff = $1 + (500 - 1) \times 0.001 = 1.5$. The implication is that we would need to increase the total sample size by 50% to achieve the same power as an unclustered (SRS-based) design. Generally, the cluster size chosen is 30–70 and is determined by how much a particular team can cover in a day or in any specified time of stay in that cluster. The number of clusters is determined by first deciding on the total sample size based on the design effect, and then the number of clusters is (total sample size)/(cluster size).

Where do we use deff? We can use this in at least four different ways. (i) Wherever feasible, form clusters in such a way that the ICC is small. This means that the units of the clusters should not be alike but instead should be different as much as possible. This will not be feasible in many practical situations. (ii) If we calculate the required sample size ignoring clustering, we must multiply it by the design effect to get the sample size required for the clustered sample. (iii) Deff can help us to balance the cluster size and the number of clusters. These two together determine the total sample size. If the total sample size is 200, you can have 5 clusters of 40 each, 10 clusters of 20 each, etc. In general, try to select as many clusters as possible of small size to get a good representation of the cross-section of the units. (iv) If we analyze the clustered data as if there were no clusters, the variance of the estimate must be multiplied by deff; hence, the SE must be multiplied by the square root of deff.

1. Bland JM. Cluster randomised trials in the medical literature: Two bibliometric surveys. *BMC Med Research Methodol* 2004;4:21. http://www.biomedcentral.com/content/pdf/1471-2288-4-21.pdf

## design of experiments, *see* experimental designs

## designs of medical studies (overview), *see also* experimental designs

The design of a study is the preplanned scheme under which the subjects are selected and the data are collected. Depending on the scheme, several names are given to the designs. An overview of various designs followed in medical studies is in Figure D.4. In place of designs, these can also be called types of medical studies.

Medical studies primarily have two designs: **descriptive** and **analytical**. Descriptive ones could be **sample surveys**, **case series**, or **census**. A **case study** is a particular form of case series. Surveys could be based on random sampling or nonrandom sampling. Each of these could be one of the several types as explained under the topic **sampling techniques**. Analytical studies could be observational or experimental. **Observational studies** are also sometimes called epidemiological as there is no human intervention. These could be **prospective** (also called follow up), **cross-sectional**, or **retrospective**. Prospective studies could be **longitudinal**, **cohort**, or other, whereas retrospective ones could be **case–control**, **nested**, or no-control studies. Thus, *retrospective* and *case–control* are not synonymous terms. **Experimental studies** in medicine are mostly carried out in the laboratory on animals or biological specimens. **Clinical trials** are experiments on human beings: the regimen under trial could be a therapeutic agent, diagnostic modality or tool, prophylactic regimen, or screening tool. **Field trials** done in communities can be for a prophylactic regimen or screening tool. The design of all these different types of "experiments" should specify whether there is a control group, whether the subjects would be **randomized,** and whether any **blinding** would be done. Also, note the possible layouts toward the bottom of Figure D.4, which could be **crossover**, **repeated measures**, **one way**, **two way**, **factorial**, etc. All these designs are explained in detail under the respective topic.

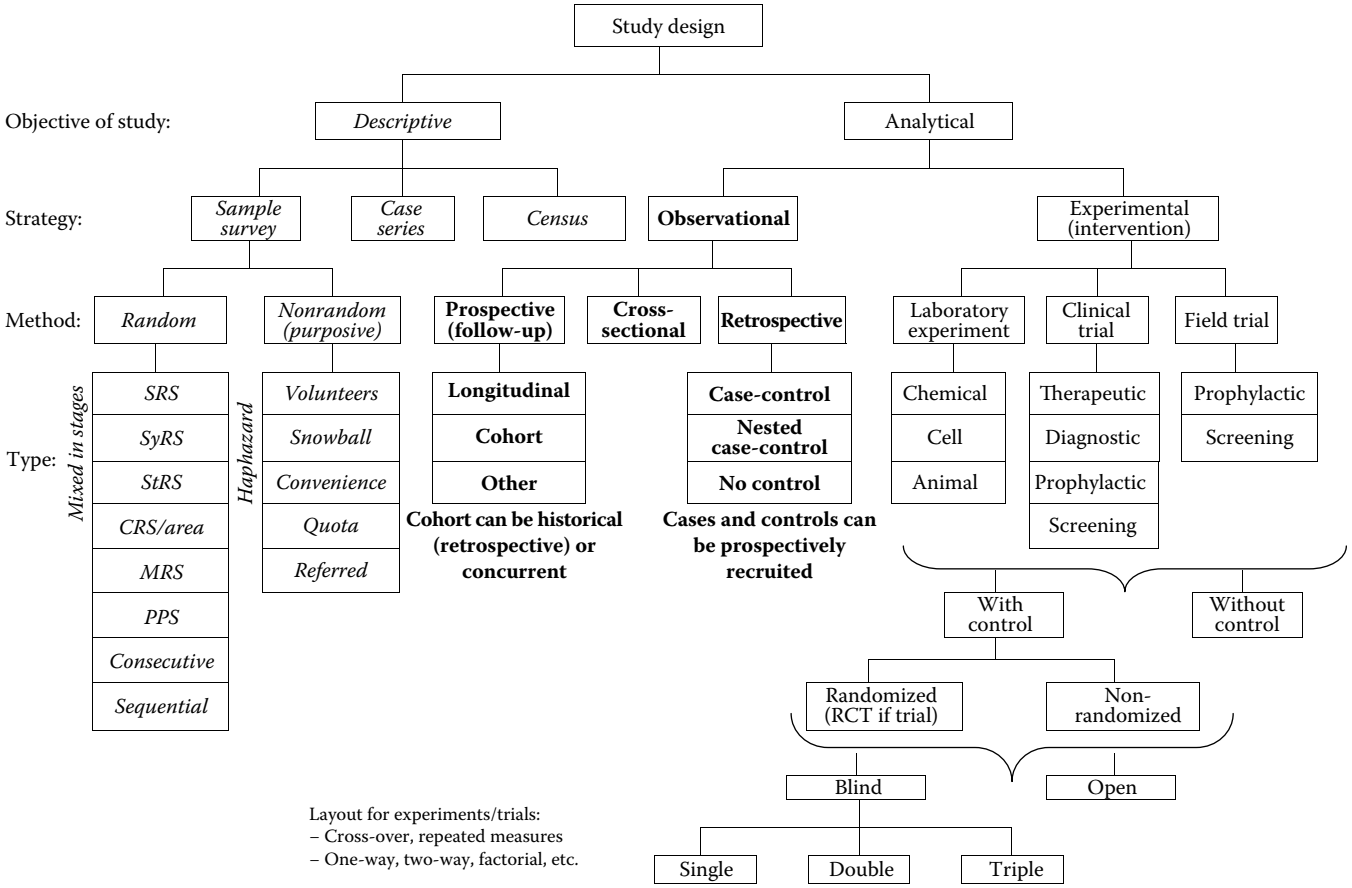



**FIGURE D.4** An overview of designs for medical studies.

## deterministic variables, *see* variables

## deviance

Deviance in statistics is the difference a variable or a set of variables makes to the log-likelihood of the sample for **qualitative** data. The **likelihood**, $L$, under a specified model is the probability of obtaining the values observed in the sample when the model is correct. A probability is necessarily a small number, i.e., less than 1, and your high school math tells you that the logarithm of a number less than 1 is negative. It is helpful to use $-2\ln L$ instead of $L$ because it is positive and the distribution of $-2\ln L$ for large $n$ has been found to follow **chi-square** under the null hypothesis $H_0$ in fairly general conditions. This can be obtained for any set of variables as in the case of **logistic regression** and again after adding or deleting one or more variables. The difference in the two values of $-2\ln L$, say between $-2\ln L_0$ and $-2\ln L_1$, is the deviance for the added variables. Since it is in logarithmic terms, the difference actually is the $-2\ln(L_0/L_1)$, where now the term in parentheses is the **likelihood ratio**. Thus, deviance is –2 times the log-likelihood ratio. This measures the contribution of the added variables to the fitness of the model—and thus can be used to assess the adequacy of the model.

In the case of logistic regression, for example, the $H_0$ is generally is that there is no relationship between the dependent and the regressor variables. With $K$ regressors, if this $H_0$ is true, all the regression coefficients $b_1, b_2, \ldots, b_K$ should be close to 0 and $\hat{\lambda} = b_0'$. Compare this with the fitted model $\hat{\lambda} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_Kx_K$. Denote the likelihood for the reduced model $\hat{\lambda} = b_0'$ by $L_0$ and for the aforementioned fitted model by $L_1$. The value of $-2\ln L_0$ for the reduced model would invariably be more than $-2\ln L_1$ for the fitted model. The difference between these two, which we are calling deviance, also follows a chi-square distribution with $K$ degrees of freedom (df's). This is also called the model chi-square. Most standard statistical software packages give the value of $-2\ln L$ for the model under consideration. For example, if $-2\ln L_0 = 83.18$ for the reduced model and $-2\ln L_1 = 61.01$ for the fitted model with $K = 3$ predictors, then model chi-square = 83.18 − 61.01 = 22.17. At $K = 3$ df's, this is highly significant ($P < 0.001$). So $x_1$, $x_2$, and $x_3$ (together) in this example are useful in understanding the model.

The following comments explain some of the implications:

- If the model is a perfect fit, then the likelihood is 1 and $-2\ln L = 0$. The higher this value, the less adequate the model. Note again that being the probability, $L < 1$, and thus, $\ln L$ is always negative, and $-2\ln L$ is always positive.
- One measure of the adequacy of a full logistic model is the extent of decrease in the value of $-2\ln L$ *relative* to the value for the reduced model. This can be calculated as follows:

$$\text{contribution of the model: } C = \frac{(-2\ln L_0) - (-2\ln L_1)}{-2\ln L_0},$$

where $L_0$ corresponds to the reduced model and $L_1$ to the fitted model as before.

- The role of $-2\ln L$ for qualitative data is similar to that of $R^2$ (square of the **multiple correlation coefficient**) in the case of a quantitative dependent, but it has a negative meaning. In the case of $R^2$, a larger value is better, but in the case of $-2\ln L$, a smaller value is better. Also, $-2\ln L$ does not fall between 0 and 1 as $R^2$ does.
- All models can be improved by adding more regressors. This will invariably decrease the value of $-2\ln L$. But this deviance may or may not be statistically significant. To test this, the difference in deviance is again referred to chi-square to obtain a *P*-value. A nonsignificant decrease indicates that adding those regressors is not helpful. Similarly, one or more regressors can be dropped in search of a more parsimonious model. A nonsignificant change in deviance justifies dropping a regressor from the model because the model still works almost just as well without those regressors.
- Also, as always in a regression setup, an increase in the number of parameters decreases $-21\ln L$ and improves fit in absolute terms, but sometimes, the df's increase relatively faster, and statistical significance declines.

Although several criteria such as generalized $R^2$, likelihood ratio, and the Wald statistic are available to check the statistical adequacy of a logistic model, the log-likelihood and the deviance seem to have better appeal and wider applicability. For other criteria, see Hosmer et al. [1].

1. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*, Third Edition. Wiley–Interscience, 2013.

## deviations, *see* variation (measures of)

## df, *see* degrees of freedom (df's) (the concept of)

## diagnosis errors

There are two types of diagnosis errors—first, misdiagnosis, which occurs when a person does not have a particular disease but is diagnosed to have one, and second, missed diagnosis, when a person has a disease but it is missed.

Clinicians these days examine a great deal of evidence before reaching a diagnosis. This evidence is in terms of clinical features, laboratory investigations, images, records, etc. Despite this, errors are not uncommon. In place of a differential diagnosis, let us restrict our discussion to the presence or absence of a specific disease. If the disease is actually not present but wrongly diagnosed as present, this is called misdiagnosis. Misdiagnosis can mean a great deal of inconvenience, cost, and side effects to a person who is actually unaffected and an unnecessary load on the health care system. Misdiagnosis is obviously a more serious error than missed diagnoses, which occur when the disease is present but is missed. Both are errors in most situations, but a missed diagnosis is likely to be detected in a subsequent encounter because the patient is likely to come back with complaints.

Tools such as laboratory tests are an integral part of modern medicine. Fine-needle aspiration cytology (FNAC) is done to detect breast cancer, Pap smear for cervical cancer, Western blot for human immunodeficiency virus (HIV) infection, and chest x-ray for tuberculosis. Often, the sign–symptom syndrome functions as a *test* to form the basis for establishing a diagnosis. Generalized maculopapular rashes and fever with cough are considered indicative of measles, and prolonged acute chest pain indicates myocardial infarction (MI). However, such tools used for evaluation and management of health and disease are seldom perfect. These examples might have convinced you that all tests are flawed to a degree. These produce correct results in many cases but fail, fully or partially, to perform well in some cases. Healthy individuals are occasionally classified wrongly as ill (giving rise to false alarm). Some individuals who are really ill may not be detected, increasing complacency. All such errors need to be controlled because there is a cost involved—cost of unnecessary treatment, cost of side effects, inconvenience, progression of disease, even death of some patients.

The ability of a tool or of a procedure to perform its assigned function correctly is called its **validity**. A valid diagnostic test would correctly detect the presence as well as the absence of the disease. Some tests are more valid than others, although they may be more expensive. Rectal sonography is considered more efficacious than prostate-specific antigen (PSA) values for detecting prostate cancer. Scintigraphy gives better results than an electrocardiogram (ECG) for MI. However, **gold standards** that give perfect results all the time are practically nonexistent, and most of them are expensive in terms of time and effort. No medical test is valid in an absolute sense. Errors in classification such as misdiagnosis and missed diagnosis occur no matter what test is used.

The true diagnosis is evaluated on the basis of more refined methods—the current gold standard—that may be far more difficult to adopt. Often, the real diagnosis emerges after the passage of time, for instance, on response to therapy or upon autopsy. Sometimes, a surrogate is used as a gold standard, such as histological evidence for cancer. If the gold itself is a bit shoddy, a good **sensitivity and specificity** may give a false sense of security, and diagnostic errors can go unnoticed.

## diagnostic tests, *see* sensitivity and specificity, predictivities (of medical tests), gain from a medical test

## diagnostic trials, *see* clinical trials (overview)

## diagrams, *see* graphs and diagrams

## dichotomous categories, *see* categories of data values

## dietary indices

These indices are used to assess nutrition intake. Most dietary studies have interest in food intake also beside nutrients. Generally, three methods are used for this purpose at an individual level. First is multiple 24-hour recall, second is 1-week diet record, and third is the food-frequency questionnaire (FFQ). All three methods elicit dietary intake and then convert that to nutrition intake, such as calories, protein, and carbohydrate, on the basis of the available standards for each intake. A large number of other questionnaires, scoring systems, and recording tools are available for this purpose. If the target is a specific group, such as adolescents, you may be able to find a method specifically tailored for this segment of the population. Patients of different diseases will require a different format altogether.

D

As the name implies, the 24-hour recall method is asking people what they consumed in the past 24 hours. This can be done for a dispersed or contiguous 2 or 3 days to get a cross-section and then averaged per day. In the second method, families are asked keep a record of consumption for one full week. These are apportioned per person considering an adult male (moderate worker) as the reference. For example, a child of age 5 years may be 0.6 unit—meaning thereby that the food needs of a child of age 5 years is 60% of that of an adult male (see **consumption units**). The third method, FFQ, typically asks about diet over the past month, generally in terms of groceries purchased and consumed. When properly adjusted, this method may be epidemiologically more relevant as it provides a long-term perspective of the effect that a diet is most likely to have on health parameters. But this may not be able to provide accurate information on occasional intakes, such as in parties and restaurants. The other two methods include this consumption.

Some biomarkers, such as urinary and plasma measurements, are more sensitive to recent intake, and their correlation with diet may be exaggerated when assessed with a 1-week record. If a 1-week record is used, the biomarkers should be assessed remotely in time. Natural week-to-week variation in 1-week dietary record lowers the validity of this method. This is true for 24-hour recall also.

Even if done several times, 24-hour recall cannot provide information on dietary intake in the long term. But 24-hour recall generally provides more accurate information as the recall lapse is minimal. This method can be focused on individuals rather than the family.

You can see that one method is advantageous in one setting and another in different setting. Some associations may be better detected by one method than the other. Choose the method that meets your objective most appropriately.

In an attempt to analyze associations of dietary indices with biomarkers of dietary exposure and cardiovascular status, Truthmann et al. [1] investigated if the dietary factors in adolescence predict cardiovascular risk marker values in adulthood in Germany. The authors concluded: "Overall, the indices, even the simpler ones, seem to have a similar general capability in predicting biomarkers of dietary exposure. To predict risk of cardiovascular disease, the dietary indices may have to be more specific."

1. Truthmann J, Richter A, Thiele S, Drescher L, Roosen J, Mensink GB. Associations of dietary indices with biomarkers of dietary exposure and cardiovascular status among adolescents in Germany. *Nutrition Metabol* 2012;9:92. http://www.nutritionandmetabolism.com/content/9/1/92

## difference-in-differences approach

This is a commonsense approach for the analysis of data arising from before–after quantitative measurements in two groups. When the same subjects are measured before the intervention and after the intervention, the quantity of interest is the change in the outcome from baseline. Values obtained before the intervention are the baseline values. This change can be measured either in absolute values or in terms of percentage over the baseline values. When this is done for two groups, such as in the control and the test group, the interest many times is to find that change in one group is different from the one seen in the other group. That is, the objective is to compare the difference (from *before* to *after* values) in group 1 with the differences observed in group 2. If the difference between these differences is nearly 0 on average, or for most subjects, the conclusion would be that the change in one group is not significantly different from the change in the other group. This naturally leads to a difference-in-differences approach.

Suppose a test group is measured before and after the treatment under test as well as a control group before and after a placebo. If the corresponding population means are $\mu_{1T}$ (before treatment), $\mu_{2T}$ (after treatment), $\mu_{1C}$ (before placebo in the control), and $\mu_{2C}$ (after placebo in the control), then the actual treatment effect is

$$\text{mean difference in differences: } (\mu_{2T} - \mu_{1T}) - (\mu_{2C} - \mu_{1C}),$$

assuming that *after* values are higher and the interest is in absolute change. If this difference in differences is found to be significant, the estimate of the treatment effect is obtained by substituting the corresponding sample means.

As for several other statistical methods, the difference-in-differences method assumes that the two groups under comparison are equivalent either by randomization or by matching. If that is not so, it would be difficult to say that the difference is the true effect.

In the usual setup, the statistical significance of this difference in differences can be checked by the **Student *t*-test** for two groups. This would utilize only the change and not the actual values. This test has the same requirement of independence of groups, Gaussian distribution, sample size, etc., as stated for a two-sample *t*-test. Otherwise, a nonparametric test can be used.

There is another useful method based on **regression**. This method uses **indicator variables** that are sometimes useful in including one or more qualitative variables in the regression. The regression approach is as follows. Consider the regression equation

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_1{}^*x_2,$$

where $x_1 = 0$ for the control group, $x_1 = 1$ for the treatment group, $x_2 = 0$ for *before* values, and $x_2 = 1$ for *after* values. These are the indicator variables. Plug in these values and obtain the following.

*Before* values: treatment group $y = b_0 + b_1$; control group $y = b_0$; difference = $b_1$
*After* values: treatment group $y = b_0 + b_1 + b_2 + b_3$; control group $y = b_0 + b_2$; difference = $b_1 + b_3$
Difference in differences = $(b_1 + b_3) - b_1 = b_3$

Thus, $b_3$ is an estimate of the actual treatment effect compared with the controls. This is the regression coefficient of the product $x_1{}^*x_2$, which is now called an **interaction**. If the interaction is significant, the treatment effect obtained by the difference-in-differences approach is significant. This significance can be tested as for any **regression coefficient**.

A word of caution is in order while working with differences. Since both *before* and *after* values are subject to sampling fluctuations and inadvertent measurement errors, the differences have higher variance. Thus, the sample size required for statistical inferences from differences is higher than the sample size required for inferences from the directly measured values.

## digit preference

It is well known that almost all of us have a special love for digits 0 and 5. This is often subconscious but quite common. Measurements are more frequently recorded ending with these digits. A person aged 69 or 71 is very likely to report one's age as 70 years. In the context of age, this is called *age heaping*. In other setups, if digit 0 is preferred, 78, 79, 80, 81, and 82 can be recorded as 80. For example, this can happen in self-reporting of weight in kilograms.

This can also happen with readings on a visual analog scale or mercury sphygmomanometer. It is quite common in recording of birth weight.

The net effect of the digit preference is that the values ending with 0 or 5 have higher frequency (spikes) than what are actually present. This can cause bias in the results. The counts at the preferred digits are composed of the actual values at these values plus the misclassified cases from the nearby values due to the preference pattern. According to Nagi et al. [1], besides 0 and 5 as the most preferred digit for reporting of age, the multiples of 2, that is, 2, 4, 6, and 8, are preferred over odd digits. The digits 1 and 9 are the least preferred. This may be due to their heaping with the nearest 0. This is depicted in Figure D.5 for reported age at death.

Another manifestation of digit preference is in forming intervals for quantitative data. Blood glucose level categories would be 70–79, 80–89, 90–99, etc., and not 64–71, 72–79, etc. Intervals 105–114, 115–124, 125–134, etc. for blood pressure (BP), for example, or 108–112, 113–117, 118–122, etc. are better to ameliorate the effect of digit preference, but the conventional intervals ending with 0 or 5 are almost invariably used. Such intervals tend to further aggravate the effect of digit preference since such categories put a value of 119 in one interval and a very close value of 120 in another interval. Actual values equal to 119 but preferred to be reported as 120 unnecessarily go in the next interval.

You can see how digit preference can distort the statistical distribution of values. If not taken into account, the results can be biased. For example, if your definition of hypertension is BP <140/90 mmHg, people with systolic BP of 138 or 139 mmHg and diastolic BP of 88 or 89 mmHg would unnecessarily go into the hypertension group in the case of digit preference. Broad et al. [2] have discussed this phenomenon of misclassification.

Nagi et al. [1] have also investigated the social and economic correlates of age heaping, and Jena et al. [3] have done this for self-reported smoking of cigarettes. In general, digit preference happens either because of approximation done by the subjects themselves at the time of inquiry, such as stating smoking 20 cigarettes a day instead of a more exact 18, or because of the observer's bias, such as in recording a systolic 130 mmHg instead of the exact 132 mmHg.

As an approximation, one can think of redistributing the excess values of the preferred digit to the neighboring values as the observed pattern of the distribution. For example, in Figure D.5, a trend can still be seen, and the frequencies on the vertical axis can be adjusted to follow the general pattern. For a more adequate solution, see Carlo et al. [4].


**FIGURE D.5** Digit preference in reporting age at death.

1. Nagi MH, Stockwell EG, Snavley LM. Digit preference and avoidance in the age statistics of some recent African censuses: Some patterns and correlates. *Int Stat Rev* 1973;41(20:165–74. http://www.jstor.org/discover/10.2307/1402833?uid=3738256&uid=2&uid=4&sid=21104660739041
2. Broad J, Wells S, Marshall R, Jackson R. Zero end-digit preference in recorded blood pressure and its impact on classification of patients for pharmacologic management in primary care—PREDICT-CVD-6. *Br J Gen Pract* 2007 Nov;57(544):897–903. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2169314/
3. Jena PK, Kishore J, Jahnavi G. Correlates of digit bias in self-reporting of cigarette per day (CPD) frequency: Results from Global Adult Tobacco Survey (GATS), India and its implications. *Asian Pac J Cancer Prev* 2013;14(6):3865–9. http://www.apocpcontrol.org/page/apjcp_issues_view.php?sid=Entrez:PubMed&id=pmid:23886198&key=2013.14.6.3865
4. Camarda CG, Paul HC. Eilers PHC, Gampe J. Modelling general patterns of digit preference. *Statistical Modelling* 2008;8(4):385–401. http://www.demogr.mpg.de/publications%5Cfiles%5C2945_1235560583_1_Camarda-Feb09-Sage.pdf

## direct standardization, *see* standardized death rates

## disability-adjusted life years (DALYs)

The difference between ideal health and achieved health is called the *health gap*. It defines the need and delineates the shortfalls. Among the many indices that measure the health gap, the one that looks important for a society is the **burden of disease** measured by disability-adjusted life years (DALYs) lost. Broadly speaking, this is computed as the sum of the *years of life lost* (YLL) due to premature mortality and the *equivalent* YLLs due to various disabilities in life. The disability can be due to disease, injury, or any other condition. The details are as follows.

Ideal health is that everybody lives for 100 years (even more!) without falling sick for a single day. Practically, this does not look feasible by any stretch of the imagination. Perhaps the next best hope is that the **expectation of life** becomes as much as the highest actually seen in a population, and this is attained without being sick. This is the basic premise on which the concept of DALYs lost is based. Note that life expectancy is the population average, and individual lives vary.

Japan has the highest life expectancy in the world. Life tables for males and females prepared by the World Health Organization (WHO) corresponded fairly well with the Japanese experience. Any death contributes to life years lost according to the remaining life expected as per their life table. Thus, a death even at the age of 90 years means 4 years lost because this is the expectancy at age 90 years in the Japanese model. Deaths of people at different ages add up to the YLLs, with a modification as shortly described.

The second component of DALYs comes from the disability arising from illnesses and impairments. Disease severity is assigned a **disability weight**, which in turn is converted to equivalent years in full health lost, by using a concept such as person trade-off. For details, see WHO's *National Burden of Disease Manual* [1]. This method of calculation has now been revised, but the basic concept remains the same. Essentially, this means that 1 year of paraplegia counts several times more than 1 year of suffering from hypertension. Death is given a weight of 1.0, and a state of complete restriction of movements (bedridden) but no other restrictions (as in the

case of fracture) can be given a weight of 0.6. Perfect health gets disability weight = 0. Lifetime duration of diseases together in terms of equivalent years forms what is called *years lost due to disabilities* (YLDs). That is, for a particular disease A in the year X,

YLD for disease A in the year X = (prevalence of the disease A in the year X) × (disability weight) × (average duration of disease A in the year X until remission or death).

This can be restricted to a particular age and sex, and adjusted for comorbidities.

The sum of YLLs and YLDs is called DALYs lost. This is considered a comprehensive measure of the health gap or burden of disease, and is generally calculated per 1000 population for any particular year. A big advantage of this measure is that it can be calculated separately for each disease or adverse health condition. The sum total of DALYs lost by all diseases must be the same as the total DALYs obtained from the age-specific death rates. This is an important self-check mechanism because otherwise, when disease-wise incidence, prevalence, and mortality are calculated, each disease overestimates itself probably to show that it is a more important health condition, and the sum total can far exceed the total burden of disease. Note that the total mortality is easy to work out on the basis of the age-sex-specific death rates and would be fairly valid as this information is largely available for each country. Disease-wise calculation is not so valid yet because of a lack of disease-wise information for different ages and sexes in most parts of the world. The **Global Burden of Disease (GBD)** study done by the Institute of Health Metrics and Evaluation (IHME) has computed DALYs lost for more than 200 health conditions for each of the 187 countries around the world. They have used statistical models to estimate the missing information. Some of the results of this study are available in their publications [2,3] and their website [4].

The earlier calculation of DALYs as done by WHO from 1990 to 2008 valued life around age 25 years much more than at childhood or at old age, i.e., nearly 1.4 compared with 1.0 at the age of 10 years and 55 years, and 0.4 at the age of 90 years. Future years lost were discounted for equivalence with the current year. This kind of adjustment has been dispensed with in GBD 2010 because of severe criticism of age weighting and discounting. Also, these calculations used incidence instead of prevalence in calculation of YLDs. That was termed *incidence perspective*, while the present is based on *prevalence perspective*. The latter more correctly reflects the current situation.

According to 2010 estimates, an average of 360 DALYs were lost in the world per 1000 population. This implies that more than one-third of life in full health is lost due to early mortality and various diseases during the lifetime. YLLs contributed nearly two-thirds of this loss, and YLDs, nearly one-third.

If the world average is considered, nearly 54% of DALYs lost in 2010 were because of noncommunicable diseases such as neuropsychiatric conditions, cardiovascular disease, and malignancy; nearly 35% because of communicable diseases and nutritional and perinatal conditions; and nearly 11% because of injuries. The spectrum differs widely from country to country.

The concept of DALYs is criticized because the unavailable information that has to be estimated by modeling, expert panels, and such other presumptive methods is substantial. Nevertheless, it remains the most comprehensive measure of burden of disease in a population. It should not be used as a summary measure of *comprehensive* health, because social and many aspects of mental health are not included in this measure.


1. Mathers CD, Vas T, Lopez AD, Salomon J, Ezzati M (Eds.). *National Burden of Diseases Studies: A Practical Guide*. Edition 2.0. Global Programme on Evidence for Health Policy. World Health Organization, 2001. http://www.who.int/healthinfo/nationalburdenofdiseasemanual.pdf
2. Wang H, Dwyer-Lindgren L, Lofgren KT, Rajaratnam JK, Marcus JR, Levin-Rector A, Levitz CE, Lopez AD, Murray CJ. Age-specific and sex-specific mortality in 187 countries, 1970–2010: A systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*. 2012 Dec 13;380:2071–94. http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(12)61719-X/fulltext
3. Vos T, Flaxman AD, Naghavi M, Lozano R, Michaud C, Ezzati M, Shibuya K, Salomon JA et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries, 1990–2010: A systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*. 2012 Dec 13;380:2163–96. http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(12)61729-2/fulltext
4. IHME. *Search GBD Data*. http://www.healthdata.org/search-gbd-data?s=world%20DALYs


## disability-free life expectancy, *see* life expectancy (types of)

## disability weight, *see also* disability-adjusted life years (DALYs)

The term *disability* can be defined as the deviation from optimal health in any domain, and disability weight is the measure of this departure. This reflects the general population's judgment about the extent of "healthlessness" rather than that of the patients. Depending on how these are calculated, the disability weights are also referred to as *quality-adjusted health weights* and *health state valuations*. These are extensively used to estimate the **burden of disease** in different segments of a population by a metric such as **disability-adjusted life years (DALYs)**.

Essentially, this means that 1 year of paraplegia counts several times more than 1 year of suffering from hypertension. Death is given a weight of 1.0, and a state of complete restriction of movements (bedridden) but no other restrictions (as in the case of fracture) can be given a weight of 0.6. Perfect health gets disability weight = 0.

Disability weights were originally worked out by a method called person-trade-off as reported by selected health professionals, but this was later judged unethical [1]. The Global Burden of Disease (GBD) study in 2010 undertook a comprehensive reestimation of disability weights, this time based on population surveys in five countries. The investigators found strong evidence of consistent results across countries with different cultural environments in this survey—providing evidence of robustness of these estimates. This study estimated disability weights for 220 health states using lay descriptions and with explicit caveats that it does not represent loss of well-being [1]. The questionnaire used in this exercise used a clear-cut choice between two states for the respondent to answer who is healthier. However, there are problems with this approach as well, particularly with long-term or permanent disabilities. For details, see Ref. [1].


1. WHO. *WHO Methods and Data Sources for Global Burden of Disease Estimates 2000–2011*. World Health Organization, 2013. http://www.who.int/healthinfo/statistics/GlobalDALYmethods_2000_2011.pdf?ua=1

## disconcordance, *see* concordance and disconcordance

## discrete variables (distribution of), *see also* continuous variables (distribution of)

It is first necessary to describe what is meant by a discrete (as opposed to **continuous**) variable. Some variables can take only a small number of values. Gender has only two possible values, and severity of disease generally has five values: none, mild, moderate, serious, and critical. Some other variables can take any of a large number of values, such as systolic blood pressure (BP) ranging from 100 to 200 mmHg. BP can be 132.7 mmHg if an instrument giving such accuracy is available. On the other hand, the parity of a woman can be 1 or 2 but never 1.6. A variable that can take only a finite, generally small, number of values in a range is called discrete. The number of deaths in a hospital in a day, blood group, and diagnosis are other examples of discrete variables.

It is common in medicine that discrete variables take only non-negative *integer* values, but the definition does not require it to be so. Shoe size can be 7, 7½, 8, 8½, etc. but is still discrete. It can take only these four values between 7 and 8½. Compare this with a variable such as age. If needed, age can be measured accurately as 7.2613 years. Although recording of age in terms of completed years is often considered adequate, particularly for adults, theoretically, age can take an infinite number of values between 7 and 8½. Thus, this is not a discrete variable.

Now for the distribution. For this purpose, discrete variables can be divided into several types, as shown in Table D.4. The last column of this table also names the corresponding discrete distribution.

All the distributions named in the last column of Table D.4 are described in detail in this volume under their respective topics. But these are not the only ones. If $r = 1$ in a negative binomial (number of "failures" before the first "success"), this leads to what is called a *geometric distribution*. You may occasionally find reference to *hypergeometric distribution* in medical literature. This is similar to a binomial distribution but arises when the probability of success changes from one occasion to another. The simplest is **uniform distribution**, when each of the possible $K$ values has the same chance of occurrence, i.e., each value has probability $1/K$.

Most discrete distributions tend to behave like a continuous distribution under limiting conditions. For example, the binomial distribution becomes nearly **Gaussian** when $n$ or $\pi$ (the probability of success in one attempt) or both are large such that $n\pi$ and $n(1 - \pi) \geq 8$. The probabilities for a discrete distribution are easily obtained by simple algebra; otherwise, statistical tables are also available. Statistical packages in any case give you the required probability for established distributions. However, **confidence intervals (CIs)** for the **parameters** of a discrete distribution are not so easy. For example, for binomial probability $\pi$, the **Clopper–Pearson interval** is used for finding the correct CI.

## discriminant analysis/functions

This section is divided into several subsections.

### *Discriminant Analysis*

The procedure to find the combinations of variables that best separate given groups is called discriminant analysis. Consider a setup where the classification structure of $n$ observations is known and this information is used to assign other observations whose classification is not known. Suppose you have information on clinical features and laboratory investigation for 120 thyroid cases. On the basis of extensive information and the response to therapy, they are divided into three groups, namely, the hyperthyroid, euthyroid, and hypothyroid patients. Assume that this division is almost infallible and there is practically no error. This is feasible in this case because the response to therapy is also known. The problem is to classify a new subject into one of these three groups on the basis of clinical features alone. This exercise would be useful when the facility for evaluation of thyroid functions is restricted. Response to therapy in any case would be available only afterward. What would be the best clinical criteria for classifying a new case with the least likelihood of error? The answer is obtained by discriminant analysis.

There are riders, though. The usual discriminant analysis is used when we have one or more continuous independent variables.

**TABLE D.4**
**Common Discrete Variables in Health and Medicine and Their Distribution**

| Features of the Characteristic | Examples | Variable | Statistical Distribution |
|---|---|---|---|
| Dichotomous (binary)—same probability of occurrence every time | Success/failure, yes/no, alive/dead, with disease/without disease, etc. | Number of "successes" in $n$ independent events | Binomial |
| Dichotomous (binary)—same probability of occurrence every time | Success/failure, yes/no, alive/dead, with disease/without disease, etc. | Number of failures one has to encounter before a prefixed number $K$ of successes are obtained | Negative binomial |
| Polytomous—mutually exclusive and exhaustive | Severity of disease (none/mild/moderate/serious/critical), blood group (O/A/B/AB), main cause of death (cardiac/cancer/kidney/etc.), parity (from 0 to a maximum, say, 7), etc. | Number of subjects in different groups | Multinomial |
| Count of cases where each occurs independently of the previous occurrence | Number of attacks of migraine in a year in one person, number of trauma deaths in a ward in 1 month, number of patients coming to an epilepsy clinic in a day, etc. There is no upper limit on this number. | Number of times an event occurs in a particular duration/place | Poisson |

The dependent, of course, is categorical and defines the groups. It is a multivariate technique that considers the latent dimensions in the independent variables for predicting group membership in the categorical dependent variable.

### Discriminant Functions

The combinations of variables that best separate the groups are called discriminant functions. They may not have any biological meaning. These functions are considered optimal when they minimize the probability of misclassification. If there are $K$ groups, the number of discriminant functions required is $(K - 1)$. The first is obtained in such a manner that the ratio of the between-groups **sum of squares** to the within-groups sum of squares is at a maximum. The second is obtained in such a manner that it is uncorrelated with the first and has the next largest ratio, and so on. If there are only two groups, only one discriminant function is needed. The nature of these functions is similar to the multiple regression equation, i.e.,

$$D_k = b_{0k} + b_{1k}x_1 + b_{2k}x_2 + \ldots + b_{Jk}x_J; k = 1, 2, \ldots, (K-1), \quad \text{(D.1)}$$

where $J$ is the number of $x$ variables on the right side. They are all considered **stochastic** in this setup, as opposed to regression, where they are considered fixed. The function in this equation is **linear**, but other forms can also be tried. The $x$ variables are **standardized** so that any particular $x$ or few $x$'s with large numerical values do not get an unfair advantage. The method used to obtain the function as represented in this equation is complex. Therefore, it is best to leave it to a software package.

If $J$ measurements $(x_1, x_2, \ldots, x_J)$ are available for each subject, it is not necessary to use all $J$ of them. Simple discriminant functions with fewer variables are preferable provided they have adequate discriminating power. This power is measured by the percentage of correct classification done by the discriminant function. Relevant variables can be selected by a stepwise procedure similar to the one explained for regression. See **stepwise methods** for details. This procedure also helps to explore which variables are more useful for discriminating among groups. Sometimes, just one variable may be enough to distinguish the groups.

### Classification Rule

When values of $x_1, x_2, \ldots, x_J$ for any particular subject are substituted in the function in equation D.1 mentioned in the preceding section, the value obtained is called the *discriminant score*. These scores for various sets of $x$'s are used in the **Bayes rule** to classify the cases. The probability required to use this rule depends on the distribution form of the variables. A **multivariate Gaussian distribution** is generally assumed. The rule also requires specification of the prior probability of a subject belonging to various groups. It is not necessarily equal. In some situations, it is known from experience that subjects come more frequently from one group than from others. This is generally estimated by the proportion of cases in different groups in the sample, provided there is no intervention that could distort sampling. If in 120 consecutive thyroid cases coming to a clinic, 30 are found to be hyperthyroid, 70 euthyroid, and 20 hypothyroid, then the estimates of prior probabilities are 30/120 = 0.25, 70/120 = 0.58, and 20/120 = 0.17, respectively. If the subjects are deliberately chosen to be in a certain proportion in your sample, such as being equal, then the prior probabilities are specified in accordance with the actual group prevalence in the target population. For example, you may wish to have 50 cases each of hyperthyroidism, euthyroidism, and hypothyroidism for your discriminant analysis, but the prior probabilities would continue to depend on their respective proportion in all the incoming cases that are slated to be the subjects for future classification. Many published reports seem to ignore this aspect and assume equal probabilities. Then the results can be fallacious. Also, do not try to temper prior probabilities to get a better classification.

In the case of two groups, the threshold for classification is

$$d = \frac{D_I + D_{II}}{2} + \frac{\ln(\text{prior probability of group II/prior probability of group I})}{D_I - D_{II}},$$

where $D_1$ is $D$ in Equation D.1 evaluated at the mean of the $x$'s in group 1 and $D_2$ is $D$ evaluated at the mean of the $x$'s in group 2. The labeling is such that $D_1 > D_2$. If $D > d$ for a particular subject, then that subject is assigned to group 1; otherwise, to group 2. This is called the classification rule. If the prior probabilities are equal, the second term on the right-hand side of this equation becomes 0.

### Classification Accuracy

The exercise of classification is first used on existing cases. For each case, a predicted class is obtained on the basis of the discriminant score for this case. Its actual class is already known. A cross-classification of cases by actual and predicted classification thus obtained is called a *classification table*. (See the example given later in this section.) This is used to find the percentage correctly classified, called the *discriminating power*. For discriminant analysis to be successful, this power should be high, say exceeding 80%. If a set of discriminant functions cannot satisfactorily classify the cases on which it is based, it certainly cannot be expected to perform well on new cases. When the percentage correctly classified is high, it is still desirable to try the discriminant functions on another set of cases for which the correct classification is known. It is only after such external validation that the hope for its satisfactory performance on new cases is high.

Introna et al. [1] studied the right patella of 40 male and 40 female Italian skeletons with respect to seven measurements: maximum height, maximum width, thickness, height and width of external facies articularis, and height and width of the internal facies articularis. Through a stepwise procedure, they found that only two measurements, maximum width and thickness, could discriminate gender correctly in 83.3% of cases. It was concluded that gender can be predicted by patellar dimensions when no other suitable remains of a human skeleton are available for gender determination. The prior probabilities in this example are 0.5 each because the two genders have the same proportions in the population. The following example gives the full details of the discriminant analysis in another context.

There is a considerable overlap between discriminant functions and logistic regression. The outcome in both is qualitative and must be mutually exclusive. Both represent the outcome as a function of the independent variables. Logistic regression can be used with a wide variety of data, but it serves best when the outcome is binary. Discriminant functions have no such restriction. But they work best when the independent variables have Gaussian distribution. Logistic regression has no such requirement.

### Example

Consider 5-year survival in cases of cervical cancer investigated as depending on age at detection (AGE) and PARITY. Suppose data are available for $n = 53$ cases, of whom 33 survived for 5 years or

more but 20 died before this. Can AGE and PARITY be effectively used to predict whether the duration of survival is going to be at least 5 years? Suppose the data summary is as given in Table D.5.

The difference between the groups in mean age at detection of cancer is not significant ($P > 0.50$), but the difference in mean parity is ($P < 0.05$). The ratio of survivors to nonsurvivors in this group is 33/20 = 0.623/0.377. Assuming that the same ratio will continue in the future, these can be used as prior probabilities for running the discriminant analysis. Then the following discriminant function is obtained:

$$D = -0.327 - 0.0251(\text{AGE}) + 0.774(\text{PARITY}).$$

For classification of subjects, we need the threshold $d$. This is obtained as follows from the equation given earlier. Substitution of group means in the discriminant function $D$ gives the following:

For nonsurvivors: $D_1 = -0.327 - 0.0251 \times 43.2 + 0.774 \times 2.6 = 0.6011$

For survivors: $D_2 = -0.327 - 0.0251 \times 42.6 + 0.774 \times 1.4 = -0.3127$

Thus,

$$d = \frac{0.6011 - 0.3127}{2} + \frac{\ln(0.623/0.377)}{0.6011 + 0.3127} = 0.1442 + \frac{0.5023}{0.9138}$$

$$= 0.69.$$

If $D > 0.69$ for a particular subject, then assign the subject to group 1 (surviving less than 5 years); if not, to group 2. The discriminant function in this case can be represented by a line, as shown in Figure D.6. This shows that the chance of survival for at least 5 years is less if age at detection and parity are high. Some points in this figure overlap.

Variations and uncertainties play their role, and many subjects are misclassified. The actual numbers are shown in Table D.6. Only 40 (75.5%) out of 53 could be correctly classified. As many as 8 (40%) out of 20 nonsurvivors are classified as survivors by this discriminant function. When the search is restricted to linear functions, no better discrimination criterion can be achieved. For better discrimination, either look for nonlinear functions or include more variables in addition to AGE and PARITY. Perhaps a more plausible conclusion is that AGE and PARITY by themselves are not sufficient to predict 5-year survival in cervical cancer cases. You can take the view that correct classification in three-fourths of the cases on the basis of such mundane variables as AGE and PARITY is good enough for practical applications. If this view is accepted, the discriminant function must be externally validated before it is used on new cases.

**TABLE D.5**
**Age and Parity in Cases of Cervical Cancer by 5-Year Survival**

| 5-Year Survival | Number of Subjects | AGE, Mean (SD) | PARITY, Mean (SD) |
|---|---|---|---|
| No | 20 | 43.2 (6.55) | 2.6 (1.43) |
| Yes | 33 | 42.6 (6.84) | 1.4 (1.27) |

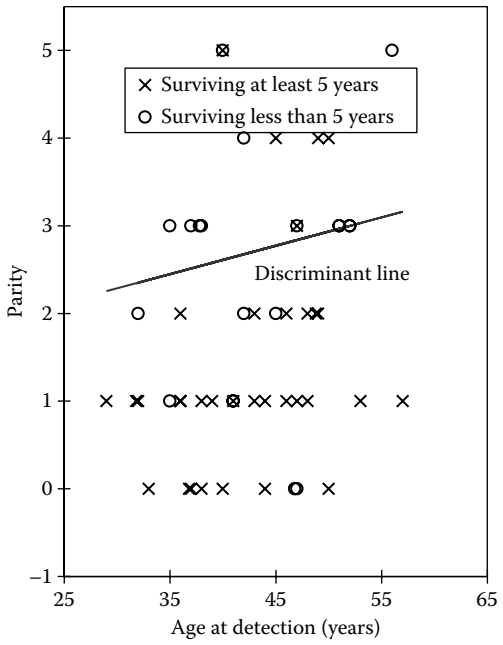



**FIGURE D.6** Scatter and discriminant line for surviving less than 5 years and at least 5 years.

**TABLE D.6**
**Classification Table Based on Discriminant Function in Our Example**

| | Predicted Group | | |
|---|---|---|---|
| **Observed Group** | **Survivors[a]** | **Nonsurvivors** | **Total** |
| Survivors[a] | 28 | 5 | **33** |
| Nonsurviors | 8 | 12 | **20** |
| **Total** | **36** | **17** | **53** |

[a] For 5 years or more.

### *Additional Points*

Note the following regarding discriminant functions:

- Discriminant functions divide the universe of subjects into sectors corresponding to different groups. Each sector defines the group to which a subject is most likely to belong.
- You might be aware that **logistic regression** is used when the target variable is binary. We have included a comment on this earlier in this section. Logistic regression for polytomous dependent variables exists, but that also generally considers two categories at a time. Discriminant analysis can be used for all categories together when the target variable is polytomous. The interpretation of a discriminant function, even when there are two groups, is not exactly the same as that of a logistic regression, but it serves the purpose of predicting the category. The associated probability can also be calculated. For the calculation of these probabilities, see Everitt and Dunn [2]. Most statistical software packages calculate these probabilities.
- Discriminant analysis is sometimes used to find whether or not a particular variable or a set of variables has sufficient

discriminating power between groups. Kilcoyne et al. [3] studied the renin–angiotensin system in 146 Black patients with essential hypertension in the United States. They divided these patients into low-, normal-, and high-renin groups. It was observed that none of the variables they studied, including incidence of cerebrovascular and cardiovascular events or age, had adequate discriminating power.

- As stated earlier, the foregoing discriminant analysis assumes that the variables are jointly multivariate Gaussian. This necessarily implies that the distribution of each variable is Gaussian. Mild deviations do not do much harm. In our example, the parity distribution is not Gaussian, yet the result may still be valid for large $n$. If many variables are binary, then the foregoing procedure is questionable. In that case, use a logistic discriminant function. This is briefly discussed by Everitt [4].
- The other important requirement for discriminant analysis is equality of the **dispersion matrices** in various groups. This is checked by the **Box *M* test**. If the dispersion matrices are really very different, the linear discriminant function is not adequate in separating the groups. A more complex, quadratic discriminant function [4] may be helpful in this case.


1. lntrona F Jr, Di Vella G, Campobasso CP. Sex determination by discriminant analysis of patella measurements. *Forensic Sci lnt* 1998; 95:39–45. http://www.ncbi.nlm.nih.gov/pubmed/9718670
2. Everitt BS, Dunn S. *Applied Multivariate Data Analysis*. Hodder Arnold, 2001.
3. Kilcoyne MM, Thomson GE, Branche G, Williams M, Garnier C, Chiles B, Soland T. Characteristics of hypertension in the black population. *Circulation* 1974;50:1006–13. http://circ.ahajournals.org/content/50/5/1006.full.pdf, last accessed December 3, 2014.
4. Everitt BS. *Statistical Methods in Medical Investigations*. Second Edition. Edward Arnold, 1994.


## disease spectrum

The term *disease spectrum* is used in at least three ways. This can be illustrated by considering pain as an example. First, one can think of characteristics of pain, such as the nature of pain (e.g., throbbing pain), regularity (e.g., intermittent or continuous), the extent of pain (e.g., local or generalized), etc. Second, one can consider the magnitude of pain as mild, moderate, severe, etc., or 7 on a 10-point scale. Prognostic severity, such as stage of cancer, also comes under this category. A third possible use of disease spectrum can be the description of the progression of infection from susceptibility to virulence. This is illustrated in Figure D.7.

This figure makes it seem as though a large percentage of exposed cases are infected and a substantial portion manifest the disease, but actually, these proportions are small in practice for most diseases.

Susceptibility is proneness to a disease. An effectively immunized person against diphtheria, pertussis, and tetanus (DPT) is not susceptible to these diseases for at least 1 year. A child is not susceptible to myocardial infarction. Susceptibility is a property of the host. Infectiousness is the property of the agent that causes infection. The measles virus is highly infectious for a susceptible person, but HIV is not. Infectivity is the actual performance in terms of the percentage infected when exposed. An infection may or may not manifest as disease. The percentage of susceptible persons who get the actual disease after the exposure can be called the pathogenicity of the disease. Out of those diseased, many will have a mild episode and recover easily. The percentage that have the severe form of disease requiring hospitalization and have a risk of death can be called the virulence of the disease. Cholera is highly pathogenic but less virulent, whereas rabies is less pathogenic but more virulent. AIDS is less infective but very pathogenic and highly virulent. At the end of this spectrum is mortality, which is the ultimate for virulence. Mortality can be measured per 1000 population or as percent affected. The latter is called the **case fatality**.

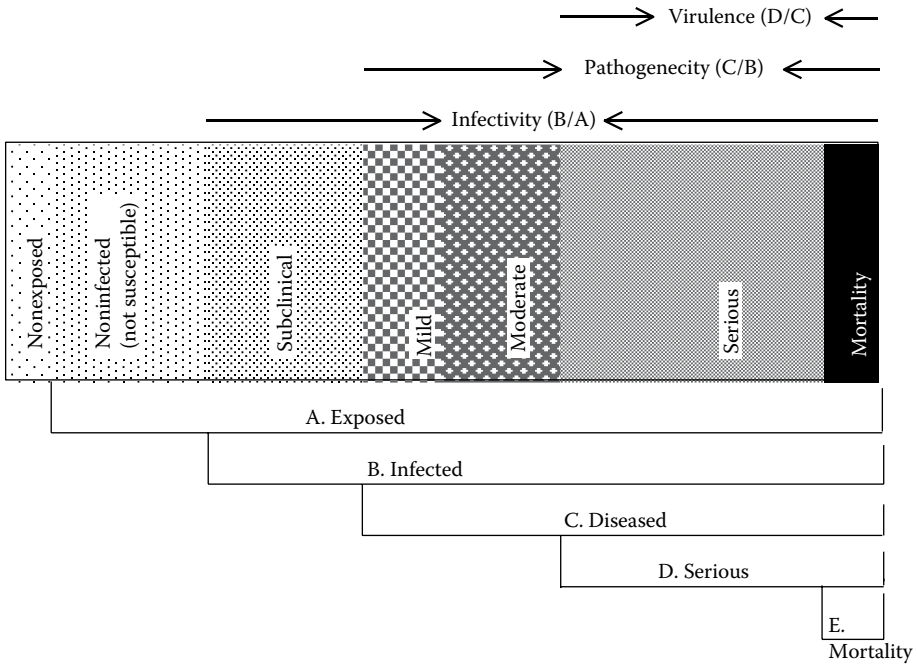



**FIGURE D.7** Disease spectrum.

The following indicators can be calculated to delineate the spectrum of a disease:

$$\text{Infectivity} = \frac{\text{number of subjects infected}}{\text{number of subjects exposed}} * 100$$

$$\text{Pathogenicity} = \frac{\text{number of subjects manifesting the disease}}{\text{number of subjects infected}} * 100$$

Virulence

$$= \frac{\text{number of subjects with serious disease (including mortality)}}{\text{number of subjects with disease}} * 100.$$

The terms *pathogenic* and *virulent* are borrowed from the field of infectious diseases but can be used for chronic diseases as well, as long as the meaning is explained. Division of cases into such a spectrum can help in choosing treatment strategies and in prognostic assessments. This also helps in their biostatistical classification.

The disease spectrum is likely to be very different during times of epidemics than in normal times. Another related concept is transmissibility. If an infected person is able to infect at least one person on average during the entire period of infectivity, then the infection will be sustained or increase. This is called the **reproductive number** of the infection. In countries where hepatitis B infection is on the rise, this rate is more than 1. If the reproductive rate is less than 1, expect that the infection will die down or stabilize at a low level.

## DisMod, *see* epidemiologically consistent estimates

## dispersion, *see* variation (measures of)

## dispersion matrix

In a multivariate setup, the variables have not just variances but also **covariances**. If these variables are ($x_1$, $x_2$, …, $x_K$), the covariance will be between $x_1$ and $x_2$, between $x_1$ and $x_3$, etc. A dispersion matrix (also called the **variance–covariance matrix** or just *covariance matrix*) is the arrangement of these variances and covariances in a square format such that the element in the $i$th row and $j$th column is the covariance between $x_i$ and $x_j$. This is denoted by $\sigma_{ij}$. Note that the "covariance" between $x_i$ and $x_i$, $\sigma_{ii}$, is the same as the variance of $x_i$. Thus, a dispersion matrix takes the following form:

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1j} & \cdots & \sigma_{1K} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2j} & \cdots & \sigma_{2K} \\ \vdots & \vdots & & \vdots & & \vdots \\ \sigma_{i1} & \sigma_{i2} & \cdots & \sigma_{ij} & \cdots & \sigma_{iK} \\ \vdots & \vdots & & \vdots & & \vdots \\ \sigma_{K1} & \sigma_{K2} & \cdots & \sigma_{Kj} & \cdots & \sigma_{KK} \end{pmatrix}$$

You can see that the dispersion matrix is the multivariate analog of the variance. Elements in the matrix give an indication as to how much dispersed values of each variable are and how much these are dependent on one another. Covariance is the measure of their linear dependence. The dispersion matrix is estimated by the sample analog of the variances and the covariances. When the variables are uncorrelated, the off-diagonal elements in the dispersion matrix are 0. In samples, they may be small values. A special form of this is the correlation matrix. A Pearsonian **correlation coefficient** is obtained when these variances and the covariances are divided by the multiplication of the respective standard deviations. In that case, all the diagonal elements of this matrix are equal to 1, and the off-diagonal elements are the Pearsonian correlation coefficients. For more details, see the topic **correlation matrix**.

Dispersion matrices are required in several multivariate methods such as **multivariate regression**, **multivariate analysis of variance (MANOVA)**, and **discriminant analysis**. Specifically, equality (homogeneity) of dispersion matrices in different groups is a requirement for a valid MANOVA test. Methods such as the **Box *M* test** are used for testing their homogeneity.

## distal measures, *see* proximal and distal measures of health and disease

## distribution-free methods, *see* nonparametric methods/tests

## distributions (statistical)

Statistical distribution of a variable describes which values are more common than others, which values are unlikely, which values do not occur, how much their variability is, whether or not they are symmetric with respect to any particular value, how much the asymmetry is, etc. For example, see the distribution of hemoglobin (Hb) values in women in Figure D.8. This shows that the most common values seen in these women are 12–13 g/dL. Values like 7–8 and 14–15 g/dL are least common; perhaps not even 1 in 50 women has such low or high values. None has a value less than 7 g/dL or more than 15 g/dL. It also shows that values below 12–13 g/dL are more common than values above, etc. Thus, the distribution is not symmetric.

The statistical distribution characterizes all the features of the data and helps us to determine what type of analysis should be done and which statistical methods should be used to derive valid inference. In short, the distribution is the backbone of the statistical methods. Mathematical formulation of these distributions helps in accurately working out the theoretical mean, variance, and such other parameters, as also the probabilities of any value bigger than or less than a specified value. Such probabilities are required primarily

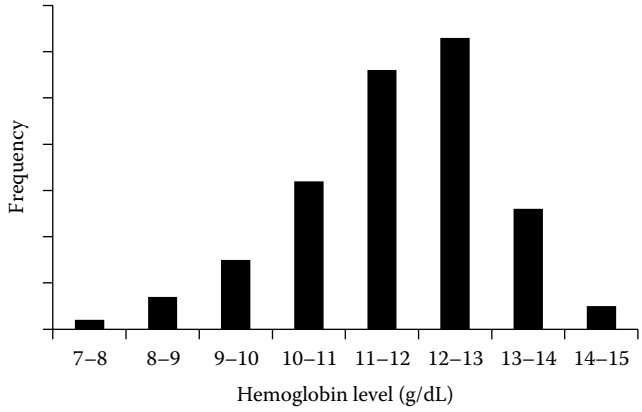



**FIGURE D.8** Illustrative distribution of Hb values in women in a developing country.

to find ***P*-values** for tests such as chi-square and Student *t*. Also, the properties of distributions of statistical summaries help in working out **confidence intervals** on the concerned parameter.

Because of the two broad natures of the variables—discrete and continuous—the distributions also have these two broad categories. Distributions of the discrete variables are called discrete distributions. These are described under the topic **discrete variables (distribution of)** in this volume. There you will find mention of hypergeometric, negative binomial, and uniform distributions. Major discrete distributions, namely, the **binomial**, **Poisson**, and **multinomial**, are presented separately under the respective topics. Similarly, major **continuous variables (distribution of)** are **Gaussian**, **chi-square**, **exponential**, **Weibull**, **Student *t***, and ***F***. All these also are discussed in detail in this volume under the respective topics.

Besides whether the distribution is discrete or continuous, other considerations while looking at any distribution for a statistical method are **skewness** and sometimes **kurtosis**. These features are measured as deviations, if any, from the **Gaussian** (normal) distribution—a central distribution for statistical theory and practices. When there are significant deviations, the distribution is generically termed non-Gaussian. When the distribution is highly skewed and sample size is not large, generally, **nonparametric methods** are preferred over Gaussian distribution–based methods. Many popular statistical methods such as Student *t* and ANOVA ***F*** are Gaussian based; so are the statistical tests and confidence intervals for many parameters. These would not be applicable if the distribution were highly skewed and the sample size were small. For conditions that allow use of a Gaussian distribution, see the topic **Gaussian conditions**.

The preceding discussion is restricted to what are called univariate distributions, since only one variable is being considered, but the distributions can be multivariate also where several variables are considered together. See the topic **multivariate distributions** for details.

## diurnal variation

These are the variations that occur at different times of the day in some medical parameters. These generally occur due to the dramatic effect of the day-and-night cycle on our physiological processes, as evidenced by the alternation of duration of activity and sleep. Theoretically, one might argue that there is a distinction between **circadian rhythm** and diurnal variation, but practically, there is no difference.

Thyroid-stimulating hormone (TSH) levels show a strong diurnal variation with low levels in the early morning and high levels in the evening. The variation could be on the order of 50%. Loss of this variation in TSH levels and other such parameters may be the expression of gradual alteration of the integrated function of the neuroimmune–endocrine system in subjects suffering from neoplastic disease [1]. The clinical onset of both myocardial infarction and stroke occurs more frequently in the early morning than at other times of day [2]. Diurnal variation in mood is a prominent symptom of depression [3].

Statistically, diurnal variations contribute to the uncertainties in interpretation of values, as illustrated by Goede et al. [4] for TSH. Indeed, many studies seem to ignore or forget about such variation, possibly leading to less valid conclusions. For measurements that show diurnal variation, the assessment in a research setup should be done at the same time of the day in all the subjects, and this time should be clearly mentioned in the article. In clinical practice, too, the assessment that a value is decreased or increased should also be done keeping diurnal variation in mind.


1. Mazzoccoli G, Vendemiale G, De Cata A, Carughi S, Tarquini R. Altered time structure of neuro-endocrine-immune system function in lung cancer patients. *BMC Cancer* 2010 Jun 21;10:314. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2910689/
2. Muller JE. Circadian variation in cardiovascular events. *Am J Hypertens* 1999 Feb;12(2 Pt 2):35S–42S. http://www.ncbi.nlm.nih.gov/pubmed/10090293
3. Murray G. Diurnal mood variation in depression: A signal of disturbed circadian function? *J Affective Disord* 2007;102(1–3):47–53. http://www.jad-journal.com/article/S0165-0327(06)00530-1/abstract
4. Goede SL, Leow MK. General error analysis in the relationship between free thyroxine and thyrotropin and its clinical relevance. *Comput Math Methods Med* 2013;831275. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3780511/


## doctor–population ratio, *see* health infrastructure (indicators of)

## donut diagram

Make a hole in a **pie diagram** and get a donut. Both have the same functionality and the same requirement. The total must be a meaningful quantity just as for a pie diagram. A donut comparison of age distribution of males and females in cases of mild hypertension is shown in Figure D.9. This shows that females are nearly one-half of males, as depicted by the size of the donuts in this figure. Note how the large percentage of those aged 35–44 years in females is shown relative to males.

The Slide Team [1] gives an excellent description of how donut diagrams can help in effective PowerPoint presentations.


1. Slide Team. *All Division Donut PowerPoint Diagram Slide*. http://www.slideteam.net/business_powerpoint_diagrams/division/division-donut.html, last accessed May 7, 2015.


## dose–response (type of) relationship

It may not be exactly true for a treatment regimen, but most toxic substances (such as insecticides) exhibit the property that higher dose results in higher response—in the case of insecticides, the response is mortality. The same property is exhibited by most anesthetic agents, where the response is the duration or effectiveness in causing sensory loss. While these may sound like extreme examples, many interventions show some pattern of change in response as the

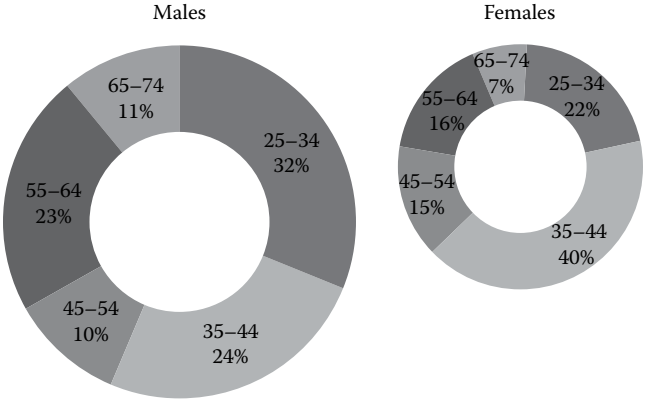



**FIGURE D.9** Donut diagrams showing age distribution of male and female cases of mild hypertension.

intensity or the amount of intervention is increased or decreased. This pattern of change is called the dose–response relationship. This could be positive (response increases with higher dose), negative (response decreases with higher dose), or none at all (response remains the same with changed dose). The relationship could be linear (each unit of additional dose brings about the same additional response) or multiplicative (each unit of additional dose has, say, twice as much effect as the previous dose), or can have any other pattern. For example, Ekwaru et al. [1] found that the dose–response relationship between vitamin D supplementation and serum 25(OH) D followed an exponential curve, and this is characterized by a multiplicative effect.

As always, for simplicity and parsimony, attempts are made to linearize the relationship between dose and response if it is not already linear, and transformations are used many times for this purpose. For example, log(dose) is used in place of dose, and square root of response is used in place of response. These transformations are called dose and response metameters, respectively. The relationship is examined by the methods used in **bioassays**, such as **parallel-line assays** and **slope-ratio assays**.

A dose–response relationship occurs in many other medical setups. This kind of relationship is known for the effect of radiation exposures and the effect of noise or air pollution. McCarthy-Jones and McCarthy-Jones [2] have observed mediation by body mass index and anxiety/depression in a dose–response manner for the majority of physical health disorders in women. Feng et al. [3] found a dose–response relationship between maternal parity and risk of congenital heart defects in offspring. A large number of such examples are available in the medical literature.

When investigating whether an association or a correlation indicates a cause-and-effect kind of relationship, there are a good number of criteria to consider, as enumerated by Indrayan [4]. One of these criteria is a dose–response kind of relationship in the sense that if the cause is present in a higher amount or at greater intensity, then the chance of an effect should also be high. The more you smoke, the higher your risk of lung cancer.


1. Ekwaru JP, Zwicker JD, Holick MF, Giovannucci E, Veugelers PJ. The importance of body weight for the dose response relationship of oral vitamin d supplementation and serum 25-hydroxyvitamin d in healthy volunteers. *PLoS One* 2014 Nov 5;9(11):e111265. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4220998/
2. McCarthy-Jones S, McCarthy-Jones R. Body mass index and anxiety/depression as mediators of the effects of child sexual and physical abuse on physical health disorders in women. *Child Abuse Negl* 2014 Nov 8. pii: S0145–2134(14)00345-7. http://www.sciencedirect.com/science/article/pii/S0145213414003457
3. Feng Y, Yu D, Chen T, Liu J, Tong X, Yang L, Da M et al. Maternal parity and the risk of congenital heart defects in offspring: A dose–response meta-analysis of epidemiological observational studies. *PLoS One* 2014 Oct 8;9(10):e108944. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4189919/
4. Indrayan A. *Medical Biostatistics*. Third Edition. Chapman & Hall, 2013: pp. 847–8.


## dot plot

Generically, a dot plot is a graphical display of data using dots in place of lines, bars, or curves. In this sense, all **scatter diagrams** are dot plots. But the most common use of dot plots in health and medicine is in representing a frequency **distribution**, particularly of a **discrete** variable (Figure D.10). In Figure D.10, we have used diamond as a dot, but you can have any other shape. Most common

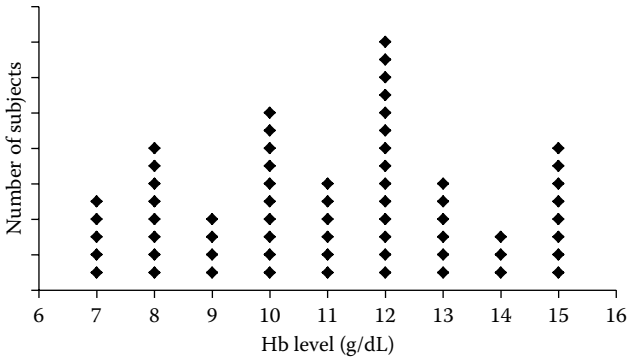



**FIGURE D.10** Dot plot of Hb level.

is the circle. Each observation is represented by a dot against the appropriate point on the horizontal scale. Thus, there are as many dots as the number of subjects. If the total number of subjects is large, each dot can represent, say, 10 subjects to avoid congestion. When a dot is representing more than one subject, this should be clearly specified. In this case, you can have a proportionate dot at the top to represent, say, seven subjects.

In Figure D.10, the hemoglobin (Hb) level actually is **continuous** but shown as discrete, where values between 6.5 and 7.4 g/dL are represented by 7 g/dL, between 7.5 and 8.4 g/dL by 8g/dL, etc. For such data, a dot plot is an alternative to a **histogram**, and the same interpretation applies. For example, you can talk about variability and **skewness** of the distribution on the basis of a dot plot of a continuous variable. But the dot plot suits well when the "values" on the $x$-axis are **qualitative**. These can be values such as site of injury, blood group, and state of residence. For such qualitative data, a dot plot is an alternative to a **bar diagram**.

## double blind trials, *see* blinding

## double sampling, *see* two-phase sampling

## dummy variables, *see* variables

## Dunnett test

Many times, the interest in multigroup experiments is in comparing each group with a particular reference group, and comparison between other groups is not required. Quite often, the reference group is the control group. For example, you may want to compare maternal serum copper concentration in normal pregnancies with pathological conditions such as threatened abortion, blighted ovum, and pyelonephritis. The Dunnett test is used for multiple comparisons where all the comparisons are with a particular reference group. This is given by

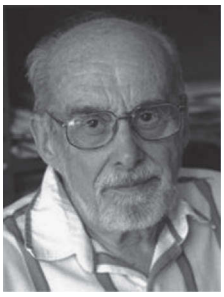
$$\text{Dunnett test: } t_d = \frac{\bar{x}_k - \bar{x}_0}{\sqrt{2\text{MSE}/n_h}},$$

where $\bar{x}_k$ is the mean of the $k$th group under comparison ($k$ = 1, 2, …, $K$) and $\bar{x}_0$ is the mean of the reference group. Thus, under this notation, there are a total of $(K + 1)$ groups in the experiment, including the reference group. Here $n_h$ is the **harmonic mean** of $n_0$ and $n_k$, the respective sizes of the reference and the comparison group.

The harmonic mean arises from $1/n_1$ $1/n_2$, occurring in the denominator of a two-sample **Student *t***. MSE is the **mean square error** that comes from the ANOVA table. The test is valid only under **Gaussian conditions**. Thus, either the sample size in each group should be large, or the values should follow a Gaussian distribution. The value of Dunnett $t_d$ is compared with its distribution at $(n - K)$ degrees of freedom (df's). You may find the value of $t_d$ for different df's in tables in some statistical books; otherwise, good software will provide the *P*-value directly.

In such comparisons, the result of the reference group is expected to be much more reliable since all the comparisons are with this group. Statistically, this means that the reference group should have larger $n$, generally, $n_0 = \max(n_k\sqrt{K})$ for all $k$. If not, the reference group values should have less variation from person to person than in other groups under comparison.

The Dunnett test was developed by Charles Dunnett in 1955 [1].



Charles Dunnett

(Copyright: Institute of Mathematical Statistics. Source: Archives of the Mathematisches Forschungsinstitut Oberwolfach.)

1. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *J Amer Stat Assoc* 1955;50:1096–1121. www.jstor.org/stable/2281208

## Durbin–Watson test

The assumption of **independence** of **residuals** is the most serious requirement for the validity of the analysis of variance (ANOVA) ***F*-test**. This is checked by the Durbin–Watson test. Residuals are the values unaccounted for by the model under consideration. The test was proposed in the years 1950–1951 [1,2].

Independence is threatened particularly in situations where observations are taken and the value of an observation depends on what it was at the preceding time. This is called autocorrelation or **serial correlation**. This happens in almost all repeated measures where, for example, a patient is assessed repeatedly after surgery. This can happen even after the time factor is properly accounted for, because of other factors that may also change results but are not accounted for.

The Durbin–Watson test checks that the residuals from a regression are statistically independent [1]. The residual at time $t$ is denoted by $e_t$ ($t = 1, 2, \ldots, T$). Generally, $T$ is the same as the number of observations $n$. The test statistic is

$$\text{Durbin–Watson } d = \frac{\Sigma_{t=2}^{T}(e_t - e_{t-1})^2}{\Sigma_{t=1}^{T} e_t^2}.$$

You can see that this statistic considers only the autocorrelation of lag 1. That is, this investigates the correlation between $e_2$ and $e_1$, $e_3$ and $e_2$, etc. If you suspect a correlation of $e_3$ with $e_1$, of $e_4$ with $e_2$, etc. (i.e., of lag 2), then the formula will accordingly change.

The rule of thumb for the Durbin–Watson test is that if $d$ is substantially less than 2, there is evidence of a positive serial correlation. If $d$ is less than 1.0, there could be a real problem. Small values of $d$ indicate that successive error terms are, on average, close in value to one another, or positively correlated. The value of $d$ becomes smaller as the serial correlation increases. If a test for negative autocorrelation is required, the test statistic to be used is $(4 - d)$ [3].

Most statistical packages routinely perform the Durbin–Watson test and give a *P*-value. If $P < 0.05$, you should reanalyze the data after controlling the factors that might be causing serial correlation. A strategy such as that of working with differences of successive values might be adopted in some cases. If $P \geq 0.05$ for Durbin–Watson, the *F*-test can be used as usual to test differences among means in different groups or for regression.

1. Durbin J, Watson GS. Testing for serial correlation in least squares regression, I. *Biometrika* 1950;37 (3–4):409–28. www.jstor.org/stable/2332391
2. Durbin J, Watson GS. Testing for serial correlation in least squares regression, II. *Biometrika* 1951;38 (1–2):159–79. www.jstor.org/stable/2332325
3. Neter J, Wasserman W, Kutner MH. *Applied Linear Regression Models*, Third Edition. CRC Press, Boca Raton, FL, 1990.

# E



## ecological fallacy

Ecological fallacy is interpreted in different ways by different researchers. This fallacy basically arises from wrong interpretation of group-based results to the individuals of these groups. Thus, this has three essential ingredients, and all of them must be present for an ecological fallacy to occur. These are the following: (i) the data must belong to groups and not to individual units, (ii) results are inferred for individuals, and (iii) when individual results become available, they contradict the group results [1]. If there is no contradiction, there is no fallacy. For example, it is generally believed that rice-eating vegetarian populations generally have better brains, whereas wheat-eating populations have better brawn. Extending this to the comparison of a rice-eating person with a wheat-eating person and saying that one will have a better brain and the other a better body, even in the sense of probability, can be fallacious. If some ancillary information on these two types of people is available to support this claim, then there is no problem. The problem arises when group results are directly extended to individuals *without* supportive evidence for the individuals, and the individual results are found to be contradictory.

This fallacy is primarily a fallout of **ecological studies**, where the units of measurements are at macro levels and not individuals, although, as explained shortly, this fallacy can also occur in other group studies. There are several variables that can be measured only at a macro level. Examples are environmental pollution and fluoride content in water supply at, say, a city level, and the type of cooking oil at a household level. These can seldom be measured at an individual level. Various community health indicators such as life expectancy and incidence of cardiovascular diseases (CVDs) are macro-level measures. When studied for many areas, these may have significant correlation. The units of study in this case are the areas and not individuals. A fallacy occurs when, for example, the results based on age at death and presence of CVDs in individuals do not provide the same kind of correlation.

In health and medicine, ecological fallacy is primarily considered a biostatistical error in interpretation when group results are used for individuals. Shah [2] studied suicide rates in the elderly (age 65+) and the prevalence of obesity (body mass index [BMI] ≥ 30 kg/m$^2$) in this age group in different countries of the world. They concluded that suicide rates in elderly females are independently associated with prevalence of obesity in different countries. The unit of study is country. They have cautioned against any extension to individuals, but if we do so, this can result in ecological fallacy.

Ecological fallacy is sometimes erroneously said to have occurred because individuals do not agree with the group means. This happens all the time and should not be termed any kind of fallacy. The nature of the mean is such that some values will always be higher, some always lower, and few, if any, will be exactly equal to the mean. In addition, the pattern in means in many cases will not match with the pattern in individuals. There are two glaring examples of this—one is the **Simpson paradox**, and the other is **regression**. Both are discussed in detail separately, but the following brief may be relevant to understand why these should not be termed *fallacies*.

Let us first understand regression. **Regression** is the relationship between the *mean* of the dependent $y$ and the given set of regressors $x$'s. Many times, it is forgotten that this relationship is for averages, and some researchers unnecessarily wonder why the results are not applicable to individuals. For individuals, something like **prediction interval** may be a better indicator of the uncertainties in the relationship. Now the Simpson paradox. This occurs when the picture obtained from the whole is different from what you get from the subgroups. This can happen when, for example, the case mix in the two groups is very different—one group has more serious cases, and the other group has more mild cases. For an example, see the topic **fallacies (statistical)**. This imbalance in the subgroups tends to provide a distorted picture of the whole. Marang-van de Mheen and Shojania [3] have discussed this in the context of standardized mortality ratios. **Standardization** of the two groups with respect to the case mix is a good remedy in such situations.


1. Idrovo AJ. Three criteria for ecological fallacy. *Environ Health Perspect* 2011 Aug;119(8):A332. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3237367/
2. Shah A. The relationship between obesity and elderly suicide rates: A cross-national study. *J Inj Violence Res* 2010 Jun;2(2):105–9. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3134913/
3. Marang-van de Mheen PJ, Shojania KG. Simpson's paradox: How performance measurement can fail even with perfect risk adjustment. *BMJ Qual Saf* 2014 Sep;23(9):701–5. http://qualitysafety.bmj.com/content/23/9/701.full.pdf+html


## ecological study, *see* ecological fallacy

## $ED_{50}$

$ED_{50}$ stands for "effective dose 50" and is used in pharmacology for the dose of a drug that is effective in 50% of cases or yields 50% response. Biostatistically, this is the **median effective dose** since this works in half of the cases and does not work in the other half. The concept of $ED_{50}$ works for a setup where increased dose is more effective, although it can also produce more side effects. Examples are anesthetic agents and other pain relievers. $ED_{50}$ can be used to compare the potency of one drug with another or to calibrate a drug for wider use. However, for this purpose $ED_{50}$ must be based on a trial on a random sample of the specified population of patients, and the trial must be conducted in standard conditions that can be fully specified and replicated. The "effect" should also be properly defined—it could be achieving the desired level of cholesterol, the ability to do some work, specified relief in pain, or any other such measurable end point.

Note that estimation of $ED_{50}$ is a kind of reverse biostatistical process. The general setup is to estimate the efficacy and its confidence interval of a given dose of a drug. Here, different doses are tried to find the dose that gives 50% efficacy. The confidence interval is obtained for the dose level and not for efficacy. This requires studying the dose–response relationship (Figure E.1) (sometimes after converting dose to log-dose and even response to log-response),

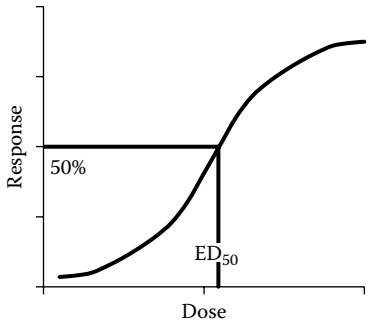



**FIGURE E.1** Dose–response curve and $ED_{50}$.

and the methods of **bioassays** are used. In Figure E.1, the response is symmetric on either side of $ED_{50}$, but for some drugs, this may not be so. In some cases, estimation of $ED_{50}$ requires that the factors influencing this also be considered, such as body mass index and severity of disease. This can also be done when needed.

You can see that $ED_{50}$ is not directly useful to clinicians for patient management. No clinician would want to use median effective dose on his/her patients. Clinicians probably will want to use the dose that is effective in 95% of cases if the side effects are negligible. This is called $ED_{95}$. Hang et al. [1] have described a study on estimation of $ED_{95}$, in addition to $ED_{50}$, of ketamine for prevention of postoperative hyperalgesia in patients undergoing laparoscopic cholecystectomy. $ED_{95}$ is worked out keeping in view its use in actual conditions.

$ED_{50}$ is primarily used to assess the potency of a drug in producing a change. Ramboer et al. [2] studied $ED_{50}$ of cetirizine and loratadine for inhibiting the wheal and flare response of the skin to the histamine pick test and concluded that on average, cetirizine is seven to nine times more potent than loratadine. This illustrates how $ED_{50}$ is used for comparing potency. Statistical methods are available to compare $ED_{50}$ of two regimens. These are based on **bioassays**.

A corresponding concept for poisons such as insecticides is the lethal dose 50 ($LD_{50}$)—the dose that kills half of the organisms or kills half of the subjects, such as rats in a laboratory.

1. Hang LH, Shao DH, Gu YP. The ED50 and ED95 of ketamine for prevention of postoperative hyperalgesia after remifentanil-based anaesthesia in patients undergoing laparoscopic cholecystectomy. *Swiss Med Wkly* 2011 May 10;141:w13195. http://www.smw.ch/content/smw-2011-13195/
2. Ramboer I, Bumtbacea R, Lazarescu D, Radu JR. Cetirizine and loratadine: A comparison using the ED50 in skin reactions. *J Int Med Res* 2000 Mar–Apr;28(2):69–77. http://www.ncbi.nlm.nih.gov/pubmed/10898119

## education indicators

Education is among the core indicators of **social health**. Many parameters of physical health are directly or indirectly affected by the level of education. Education creates awareness about health issues and thus affects health status. In clinics, the education level of the patient decides the level of interaction. In medical research, the education profile of the subjects provides the demographic structure of the subjects, which helps to assess whether or not the results can be applied to the other populations.

Education can be measured at the individual level as well as at the community level. At the individual level, years of schooling is a useful indicator as it incorporates different types of schooling into a single index. Sometimes, information on the kind of education (engineering, medical, law, etc.) is useful, and sometimes, occupation is used as a surrogate. However, our focus in this section is on community indicators.

At least four different types of indicators are used to assess the level of education in a community. For developing countries, where a substantial segment of the population is illiterate, the literacy rate can be a useful indicator. This is the percentage that is literate among people of age 6 years and above. Children younger than 6 years are supposed to be neither literate nor illiterate and thus are excluded from calculation. This gives

$$\text{literacy rate} = \frac{\text{literates among population of age 6 years and above}}{\text{total population of age 6 years and above}} * 100.$$

When calculated for age 15 years and above, this is called the **adult literacy rate**. When calculated for age 15–24 years, this is called the youth literacy rate. In the block years 2008–2012, the youth literacy rate among males of Chad was 53.6%, and among females, 42.2% [1]. This rate is showing a rapid rise all across the world, but in almost all countries, all education indicators are lower in females than in males.

The second indicator of education is the average years of schooling in the adult population.

$$\text{Average years of schooling} = \frac{\text{sum total of schooling of all individuals in a population (25+ years)}}{\text{total population (25+ years)}}.$$

To maintain validity, the age is restricted to 25 years and above for the numerator as well as for the denominator. It is presumed that all schooling will be over in practically all individuals in the population of this age. The years of schooling for each individual in a community are rarely available, and thus, this average is difficult to compute. As a middle path, the third indicator is

$$\text{percentage of adults with high school diploma} = \frac{\text{adults with high school diploma}}{\text{total adults}} * 100.$$

Adults can be defined according to local conditions. Information on the addition of high school diploma holders each year can be obtained from schools, boards, or councils awarding such a diploma. But keeping track of exits, by way of migration or death, can be difficult. To overcome this problem, the fourth indicator used is

$$\text{net enrollment ratio} = \frac{\text{children of age 6–16 years enrolled in a school}}{\text{total children of age 6–16 years in the population}} * 100.$$

*Net enrollment ratio* measures the extent to which children are using educational facilities. In place of 6–16 years, any other age group can be used. The enrollment figures can be easily obtained from schools, and the age structure of the population is generally available from other sources. Thus, this indicator can be easily computed.

Net enrollment ratio can be computed separately for primary, secondary, and tertiary schools by adjusting the age accordingly. There might be children in schools who are younger than 6 years and older than 16 years. When the numerator of this equation is relaxed to include such children also, it is called the *gross enrollment ratio.*

Since the age group in the denominator cannot be relaxed, the gross enrollment ratio can exceed 100, as occurring in Australia (104 in 2011) for primary education [2].

1. UNICEF. *Chad: Statistics*. http://www.unicef.org/infobycountry/chad_statistics.html
2. The World Bank. Data. *School Enrolment, Primary (% Gross)*. http://data.worldbank.org/indicator/SE.PRM.ENRR

## effect modification, *see* interaction

## effect size, *see also* medically important effect (the concept of)

The effect size of an intervention is the measure of the change in outcome after that intervention. This is a quantitative measure of the magnitude of the effect of the intervention—thus, it is more than just statistical testing of a hypothesis that finds whether an intervention is effective or not. This answers "how much" the effect is, and thus, a meaning is attached to the effect.

If a regimen has increased the response from 36% in controls to 54% in the test group, the effect size is 18%. One can also say that the effect is 1½ times as much in the test group as in the controls. If an intervention has decreased a sickness in a community from 15% per year to 12% per year, the effect size is 3% per year. If a supplement has increased the hemoglobin level from an average of 8.3 g/dL in anemic subjects to 10.5 g/dL after 1 month, the effect size is 2.2 g/dL. The unit of measurement changes depending upon the kind of outcome. It can be measured in terms of odds ratio, correlation coefficient, or even the change in the value of the regression coefficient.

In cases where the effect size is measured by the difference in means in two independent groups, it is customary to make it independent of units by dividing this difference by the corresponding standard deviation (SD) if this can be assumed same for the two groups. Thus, in this case, effect size = $(\mu_1 - \mu_2)/\sigma$. This is estimated by $(\bar{x}_1 - \bar{x}_2)/s$, sometimes known as *Cohen d*. In cases where the SDs are different in the two groups, you can use either the average of the two SDs or **pooled SD**, as used in Student *t* for testing the difference between two independent samples. No standard guidelines are available. One can subjectively say that the effect size, when computed in SD units in this manner, is small if around 0.2, medium if around 0.5, and large if around 0.8. This would, of course, depend on the context and can exceed 1. Since this measure is independent of units, you can easily compare the effect size of one intervention with another intervention. The usual precautions, such as adequate sample size, unbiased averages, and comparable group composition, are applicable to the estimation of effect size.

For correlation also, the same cut points (0.2, 0.5, and 0.8) can be used to subjectively interpret the effect size. No such guidelines are available for proportions because sometimes an increase of just 1% from 0.01 to 0.02 can be substantial, as in the case of the chance of death becoming twice as much in a disease that can have severe implications, and sometimes even a difference of 10%, as in the case of opinion in favor of a regimen increasing from 60% to 70%, may not have much health implication.

Effect size and its confidence interval is the core quantity in **meta-analysis**. This is because in meta-analysis, results from diverse studies meeting the preset criteria are combined to provide a pooled estimate. For this, it is necessary that the effect size be measured in comparable units.

## efficacy and effectiveness

The efficacy of a regimen is its positive response rate in ideal conditions, and effectiveness is the positive response rate in actual conditions. Much of the medical literature fails to distinguish between the two, but these can be very different.

**Clinical trials** are generally done in ideal conditions that do not exist in practice. The subjects are carefully chosen with strict inclusion and exclusion criteria, administration is done in standard conditions, efforts are made for full compliance, patients get full attention, the results are adjusted for dropouts and other missing observations, and the response is carefully assessed by experts. The actual performance of the regimen in practice may differ. Efficacy of a treatment is what is achieved in a trial that simulates optimal conditions, and effectiveness is what is achieved in practical conditions when the treatment is actually prescribed. For clarity, the latter is sometimes called use effectiveness.

Effectiveness could be lower than efficacy because of lack of compliance to the regimen due to cost or inconvenience, inadequate care, nonavailability of the drugs, etc. These rarely occur in a trial. Experience suggests that nearly three-fourths of patients do not adhere to or persist with the full prescriptions. Thus, the results for patients and maneuvers adopted during a trial do not translate to patients at large. Consequently, such external validity of trial results is not high. But clinical trials do establish the potential of a regimen to effect a change. Effectiveness, on the other hand, is a suitable indicator to decide whether or not to adopt that regimen in practice, or what to expect. For further details, see Singal et al. [1].

The concept of effectiveness as opposed to efficacy has given rise to what are called **pragmatic trials**. As explained under that topic, pragmatic trials are conducted under actual field conditions.

1. Singal AG, Higgins PD, Waljee AK. A primer on effectiveness and efficacy trials. *Clin Transl Gastroenterol* 2014 Jan 2;5:e45. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3912314/

## empiricism

Empiricism can be roughly equated with evidence available from sensory experience. Inherent in this is just not the experience but also the change that comes with time as more experience is accumulated. A hypothesis, for instance, better health leads to more happiness, remains just a thought till such time that it is really seen or experienced by at least a segment of the population. Thus, evidence is an integral part of empiricism. In contrast, mathematics and some other physical sciences are based on theories and lemmas. They are deductive and not empirical. Deductive science holds that the mind can directly perceive truth without going through the process of sensual experience. Empiricism is based on induction from sensual learning.

Empirical evidence could arise from experiments, trials, natural occurrences, observations, records, etc. It refers to the actual facts as currently present or occurred in the past. Empiricism emphasizes the tentative and probabilistic nature of knowledge. Thus, probability, and hence statistics, plays a key role in empirical conclusions.

Empiricism is sometimes contrasted with rationalism, although there is no conflict. The observations must stand up to reason and should have an adequate rational explanation. After all, it is the logic of reasoning that separates humans from other species. Research results are more acceptable when the accompanying evidence is compelling, stands to reason, and inspires confidence. Without logic, research is reduced to storytelling.

### *Medical Empiricism*

Medicine is a delicate science. It is concerned with vital aspects of life such as health, disease, and death. Thus, it brooks no error. Ironically, no theories are available that can make it infallible, and there are no lemmas and no theorems. It must depend on evidence provided by observations and experience. Medicine is largely an inductive science and has very little space, if any, for deductive methods. It is individualized yet participatory in the sense that if a treatment regimen has worked in Mr. Somebody and nine others of his clan, there is a high likelihood that it would work in the eleventh also of that type. Past experience and present evidence provide insight into the future. Such empiricism is the backbone of medical science. Empirical research helps in the quantitative assessment of the plausibility of complex issues. In dealing with a new case, or an old case with a new set of conditions, past knowledge and experience are applied, and it is hoped that they will work in the new setup also. Very often, they do, but sometimes, they do not. There is no assurance, and miscues occur.

Much of medical research is on cause and effect. Mental illumination, stipulated in theories, may provide a clue on what really may be going on, but this is not accepted in medicine without support from actual observations. Only then can you hope to convince colleagues to accept your theory. Thus, there is no escape in medicine from the evidence base and empirical process. For example, even when genomics has a wide role, the decision of what treatment to administer in a given gene structure would still depend on gaining experience.

## empty cells, *see* contingency tables

## endogenous and exogenous factors and variables

Factors working within the system under study are called endogenous, and those working from outside are exogenous. If you are studying the cardiovascular system, kidney functions are exogenous, whereas heart rate, cardiac output, and arterial pressure are endogenous. Exogenous variables are not necessarily independent—they can affect the endogenous variables, albeit in an indirect manner. They do not have a direct causal link.

Treatment for a disease will generally be directed to controlling the endogenous factors, and separate regimens may be needed to control the effect of exogenous factors. For example, for vaginal candidiasis, the oral cavity and anus could be endogenous sources of reinfection as they are from the same woman, while infection coming from sex partners is exogenous [1]. For prostate and breast cancer, genomic variations are endogenous, and lifestyle is exogenous [2]. Thus, the terms have wide and varied usage. Biostatistical usage is based on entirely different considerations, as explained next.

Classification of factors as endogenous and exogenous as stated in the preceding paragraph is for medical purposes. For biostatistical modeling, both medically endogenous and exogenous variables can be included in the same model. Endogenous and exogenous classification in biostatistical modeling depends on the model you are working on. When working on how total bilirubin level is affected by age, sex, body mass index, and dietary content in healthy people, all these are endogenous to the study. If later on it is realized for some reason that the pulse rate should also be included, this is exogenous since this is not related to any of the variables already under investigation. In a **regression** of an outcome $y$ on a set of independents $(x_1, x_2, \ldots, x_K)$, all these variables are endogenous to this model. However, $y$ may be affected by another variable $x_i$; this is exogenous for this model if it is not causally related to any of $(x_1, x_2, \ldots, x_K)$.

Whether the variables are endogenous or exogenous has important implications for some statistical methods, such as **path analysis** and **structural equation models**. Both investigate how exogenous variables affect endogenous variables, although in structural equation models, the relationship between latent factors and their indicators is also investigated. In these analyses, particularly in structural equation models, the observed variables are endogenous, and the underlying constructs that cannot be directly observed are exogenous.


1. Polanowska MM, Koszko IW, Klimowicz B et al. Endogenous or exogenous origin of vaginal candidiasis in Polish women? *Pol J Microbiol* 2013;62(3):311–7. http://www.pjm.microbiology.pl/archive/vol6232013311.pdf
2. Blein S, Berndt S, Joshi AD, Campa D, Ziegler RG, Riboli E, Cox DG. NCI Breast and Prostate Cancer Cohort Consortium. Factors associated with oxidative stress and cancer risk in the Breast and Prostate Cancer Cohort Consortium. *Free Radic Res* 2014 Mar;48(3):380–6. http://www.ncbi.nlm.nih.gov/pubmed/24437375?report=abstract


## enrollment ratio, *see* education indicators

## ensemble methods

Ensemble methods help in scientific aggregation of predictions of the same outcome by multiple models. Models can differ with respect to the variables used for prediction; case mix; statistical method (logistic, discriminant functions, regression tree, etc.); and the algorithm (agglomerative or divisive in the case of cluster analysis, backward elimination or forward selection in the case of regression, etc.). It is natural to expect that this aggregation many times would yield a prediction that is better than any of the constituent models when the right ensemble method is used. "Better" would be mostly in terms of **reliability** and, in some cases, even **validity** of the prediction, perhaps also generalizability. But the ensemble methods are mostly used for models based on huge data sets. It is desirable that the ensemble results on a new data set are tested to confirm that they really work better than any of its constituents.

A simple ensemble method is to take average of the predictions by specific inputs to diverse models. This has the same merits and demerits as any other mean. Different models may be based on different sample sizes and different methods, and it is necessary to carefully examine whether or not the average can be used. The other method is *bootstrap aggregation*, forming the acronym *bagging*. As explained for the topic **bootstrap**, this generates replicates of the data set by repeated sampling from the original data set with replacement. In this case, the bootstrap samples have the same size as the original data set. Then the model is built on each of these bootstrap samples. Suitable aggregation methods are used to get the ensemble estimates. Ghafourian and Amin [1] found boosted regression trees to be the most appropriate ensemble method for prediction of plasma protein binding of drugs. These methods are too mathematically complex for inclusion here. For more details on ensemble methods, see Rokach [2].


1. Ghafourian T, Amin Z. QSAR models for the prediction of plasma protein binding. *Bioimpacts* 2013;3(1):21–7. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3648908/
2. Rokach L. Ensemble methods for classifiers. Chapter 45, In: *Data Mining and Knowledge Discovery Handbook*. http://www.ise.bgu.ac.il/faculty/liorr/hbchap45.pdf, last accessed June 15, 2015.

## epidemic curve

A graph showing the number of new cases occurring each day or any other time period from the beginning to the end of an epidemic is called an epidemic curve. Time is plotted on the horizontal axis and the number of cases on the vertical axis. Even though this could be a bar diagram, it is interpreted as a curve because it shows trend over time.

In the case of an epidemic of infectious diseases, the number of cases gradually rises, reaches a peak, and then starts to decline. The shape of the curve depends on whether it is a common-source epidemic or a propagated epidemic. An infectious disease epidemic in a population that has a common source (such as contaminated water for cholera) would look like Figure E.2a. A sharp peak occurs a couple of days later than the incubation period, but tapering off may not be as sharp, because of varying time taken in recovery. The Bhopal gas tragedy in India and the Minamata disease outbreak in Japan caused by eating fish containing high concentrations of methyl mercury are well-known examples of common-source epidemics.

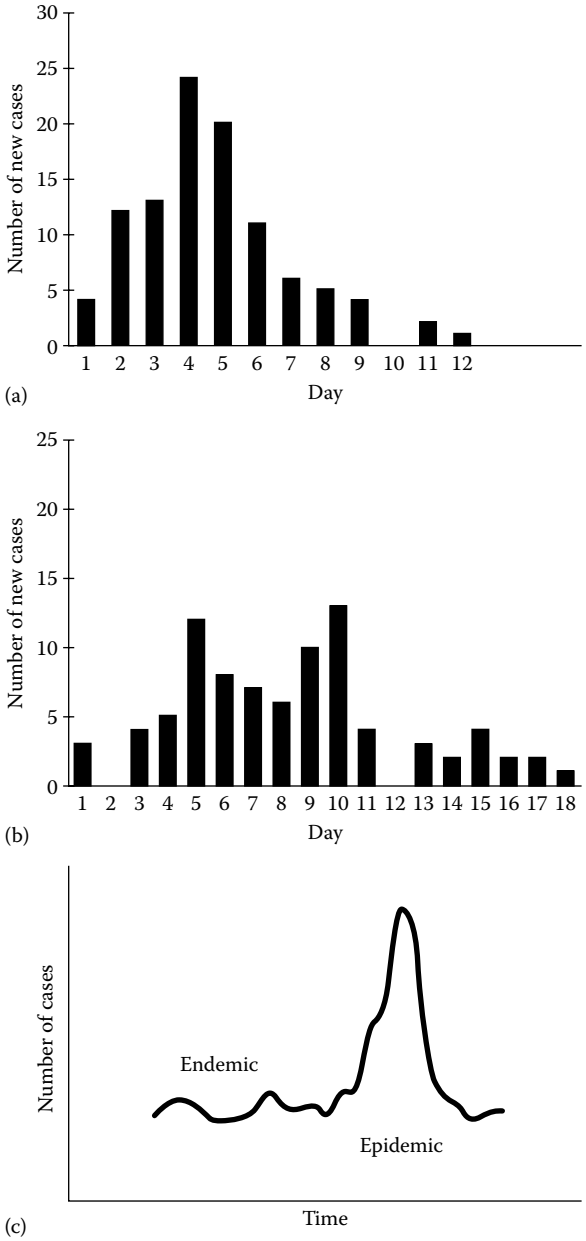



**FIGURE E.2** Epidemic curves: (a) common source; (b) propagated; (c) endemic versus epidemic levels.

A propagated epidemic results from person-to-person transmission (e.g., of hepatitis A and polio). It shows a gradual rise, sometimes intermittent peaks, and a very gradual fall (Figure E.2b). The speed of spread depends upon herd immunity, opportunities for contact, and the **secondary attack rate**. If you want to see actual epidemic curves, see Chen et al. [1] for the epidemic of shigellosis in China and Barnea et al. [2] for influenza-like illness in Israel.

An epidemic of a disease is said to exist when the occurrence is clearly in excess of the usual occurrence. The term was originally used for infectious diseases such as influenza and cholera but is now used for conditions such as deaths in vehicular accidents in vacation times. A steep rise in cases over the endemic level is shown in Figure E.2c. It helps to identify the time of start and finish of an epidemic. This is the real epidemic *curve* and not a bar diagram.

1. Chen T, Leung RK-k, Zhou Z, Liu R, Zhang X, Zhang L. Investigation of key interventions for shigellosis outbreak control in China. *PLoS One* 2014;9(4): e95006. http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0095006
2. Barnea O, Huppert A, Katriel G, Stone L. Spatio-temporal synchrony of influenza in cities across Israel: The "Israel Is One City" hypothesis. *PLoS One* 2014;9(3): e91909. http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0091909

## epidemic models, *see also* infectious disease models

In the context of biostatistics, the epidemic model is the equation that describes the number of new (incident) and/or existing (prevalent) cases at various time points during an epidemic. It is a mathematical formulation of the **epidemic curve**. This formulation smoothens the random fluctuations and can be used to estimate the precise time of the start and end of the epidemic, the time of peak incidence, and the rate of increase and decline. Any cyclical or seasonal components can also be included in the model. If the disease reaches an endemic stage, that also can be identified. More practical use of such models is in projecting the number of cases while the epidemic is still going on and in estimating the effect of interventions undertaken to modify the course of the epidemic.

While there is quite some overlap between epidemic models and **infectious disease models**, they must be regarded as distinct entities. Infectious disease models incorporate the number or proportion of the population immune, susceptible, and exposed; infectives; infectivity; recoveries; etc. On the other hand, in our opinion, epidemic models are simply for the number of cases at different stages of an epidemic that can possibly be estimated by empirical processes where some of these parameters can also be used. For example, Chin and Lwanga [1] used the following **gamma function** as an epidemic model (they called it an epidemiological model) for estimation and projection of adult AIDS incidence:

$$y = \frac{1}{(p-1)!} t^{(p-1)} e^{-t},$$

where $y$ is the incidence at time $t$ and $1/p$ is the steepness of the curve and $(p-1)$ is the duration to reach peak ($p \geq 1$). The larger the $p$, the slower the progress of the epidemic. On the basis of data from some countries, they estimated it as $p = 5$ years (Figure E.3). This estimate is based on empirical observations and not on the incubation period, force of infection, number of susceptible, etc. Time $t = 0$ is the time

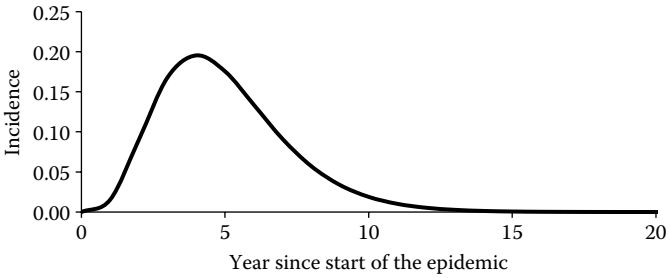



**FIGURE E.3** Graphical representation of an epidemic model for AIDS cases.

when the epidemic started. This model is fit for person-to-person transmission and assumes that the epidemic will ultimately relent. The peak of the epidemic is at time $(p - 1)$, in this case, at 4 years. If the data suggest that the peak will be reached in 8 years, you can choose $p = 9$ in this model.

The model presented in the preceding paragraph shows that the infections steeply increase in the beginning, reach a peak, and then gradually decline as the susceptible pool is depleted. Another approach for epidemic modeling, distinct from infectious disease models, could be time series fitting, as done by Lee and Wong [2] for pandemic influenza A in Hong Kong and surrounding areas. Both these approaches exploit the time trend instead of the susceptible–infections–recoveries route of the infectious disease modeling.

1. Chin J, Lwanga SK. Estimation and projection of adult AIDS cases: A simple epidemiological model. *Bull World Health Organ* 1991;69(4):399–406. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2393240/
2. Lee SS, Wong NS. Reconstruction of epidemic curves for pandemic influenza A (H1N1) 2009 at city and sub-city levels. *Virol J* 2010 Nov 16;7:321. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3003267/#!po=50.0000

## epidemiologically consistent estimates

You may be aware that the number of cases of any disease present (prevalence) in a community is affected by how many new cases are occurring (incidence), what the duration of illness is, at what rate the cases are getting cured (remission), and how many are dying (case fatality). This is explained in Figure E.4. All these can be woven into an equation under certain stability conditions, and one can be determined by the others. Also, age at onset, remission, and duration are related to age-specific mortality for that disease. When all these fall into the expected pattern, they are called epidemiologically consistent.

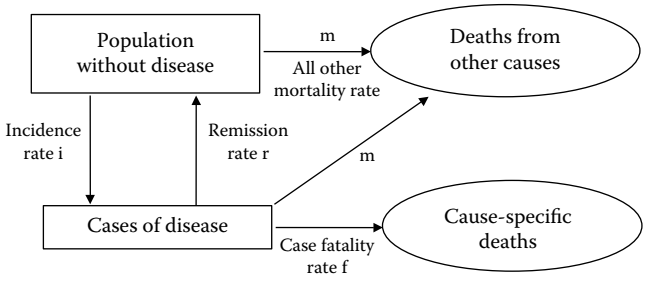



**FIGURE E.4** Interrelations for epidemiologic consistency of data. (From Mathers CD et al. [Eds.]. *National Burden of Diseases Studies: A Practical Guide*, Edition 2.0. Global Programme on Evidence for Health Policy. World Health Organization, available at http://www.who.int/healthinfo/nationalburdenofdiseasemanual.pdf, 2002.)

Quite often, different rates are estimated from disparate sources, such as prevalence and age-specific mortality from a cross-sectional survey; incidence from one cohort study; and remission, duration, and case fatality from another longitudinal study. There is a great likelihood in this case that various rates are not internally consistent. Age-specific mortality may not be the same as expected on the basis of age at onset, duration, and case fatality, or prevalence may not be the same as expected from the incidence, duration, remission, and mortality estimates. A software package called **DisMod** II, available from the World Health Organization (WHO) website [2], can be used to check the internal consistency of these estimates with some limitations. In case they are not internally consistent, more reliable rates should be used to generate consistent estimates of the other rates. This package can also be used to generate estimates of rates that are not available at all. The generated rate would be epidemiologically consistent but may not be plausible in terms of the knowledge of the experts. In that case, iterations may be needed in terms of reentering a new set of known rates so that the final estimates are not only internally consistent but also plausible.

A new version of DisMod has arrived, called DisMod-MR. This is based on a meta-regression approach and uses the systematic review of different studies to come up with a set of more valid estimates. Details of this method have been provided by Vos et al. [3].

1. Mathers CD, Vas T, Lopez AD, Salomon J, Ezzati M (Eds.). *National Burden of Diseases Studies: A Practical Guide*. Edition 2.0. Global Programme on Evidence for Health Policy. World Health Organization, 2002. http://www.who.int/healthinfo/nationalburdenofdiseasemanual.pdf
2. WHO. *Health Statistics and Information Systems: Software Tools*. http://www.who.int/healthinfo/global_burden_disease/tools_software/en/, last accessed June 26, 2015
3. Vos T, Flaxman AD, Naghavi M. Lozano R, Michaud C, Ezzati M, Shibuya K et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: A systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 2012;380:2163–96. http://www.ncbi.nlm.nih.gov/pubmed/23245607

## epidemiological studies

Although epidemiology is concerned with distribution, dynamics, and determinants of health status (including disease) in a population, all **observational studies** are customarily termed epidemiological studies. This is in contrast to **experiments**, where a human intervention is deliberately introduced to study its effect on a group of subjects. Experiments include all kinds of clinical and field trials whether for diagnostic, screening, prophylactic, or preventive purposes. Epidemiological studies are sometimes understood as nature's experiments in situations where intervention is not deliberate. If somebody is or was a smoker, you can study the effect of smoking without deliberately introducing smoking. It is generally assumed that epidemiological studies would be analytical since they would be concerned with an antecedent–outcome relationship; that really is not so. **Descriptive studies** are also considered epidemiological as they are concerned with the distribution of various factors in a group of subjects. Disease surveillance also falls under this term.

The primary format of epidemiological analytical studies is **prospective**, **retrospective**, or **cross-sectional**. A prospective study could be a cohort study, a longitudinal study, or a repeated measures

study. A retrospective study could be case–control, nested case–control, or a study without any control group. **Ecological studies** also come under the umbrella of epidemiological studies. In addition, as just mentioned, descriptive studies and disease surveillance are also epidemiological.

Because of a lack of controlled conditions, the evidence available from epidemiological studies is considered indicative rather than definitive. Various biases can occur that could make any epidemiological study result suspect. In biostatistical terms, they provide evidence on associations but not cause–effect relationships. For concluding cause–effect, a large number of other rather strict conditions must be satisfied (see **cause–effect relationship**).

There are instances in the medical literature where trials are also called epidemiological studies. If that is so, almost any medical study would be epidemiological. We advise you to guard against such "generalizations." In fact, clinical trials and laboratory experiments are the most common example of nonepidemiological studies.

## epistemic uncertainties, *see also* aleatory uncertainties

Uncertainties arising from our limitations are called epistemic. Knowledge gaps are wider than generally perceived. One paradigm says that what we do not know is more than what we know. Unfamiliarity breeds uncertainty, and this is part of the epistemic uncertainties. A long way down from the art of healing based on esoteric knowledge, the realization of epistemic gaps in medicine is recent. In the context of health and medicine, this type of uncertainty was first highlighted by Indrayan [1] in 2008.

Besides incomplete knowledge, epistemic uncertainty also includes (i) ignorance, e.g., how to choose one treatment strategy when two or more are equally good or equally bad, such as between amoxicillin and cotrimoxazole in nonresponsive pneumonia; (ii) parameter uncertainty regarding the factors causing or contributing to a particular outcome, such as etiological factors of vaginal and vulvar cancer; (iii) speculation about unobserved values, such as the effect of high levels of $NO_2$ in the atmosphere; (iv) nonavailability of the exact quantitative effect of various factors, e.g., diet, exercise, obesity, and stress on raising blood glucose level; and (v) confusion about the definition of various health conditions, such as hypertension—whether the blood pressure cutoff should be 130/85 or 140/90 mmHg.

Another kind of epistemic uncertainty arises from nonavailability of the proper instrument. How do you measure blood loss during a surgical operation? Swabs that are used to suck blood are not standardized. In some surgeries, blood can even spill onto the floor. Even a simple parameter such as pain is difficult to measure. A visual analog scale (VAS) and other instruments for this are just approximations. Stress defies measurement, and behavior/opinion-type variables present stiff difficulties. If the measurement is tentative, naturally, the conclusion, too, is tentative. The following is a brief description of some sources of epistemic uncertainties.

### *Inadequate Knowledge*

Notwithstanding claims of far-reaching advances in medical sciences, many features of the human body and mind, and their interaction with the environment, are not sufficiently well known. How the mind controls physiological and biochemical mechanisms is an area of current research. What specific psychosomatic factors cause women to live longer than men is still shrouded in mystery. Nobody knows yet how to reverse hypertension, which can obviate the dependence on drugs. Cancers are treated by radiotherapy or excision because a procedure to regenerate aberrant cells is not known. Treatment for urinary tract infections in patients with impaired renal function is not known. Such gaps in knowledge naturally add to the spectrum of uncertainty.

The preceding paragraph discusses universal epistemic gaps. In addition to these is the incomplete knowledge of a particular physician. This can arise at two levels. First is that the physician does not know enough, although medical science knows. Second is that the physician knows but is not able to recollect at the time of facing a patient. Both can result in misdiagnosis or missed diagnosis, or improper prescriptions.



### *Incomplete Information on the Patient*

When a patient arrives in a coma at the casualty department of a hospital, first steps for management are often taken without considering the medical history of the patient or without waiting for laboratory investigations. An angiography may be highly indicated for a cardiac patient, but initial treatment decisions are taken in its absence if the facility is not available in that health center. Even while interviewing a healthy person, it cannot be ensured that the person is not forgetting or intentionally suppressing some information. Suppression can easily happen in the case of sexually transmitted diseases (STDs) because of the stigma attached to them. An uneducated person may even fail to understand the questions or may misinterpret them. Some investigations such as computed tomography (CT) and magnetic resonance imaging (MRI) are expensive, and lack of funds may sometimes lead to proceeding without these investigations even when they are highly indicated. Thus, information remains incomplete in many cases despite best efforts. Clinicians are often required to make a decision about treatment based on such incomplete information.

### *Imperfect Tools*

A clinician uses various tools during the course of practice. Examples are signs–symptoms syndrome, physical measurements, laboratory and radiological investigations, and intervention in the form of medical treatment or surgery. Besides his/her own skills in optimally using what is available, the efficiency of a clinician depends on the validity and reliability of the tools he/she uses. Validity refers to the ability to measure correctly what a tool is supposed to measure, and reliability means consistency in repeated use. Sensitivity, specificity, and predictivities are calculated to assess the validity of a tool. Reliability is evaluated in terms of measures such as **Cohen kappa** and **Cronbach alpha**. In practice, no medical tool is 100% perfect, so much so that even a CT scan can give a false-negative or false-positive result. A negative histologic result for a specimen is no guarantee that proliferation is absent, although in this case, positive predictivity is nearly 100%. The values of measurements such as creatinine level, platelet count, and total lung capacity are indicative rather than absolute, i.e., they mostly estimate the *likelihood* of a disease, not establish or deny its existence. Signs and symptoms seldom provide infallible evidence. Because all these tools are imperfect, decisions based on them are also necessarily probabilistic rather than definitive.

### *Chance Variability*

Let us go a little deeper into the factors already listed. Aging is a natural process, but its effect is more severe in some than in others. When exposed equally to heavy smoking for a long duration, some people develop lung cancer, and others do not. Despite consuming the same water deficient in iodine, some people do not develop goiter, whereas some do—that, too, of varying degrees. The incubation

period differs greatly from person to person after the same exposure. Part of such variation can be traced to factors such as personality traits, lifestyle, nutritional status, and genetic predisposition, but these known factors fail to explain the entire variation. Two patients apparently similar, not just with regard to disease condition but also for all other known factors, can respond differently to the same treatment regimen. Even susceptibility levels sometimes fail to account for all the variation. The unknown factors are called **chance**. Sometimes, the known factors that are too complex to comprehend or too many to be individually considered are also included in the chance syndrome. In some situations, chance factors could be very prominent contributors to uncertainties, and in some situations, they can be minor.

### *Epistemic Gaps in Research Results*

Most medical research is an attempt to fill in epistemic gaps. Descriptive studies tell us prevalence rates of health and disease in various segments of the population and their trend, which otherwise would not be known. The objective of analytical studies is to find antecedent–outcome relationships. However, realize that only those factors that are known or suspected to affect the outcome. Others are excluded or included as independents in the study. For example, blood group could be a contributory factor for a particular health condition but will not be included till such time that some evidence comes forth implying its possible role. Including limited factors in the study is pragmatic too, as nobody can include all factors. Genomic information is not included, as it is rarely known for the subjects of research. Thus, research remains incomplete, and we wonder why results are not applicable in practical conditions. The search for truth does not relent, although goal posts are continuously shifted upward as the new results appear.

Statistical models are always developed under certain conditions. First is the limited number of independent variables and their choice, which we have mentioned in the preceding paragraph. The most commonly ignored limitation is the linearity of the effect of the factors on the outcome. This is used for simplicity as the study of curvature is not only intricate, but also, the model loses its parsimony. After all, the purpose of generating models is to explain the phenomenon in an easily understood manner in the hope that the left-out portion is not high enough to cause much damage to the explanation. In some cases, this turns out to be too much to expect. Second is the measurements. Statistical methods assume that the values of each variable are exactly known without any error. This is hardly ever true. Thirdly, it is also assumed that the measurements available are indeed valid markers for the phenomenon under study. For malaria, you may include palpable spleen without considering its false positivity and false negativity. S large number of such epistemic gaps can be cited.

Some of the impact of epistemic uncertainties can be studied by **sensitivity analyses**. These consider alternative scenarios and see how the results are affected. The big question is what, if anything, can we do to reduce the impact of epistemic uncertainties on our medical decisions. Partial answers are provided by tools such as **etiology diagrams**, **expert systems**, and **scoring systems**. However, these tackle some aspects of epistemic uncertainties and not others. The solution for others is research and to realize that we are in imperfect world.

The counterpart of epistemic uncertainties is **aleatory uncertainties**. As described under that topic, aleatory uncertainties are mostly due to biologic, environmental, instrumental, and other variations, and to biases and errors. Biostatistics seems to be the best science to deal with aleatory uncertainties. For a discussion of aleatory and epistemic uncertainties, see the work of Sandomeer [2], although this is in the context of engineering applications.

1. Indrayan A. *Medical Biostatistics*, Second Edition. Chapman & Hall/CRC Press, 2008.
2. Sandomeer MK. Aleatoric or epistemic? Does it matter? Swiss Federal Institute of Technology Zurich. http://www.ibk.ethz.ch/emeritus/fa/education/Seminare/Seminar08/PhD_Seminar_Sandomeer.pdf, last accessed April 19, 2014

## equipoises

Equipoise is the balance among the regimens under study so that none has an initial advantage over the others. This term is used in the context of clinical trials. As you can appreciate, equipoise breeds uncertainty, and this is considered a moral prerequisite for conducting any clinical trial. Equipoise is espoused as the essence of the **uncertainty principle** under which clinical trials are conducted. Uncertainty does not imply equipoise, but equipoise implies uncertainty. Uncertainty has much wider applications, and equipoise is just one source of uncertainty.

Among various equipoises discussed in the context of clinical trials, medically, the most important is **clinical equipoise**. This is the collective uncertainty among clinicians about the efficacy of the regimen under trial. For example, this exists for the use of vasopressin for management of septic shock. Lilford [1] cites the example of amniocentesis and chronic villous sampling for such clinical equipoise, although in his opinion, the latter is twice as risky for miscarriage. These two are suitable candidates for comparison in a trial. Similar equipoise exists between stenting and endarterectomy for carotid restenosis. The optimal route (intramuscular versus subcutaneous) of administration of influenza and pneumococcal vaccines in elderly patients is also under debate. However, the window of clinical equipoise is generally small as it is grabbed up fast for research, and the results one way or the other are available sooner than later.

When the concept is stretched further, it means not only uncertainty but also that the two arms of a trial are likely to result in *equal* efficacy on an a priori basis. It connotes equal uncertainty for the positive and negative outcomes of the trial. Practically, though, genuine uncertainty is enough without pressing for equal uncertainty. Clinical equipoise is the condition under which clinicians as a group would not object to their patients participating, and patients may rationally accept randomization. This equipoise *also* insulates against prejudiced assessment of the patients by the investigators. Previous evidence of benefit of a treatment may be flawed but can disturb the equipoise [2]. Sometimes, a trial is terminated early when overwhelming evidence emerges and the equipoise is disturbed.

In addition to the regimen, the groups of subjects also should be such that there is a priori uncertainty about the efficacy of the test therapy in them. This is one of the criteria for selection of cases for a clinical trial. This is called *patient equipoise* and helps to ensure that the patients are homogenous. Patient equipoise implies guarding against an unwitting tendency to include subjects who are likely to benefit from the drug under trial without declaring that the trial is restricted to such specific groups. Sometimes, health-conscious people agree to enter a trial, while in fact, participants should be fair representatives of the class of patients that are finally targeted to benefit from the regimen in the case that the trial is successful. The participants should be uncertain about the outcome of the trial so that the results are unbiased. Primarily, this is a statistical requirement for psychological homogeneity of the subjects at baseline.

The third is the *personal equipoise* of the clinician so that he/she does not feel uncomfortable about his/her own views and about the patients. A particular clinician may have a strongly positive or a very bitter feeling about a regimen even though clinical equipoise in terms of collective uncertainty may exist. A clinician who is convinced that

one treatment is better than another for a particular patient cannot ethically agree to randomization. Personal equipoise may be difficult to achieve, but efforts can be made by discussing evidence regarding the underlying uncertainties and trying to convince the particular physician that equipoise indeed exists. The term can also be used for an individual patient if he/she is undecided on which treatment to choose.

Cheng et al. [2] have discussed these three kinds of equipoise in the context of a trial for melioidosis, but they explain the concepts very well. Cook and Sheets [3] have also provided some details.


1. Lilford RJ. Equipose is not synonymous with uncertainty (Letter). *BMJ* 2001;323:574. http://www.bmj.com/content/323/7312/574.3
2. Cheng AC, Lowe M, Stephens DP, Currie BJ. Ethical problems of evaluating a new treatment for melioidosis. *BMJ* 2003;327:1280–2. http://www.bmj.com/content/327/7426/1280
3. Cook C, Sheets C. Clinical equipoise and personal equipoise: Two necessary ingredients for reducing bias in manual therapy trials. *J Man Manip Ther* Feb 2011;19(1):55–7. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3172958/


## equivalence and noninferiority trials, *see also* equivalence, superiority, and noninferiority tests

Equivalence trials aim to show that the effects of interventions differ by no more than a medically unimportant specified margin. For efficacy in clinical trials, these trials consider the possibility of lesser efficacy as well as of higher efficacy. If the specified margin is 3% and if the existing therapy has 70% efficacy, for equivalence, the efficacy of test therapy should be between 67% and 73%. If the efficacy is either more than 73% or less than 67%, the test regimen is not equivalent. Statistically, this is assessed in terms of probability. The chances of efficacy outside 67–73% should be negligible. Equivalence trials need two-tailed statistical tests and confidence intervals as discussed under the topic **equivalence**, **superiority**, **and noninferiority tests**.

Equivalence trials need far more care than the usual comparative trials. The kind of incentive available in showing that a difference exists probably is not available in equivalence trials. Thus, equivalence trials may lack natural internal checks. Inclusion and exclusion criteria should be stricter in such trials; otherwise, something like concomitant medication can tilt results toward equivalence. Also, make sure that equivalence is not due to mischievous execution of the study, such as a lot of dropouts and noncompliance to the regimen by either group, particularly in the group with the standard regimen.

As just stated, equivalence of a regimen can be concluded when its efficacy is different by not more than a specified margin. This can have serious implications. If an existing regimen has 79% efficacy and if a 3% difference is your tolerance, a regimen with 76% efficacy can be considered equivalent. This, however, means that the standard can slip over time. Now a new regimen could be evaluated for 76% efficacy in place of 79%. You may want to guard against such a fallacy.

In a therapeutic setting, the interest generally is in noninferiority rather than equivalence. Also, if the existing regimen already has 95% efficacy, superiority can be a faraway dream—only noninferiority is practical. If a new regimen has the potential to be almost as effective as the existing regimen but is cheaper or more convenient, the interest would be in noninferiority. Noninferiority trials aim to show that the effect of the new regimen is not worse than the existing regimen by more than a specified margin, and therefore, the new regimen can be advocated. Statistically, they need one-tailed procedures just like **superiority trials**, although the critical tail in noninferiority is on the right side. For superiority, it is on the left side of the distribution. For a short and crisp discussion about noninferiority trials, see Snapinn [1]. He argues that noninferiority and superiority can be assessed in the same clinical trial without statistical penalty. For further details of the design and analysis of noninferiority trials, see Rothman et al. [2].


1. Snapinn SM. Noninferiority trials. *Curr Control Trials Cardiovasc Med* 2000;1:19–21. http://link.springer.com/article/10.1186%2Fcvm-1-1-019#page-1
2. Rothmann MD, Wiens BL, Chan ISF. *Design and Analysis of Noninferiority Trials*. Chapman & Hall/CRC Press, 2011.




## equivalences (types of) in clinical trials, *see also* equivalence and noninferiority trials

Equivalence in clinical trials is similar performance of different regimens under consideration in identical conditions. Performance can be measured in a variety of ways, and this gives rise to terms such as therapeutic equivalence and bioequivalence. These equivalences are different from group equivalences at baseline, which we try to achieve through randomization and matching.

**Therapeutic equivalence** considers only the success rate as the end point, such as efficacy. Many therapeutic equivalence trials are done for an alterative dose schedule or method of administration. For example, alendronate 35 mg once weekly may be therapeutically equivalent to 5 mg daily for prevention of osteoporosis. Innovative and generic formulations of beclomethasone dipropionate may be therapeutically equivalent in adult patients with moderate to severe asthma. Both give nearly the same efficacy, i.e., the efficacy does not differ by more than a clinically unimportant margin. Baker [1] argues that different mesalamine products may not be therapeutically equivalent, because they differ with regard to where the drug is released in the intestinal tract, and this may affect the outcome. Therapeutic equivalence is also called **clinical equivalence**. Note that this is equivalence on average and not individual equivalence.

**Bioequivalence** considers the entire course of the absorption process of the drug rather than just the end point. When the course of the disease or the improvement pattern over a period of time is the same for two regimens, they are considered bioequivalent. In the context of drug trials, this requires pharmacological studies, and the comparison may be in terms of peak concentration, time to reach their peak, half-life, area under the curve, etc. These are called **pharmacokinetic parameters**. Popovic et al. [2] studied spline function for the bioequivalence of a single 240 mg orally standard retard tablet of verapamil with single 5 mg intravenous dose. Bioequivalence studies can certainly be done on sick subjects using conventional designs but can be done on healthy subjects as well for some regimens using crossover designs. For statistical details of bioequivalence, consult Patterson and James [3].

Many consider bioequivalence to imply therapeutic equivalence. This really may be so in most cases. However, the reverse is obviously not true. A distinction should also be made between bioequivalence at the individual patient level and average bioequivalence in groups of patients. If the two regimens under comparison produce different responses in individuals, this is ignored in average bioequivalence, whereas in individual bioequivalence, this interaction is an important consideration. *Average bioequivalence* would imply that either of these regimes can be prescribed to newly incoming patients. Individual bioequivalence would mean that the patient can be switched from one regimen to the other in the midst of the ongoing treatment. For example, drug switchability can be from a brand-name drug to a generic drug while the patient is still under treatment, say for cost considerations.

Related terms are *equivalence trials*, *superiority trials*, and *noninferiority trials*. These are separately discussed in this volume under the topic **equivalence and noninferiority trials**. Statistical tests for these are discussed under the topic **equivalence**, **superiority**, **and noninferiority tests**.


1. Baker DE. Therapeutic equivalence of mesalamine products. *Rev Gastroenterol Disord* 2004;4:25–8. http://www.ncbi.nlm.nih.gov/pubmed/15029108
2. Popovic J, Mitic R, Sabo A, Mikov M, Jakovljevic V, Dakovic-Svajcer K. Spline functions in convolutional modeling of verapamil bioavailability and bioequivalence II: Study in healthy volunteers. *Eur J Drug MetabPharmacokinet* 2006;31:87–96. http://link.springer.com/article/10.1007%2FBF03191124#page-1
3. Patterson S, James B. *Bioequivalence and Statistics in Clinical Pharmacology*. CRC Press, Boca Raton, FL, 2006.


## equivalence, superiority, and noninferiority tests

Equivalence tests are statistical procedures to find whether two regimens under trial are equivalent or not in their performance on average. Thus, these are for equivalence for prescription and not switching one treatment to the other midcourse. Such switching requires individual equivalence, which we are not considering here. Between the two types of **equivalences in clinical trials**, in this section, we mostly concentrate on therapeutic equivalence and later briefly mention **bioequivalence** also, which is for the entire course of the treatment. The thrust here is to test whether two groups are essentially equivalent with respect to a particular end point and the difference, if any, is of no medical consequence. Superiority and noninferiority are also discussed. To understand this topic, you should be familiar with the usual statistical **tests of hypothesis** and **confidence intervals**. The following procedures are valid under **Gaussian conditions** for both means and proportions.

### *Superiority, Equivalence, and Noninferiority*

One regimen can be considered medically equivalent to another regimen when the difference in efficacy and side effects between the two does not exceed a prespecified medically unimportant margin (see **medically important effect**). The margin could be in terms of percentage or in terms of means. Group 1 is superior to group 2 when the response in group 1 is higher by at least the specified margin, and group 1 is noninferior to group 2 when the response is lower by not more than the specified margin. This is illustrated in Figure E.5. This assumes that one is a test group and the other is a reference group. Both the groups are under trial. Horizontal bars are the 100(1 − 2α)% confidence intervals (CIs) for difference in efficacy in 10 different experiments, and the solid dot in the middle of these bars is the **point estimate**. Alpha (α) is the **level of significance**, and Δ is the prespecified unimportant margin. In this figure, experiments A and H provide unambiguous results—A has statistical significance and superiority, and H has statistical significance and inferiority. In experiment E, the interval contains 0, and the entire interval is within the limits of medical indifference. Thus, the difference is neither statistically significant nor medically significant. In experiments D and F, the interval does not contain 0, so that the difference is statistically significant but the interval again lies entirely within the limits of medical indifference; thus, the results are medically equivalent. These are the examples that illustrate that statistical significance does not imply medical significance. In experiments I and J, the CIs overlap with the limits of equivalence. In these two cases, since the CI is relatively large, a restudy with a larger sample is advisable.

Comments on the left side of Figure E.5 assume that the same margin Δ can be used for noninferiority, equivalence, and superiority, and also assume that higher values are in favor of the test regimen. A trial gives evidence of superiority of the test regimen when the lower limit of the CI for $(\pi_{test} - \pi_{ref})$ exceeds $+\Delta$, where $\pi$ is the efficacy; of noninferiority when the lower limit of the CI is

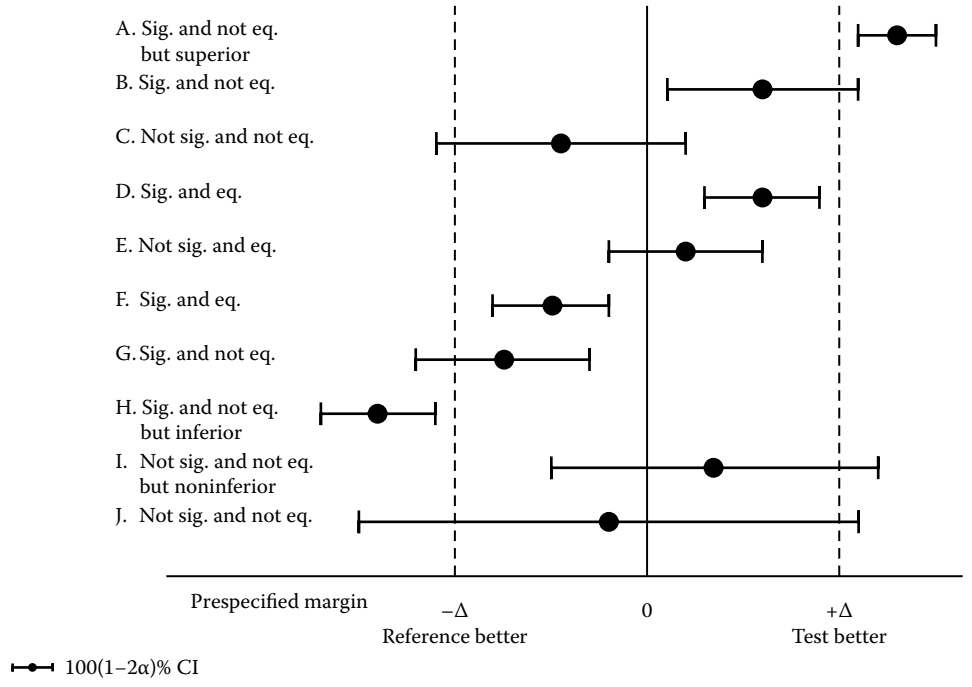



**FIGURE E.5** Statistical significance and medical equivalence. Eq., equivalent; sig., statistically significant.

more than $-\Delta$; and of equivalence when the entire CI is between $-\Delta$ and $+\Delta$. In case the efficacy of the reference group is already fairly known, there is no need to include that in the trial. In that case, the CI will be based on one-sample $100(1 - 2\alpha)\%$ confidence level and not on the difference. This may happen particularly in noninferiority trials since noninferiority is mostly tested against a known standard.

You can easily see that superiority or noninferiority is tested by specifying the medically relevant difference and conducting the usual one-tailed test. However, equivalence within a specified margin needs a little more consideration.

### *Equivalence*

For equivalence, the margin is specified for either side, generally saying that the difference does not exceed $\varepsilon\pi$, where $\varepsilon$ is the proportion of $\pi$ for equivalence. If the proportion of success in one group is $\pi = 0.50$ and 4% of $\pi$ on either side is considered to arise due to natural factors, then $\varepsilon = 0.04$ and $\varepsilon\pi = 0.04 \times 0.50 = 0.02$. The limits of tolerance for equivalence thus are 0.48 and 0.52. This is for therapeutic or clinical equivalence and not for bioequivalence.

In equivalence testing, the null hypothesis has a reverse nature: the difference is a priori specified $\Delta$ or more. Rejecting this null would mean equivalence, and the burden of rejecting the null is on the data as before. **Type I error** in this case is concluding that the difference is less when, in fact, it is $\Delta$ or more, and **Type II error** is concluding that the difference is at least $\Delta$ when, in fact, it is less. These are a kind of inverse of the conventional setups.

An approach to test equivalence is to perform **two one-sided tests (TOSTs)**. But an easier procedure is to obtain the CI for the difference between two groups at $100(1 - 2\alpha)\%$ confidence level instead of the usual $100(1 - \alpha)\%$ level. Since the difference on either side is admissible in an equivalence setup, the limits for confidence level $100(1 - 2\alpha)\%$ provide the lower and upper bound at $100(1 - \alpha)\%$ confidence. If this CI contains 0, the difference is not statistically significant at $2\alpha\%$ level of significance. If this interval is completely contained in $-\Delta$ to $+\Delta$, equivalence is concluded. Any difference less than $\Delta$ is considered trivial and within medical indifference. In exceptional cases, the difference for the lower bound can be different than for the upper bound. Thus, $\Delta_1$ and $\Delta_2$ can be separately specified. There is a convention to consider efficacy between 80% and 125% of the expected as the equivalence margin, i.e., a margin of one-fifth on either side. In some situations, this may be too wide for clinical decisions. Sample sizes with this threshold are provided by Liu and Chow [1].

Under this procedure, the difference between two proportions may be statistically significant, yet the two could be equivalent in terms of the difference not exceeding the medical tolerance. The reverse can also happen. Two proportions may not be statistically significantly different yet not equivalent.

The main difficulty in equivalence tests is not the statistical procedure but the specification of tolerance $\Delta$. This is primarily a medical decision, where biostatistics plays a secondary role. This limit depends on how much latitude can be given without compromising patient management and what variation is expected due to unforeseen and unknown factors, such as in obtaining exact measurements. By its very nature, this has to be fairly small and trivial so that it does not alter patient management. This is what makes two regimens clinically indistinguishable. Identifying such a limit could be a challenge in some situations and can force one to be inexact. See also our comments on determining the noninferiority margin later in this section.

Typically, tests for establishing equivalence need a larger $n$ than tests for establishing difference. This is because the clinically unimportant $\Delta$ chosen for equivalence is generally much smaller than the target clinically important difference in the usual comparative trials, and the formula also is slightly different.

There are other *if*s and *but*s attached to equivalence: (i) equivalence can be achieved also when both treatments are ineffective—thus, the reference under comparison must have established efficacy; (ii) unless special care is taken, equivalence in efficacy is oblivious of differences in patient compliance, side effects, and losses, which can be greater with one regimen than the other; (iii) if equivalence is found for a particular dose, it may not carry over to the other doses; and (iv) if both regimens have high efficacy, the difference could be masked, and equivalence can be fallacious.

Equivalence trials are affected much more than the usual comparative trials by violation of the **protocol**. Equivalence should be concluded only when both the per-protocol analysis and the **intention-to-treat analysis** reach to the same conclusion.

The following example illustrates equivalence testing in a simple situation. Wolf et al. [2] conducted an equivalence study in the United Kingdom comparing clonidine and midazolam as intravenous sedative agents in critically ill children. They had specified the dosages. The end point was adequate sedation at least 80% of the time. The margin of equivalence was prefixed at ±0.15. The difference in efficacy was 0.04 (clonidine had higher efficacy), and the 95% CI for the difference in efficacy was −0.13 to +0.21. Since the upper limit exceeds the equivalence margin, the two drugs do not provide equivalent sedation. Since the lower limit is within the equivalence range and upper limit higher, the authors took this to suggest that clonidine is noninferior (noninferiority $P$-value = 0.01) and possibly superior (superiority $P$-value = 0.10) to midazolam.

### *Determining Noninferiority Margin*

Noninferiority trials are done to find whether the test regimen can be as good as the one with established efficacy. The procedure for noninferiority testing is essentially the same as for equivalence trials, but the lower limit of the 95% CI should be more than $-\Delta$ in this case, and the upper limit can be any value (experiment I in Figure E.5). Interpret noninferiority with caution. It does not mean that the regimen is not inferior—only that it is not worse by more than the prespecified clinically unimportant margin. That is, any loss of efficacy has no clinical relevance. Such a regimen may have other benefits, such as convenience, cost, and safety. There are other implications too. If the clinically unimportant margin is 2%, a regimen with 86% efficacy is noninferior to one with 88% efficacy, and one with 84% efficacy is noninferior to one with 86% efficacy. Thus, the standard can progressively slip down. Thus, it is important to be judicious in choosing $\Delta$, especially for noninferiority trials.

The noninferiority margin is specified in advance and must always be justified for the specific regimen and disease you are considering. This margin does not depend on the size of the trial nor on the statistical power of the study. It primarily depends on natural variability in the difference between efficacies of the test and reference regimen that can arise from trial to trial. It also depends on what sacrifice in efficacy can be made in exchange for lower cost or fewer side effects that the regimen under test may have. A review of literature or experience might suggest what difference is clinically irrelevant. The following may also help in deciding on the noninferiority margin:

- Survey the practitioners who deal with that disease and find the range they consider unimportant considering other advantages of the regimen under investigation. This will take care of cost, convenience, acceptability, etc.
- If there are many regimens that are interchangeably used at present for treating the same condition, the difference

in their efficacies as reported in the literature can give a fairly good idea of what can be a clinically unimportant margin.
- Consider the natural variation you expect due to unavoidable errors in measurement and other assessments.
- If there is a definite safety advantage, a larger Δ can be chosen.

If the outcome of interest is death, it could be ethically difficult to specify clinically unimportant deaths. You may want to plan a superiority trial in this case, possibly with a relaxed significance level.

We have discussed equivalence and related tests for efficacy, which is measured in terms of proportion or percentage. Most clinical trials will require this kind of equivalence. However, in some rare cases, the outcome can be quantitative, such as the mean reduction in glycated hemoglobin (HbA1C) level by two different regimens. One regimen may be based on drugs and the other based on changes in lifestyle. The basic procedure remains the same, namely, finding the CI for a difference in differences in this case and examining if the lower and upper limits reach the specified threshold for medically unimportant margin. This is mostly based on Gaussian approximation and requires that the differences follow a **Gaussian distribution** at least approximately or that the sample size be large enough for the **central limit theorem** to operate.

In the case of bioequivalence, the parameter under consideration is mostly the **area under the curve**, although this is not a perfect measure. This parameter is quantitative, and the test can be done using the procedure just outlined in the preceding paragraph.

1. Liu JP, Chow SC. Sample size determination for the two one-sided tests procedure in bioequivalence. *J Pharmacokinet Biopharm.* 1992 Feb;20(1):101–4.
2. Wolf A, McKay A, Spowart C, Granville H, Boland A, Petrou S, Sutherland A, Gamble C. Prospective multicentre randomised, double-blind, equivalence study comparing clonidine and midazolam as intravenous sedative agents in critically ill children: The SLEEPS (Safety profiLe, Efficacy and Equivalence in Paediatric intensive care Sedation) study. *Health Technol Assess* 2014 Dec;18(71):1–212. http://www.journalslibrary.nihr.ac.uk/hta/volume-18/issue-71#hometab0

## errors (statistical), *see* measurement errors, Type I error, Type II error

## error sum of squares

Error in the context of error sum of squares (ESS) is the difference between the observed value and the value estimated or **predicted** by any model. The observed value for the *i*th subject is denoted by $y_i$ and the predicted value by $\hat{y}_i$. Thus,

$$\text{Error sum of squares: ESS} = \Sigma_i (y_i - \hat{y}_i)^2.$$

Note again that this is based on differences of the predicted from the observed value, whereas the usual sum of squares is based on the difference from the mean. ESS is a measure of the overall goodness of fit of the model. High ESS indicates worse fit relative to the model with low ESS. Many models, particularly **regression** models, are fitted in a manner such that ESS is the minimum. That the minimum may still be large is another issue. We explain this later in this section with the help of an example.

ESS arises in almost all statistical models. Models, by definition, are simplified versions of the intricate biological process—thus, the predicted value will seldom match exactly with the observed value. Errors will occur. Most models are fitted in a manner such that errors for some subjects are positive, for others negative, and their sum is 0. Thus, their sum fails to quantify the overall error. In place of the sum of their absolute value, it is mathematically easier to work with squares. ESS is the sum of these squares. This is also called **residual sum of squares** since the error $(y_i - \hat{y}_i)$ is also called residual, particularly in the context of the sample values.

ESS is easy to understand with the help of a simple example. Extension to complex models is immediate. Consider a study on total bone mineral density (BMD) in adult males affected by age (years) and serum C-peptide level (nmol/L). Let the **regression equation** be

$$\text{total bone mineral density (g/cm}^2\text{) } \hat{y} = 1.98869 - 0.28652*\text{serum C peptide} - 0.01155*\text{age}. \quad \text{(E.1)}$$

The observed values of total BMD for nine persons is in the first column of Table E.1, their serum C-peptide level is in column 2, and age is in column 3. When these values are substituted in the regression equation, the predicted values of total BMD are in column 4. The difference between the observed and the predicted value is the error, which is in column 5, and the square of this error is in column 6. The sum of these squares is the ESS, which is at the bottom of column 6. In this example, ESS = 0.057. You can see that the ESS will be high when the predicted values are very much different from the observed values. Thus, ESS measures the goodness of fit. However, ESS increases without bound as the number of observations increases—thus, a better measure is the **error variance**, where ESS is divided by its degrees of freedom (df's), which is based on *n*.

ESS is so important in modeling that the entire **least squares method** is based on ESS. This method finds estimates of the parameters (such as the **regression coefficients**) of the model such that the ESS is minimum. Least squares is the most popular method for this purpose. The regression coefficients in Equation E.1 are also based on the least squares method. You may want to try other values of the regression coefficients in Equation E.1 and see for yourself that ESS will increase. However, ESS will decrease if you include more predictors, say, body mass index (BMI) and creatinine level. ESS can also decrease if the form of regression is not restricted to

**TABLE E.1**
**Calculation of Error Sum of Squares**

| Total BMD ($y$) | Serum C-Peptide | Age | Predicted Value ($\hat{y}$) | Error ($y - \hat{y}$) | Error Square |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 |
| 1.32 | 0.83 | 32 | 1.3811 | −0.06108 | 0.003731 |
| 1.45 | 0.72 | 45 | 1.2624 | 0.18763 | 0.035205 |
| 1.15 | 0.82 | 56 | 1.1066 | 0.04340 | 0.001884 |
| 1.32 | 0.75 | 37 | 1.3462 | −0.02622 | 0.000687 |
| 0.99 | 0.64 | 63 | 1.0773 | −0.08728 | 0.007618 |
| 1.34 | 0.87 | 38 | 1.3003 | 0.03972 | 0.001578 |
| 1.43 | 0.95 | 26 | 1.4160 | 0.01397 | 0.000195 |
| 1.29 | 0.98 | 31 | 1.3497 | −0.05966 | 0.003559 |
| 1.18 | 0.67 | 49 | 1.2305 | −0.05047 | 0.002547 |
| | | | **Sum** | **0** | **0.057004** |

linear terms and is extended to include curvilinear and nonlinear terms. Such changes can indeed provide a better model in terms of better prediction, although there are two problems: (i) when more predictors or nonlinear terms are included, the parsimony is lost, and the model becomes difficult to understand, and (ii) one has to consider the statistical significance of "improved" prediction by such changes.

ESS is one of the several sums of squares used in statistics. Others are regression sum of squares, between-groups sum of squares, within-groups sum of squares, sum of squares due to a particular predictor, Type III sum of squares, etc. All these are explained under the topic **sum of squares**.

## error variance

Error variance is the **error sum of squares** (ESS) for a statistical model divided by the corresponding **degrees of freedom** (df's). ESS is the sum of squares of differences of model-based **predicted values** from the corresponding observed values. This measures how bad the model is at predicting values close to the observed values or how much of the total variation the model has failed to account for. The smaller the error variance, the better the model for prediction. Another popular name for error variance is the **mean square error (MSE)**.

The square of any real difference can never be negative and will be more than 0 for most values. For this reason, ESS increases without bound as the number of observations increases. ESS for 8 values is always more than or equal to ESS for 7 values, and ESS for 20 values is always at least as much as ESS for 19 values. Thus, a meaningful measure is obtained when it is divided by the number of observations. Since predicted values use the data to estimate the parameters of the model, such as regression coefficients, the number of independent observations is reduced to the df. Thus,

$$\text{error variance} = \frac{\Sigma(y - \hat{y})^2}{df}.$$

This, in fact, is the sample estimate of the error variance. The numerator is the ESS and explained under the topic **error sum of squares**, including the method of computation in a simple case. Two crucial differences between error variance and the usual variance are that (i) the error variance is based on the deviation from the predicted values, whereas the usual variance is based on the deviation from the mean; and (ii) the denominator in the error variance is the df's, and the denominator in the usual (sample) variance is $(n - 1)$. Because of these two differences, error variance varies from model to model, but the usual variance remains the same.

Error variance (or, call it MSE) has some very useful applications. Among others, it is used in constructing the **confidence interval** for the predicted value and for testing of the hypothesis on the **regression coefficients**.

## estimate

In statistics, an estimate is a sample summary value that can be used for an unknown value in the population. The sample mean $\bar{x}$ is an estimate of the population mean $\mu$, and sample standard deviation (SD) $s$ is an estimate of population SD $\sigma$. Simply stated, sample summaries are called statistics (as plural), and population summaries are called **parameters**. Whether these estimates are good or not is a different issue, and we will come to this slightly later in this section.

Two kinds of estimates are used in statistics. First is the point estimate, and second is the interval estimate. **Point estimate** is a single value, whereas **interval estimate** is the range within which the parameter value is likely to lie. The second is popularly called **confidence interval** and is discussed under that topic. We will not discuss it here. This section is restricted to the point estimate.

There is another term, *estimator*. This term is used when we use notations and not values. Although we mentioned that $\bar{x}$ is an estimate of the population mean $\mu$, actually, it is an estimator. *Estimator* is a generic term, whereas *estimate* is specific to a particular sample. When the value is used, an estimator becomes an estimate.

The term *estimate* itself signifies that there is uncertainty around it. Then what is a good estimate? When talking of quality, we turn to an estimator instead of an estimate so that we can talk in general terms.

The most important property of a good estimator is that its average over all possible samples of that kind is the parameter itself. In statistical terms, this average is called the *expected value*—expected in the sense of long-term perspective. For most estimators, the expected value can be theoretically obtained and examined whether or not it is the same as the parameter. When the expected value is the same as the concerned parameter, it is called an **unbiased estimator**, signifying that this has no bias in the long run. For example, $\bar{x}$ is an unbiased estimator of the population mean $\mu$ for almost any distribution provided that the sample is randomly drawn. It does not have to be Gaussian. This also underscores the importance of random sampling. On the other hand, the sample variance $s^2$ is an unbiased estimator of population variance $\sigma^2$ only when the divisor is $(n - 1)$. If the divisor is $n$, it is biased. That is the reason that $(n - 1)$ is used for sample variance. Also, beware that sample SD $s$ is not an unbiased estimator of $\sigma$ but is still accepted because the variance is considered the actual parameter, and not the SD. By the way, the sample median is not an unbiased estimator of population median if the distribution is **skewed**.

Just as a variable has a distribution, so does the estimator. When a large number of samples are drawn, each will provide an estimate, and those values will have a distribution in the sense that some values will be high, some low, and most around the middle in many situations. These values will have a variance and SD, although this SD is now called **standard error (SE)**. The second property of a good estimator (the first is that it is unbiased) is that it has minimum variance among all possible estimators. When there is only one such estimator, it is called unique. The estimator that is unique, is unbiased, and has minimum variance is called the *unique minimum variance unbiased estimator* (UMVUE). The sample mean is the UMVUE of population mean $\mu$ for most populations.

As a medical professional, you really need not worry about such technicalities. Statisticians take care of these properties, such as advising us to use $(n - 1)$ in the denominator for sample variance. The statistical estimators we use in medical and health sciences generally have these properties.

## estimator (unbiased), *see* unbiased estimator

## ethics of clinical trials and medical research

Three cardinal ethical considerations in medical research are that (i) the subject must be made fully aware of his/her role in the research and the possible consequences should be clear to all the parties involved; (ii) all possible steps are taken to safeguard the interest of the subjects; and (iii) expected gains must far outweigh the cost in terms of time, resources, and inconvenience. These, in brief, can be described as follows.

### Informed Consent

All subjects under investigation, even if it is just questioning, should be accurately told about the purpose of the research; what they would be asked; what invasive procedure, if any, would be used, as for taking a blood sample; what the interventions are, if any; what harm can possibly occur to the subjects; what compensations are available; what the safeguards are; what alternatives are available; how much time it will take; what the competency of the investigators is; how the results would help the subject, the researcher, and society; freedom to leave midway in case of follow-up; etc. After this explanation, the subjects are expected to sign the consent if they agree to participate. The consent must be completely voluntary, with no pressure. Consent is required even for surveys because they, too, invade private information and encroach on time. In the case of children and differently abled persons who cannot decide for themselves, the consent can be provided by the guardian or the next of kin. In the case of community studies, the consent can come from the appropriate representatives. For further details, see Ref. [1].

Note that informed consent to participate in medical research is very different from the user agreement signed when you create an account (as on Google or at a bank). That can seldom be called informed consent.

Realize that informed consent can preselect a biased group. In the case of clinical trials, they may be those who are inclined to take risks or those that are hopeless cases. Some patients or some clinicians may have strong preference for a particular therapy, and they may refuse randomization. Some eligible patients may refuse to participate when they are told that they could be randomized to placebo, the existing therapy, or an untested therapy. Some may refuse because it is a trial and not treatment per se. Considerable efforts may be needed to keep such refusals at a minimum.

Informed consent is taken after explaining the researcher's commitments. But how to ensure that the commitments made will actually be translated into actions? Most important of these is safeguarding the interest of the subjects. For this, it is necessary to properly study the possible consequences of the study on the subjects and other participants, such as those handling infectious and hazardous material. Before starting the study, the complete mechanism should be in place to deal with each of such possible consequences. Also, plan to compensate the subjects for their time and inconvenience. In some cases, free laboratory investigations, free checkups, etc. could be adequate, but in some cases, additional financial or other kinds of compensation can be provided. However, the compensation should be proportional to the expected discomfort and not excessive, which could be frowned upon as coercive or as unnecessary inducement.

In the case of clinical trials, ethical issues also exist regarding recruiting and exposing subjects to the control regimen. Control can be either on the existing regimen or on a placebo. Since a new regimen is under trial, that itself is a testimony that the existing regimen is not the ideal. Thus, legitimate questions can be raised about exposing subjects to this regimen. Placebo, in any case, is advised in certain specific situations, as described under the topic **placebo**. For example, placebo rarely can be used on subjects who are sick. Historical controls obviate this problem but are advised for a disease that has a relatively stable natural history and about which the understanding of prognostic aspects has not changed. In surgical trials, sham surgery is sometimes used as a control. This may be unethical many times because it exposes a patient to surgical risks.

It is obligatory for institutions carrying out research on human beings to have a functional *ethics committee* with powers to approve or disapprove a research project depending upon meeting the ethical requirements. This committee has at least some members who do not have stakes in that research. Human rights activists and legal experts also are included. These days, no research results are accepted unless the research has gone through the process of vetting by the institutional ethics committee. The funding agencies also are expected to pass every research project through their own ethics committee. Ethics of research on animals is also expected to be examined by an animal ethics committee.

### Ethical Cautions in Clinical Trials

Some important considerations in the ethics of conducting *a trial* on subjects, some of whom could be sick, can be reiterated as follows:

- Research should be carried out when sufficient reasons exist for that kind of research. This must be preceded by enough groundwork.
- Is the treatment regimen under test reasonably safe?
- Is there sufficient information that the treatment is likely to be beneficial?
- Have the subjects been informed about the potential benefits and possible side effects, and their consent obtained?
- Is it ethical to use a placebo on some subjects who are sick?
- Is it ethical to allocate the subjects randomly among various groups to receive different treatments? Are sufficient precautions built in to take immediate action in case an adverse reaction develops?
- Is it proper for a trial to be blind in any way?

The basic theme of all these considerations is that the science cannot compromise the interest of the individual subjects without taking them into full confidence.

It is now widely realized that researchers have an ethical obligation to make the full results of human research public. Several innovations are quickly occurring to meet this need. The World Health Organization is leading an international effort to promote registration of clinical trials at the time of initiation. The International Committee of Medical Journal Editors (ICMJE) has issued its own registration requirements [2]. These requirements include submission of the complete protocol, which would make it obligatory not just to publish but also not to miss out on any inconvenient findings.

### Biostatistical Ethics for Clinical Trials

Trials have the potential to harm the participants and society since an unproven modality or intervention is used, although the perception may be that it is beneficial. In an eagerness to complete the trial soon after problems surface, the findings may become statistically biased. Then the entire trial is questionable. When undetected inappropriate biostatistical methods are used, erroneous results are accepted, and this can cause irreversible damage to the patients and society. This can cause avoidable deaths of a number of people. An illustration of this is given by Baigent et al. [3]. Surgeons, for example, are trained for years, and even one death by mistake can be fought for millions in compensation. Biostatisticians are not so well trained—many studies are available that show that inappropriate methods have been used, and to save them from damage claims, they take umbrage under the probabilistic nature of the conclusions. Awareness is increasing to make biostatisticians accountable for what they advise or do. The following precautions can be advised.

First and foremost is equipoise. This includes clinical equipoise, patient equipoise, and personal equipoise. All these are

discussed under the topic **equipoise**. This is to ensure that there is no bias in conducting the trial and in the results. As mentioned earlier, the second is proper use of biostatistical methods. The entire effort of conducting a well-designed trial can go waste, including inconvenience to the patients, if the analytical methods are substandard. This, in fact, amounts to misuse of resources. Third is conducting a study on an adequate sample size. A smaller sample could mean wastage of resources since no reliable results can be obtained, and an unduly large sample also means wastage of part of the resources since results with adequate reliability could be obtained with a smaller sample. In the context of clinical trials, ethical issues are attached to interim appraisals also. These appraisals can help to stop the trial early either due to futility or due of availability of convincing results on efficacy early in the trial. Many trials do not follow this, although the trend is positive. Among various adaptations that can be undertaken midway, statistically most relevant is the **reestimation of the sample size** on the basis of the actual effect size found at interim stages. Reestimation requires intricate statistical inputs as this is done to preserve the level of significance and the power. Such adaptation does not cause much of ethical problems; rather, it seems to enhance ethics by keeping a provision to stop the trial early in case convincing evidence of efficacy or of futility appears. Care is taken that such appraisal does not undermine the integrity and validity of the trial.


1. Committee on Ethics. Informed consent. http://www.acog.org/Resources_And_Publications/Committee_Opinions/Committee_on_Ethics/Informed_Consent
2. ICMJE. *Clinical Trial Registration*. http://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html
3. Baigent C, Collins R, Peto R. Article makes simple errors and could cause unnecessary deaths. *BMJ* 2002 Jan 19;324(7330):167. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1122076/


## etiologic factors, *see also* causal diagrams

Etiologic factors are those that point to the origin. They are generally a set of factors or collection of factors that work together in tandem to produce a particular effect. In the context of health and medicine, they together are the contributors to a particular health outcome and can range from prenatal factors to current stresses in life. It is generally believed that a health condition is a by-product of host–agent–environment interaction. Host factors pertain to the affected person, such as genetic makeup, age, sex, parity, blood group, nutrition status, physiological and biochemical measurements, radiological investigations, and socioeconomic condition. Agent factors pertain to the causes, such as infections, injury, and side effects of regimens. The environment factors work as catalysts, such as water, pollution, traffic, availability of food, distribution mechanisms, health infrastructure, and social infrastructure.

Clearly, each health condition has its own etiologic factors. For example, Bektas-Keyhan et al. [1] found alcohol abuse, family history of cancer, smoking, and chronic mechanical traumas to be significant etiologic factors of tongue cancer, and Guagnozzi and Lucendo [2] discussed iron deficiency, vitamin $B_{12}$, and folic acid deficiencies, along with the effects of proinflammatory cytokines, hemolysis, drug therapies, and myelosuppression, as etiological factors in inflammatory bowel disease.

Identification of etiologic factors sometimes requires sustained investigations on patients in different setups, and statistical methods such as **regression analysis** are used to select a few significant ones out of many suspected factors. This kind of analysis also helps in exactly quantifying the effect on the outcome—what percentage is contributed to the "causation" by each. However, it may not be easy to attribute them as causes unless the much more restricted conditions mentioned for a **cause–effect relationship** are also satisfied. Perhaps the choice of variables to begin with and their correct assessment is the key to the identification of etiologic factors. Nonetheless, the etiology for a health condition may remain obscure till such time that we know enough about suspected contributors. This limitation falls under **epistemic uncertainties**: the etiology is still unknown for some diseases, such as fibromyalgia syndrome [3] and spontaneous coronary artery dissection [4].


1. Bektas-Kayhan K, Karagoz G, Kesimli MC, Karadeniz AN, Meral R, Altun M, Unur M. Carcinoma of the tongue: A case–control study on etiologic factors and dental trauma. *Asian Pac J Cancer Prev* 2014;15(5):2225–9. http://www.apocpcontrol.org/page/apjcp_issues_view.php?sid=Entrez:PubMed&id=pmid:24716961&key=2014.15.5.2225
2. Guagnozzi D, Lucendo AJ. Anemia in inflammatory bowel disease: A neglected issue with relevant effects. *World J Gastroenterol* 2014 Apr 7;20(13):3542–51. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3974521/
3. Okifuji A, Hare BD. Management of fibromyalgia syndrome: Review of evidence. *Pain Ther* 2013 Dec;2(2):87–104. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4107911/
4. Alfonso F, Bastante T, Rivero F, Cuesta J, Benedicto A, Saw J, Gulati R. Spontaneous coronary artery dissection. *Circ J* 2014 Aug 25;78(9):2099-110. https://www.jstage.jst.go.jp/article/circj/78/9/78_CJ-14-0773/_pdf


## etiologic fraction, *see* attributable risk (AR) fraction

## etiology diagrams, *see* causal diagrams

## Euclidean distance

Euclidean distance is just about the most popular measure of the difference between two observations. In its most simple form, the distance between a value $x$ and a value $y$ is $(x - y)$. If you have two sets each of $K$ values, namely, $(x_1, x_2, \ldots, x_K)$ and $(y_1, y_2, \ldots, y_K)$, the

$$\text{Euclidean distance} = \sqrt{\sum_{k=1}^{K} (x_K - y_K)^2} \ .$$

Note two features for the Euclidean distance. First, it is applicable to quantitative variables only. Second, the number of elements in each set must be the same. If one set has five values and the other set has eight values, Euclidean distance cannot be calculated in this format, although some variations are available. As you can see, this measures how values in one set are different from the values in the other set. This is believed to have been first proposed by Greek mathematician Euclid in the third century BC.

Euclidean distance has a big limitation. It is disproportionately affected by the units of measurement. Consider values of hemoglobin level (g/dL), body mass index (kg/m$^2$), total cholesterol level (mg/dL), and plasma creatinine level (mg/dL). Let

these be (12.4, 23.7, 189, 0.6) for one person and (13.6, 21.8, 211, 0.8) for the second person. The Euclidean distance between these two is $\sqrt{(12.4-13.6)^2+(23.7-21.8)^2+(189-211)^2+(0.6-0.8)^2}$ = $\sqrt{489.09}$ = 22.12. Out of 489.09, as much as 484 (square of 189 – 211) is contributed only by the difference in total cholesterol level. This highlights how one value can severely affect the Euclidean distance. Thus, it is desirable that the values are **standardized** before calculating Euclidean distance. For statistical applications, when a series of values are available for many subjects, standardized values are (value − mean)/SD, where SD is the standard deviation. Each value is standardized using its own mean and SD. When such values are used, we can get what is called *standardized Euclidean distance*. This helps to give equal weight to different values in the set, and the distance becomes scale invariant. Beside internal homogeneity, the standardized distance can be compared even when the units are different. For other variants of Euclidean distance, see Greenacre [1].

Euclidean distance is one of the many measures of distance, and indeed the most popular one. Among others are Mahalanobis distance, which uses a **dispersion matrix** also in computing the distance, and Minkowski distance, which uses the $p$th power and $p$th root. Euclidean distance is its special case for $p = 2$. For these and other distances, see Ref. [2].

All these are for quantitative data. For binary data and ordinal and nominal data, other kinds of measures can be used [2].


1. Greenacre M. *Correspondence Analysis and Related Methods*. Chapter 4—Measures of distance between samples: Euclidean http://www.econ.upf.edu/~michael/stanford/maeb4.pdf
2. MathWorks: *Documentation Center, pdist2*. http://www.mathworks.in/help/stats/pdist2.html


## evaluation of health programs/systems

Evaluation of a health program is a systematic examination of its strengths and weaknesses, to identify the factors for its success or failure. The purpose is to find whether it is worth it to continue the program if it is ongoing and in what aspects it needs improvement. Thus, evaluation is not just the assessment of its present status against the initial objectives but, more importantly, the assessment of the contributors to its present status. The process of evaluation is expected to help put the program back on track if derailed and to provide guidance for improving it by developing better strategies. It also ensures that the conclusions are shared and acted upon. This can help in justifying further support and funding. In its comprehensive form, evaluation involves needs assessment, process evaluation, outcome measurement, and impact appraisal. For more details on meaning and steps in evaluation of health programs, see the Centers for Disease Control and Prevention (CDC) document in Ref. [1]. The World Health Organization (WHO) manual also provides the details, although in the context of road safety.

Successful evaluation engages stakeholders, gathers credible data from all parties involved (the recipients, the providers, and the executors), processes the data with immaculate care using the right methods, and reaches evidence-based conclusions. The input and output variables must be properly specified and adequately measured. You can see that the statistical methods play a crucial role in estimating the present status as well as in delineating the exact contribution of various factors.

Most evaluations follow a defined protocol just as research does so that the results obtained are objective, consistent, valid, and comprehensive. Many of these follow the protocol of before–after study, although that has limitations. Some may follow the method of qualitative research. However, the popular saying that "research seeks to prove and evaluation seeks to improve" characteristically defines the difference between research and evaluation. Thus same methods are not necessarily applicable.


1. Centers for Disease Control and Prevention. Office of the Director, Office of Strategy and Innovation, & National Center for Infectious Diseases, Division of Bacterial and Mycotic Diseases. *Introduction to Program Evaluation for Public Health Programs: Evaluating Appropriate Antibiotic Use Programs.* Atlanta, GA: Centers for Disease Control and Prevention, 2006. http://www.cdc.gov/getsmart/program-planner/Introduction.pdf
2. WHO. *Drinking and Driving: A Road Safety Manual*: 4 How to evaluate the programme. http://www.who.int/roadsafety/projects/manuals/alcohol/4-How%20to.pdf?ua=1


## evidence-based medicine

Replacing clinical judgment with solid scientific evidence is considered a major step forward. Evidence-based medicine is the conscientious, explicit, and judicious use of current best **evidence** in making decisions about the care of individual patients [1]. It is the net result of individual clinical experience; best external evidence including **systematic reviews**; expectations of the patient; and the results of interview, examination, and investigations. Evidence-based medicine is the essence of "turning research into practice" as the attempt is to convert the information available from the research to answers relevant for patient welfare. The attempt is to effectively counter eminence-based decisions, which sometimes threaten the science of medicine. For a good critique of evidence-based medicine, see Jenicek [2].

Interest may have slowed down now in the middle of the second decade of the twenty-first century, but evidence-based medicine captured a lot of imaginations in the early 1990s. This paradigm requires medical decisions to be based on critical appraisal of documented risks and benefits of various aspects of the decision-making process, including interventions. Under this paradigm, we strive to critically appraise the best evidence available about diagnostic methods, effectiveness of treatment prognosis, or magnitude and causes of important health problems [3]. The underlying evidence is categorized from level 1, for strong evidence based on properly conducted **randomized controlled trials**, to level 7, for opinions based on experience [see Figure E.6 under the topic **evidence (levels of)**]. Note the low priority accorded to opinion, even if belonging to the experts. Experts, too, are supposed to back up their opinions with evidence under this paradigm. This is for unfiltered evidence. Filtered evidence is provided by systematic reviews, particularly **Cochrane reviews**.

Besides clinical acumen, evidence-based medicine requires expertise in retrieving, collating, and interpreting the evidence available in literature or records. It requires assessment of the validity and reliability of diagnostic procedures, efficacy and safety of medical intervention, and sensitivity–specificity of diagnostic markers. Converting these to delineate risks and benefits of different actions may be even more daunting, particularly since the available evidence may not be the best you would like to have. Developing such skills is not easy, and resources required for instant access to evidence may be woefully inadequate in most medical settings. This might be one reason that the interest in evidence-based medicine has not increased. **Epistemic uncertainties** due to a lack of reliable and valid evidence for many clinical situations may be another bottleneck. Also, evidence that evidence-based medicine works better than opinion-based medicine is slow to surface.

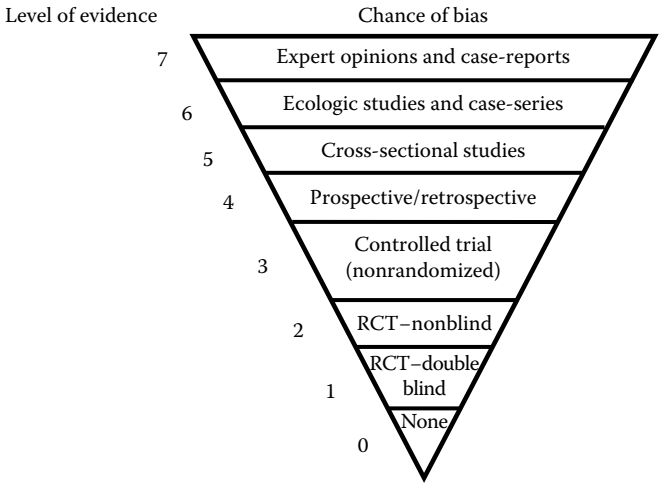



**FIGURE E.6** Level of evidence by different kinds of analytical research studies.

The intent is laudable, but the delivery so far has not been very convincing.

Although evidence is indeed the sheet anchor, patient preferences, expectations, and dilemmas get due consideration. Thus, many times, hard evidence does not translate into medical guidelines. External evidence may not be applicable to a particular patient. Consideration of a patient's needs highlights that evidence is used as a supplement and not as a replacement to clinical judgment. The purpose is to assist in arriving at a decision that is in the best interest of the patient considering the known risks and benefits. The accompanying argument must be convincing and flawless. The key is a balanced approach with the ultimate aim to reduce the areas of uncertainty and identify medical management steps that have the best likelihood of success. A glaring example often cited for non-evidence-based medical advice is that the babies should be laid to sleep on their stomach. It turned out that this carries increased risk of death while asleep. Skepticism about established beliefs is good for science, and these must be challenged as soon as new evidence emerges.

While it is true that no evidence is required for the evident, the practice of evidence-based medicine generally requires several steps, such as transforming your information requirements to framing questions, searching the literature and other evidence, evaluating the evidence and assessing its clinical applicability, and realizing its limitations. Reliability of evidence is judged by statistical methods. Statistical evidence in terms of the tests of hypotheses and confidence intervals is not the core but is ancillary in reaching an evidence-based decision. Also, our discussion is focused on the clinical setting but applies equally well to the community setting.

Among the biostatistical tools for evidence-based medicine, the best known is **decision analysis**. This takes care of the likelihoods at each step as well as the value judgments. The other is **classification and regression trees**. Details of both are provided in this volume under the respective topics. For further details, see Hunink et al. [4].

1. Sackett DL. Evidence based medicine. *Seminars in Parinatalogy* 1997;21:3–5. http://www.seminperinat.com/article/S0146-0005(97)80013-4/abstract
2. Jenicek M. Evidence-based medicine: Fifteen years later. Golem the good, the bad, and the ugly in need of a review? *Med Sci Monit* 2006;12:RA241–51. http://www.esdi.us/research/models/ebm-and-logic.pdf
3. Jenicek M. Towards evidence based critical thinking medicine? Uses of best evidence in flawless argumentations. *Med Sci Monit* 2006;12:RA149–53. http://www.medscimonit.com/download/index/idArt/452871
4. Hunink M, Glasziou P, Siegel J et al. *Medical Decision Making in Health and Medicine: Integrating Evidence and Values*. Cambridge University Press, 2002.

## evidence (levels of)

Scientific evidence is the objective information on a hypothesis. A large number of unidentified flying objects (UFOs) have been reported as cited, yet the "evidence" of this occurrence is considered weak despite some of these instances providing graphic details that are hard to deny. Physical plausibility requires that stronger evidence must emerge. So paramount is the role of evidence in science. However, various hues exist for evidence too, just as for many other concepts. For example, evidence could be against or in favor of the hypothesis, could be weak or strong, could be voluminous or scanty, and could be biased or unbiased.

In health and medicine, particularly in research, evidence typically is considered strong when obtained from experiments and trials. Properly conducted double-blind **randomized controlled trials** are supposed to provide almost infallible evidence regarding the efficacy of an intervention. This is called unfiltered evidence. Filtered evidence is available from **systematic reviews** that combine several studies of the same nature and come up with a joint conclusion. **Cochrane reviews** are regarded as the best resource for this purpose.

Other unfiltered evidence such as that provided by **observational studies** is relatively weak because of uncontrolled conditions. Expert opinion and case reports provide the poorest evidence among the various kinds of research studies. Level of evidence is fairly well depicted by Figure E.6. This is called a **bias pyramid** as it also depicts the likelihood of bias in different formats of **analytical studies** in health and medicine. This pyramid is an excellent educational tool for explaining how various designs stack up for controlling bias. This helps in fixing ideas and can provide substantial help in assessing the level of evidence available from a published or unpublished analytical study.

The underlying evidence can be categorized from level 1, for strong evidence based on properly conducted randomized controlled trials, to level 5, for opinions based on experience (see Table E.2). Note the low priority accorded to opinion, even if belonging to experts. Experts, too, are supposed to back up their opinions with the evidence under this paradigm. This hierarchy could be important for evidence-based medicine.

The volume of evidence comes from replications either in the same setup or from better-controlled studies in different setups. External evidence, whether supportive or not, is given more credence. Adequacy of evidence is judged by a lack of bias on one hand and statistical significance on the other. The idea is to plug **sources of uncertainty** as much as possible. However, **epistemic uncertainties** may remain, and those continue to pose a challenge to the veracity of evidence. Only when the results convert into actions and these actions yield results as per the expectations can we say that the evidence was adequate.

Beware of *anecdotal evidenc*e. Whereas it can provide important scientific leads, anecdotal evidence can hardly be believed to be the truth. Sometimes, an interesting event may happen in isolation because of other precipitating factors that cannot be generalized. For example, David et al. [1] discussed anecdotal evidence that

**TABLE E.2**
**Hierarchy of Study Designs by Level of Evidence for Cause–Effect Relationship**

| Evidence | Type of Design | Advantages | Disadvantages |
|---|---|---|---|
| Level 1 (best) | Randomized controlled trial (RCT)—double blind or crossover | Able to establish cause–effect and efficacy | Assumes ideal conditions |
| | | Internally valid results | Can be done only for potentially beneficial regimen |
| | | | Difficult to implement |
| | | | Expensive |
| Level 2 | Trial—no control/not randomized/not blinded | Can indicate cause–effect when biases are under control | Cause–effect is only indicated but not established |
| | | Effectiveness under practical conditions can be evaluated | Can be done only for a potentially beneficial regimen |
| | | Relatively easy to do | Outcome assessment can be blurred |
| | Prospective observational study | Establishes sequence of events | Requires a big sample |
| | | Antecedents are adequately assessed | Limited to one antecedent |
| | | Yields incidence | Follow-up can be expensive |
| Level 3 | Retrospective observational study—case–control | Outcome is predefined, so no ambiguity | Antecedent assessment can be blurred |
| | | Quick results obtained | Sequence of events not established |
| | | Small sample can be sufficient | High likelihood of survival and recall bias |
| Level 4 | Cross-sectional study | Appropriate when distinction between outcome and antecedent is not clear | No indication of cause–effect—only relationship |
| | | When done on representative sample, same study can evaluate both sensitivity/specificity and predictivity | |
| Level 5 | Case series and case reports | Help in formulating hypothesis | Many biases can occur |
| | | | Do not reflect cause–effect |

*Note:* Laboratory experiments excluded.

the records created by physicians alone have fewer errors compared with collaborative recording, but their research reveals that this is not necessarily the case. Michels and Frei [2] have talked about widely accepted negative and positive effects of vitamin C that have never been substantiated.

Also, beware of conflicting evidence. One result might lead you to believe that the effect of a regimen is positive and another that it is negative, one saying that the effect is definite and the other saying that it is doubtful. Chin and Ima-Nirwana [3] quoted from the literature to assert that there is conflicting evidence regarding the effect of α-tocopherol on bone. Davis et al. [4] have mentioned studies providing conflicting evidence regarding prognostic factors for morbidity and mortality in patients undergoing acute gastrointestinal surgery.

Statistical evidence is in terms of sample observations. This is used to draw **inference** regarding the population from which this sample has come. Credibility of statistical evidence depends primarily on the sample being representative of the population. This, in turn, depends on sample size and the **sampling techniques** used to draw the sample. Random sampling at least at one stage of the process of selection helps us to legitimately assign probabilities and to be able to use statistical methods. Statistical evidence can fall flat if the observed values are not valid or not reliable, and when inappropriate statistical methods are used for inference.

1. David GC, Chand D, Sankaranarayanan B. Error rates in physician dictation: Quality assurance and medical record production. *Int J Health Care Qual Assur* 2014;27(2):99–110. http://www.ncbi.nlm.nih.gov/pubmed/24745136
2. Michels AJ, Frei B. Myths, artifacts, and fatal flaws: Identifying limitations and opportunities in vitamin C research. *Nutrients* 2013 Dec 16;5(12):5161–92. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3875932/
3. Chin KY, Ima-Nirwana S. The effects of α-tocopherol on bone: A double-edged sword? *Nutrients* 2014 Apr 10;6(4):1424–41. http://www.mdpi.com/2072-6643/6/4/1424/pdf
4. Davis P, Hayden J, Springer J, Bailey J, Molinari M, Johnson P. Prognostic factors for morbidity and mortality in elderly patients undergoing acute gastrointestinal surgery: A systematic review. *Can J Surg* 2014 Apr;57(2):E44–52. http://canjsurg.ca/vol57-issue2/57-2-E44/

## exact confidence intervals (CIs), *see also* Clopper–Pearson interval

The usual 95% **confidence interval (CI)** for a parameter θ is obtained as $h \pm 1.96*\text{SE}(h)$, where $h$ is an estimate of θ and SE($h$) stands for **standard error (SE)** of $h$. When SE is not known, as in most practical situations, this CI is $h \pm t_{\upsilon}*\text{SE}(h)$, where $t_{\upsilon}$ is the value of Student $t$ at relevant df = υ and 95% confidence level. This is valid only when the **distribution** (**statistical)** of $h$ is Gaussian and is approximately correct for other distributions under **Gaussian conditions**. When the sample size is small and the underlying distribution far from Gaussian, we need to obtain what are called exact CIs. These are also recommended when the sample values are sparse or heavily tied (many values equal to one another). Exact CIs do not use any approximation and may be desirable in some situations even when Gaussian conditions hold true. Some statistical packages have

a provision to calculate exact CI for odds ratio, difference in proportions, and correlation coefficient, but popular exact CIs are for binomial $\pi$ and median $\tilde{\mu}$. The latter two are described next. Others are too complex for the level of this book. Besides the following methods, exact CIs are sometimes obtained by a **bootstrap** method. This method can be used for any underlying distribution.

### *Exact CI for Binomial $\pi$*

The most glaring example of a situation where exact CI is useful is for the parameter of a binary variable that follows a **binomial distribution**. In this, case the variable $x$ is the count of "successes" out of $n$ independent trials provided that the chance of success in each trial remains same. This chance is denoted by $\pi$ for the population and by $p$ for the sample. For example, $\pi = 0.5$ for the sex of an unborn child being female and $p = 34/80$ if you happen to see $x = 34$ female births out of $n = 80$ in a hospital. Quite often, $\pi$ will not be known, and you would like to find its estimate. If 17% of patients undergoing gastrointestinal surgery die within 7 days of the surgery in a hospital, what is the mortality rate in all such patients? The best point estimate is, of course, the observed $p = 0.17$ but what is the CI?

The exact CI for the binomial $\pi$ is obtained by using the exact binomial distribution in place of the Gaussian approximation. This is given as follows

$$\text{exact } 100*(1-\alpha)\% \text{ CI for } \pi\text{: [IDF.BETA}(\alpha/2,\ x+.5,\ n-x+0.5)]$$
$$[\text{IDF.BETA}(1-\alpha/2,\ x+.5,\ n-x+0.5)],$$

where IDF.BETA stands for inverse distribution function for cumulative binomial probability $\alpha/2$ corresponding to $x$ successes out of $n$ trials for the lower limit of the exact CI and probability $(1 - \alpha/2)$ for the upper limit of the CI. This is obtained by what is called a beta function. Addition of 0.5 to $x$ and $n - x$ is the **continuity correction** since a binomial distribution is a **discrete distribution**, whereas BETA function is based on a **continuous distribution**. The IDF.BETA function is available in most statistical packages, although the name can be different.

For $n = 12$ and $x = 2$, we get $p = 2/12 = 0.1667$ and the exact 95% CI is (0.0363, 0.4362) from this formula, whereas the approximate Gaussian CI is $0.1667 \pm 1.96 \times \sqrt{\frac{2}{12} \times \frac{10}{12} / 12}$, or (0, 0.3776). The lower limit is, in fact, less than 0 by this method but is stated as 0 since no $\pi$ can be negative. The exact interval just mentioned is slightly different from the popular **Clopper–Pearson interval**, which does not use continuity correction. You can see that there is a large difference between approximate Gaussian CI and the exact CI.

A similar exact CI is available for difference in proportions in independent and paired samples and for odds ratio [1].

### *Exact CI for Population Median*

Median is not a summation type of summary value, and a large sample does not help in achieving Gaussianity, as the **central limit theorem** does not generally operate in this setup. When the sample is from a Gaussian distribution, the 95% CI can still be obtained using sample median ± 1.96*SE(median), where SE(median) $\approx 1.253*\sigma/\sqrt{n}$, but not when the sample is from a non-Gaussian distribution, even when the sample size is large. Thus, the exact method for obtaining CI for a median is doubly important.

The exact CI for population median is obtained by using **nonparametric methods**. This would be more exact compared with the Gaussian CI but not fully exact. Nonparametric methods mostly require that the values observed in the sample are arranged in ascending order $X_{[1]}, X_{[2]}, \ldots, X_{[n]}$. The 95% CI for the population median (generally denoted by $\tilde{\mu}$) is in terms of the ordered values $X_{[k]}$ and $X_{[n-k+1]}$, where $k$ is largest integer such that the probability in between these two values is at least $(1 - \alpha)$. This is obtained by using binomial distribution with $\pi = ½$. The values of $k$ for different $n$ are given in Table E.3 for $(1 - \alpha) = 0.95$. With the exception of $n = 17$, the order of values in the table nearly agrees with the following:

$$\text{Lower limit} = \text{Integer part of } \left[\frac{n+1}{2} - 0.9789\sqrt{n}\right]$$

$$\text{Upper limit} = \text{Integer next to } \left[\frac{n+1}{2} + 0.9789\sqrt{n}\right].$$

In fact, these formulas work for $n = 6$ through $n = 283$ except for $n = 17$ and $n = 67$. Thus, you can use these for practically every situation—Gaussian, non-Gaussian, small $n$, and large $n \le 283$.

For illustration, the following are the numbers of diarrheal episodes (of at least 3 days' duration) during a period of 1 year in 12 children of age 1–2 years:

3 7 12 2 4 3 5 8 1 2 3 4

In this case, $\bar{x} = 4.5$ and $s = 3.12$. Of 12 observations, 8 are below the mean, and only 4 are above the mean. This indicates that there is

**TABLE E.3**
**Value of *k* for Different *n*—95% Confidence Interval (CI) for Median is $(X_{[k]}, X_{[n-k+1]})$**

| *n* | *k* | 95% CI |
|---|---|---|
| ≤5 | | 95% CI cannot be computed |
| 6 | 1 | $(X_{[1]}, X_{[6]})$ |
| 7 | 1 | $(X_{[1]}, X_{[7]})$ |
| 8 | 1 | $(X_{[1]}, X_{[8]})$ |
| 9 | 2 | $(X_{[2]}, X_{[8]})$ |
| 10 | 2 | $(X_{[2]}, X_{[9]})$ |
| 11 | 2 | $(X_{[2]}, X_{[10]})$ |
| 12 | 3 | $(X_{[3]}, X_{[10]})$ |
| 13 | 3 | $(X_{[3]}, X_{[11]})$ |
| 14 | 3 | $(X_{[3]}, X_{[12]})$ |
| 15 | 4 | $(X_{[4]}, X_{[12]})$ |
| 16 | 4 | $(X_{[4]}, X_{[13]})$ |
| 17 | 5 | $(X_{[5]}, X_{[13]})$ |
| 18 | 5 | $(X_{[5]}, X_{[14]})$ |
| 19 | 5 | $(X_{[5]}, X_{[15]})$ |
| 20 | 6 | $(X_{[6]}, X_{[15]})$ |
| 21 | 6 | $(X_{[6]}, X_{[16]})$ |
| 22 | 6 | $(X_{[6]}, X_{[17]})$ |
| 23 | 7 | $(X_{[7]}, X_{[17]})$ |
| 24 | 7 | $(X_{[7]}, X_{[18]})$ |
| 25 | 8 | $(X_{[8]}, X_{[18]})$ |
| 26 | 8 | $(X_{[8]}, X_{[19]})$ |
| 27 | 8 | $(X_{[8]}, X_{[20]})$ |
| 28 | 9 | $(X_{[9]}, X_{[20]})$ |
| 29 | 9 | $(X_{[9]}, X_{[21]})$ |
| 30+ | | Order value at $Int\left[\frac{n+1}{2} \pm 0.9789\sqrt{n}\right]$ |

a lack of symmetry, and the distribution is unlikely to be Gaussian. In ascending order, the durations are

$$X_{[1]} = 1, X_{[2]} = 2, X_{[3]} = 2, X_{[4]} = 3, X_{[5]} = 3, X_{[6]} = 3,$$
$$X_{[7]} = 4, X_{[8]} = 4, X_{[9]} = 5, X_{[10]} = 7, X_{[11]} = 8, X_{[12]} = 12.$$

Sample median = $(X_{[6]} + X_{[7]})/2 = (3 + 4)/2 = 3.5$. From Table E.3, for $n = 12$, the 95% CI is $(X_{[3]}, X_{[10]})$, i.e., in this case, (2, 7). There is a rare chance, less than 5%, that the median number of diarrheal episodes in the child population from which this sample was randomly drawn is less than 2 or more than 7.

When proceeding with the Gaussian pattern, the 95% CI for median would be $3.5 \pm 2.201 \times 1.253 \times 3.12/\sqrt{12}$ or (1.0, 6.0), where 2.201 is the value of Student $t$ at 11 degrees of freedom (df's) for 5% probability on either side. This Gaussian CI is not very different in this case but sometimes can be very different.

The CI obtained by the nonparametric method just described has *at least* a 95% confidence level. The ordered observations rarely allow the level to be exactly 95%. In some cases, it could be as high as 98%. If the CI is narrowed, the confidence level becomes less than 95%, and that is not allowed for a 95% CI. Also note that small samples can rarely provide precise information. This is illustrated by our example where the 95% CI for median is fairly wide and not able to provide much useful information.

1. Agresti A. Dealing with discreteness: Making 'exact' confidence intervals for proportions, differences of proportions, and odds ratios more exact. *Statistical Methods Med Res* 2003;12:3–21. http://www.stat.ufl.edu/~aa/articles/agresti_2003.pdf

## exact tests

Exact tests are those statistical tests of hypothesis that produce exact probability of Type I error (***P*-values**). These are in contrast to the approximate tests where the *P*-values are nearly the same as planned but not necessarily exactly the same. For example, the tests based on Gaussian approximation taking advantage of the **central limit theorem** for large $n$ are not exact. A $z$-test is exact only when the underlying distribution is exactly Gaussian. On the other hand, the **Fisher exact test** for association in a $2 \times 2$ table is an exact test for small frequencies irrespective of the underlying distribution. Exact tests generally require more intricate calculations. These tests are especially recommended when the sample size is small and the underlying distribution is far from Gaussian. Sparse data and heavily tied values (many observations have the same value) also require exact tests.

Besides the Fisher exact test and the similar **Barnard test**, this volume describes an exact binomial test for paired data in a $2 \times 2$ table under the topic **McNemar test** (where the approximate chi-square test is also described). This exact test for dichotomous data is the same as the **binomial test** for one-way tables. Among other exact tests are the **Zelen test** for equality of odds ratio in two or more independent samples, exact test for Pearson and Spearman correlations, tests for regression coefficient, tests for Cohen kappa coefficient, and Cochran test for trend in proportions. For example, Schlag [1] has discussed an exact test for correlation and slope parameter in simple linear regression, and Agresti [2] has reviewed algorithms for exact *P*-values for **contingency tables**. Most exact tests are mathematically complex and not described in this volume. Our advice is to fall back on reputed statistical software, choose the right option if available in the software package, and get the results of an exact test where needed. Some software packages provide the option to do **Monte Carlo simulations** to get exact *P*-values. Permutation tests are also considered exact tests.

1. Schlag KH. Exact tests for correlation and for the slope in simple linear regressions without making assumptions (June 25, 2008). http://ssrn.com/abstract=1162125, last accessed June 30, 2015.
2. Agresti A. A survey of exact inference for contingency tables. *Statistical Science* 1992 Feb;7(1):131–53. http://links.jstor.org/sici?sici=0883-4237%28199202%297%3A1%3C131%3AASOEIF%3E2.0.CO%3B2-A

## examination (health/medical) data/surveys

Medical data obtained after examination of a patient by a competent physician are just about the most reliable data on the condition of a patient. Data from laboratory and radiological investigation may be better in quality, but the data obtained from interviews and observations may not be so reliable. Generally, five kinds of medical examination data are available for processing, as described next. For the purpose of research, special care is taken to conceal the identity of individuals so that their privacy is not compromised.

First are the huge data available in medical records sections of hospitals. Almost any hospital these days maintains these records, particularly for inpatients. Over a period of time, they become enormous and can provide important leads on what kinds of patients are being admitted, what the findings were, how they were treated, and with what results. These records will also have investigation results as well as interview findings. Thus, you can evaluate whether or not the records are internally consistent. If the hospital is not careful, many inconsistencies may be detected, and some records may be incomplete.

Second are the data from examination of people who seek special services, such as insurance, a visa, a job in the armed forces, employment in a factory, etc. Some of these may require periodic examination. Over a period of time, these also can become enormous.

Third are the findings of examination surveys. These are not many, since it is not easy to go out in the community and examine people. The most well known of these are the National Health and Nutrition Examination Surveys periodically carried out in the United States on nationally representative samples [1]. This uses specially equipped mobile vans for examination. The data are automated right from inception and available online for follow-up action.

Fourth are the data generated by you and us through our own research efforts on targeted subjects. Many journals are now moving to a system that archives data on a website for selective access so that the communicated results can be verified. This is being done particularly for clinical trial data.

Fifth are the data available in clinics in private practice. Many clinicians are immaculate in recording and maintaining the medical data of patients, and use these to build up a longitudinal history. This helps them to look at any episode of illness in a holistic manner. But, yes, some clinicians are not that particular and waste the information they collect.

Because of the confidence one can attach to the examination data, and their ready availability to those who have access, these could be important source of information on the health status of the people, albeit for the selected group for which these databases are built up. Whereas a large number of publications appear for research data (fourth type), various useful conclusions are drawn from examination surveys (third type in our list). Singh et al. [2] analyzed data from a large number of health examination surveys across the world to conclude that total cholesterol rose more steeply with age from 30 to 54 years in high-income countries and fasting plasma glucose in Oceania, the Middle East, and the United States. The systolic

blood pressure association with age had no specific income or geographical pattern, although this also increased with age. Phillips et al. [3] analyzed examination data (second type in our list) of Vietnam War veterans and concluded that increased follicular stimulating hormone and luteinizing hormone levels are associated with higher all-cause mortality, though their effect was not independent of one another. Palmer and Stephens [4] compiled data of more than 11 million patients from more than 500 hospitals (first type of data in our list) in the United States and concluded that total costs of opioid intravenous patient-controlled analgesia generally ranged from $647 to $694. These examples illustrate that examination data of all types can be extremely useful for meaningful conclusions.


1. Centers for Disease Control and Prevention. *About the National Health and Nutrition and Examination Surveys*. http://www.cdc.gov/nchs/nhanes/about_nhanes.htm
2. Singh GM, Danaei G, Pelizzari PM, Lin JK, Cowan MJ, Stevens GA, Farzadfar F et al. The age associations of blood pressure, cholesterol, and glucose: Analysis of health examination surveys from international populations. *Circulation* 2012 May 8;125(18):2204–11. http://circ.ahajournals.org/content/125/18/2204.long
3. Phillips AC, Gale CR, Batty GD. Sex hormones and cause-specific mortality in the male veterans: The Vietnam Experience Study. *QJM* 2012 Mar;105(3):241–6. http://qjmed.oxfordjournals.org/content/105/3/241.long
4. Palmer P, Ji X, Stephens J. Cost of opioid intravenous patient-controlled analgesia: Results from a hospital database analysis and literature assessment. *Clinicoecon Outcomes Res* 2014 Jun 20;6:311–8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4073913/


## exchangeability

Two equally sized random samples from the same population are exchangeable because theoretically, both should lead to the same conclusion. In other words, any one of them can be used without altering the quality of the result. This arises from the property of random sampling. Estimates of the mean, standard deviation, proportion, etc. may differ in value from sample to sample but qualitatively are as good from one random sample as from another random sample, provided that both are of the same size and that they are from the same parent population. Exchangeability also means that the sample values can be rearranged without affecting the result. Whether the sample values are 2, 9, 3, 7, 8, and 5 or in another order 9, 8, 2, 5, 3, and 7, it does not matter. You can see that the values of mean, SD, etc. do not alter when the sequence is altered. Another application of exchangeability is illustrated in calculation of the **correlation coefficient**. The correlation coefficient between $x$ and $y$ is not affected if $x$ is replaced by $y$ and $y$ is replaced by $x$. They can be exchanged for calculation of the correlation coefficient. Exchangeability is not actually exercised; the samples are not physically exchanged—it is just the property of the sample, and we use this property to develop our procedures.

The concept of exchangeability was introduced by Haag in 1924 [1]. This concept helps in several ways, but the most obvious is the **distribution** of the difference of two exchangeable variables. When $x$ and $y$ are exchangeable, $z = x - y$ has the same distribution as the difference $-z = y - x$. This really means that the distribution of $z$ is **symmetric** about 0. This property is used in constructing, for example, a **Wilcoxon signed rank test**. Nonparametric **permutation tests** also use the exchangeable property. Procedures that combine results from several samples such as **meta-analysis** tacitly assume that the samples are exchangeable; otherwise, how can you combine them. Forecasting or prediction of values, such as by **regression**, is also based on the concept of exchangeability since the next values are considered to behave in the same manner as the existing ones we have.

Another useful application of exchangeability is in **multilevel modeling** and **generalized estimating equations**. These methods consider clustered data, such as activity at six sites in the brain (measured by electrodes). They will be correlated. In this situation, you may want to assume that the correlation between values at site 1 and values at site 2 is the same as between site 1 and site 4, or between site 5 and site 6. If so, these correlations are exchangeable and lead to relatively easily interpretable estimates of the parameters of the model. Note that this setup is different from longitudinal data on the same set of subjects, which are clustered and correlated, but it will not be realistic to assume that the correlation between values at time 1 and time 2 is the same as between time 1 and time 4. If so, these correlations are not exchangeable.


1. Bayesian Inference: The Way it Should be. *Exchangeability*. http://www.bayesian-inference.com/exchangeability, last accessed April 30, 2014.


## exclusion criteria, *see* inclusion and exclusion criteria

## exhaustive categories, *see* mutually exclusive and exhaustive categories

## exogenous factors, *see* endogenous and exogenous factors and variables

## expectation of life and the life table, *see also* life expectancy

The average number of years expected to be lived by individuals in a population is called expectation of life. This can be calculated as expectancy at birth or at any other age. The expectation of life at birth (ELB) can be crudely interpreted as the average age at death. The method used for calculating expectation of life requires building up a life table that summarizes the age at death of a group of people.

### *The Life Table*

The ideal method for computing the expectation of life is by observing age at death of a large cohort (called a **radix**) of live births as long as any individual of the cohort is alive. This may take more than 100 years and thus is impractical unless records for over 100 years are complete. Such a *cohort life table* provides summaries of those born in a particular year, and another cohort will be required for another year. This is also called a *generation life table*.

As a shortcut, a *current life table* is constructed. It assumes that the individuals at different ages are exposed to the *current* risks of mortality. Thus, the current **age-specific death rates** (ASDRs) are used on a presumed radix of, say, 100,000 persons. The average of life so obtained is the number of years a newborn is expected to live *at the current levels of mortality*. This may be characterized as cross-sectional and provides a snapshot of the current mortality experience. This is more useful, as it tells about the existing situation and can be computed immediately, obviating the need to wait for

100 years. The calculations are done as follows. The method is also used in some other applications, as described later in this section.

For constructing a life table, it is desirable to do the computation for each year of age, say at age 46 years, 47 years, etc., but the mortality rates are generally available for age groups, such as 45–49 years. When such groups are used in constructing a life table, it is called an *abridged life table*. When each single year of age is used, it is called a *complete life table*.

Table E.4 is an example of an abridged life table for urban females in a developing country for the block years 2011–2015. An example of a developing country provides an opportunity to discuss some features of a life table that do not appear in a life table constructed for developed countries. An explanation of the notations in this table is as follows:

$n_t$ = length of the age interval beginning with age $t$
$q_t$ = probability of death in the age interval $(t, t + n_t)$ when the person is alive at age $t$

The latter is the deaths occurring during the interval as a proportion of the population at the *start* of the interval $t$. Generally, the population is available for the midpoint of the interval rather than at the start. Thus, $q_t$ is estimated from the ASDR, which is based on the population at midyear. This estimate is

$$q_t = \frac{n_t m_t}{1 + (1 - a_t) n_t m_t},$$

where
$m_t$ is the ASDR for the age interval $(t, t + n_t)$, and
$a_t$ is the average length of life lived per year in that interval by those who die in the interval.

When $a_t = ½$, $q_t = \frac{n_t m_t}{1 + \frac{1}{2} n_t m_t}$. The average $a_t = ½$ in this equation works well when the risk of death is uniform throughout the interval. This is generally true for age intervals 5–10, 10–15, …, 65–70 years but not for the lower and higher age intervals. For example, it is seen that infant deaths are highly concentrated in the neonatal period and not uniformly distributed over the 0–1 interval. In low-mortality areas (developed countries), those who survive the neonatal period tend to live almost as long as anybody else. In this case, for $t = 0$, the convention is to use $a_0 = 0.1$. For high-mortality areas (developing countries), $a_0 = 0.3$ is used. Such a higher value of $a_0$ for developing countries seems paradoxical, but infant deaths in developing countries tend to be relatively more evenly distributed, as the deaths also continue to occur in the postneonatal period. This is rare in developed countries. In age interval 1–5 years, deaths are slightly more in 1–2 years than in 4–5 years. Thus, for $t = 1$, $a_1 = 0.475$ can be used for all areas, although 0.4 is advocated for developed countries. Additional adjustment is needed for the last age interval because its length is not known. Other notations are as follows:

$l_t$ = expected number of persons surviving the age $t$.

The radix is 100,000, which obviously survives age $t = 0$. The survivors at exact age $(t + n_t)$ are

$$l_{t+n_t} = \text{survivors at } t - \text{dying in the interval } (t, t + n_t) = l_t - q_t * l_t.$$

For instance, in Table E.4, expected survivors at age 35 years

$$l_{35} = 89248 - 0.01089 \times 89248 = 88276.$$

$L_t$ = Expected number of person-years lived in the interval $(t, t + n_t)$
$= (n_t/2)(l_t + l_{t+n_t})$

for all age intervals except the first two and the last. The last is denoted by $w$. For these three age intervals, $L_t$ is given in Table E.5. These are the adjustments made in appreciation of the fact that the deaths are not evenly distributed in these three intervals. Thus, these deaths cannot be averaged at the midpoint. Appropriate modifications can

**TABLE E.4**
**Abridged Life Table for Urban Females in a Developing Country for the Block Years 2011–2015**

| Age Interval (Years) | Probability of Death in the Interval | Expected Surviving Age $t$ | Person-Years Lived in the Interval | Person-Years Lived beyond $t$ | Expectation of Life at Age $t$ |
|---|---|---|---|---|---|
| $(t, t + n_t)$ | $q_t$ | $l_t$ | $L_t$ | $T_t$ | $e_t$ |
| 0–1 | 0.04989 | 100000 | 96508 | 6214374 | 62.14 |
| 1–5 | 0.02453 | 95011 | 375149 | 6117866 | 64.39 |
| 5–10 | 0.00648 | 92680 | 461897 | 5742717 | 61.96 |
| 10–15 | 0.00489 | 92079 | 459270 | 5280820 | 57.35 |
| 15–20 | 0.00737 | 91629 | 456457 | 4821550 | 52.62 |
| 20–25 | 0.00874 | 90954 | 452782 | 4365093 | 47.99 |
| 25–30 | 0.01010 | 90159 | 448517 | 3912311 | 43.39 |
| 30–35 | 0.01089 | 89248 | 443810 | 3463794 | 38.81 |
| 35–40 | 0.01168 | 88276 | 438802 | 3019984 | 34.21 |
| 40–45 | 0.01480 | 87245 | 432997 | 2581182 | 29.58 |
| 45–50 | 0.02134 | 85954 | 425185 | 2148185 | 24.99 |
| 50–55 | 0.03522 | 84120 | 413192 | 1723000 | 20.48 |
| 55–60 | 0.05783 | 81157 | 394052 | 1309808 | 16.14 |
| 60–65 | 0.14623 | 76464 | 354367 | 915756 | 11.98 |
| 65–70 | 0.15759 | 65283 | 300695 | 561389 | 8.60 |
| 70+ | 1.00000 | 54995 | 260694 | 260694 | 4.74 |

**TABLE E.5**
**Estimation of $L_0$, $L_1$, and $L_w$ for Developing and Developed Countries**

| Age Interval (Year) | Developing Country | Developed Country |
|---|---|---|
| 0–1 | $L_0 = 0.3\ l_0 + 0.7\ l_1$ for $a_0 = 0.3$ | $0.1\ l_0 + 0.9\ l_1$ for $a_0 = 0.1$ |
| 1–4 | $L_1 = 1.9\ l_1 + 2.1\ l_5$ for $a_1 = 0.475$ | $1.6\ l_1 + 2.4\ l_5$ for $a_1 = 0.4$ |
| Last | $L_w = l_w * \log_{10} l_w$ | $L_w = l_w / M_w$ |

be made for a specific country according to its pattern of mortality. You may find different adjustments in the literature.

In Table E.4, $L_w = 54995 \times \log(54995) = 260694$.

Now,

$$\begin{aligned} T_t &= \text{number of person-years lived beyond age } t \\ &= L_t + L_{t+n_t} + \cdots + L_w \\ &= L_t + T_{t+n_t} \end{aligned}$$

and $T_w = L_w$. This means that the totaling is done from the bottom upward. Thus, the last two columns are calculated after the calculations for the previous columns are complete. Finally,

$$e_t = \text{expectation of life at age } t$$

$$= \frac{\text{number of person-years lived beyond age } t}{\text{number of persons surviving age } t} = T_t / l_t.$$

Table E.5 is peculiar in the following ways: (i) For most of the intervals, as already stated, it is reasonable to assume that the deaths are uniformly distributed, and an average of half-length is lived by the persons dying in the interval. This, however, is not true for the intervals 0–1 year, 1–4 years, and 70+ years. Such a consideration leads to different values of $L_0$, $L_1$, and $L_w$, as stated in Table E.5 for developing and developed countries. (ii) In this example, the expectation of life at 1 year ($e_1 = 64.39$ years) is more than the ELB ($e_0 = 62.14$ years). Thus, a child of 1 year is expected to live for 64.39 more years, whereas at birth, the expectation is only 62.14 years. This discrepancy is due to the high infant mortality rate (IMR). This is true for many developing countries. (iii) This life table is for the block of a 5-year period, namely, 2011–2015. The annual estimates of ASDRs in many developing countries show considerable fluctuation because these are based on sample studies. The average of a 5-year period is fairly stable and is better in reflecting the trend at the midpoint of the block, which is the year 2013 in this case.

The following comments contain useful information regarding life tables:

- All life tables assume that there are no sudden deaths owing to calamities such as famine, earthquakes, and cyclones, or at most, they are minimal and do not affect the general pattern of mortality.
- Males and females in general have differential mortality patterns, and therefore, separate life tables are drawn for the two sexes. A similar distinction may also be needed between other groupings such as rural–urban, low income–high income, and laborer–executive classes.
- Age-specific mortality may change from year to year. The rates in the year 2012 are not necessarily the same as in the year 1995. Thus, a life table is prepared every year or at least once in 5 years, when the changes become noticeable.
- The ELB is severely affected by infant mortality. This is of particular concern to the developing nations, where IMR is high. For this reason, expectation of life at 1 year (EL1) is sometimes preferred as an indicator of longevity.

The expectation of life is a very popular measure of health on one hand and of socioeconomic development on the other. It is considered a very comprehensive indicator because many aspects of development seem to reflect on the longevity of people. In 2013, it ranged from a low of 46 years in Sierra Leone to a high of 84 years in Japan. It has gone up in China from nearly 40 years in 1950 to nearly 75 years in 2013. Earlier evidence suggests that the maximum ELB attainable is 85 years [1], but it is already considered to have the potential to reach 86 years [2]. Note that this is an *average* attainable by a population. Individuals have been known to live for as long as 125 years.

### *Application of Life Table Method to Other Setups*

As remarked on earlier, the life table is a fairly general method and is used in a wide variety of situations. The method can be used for any **arrival–departure process**, of which the birth–death process is a particular case. The outcome of interest can be remission or recurrence or any such event. Note that one can be flexible with regard to the arrival time, which can vary from subject to subject. In Table E.4, the population was fixed and subject to departures only. But the structure could be that subjects join the group at different chronological points in time, remain in the group for varying periods, and leave at different points in time. Tietze [3] proposed a life table–like method for acceptors of temporary methods of birth control. For example, oral contraceptive users start taking the pill at different points in time, continue taking it for different periods (say from 1 to 36 months), and then stop for a variety of reasons (accidental pregnancy, appearance of side effects, planned pregnancy, etc.). The life table method can be used to assess the expected duration of continuation of intake in this situation. Pharoah and Hollingworth [4] used the life table survival to work out the cost-effectiveness of statins in lowering serum cholesterol concentration in people at varying risk of fatal cardiovascular disease in the United Kingdom. The other important application of the life table method is in studying the pattern of survival of patients with different treatment regimens or with different risk factors. For this, see **survival analysis**.

For other forms of expectation of life such as healthy life expectancy and health-adjusted life expectancy, see the topic **life expectancy (types of)** in this volume.

1. Olshanky SJ, Carnes BA, Cancel C. In search of Methusaleh: Estimating the upper limits to human longevity. *Science* 1990;250:634–90. http://www.sciencemag.org/content/250/4981/634.abstract
2. Murray CJL, Ezzati M, Flaxman AD, Lim S, Lozano R, Michaud C, Naghavi M et al. Comprehensive systematic analysis of global epidemiology: Definitions, methods, simplification of DALYs, and comparative results from the Global Burden of Disease Study 2010: p. 14. Supplement to: Murray CJL, Ezzati M, Flaxman AD et al. GBD 2010: Design, definitions, and metrics. *Lancet* 2012;380:2063–66. http://www.thelancet.com/cms/attachment/2017336178/2037711222/mmc1.pdf

3. Tietz C. Intra-uterine contraception: Recommended procedures for data analysis. *Stud Fam Plann* 1967; No.18 (Suppl.):1–6. http://www.popline.org/node/473147
4. Pharoah PDP, Hollingworth W. Cost effectiveness of lowering cholesterol concentration with statins in patients with and without pre-existing coronary heart disease: Life table method applied to health authority population. *BMJ* 1996; 312:1443–8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2351181/

E

## experimental designs, *see also* experimental studies

An experimental design is the specification of the process under which an experiment is carried out from the beginning to the end. This is the pattern, scheme, or plan to collect evidence. *Experiment* is the term used for a study where human interventions are intentionally introduced to assess their effect. Thus, the experimental design would involve steps such as the type of subjects, what intervention would be introduced, how, when, who, etc., and then how its effect would be evaluated. The former would include the selection of subjects, the size of the sample, and the method of assigning the interventions to the subjects; and the latter would include defining the primary and secondary outcomes and their method of measurement, collection of data, and their processing. You can see that experimental designs have heavy statistical content. All these are specified in advance so that no bias creeps in. Credibility of research findings depends on the ability of the design to provide infallible evidence. It should have all the features listed under the topic **designs of medical studies**, and much more, as described here.

In an expanded form, some include the entire **protocol** under the design. In our opinion, aspects such as setting objectives and review of literature are part of the protocol but not of the design. In addition, there is a limitation too, particularly in the context of health and medicine. Broadly speaking, experiments include clinical and field trials, but it is customary in health and medicine to restrict the term to laboratory experiments that are done on animals or tissues and specimens. We also limit to this restricted usage in the following paragraphs, although many of these steps are directly applicable to clinical and field trials as well.

An experiment must be designed such that it can answer the question of interest clearly and efficiently. It includes features such as what factors to investigate, what levels of various factors are important, how many groups of units are needed, how to allocate groups to various interventions including control, etc. Validity of experimental results is directly affected by the structure of the design and its correct implementation. For this reason, attention to experimental design is extremely important. It is an important component of the study protocol also.

A design is easy to construct and a constructed design is easy to understand when a few terms used in this context are clear. First, a *factor* is an **independent** characteristic whose levels are set by the experiments. In an experiment on mice, mice can be divided by strain (Sprague–Dawley, Wistar, Donru, etc.) to see the effect in different strains, or by their weight into light, moderate, and heavy to study the effect in different weight categories. The factor is strain in the first case and weight category in the second case.

Levels of the factor connote the strata into which a factor is divided. When there are three weight categories of mice as in the preceding paragraph, weight has three levels. The weight levels light, moderate, and heavy have gradient, but that is not necessary. Strains of mice do not have any such gradient, but they are still called levels. If four strains are considered for the experiment, it has four levels. The design will greatly depend on how many levels of various factors need to be studied to adequately answer the research question.

An essential factor in an experiment is the intervention. It has a minimum of two levels—intervention present and intervention absent (control). If there are three dose levels such as 0, 5, and 10 mg/kg of body weight of mice, the intervention has three levels.

The last term in the context of experiment is the **response**. Response is the outcome over which the experimenter has no control. It can be qualitative, such as positive change, no change, and negative change, or can be quantitative, such as time taken in recovery or the magnitude of decrease in the level of pain. Nuances of **qualitative** and **quantitative** characteristics are discussed in a separate, topic but note here that the statistical analysis of data heavily depends on the nature of the response variable.

Haphazard collection of data without a design can be expensive and time-consuming, requiring many runs, and yet not provide the focused conclusions you are looking for. **Confounding** may occur, and that can make it difficult to distinguish the effect of various factors. Thus, design is important. A well-designed experiment can (i) easily spot where the differences exist, (ii) provide more reliable and focused answers to your questions, (iii) reduce cost by avoiding wastage, and (iv) provide insight regarding the patterns that otherwise would be difficult to discern.

### *Choice of Experimental Unit*

An experiment necessarily requires exposing units to an intervention whose effect is unknown. If the effect is known, there is no need to do that experiment. Any investigation originates from uncertainty—the environment in an experiment is doubly uncertain because of the intervention. A suitable design certainly helps in controlling some of these uncertainties, but the choice of experimental units, the method of selection, and the size of the experiment are also important.

As mentioned earlier, an experiment can be carried out on biological material, on animals, and later on human beings. Commonsense ethics dictates that the preference is the lowest biological entities and then progressively to higher ones when the lower entities fail to meet the requirement. Thus, the first efforts must be directed to identify a suitable material such as tissue, cell, blood sample, serum, or body fluid for experiment. For example, for the effect of temperature on erythrocyte sedimentation rate (ESR), the experimental unit for exposure to various temperatures would obviously be blood sample. Material already may be available in a laboratory, and this is collected for the investigation required for the patient but is preserved as a record. In that case, no fresh consent is required. The investigation would be unlinked anonymous, and the person concerned does not come into the picture. See the following example for such an experiment.

The twin organisms *Acinetobacter calcoaceticus–Acinetobacter baumannii* can be isolated and identified from clinical samples such as tracheal tips, wound swabs, urine, and blood. The biofilm forming ability of the complexes can be studied under different conditions, such as different media strengths, room temperatures, and pH values of the medium. Biofilm forming ability is measured by optical density value. This property may be related with the antimicrobial susceptibility pattern and the site of isolation of strains. Thus, the results of such an experiment can help devise coatings or other mechanisms that inhibit the biofilm forming ability of these bacteria in humans.

In some situations, however, caution may be needed in conducting experiments on biological material. A great debate is going on regarding research on stem cells, particularly when they are obtained from human embryos, even though such embryos pile up in some hospitals.

If biological specimens are not able to serve the purpose, the search shifts to a suitable animal model. Sometimes, stage 1 of the experiment is done on biological specimens and stage 2 on an animal model. For live animals, preference is given to small animals such as mice, guinea pigs, and rabbits. Ethics for them is slightly relaxed compared to bigger animals such as monkeys and dogs. The choice is primarily influenced by the suitability of the animal for the disease under consideration and the possibility of the implication of results for humans. Mice are a favorite because they are small, prolific, vertebrate, and mammals, and get diseases such as diabetes and cancer. They may mimic human conditions well. Whatever species and strains are used, they must be sensitive so that the effect, if any can, show up. No group should be in a different room or a different shelf, because the environment may be different.

Before embarking upon animal experimentation, consult the local ethics, which is generally quite well documented. The issues regarding administering harmful substances to animals, their torture, and sometimes sacrifice are more concerned with ethics than statistics. See books such as that by Wolfensohn and Lloyd [1] for animal experimentation ethics and care.

### *Types of Experimental Designs*

The design of experiments is a vast subject, and full books have been written on this topic (see, e.g., Refs. [2,3]). Thus, our description in this volume is necessarily short. We have also divided this into specific topics for different types of designs. See the following topics for details:

**balanced and unbalanced designs**
**completely randomized designs**
**crossover designs**
**randomized block designs**
**one-way designs**
**two-way designs**
**factorial and partially factorial designs**
**repeated measures designs**

Other aspects of the experimental design are selection of subjects and sample size. Laboratory experiments are generally conducted on homogeneous material, and the selection generally is not difficult. The only consideration is whether this material will allow you to proceed with human trials or with trials that improve health care. Because of homogeneous material and controlled conditions in a laboratory, there is no need for a big sample size for medical experiments. Many experiments are done on just about five rats in each group. For further details, see **sample size for medical experiments**. Also see the topic **experimentation (statistical principles of)**, where randomization, replication, and controls are discussed. They apply to clinical trials also.

### *Choosing a Design of Experiment*

The question of interest should be precisely defined so that a design can be devised that can adequately answer that question. An elaborately conceived design will fail if it answers a wrong question. Thus, always devote some time in brooding over the problem and converting it to an investigable question. This can be a big help in devising a right design. A precise question can help to decide what specific factors to investigate, and what and which levels of these factors should be included. Narrow them down to as small a list as you can manage without sacrificing the utility of the experiment. If there is any doubt regarding a particular factor even after consulting the literature and experts, it is better to include it in the experiment and check yourself whether it is helpful in explaining the response. Alternatively, compare the cost of ignoring a factor with the cost of inclusion, and do what looks more efficient.

**TABLE E.6**
**Design Appropriate for Various Experimental Conditions**

| Outcome of Interest | Conditions | Design |
|---|---|---|
| Effect of a stimulus or doses of stimuli in homogenous group of units | No interest in stratified analysis | One way |
| Effect of two or more treatments in the same group of subjects | No carryover effect and no natural periods of remission | Crossover |
| Average outcome for combination of levels of two or more factors, including interaction | All combinations of factor levels are relevant and feasible; interest is in stratified analysis if the units are heterogeneous | Factorial—it can be two way or multiway depending on the number of factors |
| Average outcome for each level of each factor but interactions not of much interest | Some combinations of levels of one or more factors not relevant or not feasible | Partially factorial |
| Trend over time | – | Repeated measures |

Although the appropriate conditions for each design are mentioned at the time of discussing that design, the summary in Table E.6 may be helpful to get an overall view.

1. Wolfensohn S, Lloyd M. *Handbook of Laboratory Animal Management and Welfare*, Third Edition. Blackwell Publishing Inc., 2003.
2. Montgomery DC. *Design and Analysis of Experiments*, Eighth Edition. Wiley, 2012.
3. Goos P, Jones B. *Optimal Design of Experiments: A Case Study Approach*. Wiley, 2011.

## experimental studies, *see also* experimental designs

The essential ingredients of an experiment are a purported manipulable cause and an anticipated effect. The relationship is speculative—if already confirmed, there is no need for an experiment. An experiment is a procedure to verify or falsify a causal relationship by introducing the purported cause and observing the effect. This definition requires that a cause be hypothesized and its possible outcomes visualized in advance. Observations are geared to measure this anticipated outcome. Experiments tend to provide results that transcend time and space, and in many cases beyond the population under study. Observational studies lack such generalization.

Broadly speaking, the occurrence or nonoccurrence of different maternal complications in anemic and nonanemic women is an experiment, albeit performed by nature; so is the occurrence of goiter of various grades in areas with iodine deficiency in water. When a rare or unique opportunity is available to observe or to study the effects of specific events as a result of naturally occurring changes,

it is called a **natural experiment**. For example, the tsunami of 2005 provided a rare opportunity to study the health consequences of such a disaster. John Snow's classical discovery that cholera is a waterborne disease was the outcome of a natural experiment. Snow identified two mixed populations, alike in many important respects but different in the sources of water supply to their households. The large difference in the occurrence of cholera among these two populations gave a clear indication that cholera is a waterborne disease. This was demonstrated in 1854, long before the advent of the bacteriological era. Such natural experiments generally come under the domain of **observational studies**.



This section concerns experiments with *human intervention* for changing the course of events. They are therefore also called **intervention studies**. By exercising control on the extraneous factors that can affect the outcome, experiments are the most direct method to study cause–effect relationships. Experimental evidence is generally more compelling than that available from observational studies. Science is achieved by experiments.

A medical experiment can be carried out in a laboratory, clinic, or community. The subjects for experiment in a clinic or community are human beings, and such an experiment is generally termed a **trial**. A laboratory experiment, on the other hand, may involve inanimate entities, such as physical forces or chemicals: in the context of medicine, laboratory experiments are generally conducted on biological material or animals. Laboratory experiments often provide important clues to the potential of the intervention for formulation into a therapeutic agent. When successful, they pave the way for human studies. Thus, such experiments have a special place in medical studies. Even if you do not plan to conduct an experiment yourself, the details mentioned herein will help you to better understand and interpret the results of experiments conducted by others.

## *Basic Features of Medical Experiments*

Sometimes, passive observation is not enough. You may have tried twisting the tail of your dog to see how he/she reacts. There is always a curiosity to explore the consequences of our actions. The first basic feature of any experiment is manipulation—the stimulus. The natural course of events is sought to be changed by human intervention. Although an experiment can be carried out for an intervention whose mechanism of action is still unknown, cause–effect inference is complete only when a biological explanation is available.

The term *medical experiment* is generally restricted to laboratory experiments performed on animals and other biological specimens. In the context of drug development, an experiment in the laboratory is performed in the first phase to develop a molecule that has desirable biochemical properties—and thus has propensity to be beneficial in promotion of health, prevention of sickness, or treatment of disease—and second to establish that these properties are indeed present. This exercise requires enormous inputs of theoretical knowledge about various compositions. Once such a compound is obtained, its properties are investigated, the compound is modified as needed, and the drug development starts. A formulation is prepared that could ultimately take the shape of a drug. This formulation is first tried on animals that could somewhat simulate human conditions. Trying a new drug on human subjects without establishing its efficacy and safety in laboratory and animal setups is considered unethical, and is almost never allowed.

The second basic feature of scientific experimentation is the meticulous control over the experimental conditions, which helps to draw inference on cause–effect relationships. This is done by following established **experimentation (statistical principles of)** over the units of experiment and the process. The unit of medical experiments is the biologic specimen or animal that receives or does not receive the intervention, and is observed for the anticipated changes.

The third basic feature of an experiment is that it is replicable. Since the experimental conditions are standard and the intervention well defined, it is not difficult for others to repeat the experiment and verify the results if they so want and have resources.

Although these features and other aspects are discussed in this section in the context of experiments on biological samples and animals, they are, by and large, applicable to **clinical trials** as well.

## *Advantages and Limitations of Experiments*

Despite being the most direct method of investigating cause–effect type of relationships, experiments do have some limitations. But first, the advantages.

As already stated, the basic advantage of an experiment is its ability to provide clear evidence of the effect of an intervention when properly designed and performed in standard laboratory conditions. An antecedent–outcome relationship can be obtained if present, although the underlying mechanism may not be clear—in this case, it will not be explained as cause–effect. Thus, a mere opportunity of an experiment generates awareness about complications involved in a medical relationship. An experiment is a good portal to test existing knowledge of the possible **confounders**, and it also opens up the possibility of considering unknown confounders that are in an **epistemic** domain. Thus, an experiment can broaden the possibilities of new explanations and new ways to solve a problem, howsoever limited.

Experiments on biological samples and sometimes on animals can also be done for harmful procedures and substances. Such an opportunity does not exist for human trials. Nobody would deliberately expose people to a pesticide to find how it affects health. This can be done with animals within the perimeter of ethics. Experiments that require sacrifice are also done for small-sized animals. See the following example.

To determine whether continuous or cyclic hormone replacement therapy (HRT) is better, Sun et al. [1] conducted an experiment on 142 Sprague–Dawley rats that were randomly divided into seven groups. Besides normal estrous and ovariectomized controls, the other five groups received treatments imitating clinical regimens with different combinations. The rats were sacrificed, and mitotic index and proliferating cell nuclear antigen (PCNA) were the outcome measures. Note the relevance of animal experimentation in this setup. The results suggest that the continuous regimen was better than the cyclic regimen in postmenopausal HRT.

The most serious limitation of an experiment is that it is carried out in near-ideal conditions that do not exist in the practical world. Thus, the results sometimes are not reproduced when applied in actual conditions. An experiment is necessarily context specific, and generalization is difficult; and even a fully internally valid experimental result may have low **external validity**. Experiments on biological material and animals rarely provide results that can be used on humans, although they do provide evidence one way or another to proceed to human experimentation. Thus, they indeed provide a valid base for clinical trials. A trial on nimodipine enrolled more than 7000 stroke patients, but the review of animal studies did not identify its protective role. Had animal studies properly been reviewed, this waste could have been avoided [2].

As stated earlier, experiments are sometimes done without understanding the biological processes involved in the anticipated relationship between antecedents and the outcome. In that case, the results may remain uninterpretable till such time that the basis of the relationship emerges.

1. Sun A, Wang J, Zhu P. How to use progestin in hormone replacement therapy: An animal experiment. *Chin Med J* (Eng) 2001;114:173–7. http://www.ncbi.nlm.nih.gov/pubmed/11780201
2. Champkin J. "We need the public to become better BS detectors" Sir Iain Chalmers. *Significance* July 2014; 25–30. http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2014.00751.x/pdf

## experimentation (statistical principles of)

The purpose of statistical principles of experimentation is to provide adequate control of the experimental conditions so that the effect of uncertainties is minimized and a valid conclusion can be drawn. These include control group, randomization, and replication. **Blinding** and **matching** are more relevant for human trials, and these are discussed under those topics. Nonstatistical principles include ethics. For ethics in animal experiments, see Monamy [1]. For humans, see the topic **ethics of clinical trials and medical research**, but the following also applies to human trials.

### *Control Group*

An important source of bias in all medical studies is the **Hawthorne effect**. In the context of experiments, this says that the experimental units tend to behave differently when they know that they are being observed. The same occurs when an inert procedure is used that otherwise should not produce any change. To account for this psychological effect, an experiment is almost invariably carried out by including a control group that does not receive the active ingredient of intervention but receives null stimulus or an inert substance such as saline injection or a placebo. All other procedures and maneuvers remain the same. In some experiments, the control group could receive the existing or prevalent intervention while the test group receives the new intervention. For example, in an experiment on a new drug that could relieve pain, the control could receive the existing drug after pain is induced in both groups of mice by some mechanism.

When a separate group of units receives the control regimen, this is called a parallel control group. Some experiments are done without a parallel control group. Measuring tail-flick latency in mice to heat exposure before and after an analgesic does not have a parallel control. Such an experiment is called a before–after study, where each unit serves as its own control. This is also called an uncontrolled trial, but there is a control in this setup as well, though not a separate group.

Control is required in an experiment so that the *net effect* of the intervention can be evaluated. Although the need for a control is not debated, what actually is required for control is a subject of debate. Units in the control group should be equivalent to the experimental units at baseline. Since medical experiments are on biologic material or on animals, this is not as much a restrictive requirement. For an experiment on the shape of red blood cells at various ionic strengths, various groups of cells including the control would be equivalent so long as they belong to the same genetic stock. It is because of such baseline equivalence and standardized conditions in the laboratory that experiments are able to provide compelling evidence of the presence or absence of a cause–effect type of relationship.

### *Randomization*

Randomization is a standard scientific procedure and now almost universally adopted in medical experiments. This envisages random allocation of units to different stimuli, one of which could be null—the control. Even though the units in all groups should be homogenous to begin with, randomization is advocated for providing a chance to the residual or unforeseen heterogeneity to equally divide among various groups. *The test group should be no different than the control group except for the intervention itself.* Then, any difference from control in the outcome can be legitimately attributed fairly to the intervention. The effect of chance, arising from including a limited number of units in the experiment, would be small when such equality is assured. Randomization also tends to remove the possible bias of the investigator in allocating units to different stimuli. In addition, randomization is a necessary ingredient for validity of the statistical methods needed to analyze the data.

Randomization apparently looks easy but actually can be difficult to implement. For example, assigning the even-numbered subjects in sequence to the test group and the odd to the control group has the potential to cause bias because it can be easily tempered. Assigning one group of mice to the test and the other to the control can also be biased because the first group may be of a specific type and the second group of another type. If an experiment is to be conducted on 30 mice with 15 each in the test and the control group, genuine randomization is best achieved by computer-generated random numbers. This applies to human trials also. The website http://www.randomization.com does this very efficiently and tells you which numbers will be in group 1 and which in group 2. If there are more than two groups, random allocation can be done accordingly by using the simple procedure provided at this site.

Randomization tends to work well in the long run but may fail for a particular experiment. The standard practice is to compare the baseline of the two groups after allocation to confirm that they are indeed equivalent with respect to the known factors. In the case of experiments on mice, they can be compared for age, sex, weight, or any other characteristics that might affect the outcome. Statistically, this equivalence can be confidently inferred only when each group has a reasonably large number of units. Also, beware that such equivalence is for averages or proportions in the two groups, and not for one-to-one matching. Average equivalence is enough for most experiments that aim to compare the average outcome in one versus the other group or groups.

### *Replication*

Replicability is a basic feature of any experiment, and it is through replication that considerable confidence is added when the same results are repeatedly obtained under similar conditions. Replication by another worker in a different setting helps to confirm that the results were not dependent on local conditions that were unsuspectingly assisting or hindering the outcome. Such replication also provides evidence for the robustness of the results. Consensus among the researchers and evidence from the experiments make a concoction that could hardly be denied in science.

Sometimes, the investigator himself/herself wants to replicate the experiment to be on firm footing. This is particularly desirable when the response varies widely from subject to subject. The variability across subjects remains unaltered, whereas replication reduces variability in experimental results. If the results between replicates are similar, their reliability naturally increases.

Replication fulfills an important statistical requirement. They help in quantifying the **random error** in the experimental results. Replication is also an effective instrument in reducing the effect of this error. Random errors occur due to imprecision of the process—in methods, measurements, etc., such as in staining, magnification, rotation, counting, and timing; in the environment, such as room temperature, humidity, or minor variation in chemicals and reagents; or due to the varying care adopted by different observers.

Replications help to estimate the effect of random errors on the outcome.

1. Monamy V. *Animal Experimentation: A Guide to the Issues*, Second Edition. Cambridge University Press, 2009.

## experiment-wise and comparison-wise error rate

Experiment-wise error rate is the combined probability of **Type I error** when several statistical **tests of significance** are done using the same set of data. On the contrary, the probability of Type I error committed by single test of significance is the comparison-wise error rate. The term *comparison* here is for a test of significance, and the term *experiment* is for a study. In most practical situations, several tests of significance are done in one study, and each test commits its own Type I error—thus, experiment-wise error rate becomes important.

Type I error is erroneously rejecting a null hypothesis when it is true. This is like punishing an innocent person, which can happen because of strong circumstantial evidence. You can see that this is a serious error compared with **Type II error**, which is accepting a false null (not being able to punish the guilty because of a lack of evidence). Type I error occurs because the sample happens to be such as to allow us to reject the null hypothesis. Because of its seriousness, we set the limit of the chance of Type I error to a low level, called **level of significance**, and denoted by $\alpha$. This is generally kept at 5%. When a test of significance is done on a set of data, the actual probability of Type I error is called ***P*-value**. The null is rejected when the *P*-value is less than $\alpha$. When two or more tests are done on the same data, each will have its own *P*-value. The combined *P*-value can become enormous when several tests are done on the same data. This means that the probability of Type I error can become too high. This brings in the concept of experiment-wise error rate. When several tests are done, it is advisable that the experiment-wise error rate is fixed at, say, a 5% level, which is the maximum for the combination of the tests, and the individual tests are done with a much more restricted level of significance to keep the combined error rate within the limit.

As an illustration, if the level of significance for one test is $\alpha$, the combined level of significance for $K$ tests inflates to $[1 - (1 - \alpha)^K]$. This is the experiment-wise error rate for $K$ tests. If $\alpha = 0.05$ and you have done $K = 6$ tests on the same data, the combined level of significance is $[1 - (0.95)^6] = 0.26$. This is the threshold of the probability of Type I error you are allowing in the study results. This high chance of error is not acceptable by any standard. If you want to keep this chance limited to not more than 0.05 for $K = 6$ tests, then $[1 - (1 - \alpha)^6] = 0.05$ gives $\alpha = 0.0085$. Thus, the level of significance of each test should be 0.0085. In a way, this highlights the need to keep the number of statistical tests to minimum. This also is the principle behind **multiple comparisons** in analysis of variance (ANOVA) when each group is compared with each of the others.

## expert systems

Expert systems in medicine are computer aids for diagnosis, treatment, and prognostic assessments of patients. They take signs and symptoms and other patient details as inputs and provide guidance on what possibly could be the ailment with what probability, what investigations to order, how to use these results to modify the diagnosis, what possible treatments can be advised, and what other advice can be given to the patients. They tend to take away the element of forgetting something and help in supplementing the information base. Thus, expert systems partly address **epistemic uncertainties**.

Epistemic uncertainties arise not only because of universal ignorance about some biological processes but also substantially from the failure to judiciously apply the available knowledge. The knowledge base in medicine is enormous, and it is rapidly increasing, making it increasingly difficult for a physician to remember and recall everything about a disease at the time of facing a patient. Errors while prescribing drugs are also not uncommon. Computers have tremendous capacity to store information in a systematic manner and to retrieve it selectively as needed at an instant's notice. They can be programmed to take signs and symptoms as inputs and provide prompts on likely diagnoses along with the probability of each. Also, laboratory, radiological, and other investigations can be suggested that might help in focusing on or excluding a specific diagnosis. Computer alerts on alternative strategies for treatment, as also on the prognostic indicators, can be obtained after a plausible diagnosis is identified. For this, a database is prepared containing various possibilities, and rules are devised to selectively retrieve the information after proper matching (for details, see Giarratano and Riley [1]). All this is put together in a software package and is called an expert system. The success of an expert system largely depends on (i) acquiring wide and valid information; (ii) articulated pooling of this information into a knowledge base; (iii) devising correct rules for diagnosis, treatment, and prognosis; and (iv) adequate programming that uses the knowledge base and rules effectively for practical management. If these are not fulfilled, an expert system may give a false sense of security.

An expert system can have an interface with the introduction of newly emerging diagnostic and treatment strategies so that they are automatically incorporated when approved. It can also warn against the possibility of a drug reaction, allergy, or overdose. The function of this system is no more than a reminder on the basis of the inputs, and the decision always remains with the physician. He/she may or may not agree with the investigation, diagnosis, and treatment suggestions of the expert system.

An appropriately prepared expert system can be of considerable help in reducing errors in medical decisions. The advantage of such an expert system is that it is capable of processing a large amount of information without error as per the program or the knowledge given to it by a group of experts after considerable discussion. An individual clinician may not know as much. As always, there are *if*s and *but*s. Medical care is much more than just diagnosis and treatment. Timely detection and timely intervention, adequate care of the patient, proper advice on prognostic implications, etc., are important ingredients, which may not be properly accounted for by an expert system. A patient's own preferences and perceptions also matter. Deficiency in one can cause a chain reaction and upset the whole process. Simultaneous consideration of all such factors is computationally difficult. Thus, errors can occur with an expert system also.

Expert systems are easy to talk about but are extremely difficult to develop. No expert system can be better than the expertise given to it. This implies that the system should be based on the knowledge of real experts. They are rare, and each expert is a specialist of one's own subdiscipline. To put various experts together to develop a comprehensive computer-based system has turned out to be almost an unachievable ideal. For this reason, efforts remain limited to specific diseases. Some are briefly described in the examples given as follows.

Irrespective of the expertise of the system, it is seldom able to think as critically as a human mind can. It restricts itself to the structured inputs and can incorporate only explicit knowledge. Tacit knowledge rooted in context, experience, social values, etc. resides in the human mind but is hard to deliver to an expert system. An expert system can never be helpful in a situation that has not been thought of at the time of its preparation. Such situations

can always arise because of highly individualistic interactions of various biological and environmental factors. An expert clinician can perceive alternative scenarios for a particular condition and can immediately react to an unforeseen observation, but an expert system cannot. Thus, sufficient caution should be exercised while using any expert system. They must be considered only as an aid and not as a guide. The decision at each step has to be entirely that of the attending clinician; he/she may or may not agree with the suggestion of the expert system. An expert system cannot replace a clinician but can act as a supplementary, perhaps a very useful supplementary, when prepared with a sound knowledge base and correct rules. As of now, the nature of the help of an expert system is the same as that of laboratory and radiological investigations.

For the reasons just explained, not many good expert systems are available. Most of those in the market are of dubious quality and should be used with due care. The purpose of introducing expert systems here is to explain that they can really help in minimizing uncertainties, both in clinical and research setups, when properly developed and judiciously used. But that is a big *if*!

Here are the briefs of some expert systems.

- Differential diagnosis of disorders and diseases manifested by tall stature [2]: For this the diagnostic criteria were developed by a panel of seven experts. In addition, manuals and textbooks, databases, and online resources were also consulted. Linguistic terms were also studied. An interface was made up by a set of slides. The major sources of information were the London Dysmorphology Database and Orphanet. The expert system produces the five most probable diagnostic possibilities and ranks them in order of likelihood depending on the inputs provided to it. These inputs are based on the information provided by the patient and the initial assessments by the clinician.
- Automated visual fields: Feldon et al. [3] used visual fields from eyes with 189 nonarteritic anterior ischemic optic neuropathy to develop a computerized classification system using nonischemic optic neuropathy decompression. The expert panel had six neuro-ophthalmologists, who described definitions for visual field pattern defects using 19 visual fields and several levels of severity representing a range of pattern defect types. The expert panel subsequently used 120 visual fields to revise the definitions. These were converted to a rule-based computerized classification system. The system was subsequently used to categorize visual field defects for an additional 95 nonarteritic anterior ischemic optic neuropathies. Agreement with the experts was not really high, and further modification was done.

The two expert systems cited are not for complex conditions and are focused on specific conditions. Yet, they illustrate the kinds of problems encountered in developing an expert system.

Among others that you may want to review are (i) a decision support system to improve clinicians' interpretation of an abnormal liver function test [4], (ii) a three-stage expert system based on support rector machines for thyroid disease diagnosis [5], and (iii) a medical expert system for the diagnosis of ectopic pregnancy [6].

1. Giarratano JC, Riley GD. *Expert Systems: Principles and Programming*, Fourth Edition. Course Technology, 2004.
2. Paghava I, Tortladze G, Phagava H, Manjavidze N. An expert system for differential diagnosis of tall stature syndrome. *Georgian Med News* 2006;131:55–8. http://www.ncbi.nlm.nih.gov/pubmed/16575134
3. Feldon SE, Levin L, Scherer RW, Arnold A, Chung SM, Johnson LN, Kosmorsky G et al. Development and validation of a computerized expert system for evaluation of automated visual fields from the Ischemic Optic Neuropathy Decompression Trial. *BMC Ophthalmol* 2006;6:34. http://www.biomedcentral.com/1471-2415/6/34
4. Chevrier R, Jaques D, Lovis C. Architecture of a decision support system to improve clinician's interpretation of abnormal liver function test. *Stud Health Technol Inform* 2011;169:195–9. http://www.ncbi.nlm.nih.gov/pubmed/21893741
5. Chen HL, Yang B, Wang G, Liu J, Chen YD, Liu DY. A three-stage expert system based on support vector machines for thyroid disease diagnosis. *J Med Syst* 2011; Feb 1. http://link.springer.com/article/10.1007%2Fs10916-011-9655-8#page-1
6. Kitopporntheranunt M, Wiriyasttewong W. Development of medical expert system for the diagnosis of ectopic pregnancy. *J Med Assoc Thai* 2010; 93(Suppl 2):S43–9. http://www.ncbi.nlm.nih.gov/pubmed/21302398

## explanatory and predictive models

An explanatory model [see **models (statistical)**] is the one that describes an outcome in terms of antecedents—leading to clues on how the outcome may have occurred. In contrast to this is a predictive model, which also applies to antecedent–outcome relationships, but in in this model, the choice of the predictors is not important—the only objective is to be able to predict the outcome. In the case of an explanatory model, the choice of antecedents is important as they reflect the **etiological factors** and not just any associative or correlative factors. Etiological factors can also be called determinants. These are not required for predictive models.

You may notice a considerable overlap in usage of these terms in medical literature. A good explanatory model is always a good predictive model, but a good predictive model is not necessarily a good explanatory model. This is because the antecedents used for prediction may not be etiological factors. Blood pressure level in nonhypertensive subjects can be predicted by age, sex, and socioeconomic status, but also, almost equally well, by fat intake, exercise, and genetic history. Predictive models are not unique—two models with different antecedents can be equally good for prediction. This is not likely to happen with an explanatory model since the antecedents would be those that determine the outcome.

Chi et al. [1] used a model to find that decayed, missing, and filled surfaces (dmfs), not having a dental home, low caregiver education, and living in a nonfluoridated community are the factors for development delays in low-income preschoolers in the United States. If it is an explanatory model, these factors are believed to be the determinants and not just associated factors. On the other hand, Wopken et al. [2] developed a prediction model of tube feeding dependence after curative radiation in head and neck cancer. The most important predictors were weight loss prior to treatment, advanced T-stage, positive N-stage, bilateral neck irradiation, accelerated radiotherapy, and chemoradiation. Some or all of these factors may be determinants, but since this is not investigated, it is prudent to call it a predictive model.

It is difficult to hypothesize factors as determinants and not just correlates. For blood pressure, income could be a good correlate, but it is difficult to say that it determines, even partially, the blood pressure level. For a factor to qualify as a determinant, you may have to go a step ahead of association and establish a **cause–effect relationship**. Thus, the criteria for this kind of relationship should be largely fulfilled. When this is done, the model with these antecedents could be called an explanatory model. Apparently, this kind of exercise was not undertaken by Chi et al., quoted in the preceding paragraph.

Thus, it looks unfair to call it an explanatory model. This illustrates our assertion that there is confusion in the literature about the terms predictive model and explanatory model.

Automatic selection of variables such as by stepwise methods may not be a good strategy for explanatory models because then the effect is adjusted only for those regressors that remain in the model after such statistical selection. However, these methods are generally adequate for predictive models.

1. Chi DL, Rossitch KC, Beeles EM. Developmental delays and dental caries in low-income preschoolers in the USA: A pilot cross-sectional study and preliminary explanatory model. *BMC Oral Health* 2013 Oct 12;13:53. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3906997/
2. Wopken K, Bijl HP, van der Schaaf A, Christianen ME, Chouvalova O, Oosting SF, van der Laan BF et al. Development and validation of a prediction model for tube feeding dependence after curative (chemo-) radiation in head and neck cancer. *PLoS One*. 2014 Apr 15;9(4):e94879. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3988098/

## explanatory variables, *see* variables

## exploratory data analysis, *see also* exploratory studies

As the term implies, exploratory analysis is just to explore the data and not to come to any conclusion. The term was first used by John Tukey of Tukey test fame. Finding out the characteristics of a data set is exploratory analysis. This includes calculating the mean and standard deviation (SD), finding out the shape of the frequency **distribution**, locating the **outliers** or wrongly entered values, making two-way or complex tables to explore any relationship, making some graphs for visual exploration, etc. All this comes under the rubrics of **descriptive analysis**. This is an open-ended exercise as anything that comes to mind can be explored. This not only helps to understand the data but also can help in cleaning and filtering the data, thus prepare them for the analysis. Exploratory data analysis sets the tone for the real statistical analysis with methods such as regression, analysis of variance (ANOVA), discriminant analysis, and tests of significance. In some situations, though, exploratory analysis would include so-called advanced analysis with some tentative outcome or tentative method to see if that gives any result worth pursuing.

Most biostatistical methods are valid under certain conditions. For example, ANOVA is valid under **Gaussian conditions**, which means that either the distribution of the outcome variable must be Gaussian or the sample size must be large, and the values must be independent of each other. This requires that the data be first *explored* to confirm that these conditions are fulfilled. **Logistic regression** analysis requires that at least 10 values are available for each cross-section of the independent variable categories. Exploratory data analysis will tell if this really is so in your data. If not, you may want to reduce the number of independent variables or the number of categories. Otherwise, exploratory analysis can also indicate which variables could be appropriate for inclusion among the independents in a regression if that is not already committed in the **protocol**.

There seems to be some confusion in the medical literature between exploratory study and exploratory analysis. For example, Liang and Wu [1] used multinomial logistic regression, probit regression, and structural equation model methods to analyze data on health-related quality of life yet call it exploratory analysis. What they possibly mean is exploratory study as an exploratory analysis rarely can be so intensive. See the topic **exploratory studies** for details.

1. Liang Y, Wu W. Exploratory analysis of health-related quality of life among the empty-nest elderly in rural China: An empirical study in three economically developed cities in eastern *China*. *Health Qual Life Outcomes* 2014 Apr 25;12(1):59. http://www.hqlo.com/content/12/1/59/abstract

## exploratory studies

Exploratory studies are those preliminary studies that are done to explore whether or not a particular topic is worth studying with a full-scale study. Such studies, by their very nature, are generally carried out on a relatively small sample, may not have a perfectly valid **design**, and may also have restricted coverage in terms of variables. Concepts such as statistical **power** and the statistical **significance** of the differences and relationships do not strictly apply to exploratory studies. An exploratory study can examine multiple end points to be able to choose a few for the final study and can also be geared to provide information on what type of subjects should be included, what exact intervention (if any) can be studied, and what type of measurements and by whom would be appropriate. In summary, this kind of study provides a basis for planning and carrying out the final study on a much firmer footing. In some situations, a large-scale study on a new concept is also called exploratory when the objective is to find the proof of concept so that this concept can be pursued further by other similar studies by different workers at different locations.

Many studies in medical journals are termed exploratory because of their own limitations. Muratori et al. [1] called their study exploratory when they studied outcome in 70 children with autism who received usual treatment probably because the sample size was small. Wan et al. [2] also called their study exploratory while studying oral mucosal colonization of human gastric *Helicobacter pylori* in 60 mice since the objective was to explore whether buccal mucosa in mice could serve as a workable animal model for further research on this topic. Some researchers use the term *exploratory* for their study as an abundant caution lest their results be challenged. This is just to express a lack of confidence in the results either because the study was not properly done or because the idea put forth is too radical to be accepted right away. Thus, the term can have varied usage.

1. Muratori F, Narzisi A, IDIA group. Exploratory study describing 6 month outcomes for young children with autism who receive treatment as usual in Italy. *Neuropsychiatr Dis Treat* 2014 Apr 8;10:577–86. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3986291/
2. Wan X, Tang D, Zhang X, Li H, Cui Z, Hu S, Huang M. Exploratory study of oral mucosal colonization of human gastric *Helicobacter pylori* in mice. *Int J ClinExp Med* 2014 Mar 15;7(3):523–9. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3992389/

## exponential curve/function, *see also* exponential distribution

Exponent is the mathematical term for the power of a number. For example, in $3^6$, 6 is the exponent. As you can see, the function of an exponent is to multiply. Thus, the series 2, 4, 8, 16 is exponential since each term is twice its previous term, whereas the series 2, 4, 6, 8 is **linear** since each term is two more than the previous. If the exponent is more than 1, the numbers rapidly rise—$20^2 = 400$ and $30^2 = 900$. If the exponent is less than 1, they gradually fall. For example, $64^{0.5} = 8$, and $81^{0.5} = 9$. The difference between 64 and 81 is large, but the difference between 8 and 9 is small because the exponent is less than 1.

In health and medicine, we will be mostly concerned with outcomes that have exponents greater than 1. Bacterial growth in the absence of any intervention is exponential. It only means that the number of bacteria multiplies and does not increase additively. The curve depicting this phenomenon is called an exponential curve, and the equation describing this is called an exponential function or exponential model. An exponential curve will have a slow rise in the beginning and steep rise later (Figure E.7). Thus, it shows accelerated growth. Scott et al. [1] reported that the volume of fetal cerebellum increased approximately sevenfold from 20 to 31 gestational weeks, and the exponential curve is a better fit than a line. This may not be as steep as shown in Figure E.7, but for example, intrauterine pressure follows nearly the same steep rise with gestational age [2] as shown in in this figure. Consider Ebola virus cases in the world, which have exponentially increased from 9 in March 2014 to 14,500 in November 2014. These are the reported cases, and the unreported ones may be much higher. The plot of the number of cases with time will be similar to Figure E.7.

Simply stated, the exponential function has the following form.

$$\text{Exponential function: } y = ae^{bx},$$

where $e$ stands for the Naperian base and $x > 0$. This base is the usual mathematical constant used in exponential functions since it helps to use natural logarithms. In terms of logarithm, this function is $\ln(y) = \ln(a) + bx$. The equation now becomes additive and not multiplicative. Logarithmic transformation converts exponential function into a linear equation. The shape and location of the exponential curve, for example, whether it will rise steeply or slowly, will depend on the values of $a$ and $b$. If the rise in the value is rapid and steep, the value of $b$ will be large. The value of $b$ is positive for exponential growth and negative for exponential decay, where decay means declining trend. These parameters may or may not have biological meaning depending on the application. For example, for the area of opacity of cataract eye lenses (in terms of pixels) at time $t$ (days), $O_t$ can be expressed as [3]

$$O_t = O_m(1 - e^{-k(1-t)}),$$

where $O_m$ is the opaque area at maturity. The constant $k$ can now be interpreted as the opacification rate. This equation has been used to study the effect of particular eye drops on opacification rate. Tumor growth in cancer is frequently studied using exponential growth models. For one such study for tumors in lung cancer, see Li et al. [4].

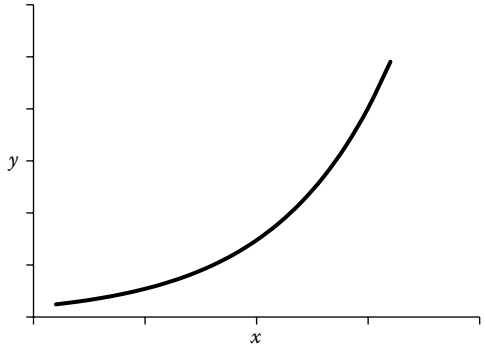



**FIGURE E.7** An exponential curve.

1. Scott JA, Hamzelou KS, Rajagopalan V, Habas PA, Kim K, Barkovich AJ, Glenn OA, Studholme C. 3D morphometric analysis of human fetal cerebellar development. *Cerebellum* 2012 Sep;11(3):761–70. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3389138/
2. Sokolowski P, Saison F, Giles W, McGrath S, Smith D, Smith J, Smith R. Human uterine wall tension trajectories and the onset of parturition. *PLoS One* 2010 Jun 23;5(6):e11037. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2890413/
3. Nagai N, Ito Y, Takeuchi N. Pharmacokinetic and pharmacodynamic evaluation of the anti-cataract effect of eye drops containing disulfiram and low-substituted methylcellulose using ICR/f rats as a hereditary cataract model. *Biol Pharm Bull* 2012;35(2):239–45. https://www.jstage.jst.go.jp/article/bpb/35/2/35_2_239/_pdf
4. Li M, Jirapatnakul A, Biancardi A, Riccio ML, Weiss RS, Reeves AP. Growth pattern analysis of murine lung neoplasms by advanced semi-automated quantification of micro-CT images. *PLoS One* 2013 Dec 23;8(12):e83806. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3871568/



## exponential distribution, *see also* exponential curve/function

A quantitative variable $x$ is said to follow an exponential distribution when its chances of occurrence show a sharp rise or sharp fall with the increasing value of $x$. For example, the probability of death after the age of 70 years steeply rises as age increases. This probability never comes down. Thus, the age at death after 70 years follows nearly an exponential distribution. One of the possible shapes of the distribution is the same as that of an **exponential curve**—the difference being that the vertical axis in the case of **distribution** is the **probability density**, which measures the proportion of subjects, whereas it can be any variable in the case of an exponential curve. The area under any statistical distribution is 1, so is with exponential distribution. As a medical professional, you may never need the following, but just for information, note that

$$\text{exponential distribution: } f(x) = \lambda e^{-\lambda(x-\alpha)},\ \lambda > 0, \text{ and } x \geq \alpha.$$

In this equation, the minimum value of $x$ is denoted by $\alpha$. In our example on age at death, $\alpha = 70$ years. The value of $\lambda$ determines the steepness of the rise in probability of the event (death in our example) with $x$ (age in our example). When $\alpha = 0$, the mean of this distribution is $1/\lambda$, and the standard deviation is also $1/\lambda$. Thus, this essentially is a single-parameter distribution, namely, the average number of events per unit interval. In the case of age at death for persons alive at age 70 years, the distribution takes a slightly different shape (Figure E.8a). A special feature of an exponential distribution is that the probability density at subsequent values of $x$ multiplies with a constant no matter what the preceding value is.

Exponential distribution is frequently used for the length of intervals between two consecutive random events, such as admission and discharge from a hospital, two episodes of diarrhea or two episodes of angina, and two births. Many survival durations also follow an exponential distribution. The probability of surviving for long, particularly after a terminal disease such as cancer, steeply declines as the duration increases and the shape reverses (Figure E.8b). Both are exponential. One is an exponential rise, and the other is an exponential fall. Leroy et al. [1] explored exponential distribution, among others, for time to onset of adverse drug reactions from treatment exposure. Closas et al. [2] observed that the incidence of influenza in nonepidemic periods follows an exponential distribution.

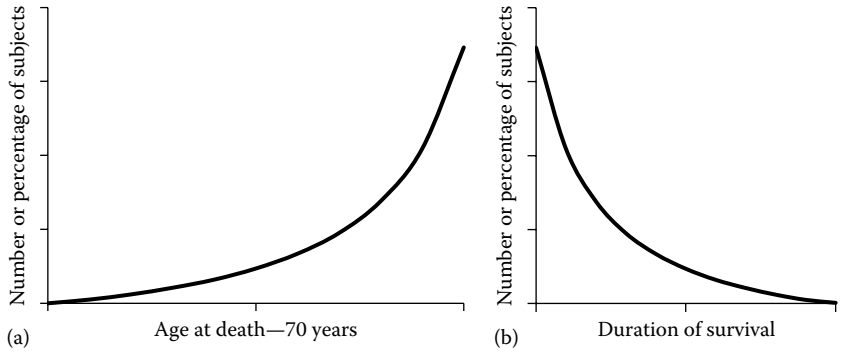



**FIGURE E.8** (a) Exponential distribution of age at death for the persons alive at age 70 years; (b) exponential distribution of survival duration after a terminal disease.

When a variable follows an exponential distribution, you cannot use **Gaussian distribution**–based methods such as Student *t* and analysis of variance. Log transformation of data may help in such cases.

1. Leroy F, Dauxois JY, Théophile H, Haramburu F, Tubert-Bitter P. Estimating time-to-onset of adverse drug reactions from spontaneous reporting databases. *BMC Med Res Methodol* 2014 Feb 3;14:17. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3923259/
2. Closas P, Coma E, Méndez L. Sequential detection of influenza epidemics by the Kolmogorov–Smirnov test. *BMC Med Inform Decis Mak* 2012 Oct 3;12:112. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3557152/

## extrapolation, *see also* interpolation

Extrapolation is the scientific estimation of a value outside the range of the actually observed values. In the context of statistics, a model such as a **regression equation** is constructed on the basis of the observed values, and this model is used to extrapolate unobserved values. It is presumed that the unobserved values outside the range of observed values will follow the same pattern as the observed values, although this can be a big *if* in some situations. In the case of time series, extrapolation refers to the prediction of the future values on the basis of the trend of the past values. Extrapolation requires the following precautions.

- The model must be good to begin with in the sense that it describes the dependent variable with reasonable **validity** and **reliability**.
- Extrapolation can be done for values slightly outside the range of observed values because values grossly outside may not follow the same pattern—this would not be known, since these values have not been studied.
- Keep in mind that it is just an estimated value and is subject to errors and uncertainties.

Wodarz et al. [1] have discussed how extrapolating the risk of low-dose radiation from a model on high doses can be wrong because the cellular response can be very different at low doses. Thus, there are hazards in extrapolation, and it cannot be indiscriminately done. Nagao [2] studied the relationship between Kawasaki disease and decrease in **total fertility rate** (TFR) in a **time series** from 1979 to 2011 in Japan and concluded by extrapolation that this disease emerged in the 1960s because TFR dramatically decreased in the 1940s through the 1950s. That sounds too far back but may still be true if the trend was the same at that time also.

Extrapolation is kind of the reverse of **interpolation**, where unobserved values within the range of observed values are estimated.

1. Wodarz D, Sorace R, Komarova NL. Dynamics of cellular responses to radiation. *PLoS Comput Biol* 2014 Apr 10;10(4):e1003513. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3983039/
2. Nagao Y. Decreasing fertility rate correlates with the chronological increase and geographical variation in incidence of Kawasaki disease in Japan. *PLoS One* 2013 Jul 8;8(7):e67934. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3704585/

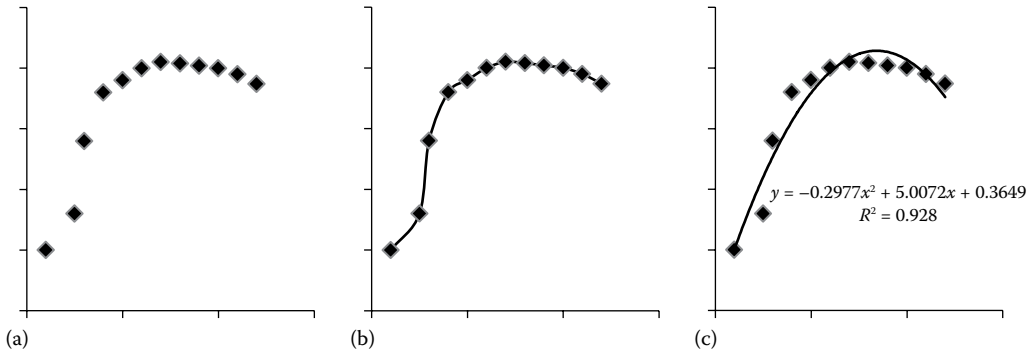



**FIGURE E.9** (a) Actual points; (b) eyeball fit; (c) second-degree polynomial fit.

## eyeball fit

The term *eyeball fit* is generally used for a graphical trend where the equation is not obtained but the **scatter plot** is used to guess the pattern of the trend using our own sensory expertise. If this expertise is sufficiently good, the eyeball fit may not be much different from the trend obtained through statistical equations. In some situations, particularly when graphical representation serves the purpose, the eyeball fit may be even better than the equation-based trend. In many situations, eyeball fit provides an insight into what equation we should be looking for. This is quick and does not require any computation. However, this is not science and is not easily acceptable.

Consider the data points in Figure E.9. The eyeball fit in Figure E.9b has a jug-handle shape with a flat top, whereas the second-degree polynomial curve in Figure E.9c does not have a jug-handle bend, and the peak is relatively sharp. Statistically, there is nothing wrong with the polynomial fit as the square of the **multiple correlation coefficient** $R^2 = 0.928$ obtained for this curve is extremely good and gives confidence in the curve. The number of points in this example is too small for any firm conclusion, but the bend in the jug handle and the flat top may be real and better for explaining the biological phenomenon under study. If knowledge and gut feeling so dictate, an equation that incorporates these two features can be obtained. If such an equation seems difficult, the eyeball fit can still be used to graphically explain how the bend and the flatter peak could be real. This is the advantage of the eyeball fit. This can be explored particularly in cases where the right equation is difficult to conceive and obtain.

# F

## face validity, *see* validity (types of)

## factor analysis, *see also* factor scores

Factor analysis is a statistical method to identify unobservable underlying factors that possibly lie behind the observed values of several correlated **continuous variables**. The term *factor analysis* was coined by Charles Spearman in 1904, and the method is also generally ascribed to him [1]. How do you assess the health of a child? You will see his/her height, weight, skinfold, eyes, nails, etc. Health is not directly observable but is measured through a host of such indirect variables. Conversely, you may have a variety of variables, and you want to know what factors are underlying those variables to cause those values. There are situations where the observed variables can be considered to be made up of a small number of unobservable, sometimes abstract, factors. These factors can also be considered as **constructs** that are latent in the data. A latent construct can be measured indirectly by determining its influence on responses on measured variables. The purpose of factor analysis is to unravel these factors.

Health in its comprehensive form (physical, social, mental, and spiritual) can be measured in adults by a host of variables, such as the kind and severity of medical complaints if any, obesity, lung functions, smoking, sexual behavior, marital relations, job satisfaction, unmet aspirations, income, and education. It is not apparent how much of, say, job satisfaction is ascribable to physical health; how much to each of the other components (social, mental, and spiritual) of health; and how much is the remainder that cannot be assigned to any of these components. For a variable such as obesity, the physical component may be dominant, and for a variable such as education, the social component may be dominant. The technique of exploratory factor analysis may sometimes be used to identify such underlying factors.

### *Exploratory Factor Analysis*

The theory of exploratory factor analysis presumes that the variables being studied share some unknown number of common factors that give rise to **correlations** between them. Nothing is considered known about the nature of the relationships. The strength of the explanatory factor analysis method is that it identifies relatively few underlying factors. The variables actually measured are considered as manifestations of these factors, so the observed variables are now better understood as **surface attributes**. On the other hand, the underlying factors are the **internal attributes**. The primary objective of factor analysis is to determine the number and nature of the underlying internal attributes, and the pattern of their influence on the surface attributes. In a successful factor analysis, a few factors would adequately represent relationships among a relatively large number of variables. These factors can then be used subsequently for other inferential purposes.

Factor analysis is easy to understand through the **principal components** approach. In this approach, the first linear component is discovered that is able to account for maximum variation among observed variables. The second linear component is obtained from the remaining variation, and so on. The first principal component may be able to take care of 62% of total variation, the second 23%, and the third only 8%.

Consider two variables—height and weight—whose essence is captured by body mass index (BMI) for many applications. This combined variable provides most of the information contained in two different but correlated variables. BMI is not a linear combination of height and weight, but there are situations where most of the information contained in two or more variables can be captured by one linear combination of them. Similarly, when really successful in capturing the essence of several variables, principal components can reduce the number of variables, called *data reduction*—or more appropriately, *dimensionality reduction*.

### *Statistical Procedure for Factor Analysis*

The statistical procedure of factor analysis is to obtain each observed variable as a combination of a few unobservable factors, i.e.,

observed value of a variable = linear combination of factors + error.

If the observed variables are $x_1, x_2, \ldots, x_K$, the factor analysis seeks the following:

$$\left.\begin{aligned} x_1 &= a_{11}F_1 + a_{12}F_2 + \cdots + a_{1M}F_M + U_1, \\ x_2 &= a_{21}F_1 + a_{22}F_2 + \cdots + a_{2M}F_M + U_2, \\ &\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ &\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ x_K &= a_{K1}F_1 + a_{K2}F_2 + \cdots + a_{KM}F_M + U_K, \end{aligned}\right\}$$

where $F_1, F_2, \ldots, F_M$ are the $M$ unobservable factors common to $x_k$s, and $U_k$s ($k = 1, 2, \ldots, K$) are called *unique factors*. A schematic representation is in Figure F.1, where the number of $x$ variables is six

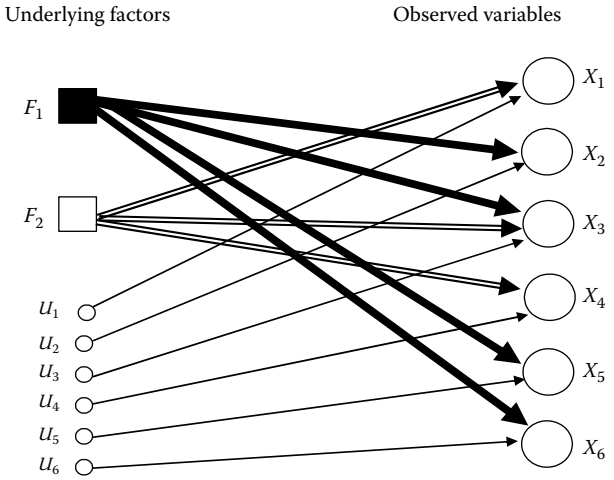



**FIGURE F.1** Schematic representation of factors.

but the number of factors is two. Unique factors represent the part that remains unexplained by the factors.

The coefficients $a_{km}$ ($k$ = 1, 2, …, $K$; $m$ = 1, 2, …, $M$; $M \ll K$) are estimated by the factor analysis procedure. These coefficients are called **factor loadings** and measure the importance of the factor $F_m$ in the variable $x_k$ on a scale of 0 to 1. When a loading is small, say less than 0.20, the corresponding factor is dropped from the previous equation. Thus, different variables may contain different sets of factors. Some factors would overlap and would be present in two or more variables. In Figure F.1, factor $F_1$ is common to $x_2$, $x_3$, $x_5$, and $x_6$, and $F_2$ is common to $x_1$, $x_3$, and $x_4$. Each $x$ has its own specific $U$. See Streiner and Norman [2] for more details.



The aforementioned set of equations mentioned differs from the usual **multiple regression** equations because the $F_m$s are not single independent variables. Instead, they are labels for combinations of variables that characterize these constructs. They are obtained in such a manner that they are uncorrelated with one another. The statistical method of principal components is generally used for this purpose. For details, see Kline [3]. Statistical software easily provides an output but might ask you to specify different aspects of the methodology. You should not undertake this exercise yourself unless the intricacies are clear. Do not hesitate to obtain the help of a biostatistician when needed. However, the following basic steps may clarify the steps required for this exercise.

Step 1. Obtain the correlation matrix of all $x$ variables you have observed. This is just a square arrangement of the correlation of these variables with one another.

Step 2. Check that correlations are sufficiently high among subsets of variables. You can use the **Bartlett test** of sphericity to test the null hypothesis that the correlation matrix is an identity. Identity means that the diagonal elements are 1 (correlation of $x_k$ with $x_k$ is always 1) and the off-diagonal elements are 0 (correlation of $x_k$ with any other $x$ is 0 for all $k$). You can expect good results from factor analysis when the Bartlett test is statistically significant.

Step 3. Check that the variables indeed share some common factors that are probably giving rise to these correlations. This can be checked as follows for linear correlations: **Multiple correlation** of each $x_k$ with other $x$'s should be high. The **partial correlations**, which are correlations between two variables after the effect of others is adjusted, should be low.

Two tests to find whether the data set you are examining is suitable for factor analysis are the **Kaiser–Meyer–Olkin** (KMO) [4] measure and the Bartlett test of sphericity. We have already mentioned the Bartlett test. The KMO measure is based on partial correlation and interpreted as follows: <0.5, unacceptable; 0.51–0.60, poor; 0.61–0.70, mediocre; 0.71–0.80, middling; 0.81–0.90, good; 0.91–1.00, excellent.

Step 4. Once convinced from steps 1–3 that an appropriate setup to try factor analysis exists, enter the correlation matrix into factor analysis software. The output will give you factor loadings and a host of other information that can help to identify the factors.

## Features of a Successful Factor Analysis

There are some steps in factor analysis that are not considered fully scientific. Despite this, the technique is popular with social scientists and is now making inroads into medical sciences as well. Beware that the technique sometimes fails to identify meaningful factors. Its success is assessed on the basis of the following considerations.

One of the steps in factor analysis is breaking down the total variation among variables into variations accounted for by different factors. The analysis is considered successful when only a few factors are able to account for a large part of the total variation, say more than 70%. A criterion for successful factor analysis is to be able to find some factors with very high loadings (close to ±1) while the others have very low loadings (close to 0). Also, the factors should largely be distinct or at least nonoverlapping. To achieve this, the technique of rotation of axes is sometimes adopted. Besides the popular **varimax rotation**, other rotations are quartimax and equamax. The rotation sometimes helps to extract factors with selectively high loadings and thus to obtain interpretable factors.

Perhaps the most important decision in factor analysis is the choice of the number of factors to be extracted. The statistical procedure theoretically is such that as many factors can be extracted as the number of variables. But only a few factors would be important. This importance is generally assessed by the *eigenvalues* for the factors. Eigenvalues depend on the correlation structure among the variables. Factors with eigenvalues greater than 1.0 can be considered important because they explain more variance than is explained by a single variable. When common factors are indeed present, only a few factors are likely to have this property. If the number of factors so identified is more than you think it should be, the threshold of the eigenvalue can be raised from 1.0 to 1.5 or any other suitable number. The other alternative is a **scree plot** of eigenvalues. The points where this plot levels off can be regarded as indicative of the number of factors.

Thanakwang et al. [5] carried out factor analysis of a 36-item scale for active ageing in Thai adults. The KMO value was 0.933, and Bartlett's test of sphericity was also highly significant. Both indicated the appropriateness of the data for further factor analysis. Seven factors with eigenvalues greater than 1 were identified. These accounted for 68.53% of the total variance.

## Confirmatory Factor Analysis

The method described in the preceding paragraphs is mostly for exploratory purposes. This method does not assume any preconceived structure of the factors. If the objective is to assess whether a set of variables conforms to a *known* structure of the factors, then confirmatory factor analysis is needed. This requires that the model is fit to the data and its adequacy is evaluated. Confirmatory factor analysis requires a much bigger sample size and is much more difficult than exploratory factor analysis. This can be done as a part of **structural equation modeling** [6].

For more details of both types of factor analysis, see O'Rourke and Hatcher [7].


1. Richard H, Williams RH, Zimmerman DW, Zumbo BD, Ross D. Charles Spearman: British behavioral scientist. *Hum Nat Rev* 2003(12 March);3:114–8. http://human-nature.com/nibbs/03/spearman.html
2. Streiner DL, Norman GR. *Health Measurement Scales, A Practical Guide to their Development and Use*, Second Edition. Oxford Medical Publications, 1995.
3. Kline P. *An Easy Guide to Factor Analysis*. Routledge, 1994.
4. Morrison DF. *Multivariate Statistical Methods*. Fourth Edition. Duxbury, 2004.
5. Thanakwang K, Isaramalai SA, Hatthakit U. Development and psychometric testing of the active aging scale for Thai adults. *Clin Interv Aging* 2014 Jul 24;9:1211–21. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4116362/

6. Schreiber JB, Stage FK, King J, Nora A, Barlow EA. Reporting structural equation modeling and confirmatory factor analysis results: A review. *J Educ Res* 2006;99(6):323–37. http://www.jstor.org/stable/27548147?seq=1#page_scan_tab_contents
7. O'Rourke N, Hatcher L. *A Step-by-step Approach to Using SAS System for Factor Analysis and Structural Equation Modeling*, Second Edition. SAS Institute, 2013.

## factorial and partially factorial designs

A design of an experiment that includes combinations of the levels of two or more factors under experiment that can be assigned to any unit of the study is called a factorial design. The outcome under study must be quantitative, and the two factors must be amenable to simultaneous administration. When all possible combinations of the levels of various factors are included in an experiment, it is called *fully* factorial. The *fully factorial design* answers the following question: what *combination* of levels of various factors is the most effective? Statistically, a factorial design has two or more categorical independent variable and a single quantitative dependent variable.

Figure F.2a represents a full-factorial design of 3 factors, each at 3 levels. There are a total of $3^3 = 27$ combinations, such as (0, 0, 0), (0, 0, 1), (0, 0, 2), etc., where 0, 1, and 2 are the *levels of the factors*. Note that no combination is left out. If each combination is to be given to 10 subjects each, this design will require 270 subjects. In the case of an unbalanced design, the number of subjects in different groups can vary (Figure F.2b). This figure shows 3 × 2 factorial designs where one factor has 3 levels and the other has 2 levels. The factor levels are considered nominal in a factorial design so that the distance between them is the same, a feature that could be absent if any of them is quantitative with or without categories.

A factorial design is useful in the study of potential **interactions**. This design is easy to implement when prior knowledge is available regarding the levels of various factors that should be simultaneously considered. The 3 × 2 experiment described in Table F.1 is a factorial experiment. In this experiment, factor 1 is the type of hormone replacement therapy (HRT), which has three levels, and factor 2 is the estrous status of the mice, which has two levels. All six possible combinations are included so that the design is fully factorial. This assumes that the estrous status of the mouse can be determined by the experimenter—HRT in any case can be randomly assigned.

Note for Table F.1 that mice are allocated to each level of each factor so that there are six groups. In this example, each group has the same number of mice ($n = 10$), but that is not a requirement. It is just that equal numbers make the analysis and interpretation a lot easier. When the numbers of subjects in different groups are unequal, the design becomes **unbalanced**.

**TABLE F.1**
**A Two-Way Factorial Design (3 × 2) with 10 Mice in Each Group (Number of Mice in Each Group Shown after Random Allocation)**

| | Factor 2 | |
|---|---|---|
| **Factor 1** | **Normal Estrous** | **Ovariectomized** |
| Continuous HRT | 3, 7, 8, 14, 18, 30, 31, 37, 48, 52 | 1, 5, 10, 11, 15, 20, 27, 42, 43, 49 |
| Cyclic HRT | 4, 9, 13, 21, 32, 39, 44, 47, 50, 57 | 6, 17, 26, 33, 34, 40, 51, 56, 58, 59 |
| No HRT (control) | 12, 22, 23, 25, 36, 41, 45, 53, 55, 60 | 2, 16, 19, 24, 28, 29, 35, 38, 46, 54 |

An experiment with $K$ factors, each with $C$ levels, is called a $C^K$ factorial experiment. A simpler format is $K$ factors at two levels each—a $2^K$ factorial experiment. See the following example for a $3^2$ experiment.

In the case of potentially therapeutic or anesthetic agents, it is common to try to evaluate combinations of various dosages of two or more formulations. In an experiment on anesthetic drug A (e.g., xylocaine), dosages under experiment could be 0 (none), 1, and 2 mg, and for drug B (e.g., bupivacaine), dosages could be 0, 0.5, and 1 mg for adult rabbits. The possible combinations of factor levels in this experiment are given in Table F.2.

All combinations are included in a fully factorial experiment, relevant or not. This can be implemented only when each level of one factor can be administered with each level of the other factors. In the case of drugs, for example, it is possible that a high dose of one drug cannot be given with a high dose of the other drug. In this case, a fully factorial experiment cannot be done. See the following.

**TABLE F.2**
**Combinations in a $3^2$ Factorial Experiment**

| | Group | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| Dosage of drug A (mg) | 0 | 0 | 0 | 1 | 1 | 1 | 2 | 2 | 2 |
| Dosage of drug B (mg) | 0 | 0.5 | 1 | 0 | 0.5 | 1 | 0 | 0.5 | 1 |

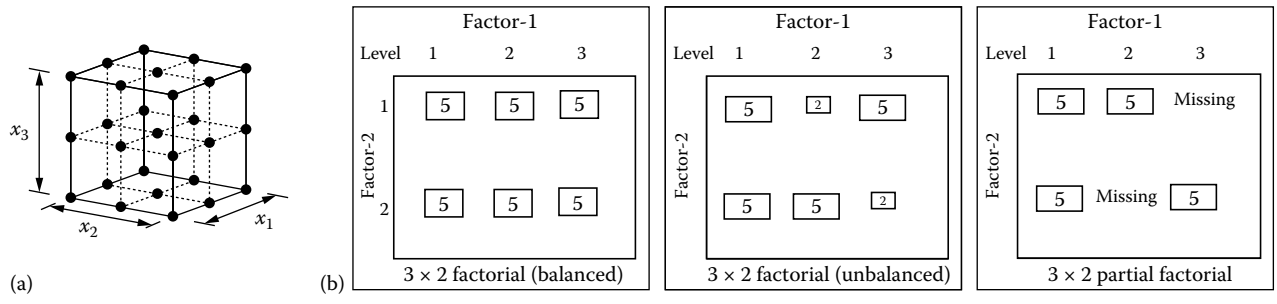



**FIGURE F.2** (a) Pictorial representation of full-factorial design of 3 factors each at 3 levels (a 3 × 3 × 3 design). (b) Factorial (balanced and unbalanced) and partially factorial designs.

### Partially Factorial Designs

A factorial design is called incomplete when one or more combinations of the levels of factors are missing. In the experiment described in Table F.1, one might be tempted to study control mice (no HRT) only for normal estrous and not for ovariectomized mice: if so, the experiment is no longer fully factorial. If there is prior information that 2 mg of drug A in combination with 1 mg of drug B in our example in Table F.2 is harmful, this combination will be unethical and omitted. Such designs are called *partially factorial*. This is also called a *quasi-factorial design*, sometimes *fractional factorial design*. This also is used in setups where the total number of combinations becomes too large to handle. Suppose you have 4 levels of factor 1, 3 levels of factor 2, and 5 levels of factor 3. This makes a total of $4 \times 3 \times 5 = 60$ combinations. This may be too expensive to carry out. Then only the most relevant combinations can be included. The right panel of Figure F.2b shows a partially factorial design where some combinations are missing. The following example illustrates an application of the partially factorial design.

Carbamazepine and lamotrigine are anticonvulsant drugs, which may have a role in treatment of epilepsy and bipolar disorders. For experiments on mice, they require a vehicle such as methylcellulose. The effect of these drugs is often studied on various parameters in mice. One among these is oxidative stress, measured by malondialdehyde (MDA), glutamine synthase (GS), etc. For an example of such an experiment, see Pavone and Cardile [1].

An experiment on carbamazepine (factor A), lamotrigine (factor B), plus the vehicle (factor C) has these three factors. With *given* and *not given* as two levels of each of these factors, a full-factorial experiment requires $2 \times 2 \times 2 (= 2^3) = 8$ groups. Although the vehicle alone can be given, A and B cannot be administered without the vehicle. Thus, the combinations available are O (placebo), C, AC, and BC. In this experiment, only four groups were feasible or relevant against the required eight for a full-factorial experiment. Thus, this would be a partially factorial experiment.

The statistical analysis of a full-factorial experiment is relatively easy, and interpretation is even easier than that of a partially factorial experiment. Thus, biostatisticians tend to advise a full-factorial design and avoid the partial setup. However, in conditions where one or more particular combinations of levels of factors are undesirable or irrelevant, as in the example in the preceding paragraph, a partially factorial experiment can be designed and executed. Statistical analysis of this design will be tough but can still be done with the help of modern software. The main casualty in such designs is the interaction of two or more factors because that rarely can be evaluated in a partially factorial design. Nonetheless, partially factorial designs can still be devised if the study of a particular interaction is important. For further details, see Mee [2].

### Analysis of Factorial Designs

Analysis of data from fully factorial designs is done by two-way, three-way, or higher-way analysis of variance (ANOVA), depending on the number of factors. The analysis has two distinct steps. One is the estimation of the main effects and interactions, and the other is testing their statistical significance. Estimation of **main effects and interaction effects** are described separately under that topic. For statistical significance of the factor effects, ANOVA is done on the same pattern, as explained for **two-way ANOVA**. This can be extended to three or more factors. The ANOVA described therein is for fixed effects, as **random effects** are discussed separately. Also, note that the method of analysis is the same whether the factor-level combinations could be randomly assigned to all the participants or the randomization is restricted to certain groups, as in a **randomized block design**.

Analysis of a partial factorial design is tough, as indicated earlier. The missing combinations generate empty cells. For example, if you have 5 levels of factor 1 and 3 levels of factor 2, you should have all the 15 combinations in your design for a fully factorial experiment. If you have just 9 of these 15 combinations, the other 6 are considered empty cells. Any hypothesis that involves any of the missing combinations cannot be tested. When empty cells are present, the usual Type III sum of squares that we generally use in ANOVA is not applicable; instead, Type IV sum of squares is used. These types are described under the topic **sum of squares (types of)**.

The method of analysis just mentioned is valid when the levels of the factors are nominal with no order. When these are ordered or quantitative, such as dose of 0, 1, and 5 mg, and the objective is to examine graded response in such ordinal categories, it is better to use a regression approach, as in a **general linear model**. Statistical software packages generally require that the command for such analyses be written with extreme care and the output be properly interpreted. Thus, this kind of analysis should be undertaken only by statistical experts.

1. Pavone A, Cardile V. An in vitro study of new antiepileptic drugs and astrocytes. *Epilepsia* 2003;44(Suppl 10):34–9. http://www.ncbi.nlm.nih.gov/pubmed/14511393
2. Mee RW. *A Comprehensive Guide to Factorial Two-Level Experimentation*. Springer Science + Business Media, LLC. 2009.

## factor loadings, *see* factor analysis

## factors (classificatory and experimental)

A factor in biostatistics is that characteristic of subjects that is under consideration for its implication on the research outcome. Obesity, drug doses received (or assigned), education, severity of disease, and type of diet are examples of factors. The term is used for **discrete variables** rather than for **continuous variables**. The convention is to call continuous variables **covariates** or **concomitant variables** in place of factors. Thus, blood pressure (BP) level can be a covariate but generally will not be considered as a factor unless it is categorized as hypertension (yes/no) or as systolic level into categories such as <110 mmHg, 110–129 mmHg, 130–139 mmHg, etc. Thus, a factor can be quantitative but not continuous. The categories of the factor, be it dose, disease severity, BP, or any other, are called *levels of the factors*. If BP is categorized into six categories, the factor BP has six levels. For BP, these categories have order, but that is not necessary. Blood group has four categories without any order, but they still would be called levels.

In analytical studies, the term *factor* is used for **antecedents** rather than for **outcomes**. Outcomes can also be construed as responses in many situations. Researchers commonly aim to find out how outcome or response is affected by the level of various factors. However, the term *factors* is also used in **descriptive studies** where there is no interest in any specific outcome, such as to find out what segment of the population is using different types of tobacco products (in this setup, the factor is tobacco, and different types of tobacco products are the levels) or what the signs–symptoms profile of cases of tuberculosis of the bone is.

For a research setup, factors can be divided in two broad statistical categories. One is classificatory, over which we do not have any control, and the other is experimental, over which we have some control. This distinction is important for interpretation of the results, though possibly not so much for statistical calculations. As an example, consider obesity as a factor. This is something that

cannot be assigned to a person in any study, as this is preexisting. When the subjects are divided by obesity (thin, normal, overweight, and obese), this factor is classificatory, as you just have to observe which category the subject belongs to. Contrast this with a setup of a clinical trial with three doses of a drug, and you can assign a subject to any of the doses you want, generally by a random method. This is what is regularly done in the random allocation of subjects to the doses of the drug. Dose in this setup is an experimental factor. The researcher has the option to allocate the level of an experimental factor to any consenting subject in the study, but this option is not available for the level of a classificatory factor.

What is the implication? The **effect** of the level of an experimental factor is interpreted to apply to the entire population of subjects from which the subjects are chosen, while the effect of the level of a classificatory factor is interpreted to apply to only that part of the population that has that level. If there is a total population of 800 subjects, and a trial on a sample of 90 subjects is conducted, of which 30 were allocated to receive a dose of 5 mg, the estimate of the effect of this dose based on these 30 would be applicable to all 800 subjects. On the other hand, if only 150 of these 800 are overweight, and 20 of these 90 subjects in the trial are overweight, the estimate of the effect of overweight on the outcome based on 20 overweight subjects will be applicable to the 150 overweight subjects in the population, and not to all 800. See how the interpretation differs depending on whether a factor is classificatory or experimental. Of course, we are assuming in our explanation that the selection and allocation is **random**, so the subjects are representative of the concerned population.

## factor scores, *see also* factor analysis

**Factor analysis** is the statistical method used to identify a small number of underlying factors in a set of a large number of correlated continuous variables. This method tries to express each variable in terms of the small number of identified factors. It is also possible in this analysis to perform a reverse process whereby factors are expressed in terms of the variables. This reverse process gives the following type of equation for each the $M$ factors underlying $K$ variables ($M \ll K$):

$$F_m = b_{m1}x_1 + b_{m2}x_2 + \ldots + b_{mK}x_K;\ m = 1, 2, \ldots, M.$$

The coefficients $b_{mk}$ ($m = 1, 2, \ldots, M$; $k = 1, 2, \ldots, K$) are called *factor score coefficients*. When the observed values of ($x_1$, $x_2$, …, $x_K$) for a subject are substituted in the aforementioned equation, the quantity obtained is called the factor score for the $m$th factor for that subject. This measures the importance of the factor for that individual. If the factor score for the third factor ($F_3$) is high for the sixth subject and relatively low for the tenth subject, it can be concluded that $F_3$ is influencing the sixth subject more than the tenth subject. Such factor scores for each factor can be obtained for each subject by using appropriate software. These scores can be used for a variety of purposes. Chandra Sekhar et al. [1] used these scores to develop an index of need for health resources in different states of India, and Jabłoński and Kozakiewicz [2] used them to develop a rule to predict death of acute mediatinitis patients.


1. Chandra Sekhar C, Indrayan A, Gupta SM. Development of an index of need for health resources for Indian states using factor analysis. *Int J Epidemiol* 1991;20:246–50. http://www.ncbi.nlm.nih.gov/pubmed/2066229
2. Jabłoński S, Kozakiewicz M. Evaluation of recovery in iatrogenic evoked acute mediatinitis. *Inflammation* 2013 Oct;36(5):1055–63. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3781308/


## fallacies (statistical), *see also* misuse of statistical tools

Fallacies are anomalies that considerably reduce the credibility of a report. Statistical fallacies are common in medical literature. This section enumerates some such fallacies, hoping to create awareness of situations when these can occur.

Statistics show that more people die in hospitals than at home. Also, there is a strong association between dying and being in bed. The absurdity of such associations is apparent. No one would advocate avoiding hospitals or beds to prolong life. These might seem to be extreme examples, but the same sort of errors of logic sometimes passes unrecognized in the medical literature [1].

Presented in Figure F.3 is one statistical fallacy regarding the use of line diagrams. It is commonplace to see the lines with ±1 standard deviation (SD) as error bars for each point on the $x$-axis with vertical lines on one or both sides of the line. This can be fallacious because the actual variation is much more, something like ±3SD. Sometimes, ±1 standard error (SE) is shown, which is even more fallacious because that applies to the mean and not to the values.

### *Fallacies in Analysis*

Biological processes are complex, and most biostatistical analyses fail to match the reality of the biological systems. For example, **models** are acknowledged as necessarily simplified versions of the intricate biological processes. Thus, fallacies are bound to arise. Many times, they present a false aura of precision due to their underlying mathematical formulation. Because of imposed simplicity, even statistically adequate models may fail to live up to the expectations in practice. In general, there is no 100% correct model: it is just that some models are better than others. Here are some more examples.

There is no doubt that hardly any relationship in medicine is linear. Yet, a **linear relationship** is the most commonly studied form of relationship in health and medicine. This simplification seems to work fairly well in some situations but can destroy an otherwise very clear relationship in others, for example, the rise and fall of lung function with increase in age from 0 to 70 years. This is aptly represented by a parabolic curve, but the relationship vanishes if only linearity is considered. Another example is the relationship between glomerular filtration rate and creatinine level, illustrated in an example by Indrayan [3]. The linear relationship in this case is medically unsatisfactory despite a high $R^2 = 0.81$. These examples

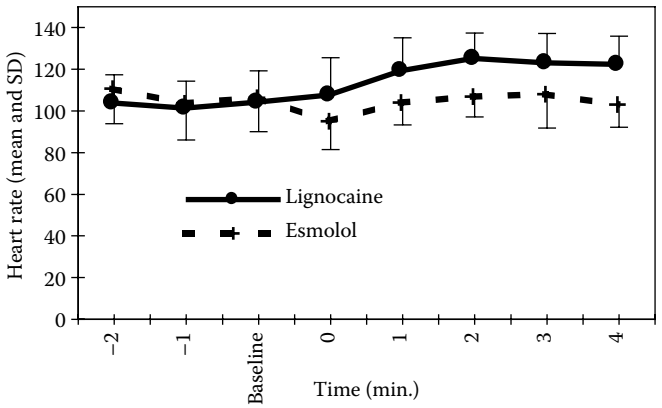



**FIGURE F.3** Heart rate in women with pregnancy-induced hypertension undergoing cesarean section—esmolol and lignocaine groups. (From Wilson A, Siskind V. *Int J Epidemiol* 1995;24:678. http://www.ncbi.nlm.nih.gov/pubmed/8550263. With permission.)

show that there is a definite need to curb the tendency to linearize a clear nonlinear relationship.

Investigators tend to look for one equation—either linear, curvilinear, or whatever—for the entire range of a measurement. Left ventricular ejection fraction (LVEF) ranges from 10% to 90%—less than 50% is considered a coronary risk, and more than 60% does not provide any benefit. If coronary risk is plotted against LVEF, it will be flat after 60%. One model over the entire range of 10% to 90% would not capture this phenomenon. For measurements that show one pattern for $x \leq a$ and another $x > a$, a **spline regression** is needed. Without this, the results could be fallacious. Realize, however, that a spline can be used only if you know, or at least suspect, that the patterns could be different over different ranges of $x$. If this is not known, the fallacy can continue for a long time.

A **Gaussian** form of distribution of various quantitative measurements is so ingrained in the minds of some workers that they take it for granted of medical measurements. For large $n$, the **central limit theorem** can be invoked for inference on means, but nonparametric methods should be used when $n$ is small and the distribution is far from Gaussian. At the same time, note also that most parametric methods, such as $t$ and $F$, are quite **robust** to mild deviation from the Gaussian pattern. Their use in such cases does limited harm so long as the distribution has a single mode. Special attention is required if the distribution looks **bimodal**.

For many statistical procedures, the assumptions of **independence** of observations and of **homoscedasticity** are more important than that of a Gaussian pattern. Independence is threatened when the measurements are serial or longitudinal. Homoscedasticity or uniformity of variance is lost when, for example, SD varies with the mean. And this is not so uncommon. Systolic levels are higher in postmenopausal females, and so are their variances. Persons with higher body mass index (BMI) also tend to exhibit greater variability in BMI than those with lower BMI. Thus, care is needed in using methods that require uniformity of variance.

### Further Problems with Biostatistical Analysis

- **Categorizing** a **continuous variable** causes loss of information and can result in bias, loss of **statistical power**, and increase in **Type I error**. Different categorizations can lead to different conclusions. In some cases, a clever investigator can choose categories after seeing the data to connive to reach a preconceived result. On the other hand, in a situation such as finding a trend with regression, a continuous scale tends to disregard faraway values as outliers, if they are few, even when these contain important information.
- Mean and SD are inappropriate for variables with highly skewed distributions. For such variables, the median and interquartile range are more appropriate. The interquartile range comprises the middle half of the subjects. The difference (median – $Q_1$) when compared with ($Q_3$ – median), where $Q_1$ and $Q_3$ are the first and third **quartiles**, gives an idea of the extent of skewness. Interquartile range also rules out the "absurd"-looking statement, such as mean lipoprotein(a) [lp(a)] level is 8 ± 20, since this cannot be negative. The distribution of lp(a) is highly positively skewed, and the SD is very high. Most values are around 5 mg/dL—thus, the mean could indeed be 8 and SD could be 20 mg/dL, but mentioning this as 8 ± 20 gives an erroneous impression.
- A real anomaly arises when too many 0s occur in a data set (such as runs by batters in baseball games). In medicine, this happens when, for example, you are studying smoking among adolescents. If the duration is recorded as 0 for nonsmokers and the actual duration on a continuous scale for smokers, neither the mean nor the median will work well because of the large number of 0s. If duration is categorized as 0 or more than 0, or in three or four categories, the analysis can be appropriately done by using proportions.
- Gaussian distribution of values is required for many statistical methods, but it can be difficult to check in some cases. Small samples would not be able to provide evidence against Gaussianity because of a lack of power. The tests such as **Kolmogorov–Smirnov**, **Anderson–Darling** (A-D) and **Shapiro–Wilk** work well with large samples and not with small samples, although the A-D test can work for relatively small samples. For really small samples, check whether some extraneous evidence (from experience or literature) is available that the distribution is a Gaussian. The other option is to examine the plot, and if this suggests a non-Gaussian pattern, it might be safe to use **nonparametric** methods for small samples.
- On the other hand, some researchers are nonparametric enthusiasts and use these methods even when the sample size is large. This could result in loss of power. So far, fortunately, the number of such enthusiasts is not high.
- In paired comparisons, change such as from preintervention to postintervention is a natural outcome of interest. Since baseline (*pre-*) values can affect the amount of change, many times, percent change is calculated. This works well provided that (i) negative and positive changes are not canceling on averaging and (ii) percent change is independent of the baseline. These limitations tend to be overlooked, causing fallacies.

### Arbitrary Variable Selection

More than 100 signs–symptoms of hypothyroidism can be identified. These include puffy face, cold intolerance, thin hair, and brittle fingernails. This probably is true for many medical conditions. Two kinds of fallacy can arise in such cases. One is due to a priori (before modeling) selection of variables. If all the variables cannot be studied because of a limitation of sample size or otherwise, there must be sufficient reason for including some and excluding others. In many situations, such reasons do not adequately exist, and subjective preference plays a role, which obviously affects the results. If all known variables are entered into the model, the sample size must be enormously large. Sometimes, one variable at a time is considered, and the statistically significant ones are included in a multivariable setup. This ignores interdependence—thus, the choice is statistically difficult, and the choice of variables remains questionable. Keep this limitation in mind while making an inference, or else fallacies will occur.

Quite often, the factors affecting the outcome are not fully known (see **epistemic uncertainty**). Obviously, only known or, at best, suspected factors can be studied. Large unexplained variation is one indicator of the inadequacy of a model that can be due to exclusion of relevant variables. Unknown factors can be important even when they account for only a small part of the unexplained variation. This can happen when an unknown factor is closely linked to one or more of the known factors.

### Other Statistical Fallacies

There are a large number of several other fallacies that can be listed. Fallacies occurring due to **data dredging**, improper use of the

**person-years** metric, misinterpretation of **intention-to-treat analysis** and **equivalence studies**, misinterpretation of **number needed to treat**, etc. are discussed under the respective topics. Discretion in the use of proportion in place of the mean in some cases or vice versa in interpreting cross-sectional as prospective outcomes (or interpreting prevalence as incidence), and lack of comprehensive adjustments for **confounders** can also cause fallacies.

It is hoped that the examples given in this section give the reader a flavor of statistical fallacies: where and how they can arise, and how destructive they can be. Many more examples could be given: see the work of Indrayan [3] for a comprehensive review. Remember that the scientific community is not free of greed, and their fraud is difficult to detect.


1. Ludwig EG, Collette JC. Some misuses of health statistics. *JAMA* 1971;216:493–9. http://jama.jamanetwork.com/article.aspx?articleid=335725
2. Wilson A, Siskind V. Coronary heart disease mortality in Australia: Is mortality starting to increase among young men? *Int J Epidemiol* 1995;24:678. http://www.ncbi.nlm.nih.gov/pubmed/8550263
3. Indrayan A. *Medical Biostatistics*, Third Edition. Chapman & Hall/CRC Press, 2012.


## false negative and false positive

The terms *false negative* and *false positive* in biostatistics are used in two separate contexts. The first is for medical tests, which include not just laboratory and radiological investigations but also narratives, such as signs–symptoms used in diagnosis and monitoring, and indices, such as pain score for clinical assessment. The second is for statistical tests of **significance**, which also can provide false-negative and false-positive results.

The result of a medical test is called false negative when the test is negative, whereas the disease is actually present. It is called false positive when the test is positive when there is actually no disease. False negativity and false positivity are evaluated after using the test on a series of homogeneous patients whose actual disease status is known. These values are used to assess the extent of the validity of the test and to identify the situations where the test can be used—a test with low false negativity can be useful for screening, and a test with low false positivity can be useful for confirming a diagnosis.

Denote the presence of disease by D+ and absence by D–, test positivity by T+ and test negativity by T–. Also, let true positives be abbreviated as TP, false positives as FP, true negatives as TN, and false negatives as FN.

Table F.3 contains the results of electrocardiograms (ECGs) performed on a total of 700 subjects with complaints of prolonged acute chest pain. Of these, 520 cases were earlier confirmed for the presence of myocardial infarction (MI) and 180 for its absence. The test under review in this case is ECG. In 104 cases, the ECGs were normal among the 520 who have had MI. Thus, the false-negative rate of ECGs in this example is 104/520 = 0.20, or 20%. This is high by any standard. Similarly, false positives are those that have shown ECG aberrations but had no MI. This rate in this example is 9/180 = 0.05, or 5%. Thus, you get a feeling of how good this test is at correctly detecting the presence or absence of disease in those whose disease status is already known. False-negative and false-positive terms for medical tests have intimate implications in the concepts of **sensitivity and specificity**, and **predictivities**. See these topics for more details.

**TABLE F.3**
**MI and ECG in Cases of Acute Chest Pain**

| | MI | | |
|---|---|---|---|
| **ECG** | **Present** | **Absent** | **Total** |
| Positive | 416 (TP) | 9 (FP) | **425** |
| Negative | 104 (FN) | 171 (TN) | **275** |
| **Total** | **520** | **180** | **700** |

The second usage of false-negative and false-positive terms is for the results of statistical tests and is more complex. In order to understand this, see the topic **tests of hypothesis (philosophy of)**. A statistical test requires that a null hypothesis be set and a test criterion appropriate for the kind of data and the design of study be calculated assuming that the null is true. The value of this criterion is used to find the ***P*-value**, which is the probability of rejecting the null when, in fact, it is true (also called the probability of **Type I error**). If you consider rejecting a null as a positive result, this error means a false-positive result, and is like punishing an innocent in a court of law and also like declaring a person without disease as sick. This can have severe implications for science too. Correspondingly, a false-negative result of a statistical test is when an effect is present but we conclude that there is no effect. This is **Type II error** and not as serious. The complement of the probability of Type II error is called the statistical **power**. See these topics for more details.

## familial aggregation

Familial aggregation is said to be present when similarity in members within a family is more than expected by chance. You may have heard that some diseases run in families—meaning thereby that if one is suffering, other members of the family also have a tendency to suffer. Diabetes and hypertension are examples of such diseases. This could be because of shared environment (diet, stress, infections, etc.) or because of genetic influence, or both. For example, 20–25% of all cases of colorectal cancer in Australia are observed to occur in families, but the contribution of genetics and shared environment is not yet known [1]. Such aggregation can happen with any characteristic or trait. For example, the dietary constituents of members of a family are more likely to be similar than across families. Tobacco and alcohol use have strong aggregation within families.

Statistically, familial aggregation is established by comparing within-families variation with between-families variation. For this, you need to obtain data on members of a large number of families. For a quantitative variable such as blood pressure, *within* variance and *between* variance can be easily obtained by the method **analysis of variance (ANOVA)**. Average levels in the members of the families of the affected persons versus the average in the family members of those not affected can give some idea of the familial aggregation. The extent of familial aggregation of a quantitative measurement can be measured by computing the **intraclass correlation coefficient**. For qualitative characteristics, such as whether an adult is hypertensive or not, proportions within families can be compared with proportion between families. In a **case–control study** setup, family members of the subjects with disease and family members of the subjects without disease are investigated for the relative presence of the condition. The difference between these two proportions can be considered a measure of aggregation. The procedure is not so straightforward in a cohort study setup. Hudson et al. [2] have provided the details of the method generally used for cohort studies also, although their paper is on aggregation of two disorders.

Segregation of the genetic effect from the environment effect in familial aggregation can be difficult. Many times, first-degree and second-degree relatives are studied to delineate the genetic effect. This could be sex linked too. The environmental effect can be studied in genetically dissimilar people living together for long, such as those who are married.

Familial aggregation has both medical and statistical implications. When an unsuspected familial aggregation is discovered for any health condition, this sends an alert to the researchers, first to confirm this with other such studies and second to examine possible causes of this. In most situations, some ascribable cause would be found, but in some cases, this could be obscure—perhaps beyond our present knowledge. The biostatistical implication is that because of aggregation, the values seen within a family are not independent—thus, the usual methods of analysis are not applicable. Statistically, a family is a **cluster** whose values are more akin to one another than values seen outside the cluster. This clustering may have to be factored into the analysis by using methods such as **hierarchical analysis** and **generalized estimating equations (GEEs)**.

A good biostatistical study that explores familial aggregation of cerebral palsy in a cohort setup in Norway is that by Tollånes et al. [3]. This study uses twins, first-degree relatives, second-degree relatives, and third-degree relatives as comparator groups.


1. Ouakrim DA, Boussioutas A, Lockett T, Hopper JL, Jenkins MA. Cost-effectiveness of family history-based colorectal cancer screening in Australia. *BMC Cancer* 2014 Apr 16;14:261. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4021190/
2. Hudson JL, Laird NM, Betensky RA. Multivariate logistic regression for familial aggregation of two disorders. I. Development of models and methods. *Am J Epidemiol* 2001;153(5):500–5. http://aje.oxfordjournals.org/content/153/5/500.full
3. Tollånes MC, Wilcox AJ, Lie RT, Moster D. Familial risk of cerebral palsy: Population based cohort study. *BMJ* 2014;349:g4294. http://www.bmj.com/content/349/bmj.g4294?utm_medium=email&utm_campaign=16149&utm_content=The%20BMJ%20%20What%27s%20New%20Online&utm_term=Familial%20risk%20of%20cerebral%20palsy&utm_source=Adestra_BMJ


## fertility indicators

The fertility indicators measure the level of fertility in a community. The term *fertility* can be used for individuals, but the term *indicators* is used for groups. Fertility is the actual performance in terms of live births and excludes stillbirths and abortions. Do not confuse fertility with the capability to reproduce, for which the right term is *fecundity*. A couple may be fully fecund but not fertile if they decide not to reproduce.

Having laid down the basics, we are now in a position to explain various indicators of fertility. There are many, but the one that is simplest and most commonly used is the crude birth rate (CBR). This is the number of live births in a population in 1 year per 1000 population counted at midyear. This can be written as

$$\text{crude birth rate: CBR} = \frac{\text{number of births in a year}}{\text{midyear population}} \times 1000.$$

CBR generally varies between 5 and 30, i.e., generally between 5 and 30 births take place per 1000 population in 1 year. It generally varies inversely with the stage of development of the population. Birth rate can be standardized using the standard age structure of women of childbearing age (generally 15–49 years) along the lines we have explained for **standardized death rate**. But this is rarely done.

Since the real focus for fertility is women of the reproductive age group, the denominator can be replaced by this segment of the population. This will give the number of live births per year per 1000 women of age 15–49 years, called the **general fertility rate**. In terms of formula, this is

$$\text{general fertility rate: GFR} = \frac{\text{number of live births in one year in a population}}{\text{total number of women of age 15–49 years in that population}} \times 1000.$$

The denominator includes women who rarely contribute to the fertility, such as unmarried women. Thus, this indicator can be further restricted to married women in areas where only married women reproduce. Both the numerator and the denominator will be counted for married women only. The name of the indicator then would be general *marital fertility rate*.

The two indicators just mentioned are overall indicators and can be made more specific. Since fertility in women varies greatly with age (remember that we are talking of actual live births and not the capability), **age-specific fertility** for age groups 15–19, 20–24, 25–29 years, etc., of women may be of some interest. This is obtained when the number of births to the women of a specific age is divided by the number of women of that age. Thus, for example,

$$\text{age-specific fertility rate for age group 30–34 years: ASFR} = \frac{\text{number of live births in 1 year to women af age 30–34 years}}{\text{total number of women of age 30–34 years in that population}} \times 1000.$$

According to the World Population Prospects of the United Nations for 2012 [1], the average age-specific fertility rate for age 15–19 years in the world was 89.8 births in the block years 1950–1955, which declined to just 48.9 births in the block years 2005–2010. Much of this is due to an overall decline in fertility, but some is also due to a shift in the fertility to higher age groups. Compared to 48.9 in the age group 15–19 in the block years 2005–2010, the rate was 151.7 in the age group 20–24 years. It was 157.2 in Malawi and 50.6 in India during this period. Thus, there is a huge difference across age groups as well as across countries.

Sometimes, the interest is in comparing the age trend of fertility in two or more populations. For example, Figure F.4 shows that fertility is highest in age group 20–24 years in India and in age group 30–34 years in Ireland. Actual values are shown in Table F.4.

There are some other fertility indicators that are not so simple. First is the **total fertility rate**. This is the average number of births

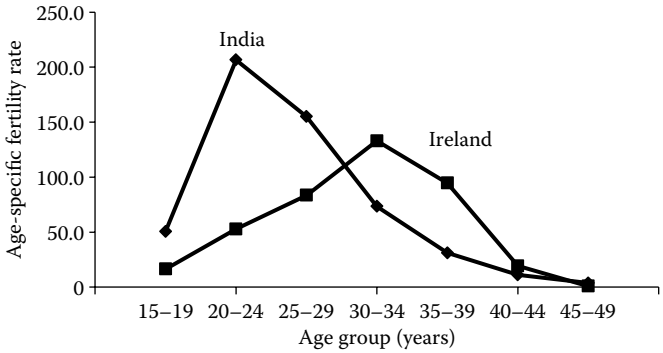



**FIGURE F.4** Age-specific fertility rates in India and Ireland, 2005–2010. (From UN. *World Population Prospects: The 2012 Revision*. http://esa.un.org/wpp/excel-data/fertility.htm.)

**TABLE F.4**
**Age-Specific Fertility Rates (per 1000 Women) in India and Ireland, 2005–2010**

| | Age Group (Years) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **15–19** | **20–24** | **25–29** | **30–34** | **35–39** | **40–44** | **45–49** | **Sum** | **TFR = 5*Sum/1000** |
| India | 50.6 | 206.6 | 155.0 | 73.4 | 31.2 | 11.2 | 3.8 | 531.66 | 2.6583 |
| Ireland | 16.5 | 52.8 | 83.5 | 133.0 | 94.6 | 19.3 | 0.9 | 400.68 | 2.0034 |

*Source:* UN. *World Population Prospects: The 2012 Revision*. http://esa.un.org/wpp/excel-data/fertility.htm.

to a woman during the entire reproductive phase. In place of follow-up for 34 years from age 15 years to age 49 years, the current age-specific fertility rates are used for calculating this rate. This makes it somewhat hypothetical but helps in getting the current status. The calculations are done as follows.

Total fertility rate (TFR)

= 5*Σ(ASFR for the 5-year age groups per woman).

Summation is over the age groups. In case the ASFR for each year of age, 15–16, 16–17, etc., is available, there is no need for the multiplier 5 in the formula. The calculations for India and Ireland are illustrated in Table F.4. Since TFR is per woman, we have to divide by 1000 in the last column (the ASFRs in Table F.4 are per 1000 women). Since TFR is independent of the age structure of women, it is comparable across populations. Thus, it would be legitimate to say that fertility in India was about one-half birth more than in Ireland during 2005–2010.

**Gross reproduction rate** (GRR) is the average number of daughters that would be born to a woman during her lifetime if she passes through her childbearing years conforming to the age-specific fertility rates of a given year. This is the same as TFR but is now restricted to female births. If these are one-half of the total births (as they nearly are in almost any population), the GRR would be nearly one-half of the TFR. Thus, GRR is nearly 1.33 for India in the block years 2005–2010 and 1.00 in Ireland.

Both TFR and GRR assume that all women remain alive all through the reproductive phase. However, some may die early and reproduce less. When an allowance for this is made in GRR, what we get is called the **net reproduction rate** (**NRR**). This rate is more practical and is used to determine whether a population has reached the replacement level or not. A rate nearly equal to 1.00 for a long time indicates that the population has reached to the stationary level. NRR > 1.00 means that the population is increasing, and NRR < 1.00 means that the population is decreasing. Many growing populations in developing countries aim at NRR = 1.00.

In case you are interested to see a report on fertility levels in a developing country, see Ref. [2] for India.


1. UN. *World Population Prospects: The 2012 Revision*. http://esa.un.org/wpp/excel-data/fertility.htm
2. Government of India. *SRS Report 2012*, Chapter 3, Estimates of fertility indicators. http://www.censusindia.gov.in/vital_statistics/SRS_Report_2012/10_Chap_3_2012.pdf


## field trials, *see* clinical trials

## Fieller theorem

The Fieller theorem, first reported by Edgar Fieller in 1954 [1], provides a **confidence interval** (CI) for the ratio of two correlated variables that jointly follow a **bivariate Gaussian distribution**. The Fieller theorem can be used for finding the CI of the ratio of two means and two proportions. It yields approximate intervals when the distributional assumptions relating to the two variables are not strictly satisfied. The CI requires that the covariance of the two variables under consideration is also available or can be estimated.

The Fieller theorem says that if $u_1$ and $u_2$ are (possibly correlated) means of two samples with population means $\mu_1$ and $\mu_2$, respectively, variances (in this case, the square of the standard errors, since $u_1$ and $u_2$ are sample means) $v_{11}\sigma^2$ and $v_{22}\sigma^2$, and covariance $v_{12}\sigma^2$, then a $100(1-\alpha)\%$ CI for the ratio $\theta = \mu_1/\mu_2$ is given by

$$\frac{1}{(1-g)}\left[\left(\frac{u_1}{u_2} - g\frac{v_{12}}{v_{22}}\right) \pm \frac{z_{1-\alpha/2}\sigma}{u_2}\sqrt{v_{11} - 2\frac{u_1}{u_2}v_{12} + \frac{u_1^2}{u_2^2}v_{22} - g\left(v_{11} - \frac{v_{12}^2}{v_{22}}\right)}\right],$$

where $g = \dfrac{z_{1-\alpha/2}^2 v_{22}\sigma^2}{u_2^2}$. The minus sign within the equation will the give lower limit of the CI, and the plus sign will give the upper limit. This complex-looking equation, in fact, is the solution for θ in the quadratic equation

$$\frac{(u_1 - \theta u_2)^2}{(v_{11} - 2\theta v_{12} + \theta^2 v_{22})\sigma^2} = z_{1-\alpha/2}^2. \tag{F.1}$$

This exploits the premise that $\theta = \mu_1/\mu_2$ implies $\mu_1 - \mu_2\theta = 0$. This is a linear combination whose estimate is $u = u_1 - u_2\theta$. Being a linear combination of Gaussian variables, $u$ will follow a Gaussian distribution with mean 0 and variance as in the denominator of Equation F.1. If you take the square root of Equation F.1, you will get the familiar result that $z = u/\mathrm{SE}(u)$ has a Gaussian distribution with mean = 0 and variance = 1.

As a medical professional, you may never have to use this complex equation. But there are some practical situations where CI for ratio of means is required. Most prominent of these is the relative potency of a drug in bioassays. This CI can be adapted to the proportions such as for the relative efficacy of one regimen compared with another, where both efficacies are in terms of percentage.

If you happen to come across a problem where the CI of a ratio of means or proportions is needed, try to get hold of a statistical package where this can be obtained. There are not many at this time, and that is the reason why we have given the formula despite it being so complex. However, keep the following in mind when applying this result:

- Variances and the covariance of $u_1$ and $u_2$ would be rarely known, and as usual, they would need to be replaced by their sample estimates. This would immediately turn a Gaussian distribution to a **Student *t*-distribution**. Thus, $z$ in the CI will be replaced by the table value of t at the

desired **confidence level**. The degrees of freedom for this would be $(n_1 + n_2 - 2)$.

- If you look at the CI formula, you will find that $(1 - g)$ is in the denominator. For CI to be sensible, $(1 - g)$ should be neither negative nor too small. This stipulates that $g < 1$ and not close to 1. The value of $g$ will be too large if $u_2$ in the denominator of $g$ is too small. There are alternative ways to find the CI in these situations, as discussed by Franz [2].

Beyene and Moineddin [3] have also presented some alternative ways to find the CI of a ratio. Another method to find the CI of a ratio estimator is bootstrap, as discussed by Campbell and Torgerson [4].




1. Fieller EC. Some problems in interval estimation. *J R Stat Soc, Series B* 1954;16(2):175–85. http://www.stat.cmu.edu/~fienberg/Statistics36-756/Fieller-JRSSB-1954.pdf
2. Franz VH. *Ratios: A Short Guide to Confidence Limits and Proper Use*, 2007. http://arxiv.org/pdf/0710.2024.pdf, last accessed July 20, 2014.
3. Beyene J, Moineddin R. Methods for confidence interval estimation of a ratio parameter with application to location quotients. *BMC Med Res Methodol* 2005;5:32. http://www.biomedcentral.com/1471-2288/5/32/
4. Campbell MK, Torgerson DJ. Bootstrapping: Estimating confidence intervals for cost-effectiveness ratios. *QJM* 1999;92(3):177–82. http://qjmed.oxfordjournals.org/content/92/3/177.long


## file-drawer effect

The file-drawer effect refers to the practice of researchers choosing not to publish their completed studies because of a negative outcome, i.e., studies that do not support the thrust of their research. Negative outcome here refers to the finding of no statistical **significance**, not to the finding that something affects us negatively. Negative outcome may also refer to finding something that is contrary to one's earlier research or to what one expects from conventional wisdom. A similar problem arises when journal editors are reluctant to publish currently unpublishable or unpopular findings. This, along with the file-drawer effect, is called **publication bias**. This bias runs behind the curtain, and one rarely knows how many studies have been lost.

The practice of reporting and publishing only positive research skews the understanding of the topic at hand toward positive findings. For example, **systematic reviews** and **meta-analyses** seek to summarize available publications on a topic. These will give a biased result if the file-drawer effect is in operation. There is no easy way to assess if it is there and how much. Thus, there is no way to adjust the result, and we many times find that the result obtained by the best of reviews is not able to yield the desired outcome in practice. Note that systematic reviews are considered the gold standard for reaching a **valid** and **reliable** conclusion, but few realize that these also can be flawed due to the file-drawer effect in particular and publication bias in general.

Because of the file-drawer effect, researchers might spend time unnecessarily researching something that in fact has already been researched (but not reported). The advent of a clinical trials registry that requires each **clinical trial** to be registered at the time of initiation may address this problem regarding trials because the researchers are bound to report their findings. But other forms of research remain unattended to.

## finite population correction

Finite population correction is the factor used in adjusting the estimates of the **standard error (SE)** of sample summaries (such as mean and proportion) when the sample size is large relative to the size of population and the **sampling is without replacement**. The usual estimates of the SE of parameters such as mean and proportion are valid when the chance of each unit being selected is the same. This happens in random sampling with or without replacement from an infinite population or in random sampling with replacement from a finite population. Statistically, a population is considered finite when sample size is relatively large, say, exceeding 5% of the population size. In this situation,

$$\text{finite population correction: FPC} = \sqrt{\left(\frac{N-n}{N-1}\right)},$$

where $n$ is the sample size and $N$ is the population size. This accounts for the added precision (reduced SE) gained by sampling close to a larger percentage of the population. The smaller the $n$, the higher the correction. When the FPC is used,

$$\text{SE}(\bar{x}) = \sqrt{\left(\frac{N-n}{N-1}\right)}\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \text{SE}(p) = \sqrt{\left(\frac{N-n}{N-1}\right)}\frac{\sqrt{p(1-p)}}{\sqrt{n}}.$$

The correction is negligible, for example, if you have a sample of 200 from a population of 10,000. Also note that a sample of size 200 out of 10,000 has nearly the same SE as a sample of 200 from 50,000. That is, the **sampling fraction** $(n/N)$ after a certain limit does not matter in the reliability of the estimate (the smaller the SE, the higher the reliability or precision). This may sound strange to those not familiar with sampling. Many would not believe that a sampling fraction of 200/10,000 = 0.02 (or, 2%) and a sampling fraction of 200/50,000 = 0.004 (or 0.4%—one-fifth of the previous) have practically no effect on the reliability of the estimate so long as the population is large. This is a remarkable result in sampling theory. It busts the erroneous belief that a sample mean based on a very small fraction of the population cannot be as precise as one with larger sampling fraction.

The correction we have stated is for **simple random sampling**. Similar corrections are available for other sampling methods such as **stratified** and **two-stage sampling**. For this, see the classical book by Cochran [1].

There are examples in the medical literature of the use of FPC. Godinho et al. [2] sampled 345 out of 803 policemen in Montes Claros, Brazil, to study periodontal disease. They used FPC because of a large sample relative to the population. Kolahi et al. [3] used FPC for **clustering sampling** for calculating the confidence interval (CI) for the incidence of diarrhea in children in Iran when they selected a sample of 2016 children out of an estimated 22,000.


1. Cochran WG. *Sampling Techniques*, Third Edition. Wiley, 1977.
2. Godinho EL, Farias LC, Aguiar JC, Martelli-Júnior H, Bonan PR, Ferreira RC, De Paula AM, Martins AM, Guimarães AL. No association between periodontal disease and GHQ-12 in a Brazilian Police population. *Med Oral Patol Oral Cir Bucal* 2011 Sep 1;16(6):e857–63. http://www.medicinaoral.com/pubmed/medoralv16_i6_p857.pdf
3. Kolahi AA, Rastegarpour A, Abadi A, Gachkar L. An unexpectedly high incidence of acute childhood diarrhea in Koot-Abdollah, Ahwaz, Iran. *Int J Infect Dis* 2010 Jul;14(7):e618–21. http://www.sciencedirect.com/science/article/pii/S1201971210000020


## Fisher–Behrens problem, *see* Welch test

## Fisher exact test

The Fisher exact test [1] is used to test the statistical significance of the association between two qualitative characteristics when the

sample size is small. The usual test for association is **chi-square**, but it fails to oblige when $n$ is small since an exact method is needed in this case. For a **2 × 2 table**, the null hypothesis can be tested by the Fisher exact test (also known as the **Fisher–Irwin test**). Here, the **null hypothesis** is a statement of "no effect" or "no difference," and the alternative hypothesis is that there is an association. As usual, the test of significance is designed to assess the strength of the evidence against the null hypothesis.


Ronald Fisher

Consider the notations given in Table F.5 for an antecedent–outcome setup. The notations are the same for any 2 × 2 table. $O$ is the notation for the observed frequency and $\pi$ for the probability, subscripts identify the cell, and the dot in the subscript is for the total. If the margins $[O_1\cdot(\pi_1\cdot), O_2\cdot(\pi_2\cdot), O\cdot_1(\pi\cdot_1), O\cdot_2(\pi\cdot_2)]$ are considered fixed, then the probability by **multinomial distribution** is

$$P = \sum \frac{O_1\cdot!O_2\cdot!O\cdot_1!O\cdot_2!}{n!O_{11}!O_{12}!O_{21}!O_{22}!},$$

where the sum is over all the configurations in a 2 × 2 table that are observed to be as or more extreme in favor of $H_1$, without altering the marginal totals. This is the exact ***P*-value** in this case. This can be easily calculated manually for small $n$ (say $n < 10$) but can become difficult for larger $n$. Most statistical software packages would give this exact $P$-value for small $n$ using the aforementioned equation. Reject $H_0$ if $P < 0.05$, where the **level of significance** is set at 5%; otherwise, be content with the assertion made in $H_0$. The calculations are illustrated in a simple example given later in this section.

The probability in the aforementioned equation gives the *one-tailed* $P$-value. This is typical for the Fisher exact test. If a *two-*tailed value is needed, proceed as follows: (i) Double the one-tailed $P$-value if any (row or column) marginal totals are equal. Equal totals imply symmetry, and this allows such doubling. (ii) Calculate a separate $P$-value from the aforementioned equation for each tail if the marginal totals are not equal.

As already mentioned, the Fisher exact test assumes fixed marginal totals. This restriction is valid for a situation where, for example, you have six persons with disease and six without disease, and the test is built in a manner such that it is constrained to give six positive and six negative results. You can see that this is an unnatural restriction and would rarely hold in practical situations. This restriction makes the test overly conservative (i.e., fails to reject $H_0$ where it should). To overcome this problem, another test called the **Barnard test** is advocated, which does not require fixed margins. This test is computationally difficult and is not popular yet. For comparison between the Fisher exact and Bernard tests, see Mehta and Senchaudhuri [2].

**TABLE F.5**
**General Structure of a 2 × 2 Contingency Table**

| Variable 2 (Outcome) | Variable 1 (Antecedent) Present | Absent | Total |
|---|---|---|---|
| Present | $O_{11}$ ($\pi_{11}$) | $O_{12}$ ($\pi_{12}$) | $\boldsymbol{O_1.(\pi_1.)}$ |
| Absent | $O_{21}$ ($\pi_{21}$) | $O_{22}$ ($\pi_{22}$) | $\boldsymbol{O_2.(\pi_2.)}$ |
| **Total** | $\boldsymbol{O._1}$ **($\boldsymbol{\pi._1}$)** | $\boldsymbol{O._2}$ **($\boldsymbol{\pi._2}$)** | $\boldsymbol{n}$ |

**TABLE F.6**
**Example of Fisher Exact Test**

| | Antecedent | | | | | |
|---|---|---|---|---|---|---|
| **Outcome** | **+** | **–** | **Total** | | | |
| + | 1 | 7 | **8** | 0 | 8 | 8 |
| – | 5 | 2 | **7** | 6 | 1 | 7 |
| **Total** | **6** | **9** | **15** | **6** | **9** | **15** |

For illustration of the calculation of the Fisher exact test, consider the data in Table F.6. In the left panel are the observed frequencies. In the right panel are the frequencies that would be more adverse to the null hypothesis of no association. No other extreme configuration is possible because marginal totals have to remain the same, and one cell frequency is already 0 in the second configuration.

From the previous equation,

$$P = \frac{8!7!6!9!}{15!1!7!5!2!} + \frac{8!7!6!9!}{15!0!8!6!1!}$$
$$= 0.0336 + 0.0014$$
$$= 0.035.$$

This is less than 0.05. Since it is a **one-tailed probability**, the conclusion too would be one-sided. In this case, there are seven subjects with positive outcome when the antecedent is absent and five subjects with negative outcome when the antecedent is present. Of a total of 15 subjects, these 12 favor the antecedent as an inhibitor to the outcome against the other 3 who do not. A small $P$-value shows that this is statistically significant. Thus, the conclusion is that the presence of the antecedent inhibits the outcome in a statistically significant manner at a 5% level of significance. For such one-sided conclusions, one-sided $\alpha$ is used.

The Fisher exact test can be extended to larger tables. Mehta and Patel [3] have described a network algorithm for performing the Fisher exact test in ($R \times C$) contingency tables. This extension was first proposed by Freeman and Halton [4].


1. Fisher RA. *Statistical Methods for Research Workers*. Oliver & Boyd, 1925.
2. Mehta CR, Senchaudhuri P. Conditional versus unconditional exact tests for comparing two binomials. Cytel Software Corporation 4 September 2003. http://www.cytel.com/Papers/twobinomials.pdf, last accessed July 20, 2014.
3. Mehta CR, Patel NR. A network algorithm for performing Fisher's exact test in ($r \times c$) contingency tables. *J Am Stat Assoc* 1983;78:427–34. http://www.jstor.org/discover/10.2307/2288652?uid=3738032&uid=2&uid=4&sid=21104320899587
4. Freeman GH, Halton JH. Note on an exact treatment of contingency goodness of fit and other problems of significance. *Biometrika* 1957; 38:141–9. http://www.jstor.org/stable/2332323

## Fisher–Irwin test, *see* Fisher exact test

## Fisher *z*-transformation, *see* comparison of two or more correlation coefficients

## fixed and random effects, *see also* mixed effects models

The terms *fixed* and *random effects* are used for the levels of **factors** that are analyzed by **analysis of variance (ANOVA)**. For example, in a trial, different doses of a drug such as 0 (placebo), 5, 10, and 20 mg are the four levels of the factor "drug." The interest in this case is restricted to these particular doses, and no inference is drawn for any other dose such as, say, 12 mg. Thus, these levels are fixed. If the outcome is the fever clearance time in malaria patients, you can exactly estimate the effect of these specific doses on this outcome. Now suppose this trial is done in 5 out of, say, 200 centers in a country. In this case, the interest is not limited to these selected five centers but is extended to find if there is any intercenter difference in the outcome in these 200 centers. Let these 5 centers be a random sample of all the centers so that a generalized conclusion about the effect of centers can be made. The 5 centers in the study could be any of those 200. In this case, the effect of centers on the outcome will be modeled as a random effect. In the case of fixed effects, no inference can be drawn about any other level except the ones under study, while for random effects, the inference would be valid for the entire population from which the factor levels are randomly drawn. If drug levels have fixed effects and centers have random effects, the model will be called a **mixed effects model** when both are studied together.

In our example of a trial at five different centers, the interest would not be in estimating the mean effect of the specific centers on the outcome but would be in the variance across such centers. This will delineate how different outcomes in different centers are. The larger the variance, the greater the effect of centers. Thus, the method of conducting the ***F*-test** is different when random effects are present. This is sometimes referred to as **variance components analysis** because of the emphasis on variances rather than the means. For details, see Searle et al. [1].

When a drug is given to, say, a sample of 10 subjects, and **repeated measures** are taken at fixed points of time for each subject to see the time effect as in **bioequivalence studies**, such data can be analyzed by two-way ANOVA considering time points as factor 1 and subjects as factor 2 and $n = 1$ (1 observation for each time and subject). The assumption in this case is that the interest is only in comparison of means at different points in time and not the time trend of the response. Time trend would mean extrapolating the effect of time to the time points not actually studied. This analysis also requires that after the time effect is extricated, the **residuals** are independent and have same variance. But this procedure violates one basic premise. The conventional ANOVA considers levels of the factors fixed and of specific interest to the problem. Subjects are almost invariably a random sample from a defined population. When such subjects are considered factor levels, as just suggested for repeated measures, these give rise to what we have called a random effect. For details of the analysis of repeated measures, see Raghavarao and Padgett [2]. We have also provided some details under the topic **repeated measures ANOVA** in this volume.


1. Searle SR, Casella G, McCulloch CE. *Variance Components*. Wiley, 2006.
2. Raghavarao D, Padgett L. *Repeated Measurements and Crossover Designs*. Wiley, 2014.


## flowchart

A flowchart is a text-based diagram that provides a sequence of steps on how to proceed with a problem or to approach a problem for reaching a defined goal. It contains text in boxes of various shapes that identify the step/s. Arrows provide the direction and lead to the next steps. Thus, a flowchart maps the process and provides an algorithm. This helps to achieve clarity that otherwise may be difficult in a description with words.

An example of a flowchart is in Figure F.5. This is on a pathologist's report in a case of testicular cancer. Note that there are different shapes of the boxes: each has its specific meaning. For example, an oval shape is for the beginning and end, rectangles are for the process at the designated step, a diamond is where your decision is required, and a rhombus (not in this figure) is for entering data. An arrow coming from one shape and ending at another shape signifies flow. Diamonds for a decision typically contain a yes/no question or true/false test. This shape is unique in that it has two arrows coming out of it, one corresponding to "yes" or "true" and the other corresponding to "no" or "false." The arrows anywhere can have labels. For more details on the shapes used in the flowchart, see the illustration by Grout [2].

Flowcharts can be used for describing the algorithm for managing patients, as done by Bruyere et al. [3] for cases of knee osteoarthritis. Such charts can also be used to describe some crucial aspects of the design of a study, such as the **CONSORT** flowchart for clinical trials.


1. RFFlow5. Medical flow charts. http://www.rff.com/medical-flowcharts.htm
2. Grout J. Using process mapping to identify patient safety hazards in micro-systems: Flow chart. http://facultyweb.berry.edu/jgrout/processmapping/Flow_Chart/flow_chart.html, last accessed May 23, 2015.
3. Bruyère O, Cooper C, Pelletier JP, Branco J, Luisa Brandi M, Guillemin F, Hochberg MC et al. An algorithm recommendation for the management of knee osteoarthritis in Europe and internationally: A report from a task force of the European Society for Clinical and Economic Aspects of Osteoporosis and Osteoarthritis (ESCEO). *Semin Arthritis Rheum* 2014 May 14. pii: S0049-0172(14)00108-5. http://www.semarthritisrheumatism.com/article/S0049-0172(14)00108-5/fulltext


## focus group discussion

A focus group discussion (FGD) is a tool to gather data on soft aspects of social concerns that can rarely be studied by surveys. Soft aspects include opinions, beliefs, feelings, attitudes, etc. These concerns can quickly change in an interactive setting in a group since people's response on soft issues is many times guided by the views of the others. Surveys tend to collect data on immediate feelings as they do not provide an opportunity to go deeper or to be affected by the opinions of others. FGDs are an ingenious way to get to the bottom of feelings of people, which some researchers feel may show up better in a group discussion. It is a tool for now expanding **qualitative research** as opposed to the regular quantitative research. There is a growing realization that feelings of people are important contributors to the success or failure of programs, rather than just physical facilities. For example, Crankshaw et al. [1] used FGD to elicit the health care provider perspective on delivering safer conception services for HIV-affected couples in South Africa.

For an FGD, a small group of, say, 6–10 of the people concerned with the issue under investigation is gathered at a convenient location and asked to discuss a specific topic or posed a specific question

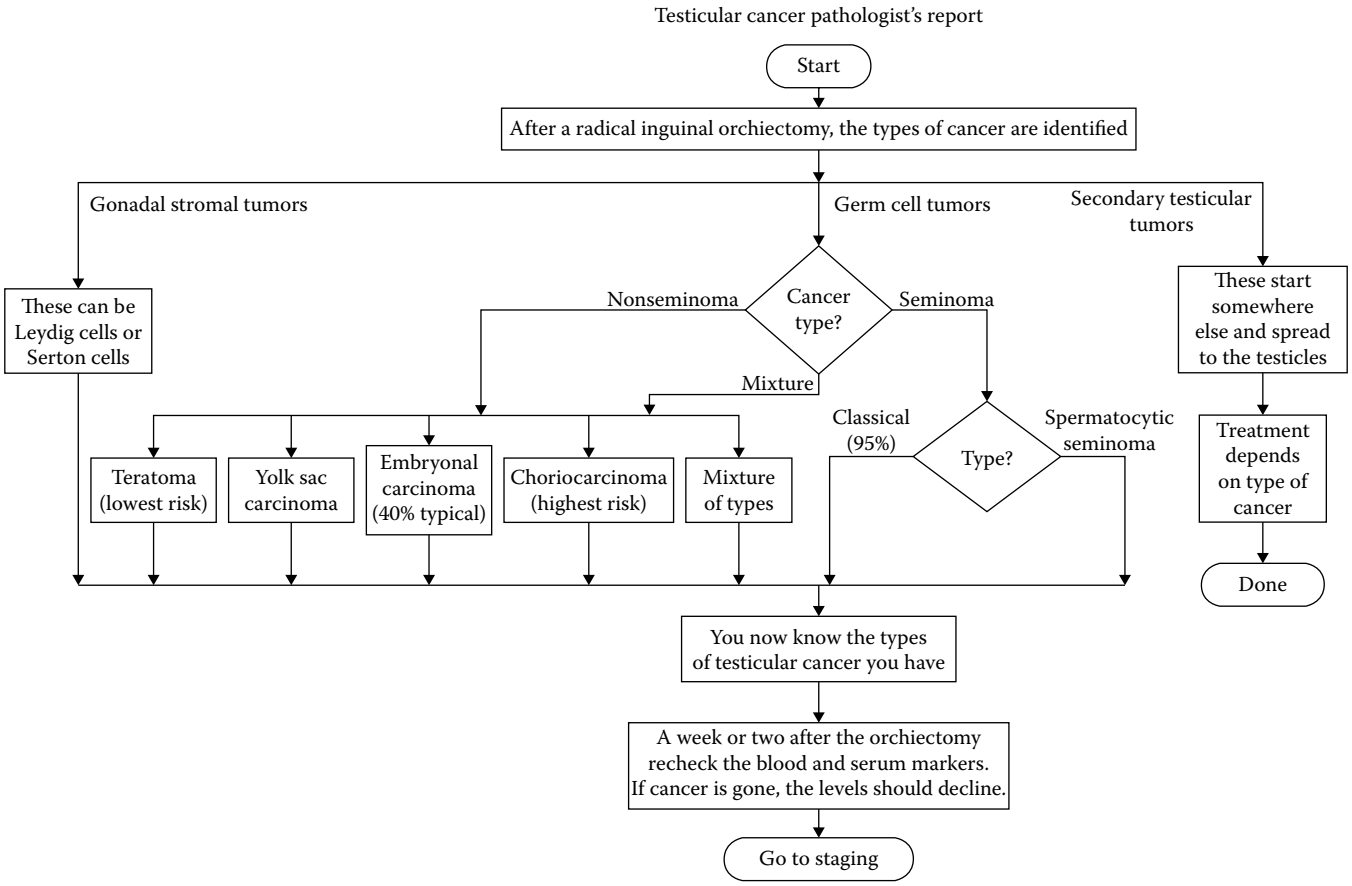



**FIGURE F.5** Flowchart of pathologist's report in case of testicular cancer. (From RFFlow5. Medical flow charts. http://www.rff.com/medical-flowcharts.htm. With permission.)

(or a set of questions). The group is subjectively chosen so that it can have diverse opinions and possibly have full knowledge of the problem or program being discussed. Generally, they are strangers to one another. They can even be the opinion makers of society.

The questions are carefully chosen and properly worded so that everybody clearly understands them with the same meaning. The agenda is carefully drawn keeping in view the objectives of the whole exercise. The members of the group are encouraged to express themselves openly and fearlessly for a range of views to emerge. The discussion is guided by a neutral moderator to ensure that the discussion does not go off track. The moderator is also assigned the task of keeping the discussion lively and ensuring that everybody participates. At the end, an expert prepares a report that summarizes salient features and the consensus arrived at from the discussion. This report is on themes and not on quantitative outcomes of the discussion. In case diverse opinions remain, sometimes, another group of people is consulted. The summary of the discussion can help in understanding the apprehensions and anxieties of the people on one hand, and appreciation and other positive feedback on the other, regarding the problem or the program under discussion. This understanding can help in implementing the process more smoothly or changing it.

FGD results carefully consider inconsistencies and variations that are bound to exist among members of such a group. Diverse opinions are welcome, although the discussion has to be kept on track. The discussion may provide an insight into varying beliefs and experiences, and how they affect attitudes and behaviors. For more details of FGD, see Hennink and Leavy [2].

Because of the substantial role of subjective opinions in this methodology, many researchers suspect the findings of such qualitative research. Indeed, another FGD may lead to another finding, but the regular qualitative research also has potential for this aberration. Our advice is to restrict FGDs to only those areas where regular surveys fail to provide a comprehensive picture, or use them to supplement the findings instead of considering FGD an independent tool.

1. Crankshaw TL, Mindry D, Munthree C, Letsoalo T, Maharaj P. Challenges with couples, serodiscordance and HIV disclosure: Healthcare provider perspectives on delivering safer conception services for HIV-affected couples, South Africa. *J Int AIDS Soc* 2014 Mar 12;17:18832. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3956311/
2. Hennink MM, Leavy P. *Focus Group Discussions (Understanding Qualitative Research)*. Oxford University Press, 2014.

## follow-up studies, *see* prospective studies

## forecasting

Forecasting is the scientific estimation of unobserved values, particularly future values. Weather forecasting and forecasting of company stocks are regular occurrences, but forecasting is frequently used in the medical and health setup as well. Prognosis of a patient is an exercise in forecasting. For example, the existence of extranodal spread in various types of breast cancer has predictive value in forecasting the number of metastatic lymph nodes and the disease

prognosis [1]. Forecasting a critical patient's extent of recovery in the next 24 hours dictates what steps are required to be taken to save the patient. In the public health domain, forecasting the trend of, say, infant mortality rate (IMR) may be helpful for insights on what planning is needed and what resources must be allocated to reach a particular target in the specified time frame. All these examples illustrate attempts to study previous experience in similar circumstances and try to extrapolate on the basis of what circumstances are likely to develop in the future. Statistically, these circumstances are the **independent variables** of the model, and the outcome under forecast is the **dependent variable**.

Statistical tools generally used for forecasting are **time series models** and **regression**-based **predictive models**. However, all statistical methods, as of now, including the two just named, consider the average behavior of a *group* of patients and hope that an individual patient will revolve around this average. This strategy can and does misfire in some situations but still works well in the long run in the sense that the results are likely to be close to those predicted in many similar cases. Admittedly, the strategy may desperately fail for a particular individual. Perhaps clinical acumen of the attending physician or surgeon is more important for individual forecasting than biostatistical modeling.

With time series models, the independent variable is time. Most time series throw up dependent values at sequential points in time even when the effect of time is removed. This happens because of the presence of other factors that move with time but are not exclusive to the time. For example, in the case of IMR, the rate in the year 2020 will depend not just on the trend in previous years but also on what other social (e.g., educational or nutritional) and health infrastructure changes are anticipated for the future that are not in the past pattern. Thus, the ordinary regression methods do not apply, and special techniques such as **moving average** and **autocorrelation** are used for time series modeling. These methods take care of other factors as well that may be changing with time. Predictive regression models can have a large number of independent variables (or, call them predictors in this case) but mostly require that the residuals are independent and not autocorrelated. This works well when time is not a factor, such as in prediction of glomerular filtration rate when creatinine level is controlled.

No matter what statistical model is used, the forecast presumes that the conditions prevailing earlier continue and the effect of those not earlier prevailing can be conjectured. That is, the relationship derived from the available data is adequate to predict the unavailable values. For example, Axelsen et al. [2] found that influenza forecasts are driven by temperature, humidity, antigenic drift, and immunity loss. All these values must be available. Another requirement is that the model must be based on a large number of observations that include varying values of the predictors. Both of these are strong requirements and should be thoroughly examined for their validity before forecasting. For this reason, statistical models are able to correctly forecast values when the conditions, such as the status of a subject, are expected to change marginally and not substantially. For a time forecast, perhaps the next few years can be predicted with precision provided you already have a sufficiently long time series.


1. Kaygusuz EI, Cetiner H, Yavuz H. Clinico-pathological significance of extra-nodal spread in special types of breast cancer. *Cancer Biol Med* 2014 Jun;11(2):116–22. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4069798/
2. Axelsen JB, Yaari R, Grenfell BT, Stone L. Multiannual forecasting of seasonal influenza dynamics reveals climatic and evolutionary drivers. *Proc Natl Acad Sci U S A* 2014 Jul 1;111(26):9538–42. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4084473/


## forest plot, *see* meta-analysis

## forward selection, *see* stepwise methods

## fractiles, *see* quantiles

## frailty models

Statistically, frailty is the unobserved individual heterogeneity in survival models. The usual **survival analysis** is valid when the hazard of the outcome is homogenous across the subjects after taking all the covariates into consideration. However, in many cases, it is not possible to measure all the covariates—sometimes due to the cost involved and sometimes because the importance of some covariates is in doubt—and considerable heterogeneity remains. In addition, our knowledge is always limited (see **epistemic uncertainties**), and some covariates are just not known. In any case, all individuals differ in their vulnerability. Ignoring heterogeneity produces an overestimate of the life expectancy and underestimates rates of individual aging. Assessment of the achievements made by way of health improvement or interventions gets affected. Frailty models contain a random effect component for studying such "hidden" heterogeneity. This is considered to have a multiplicative effect on the baseline **hazard function**. Thus, frailty models are kind of an extension of the **proportional hazards model**—popularly known as the Cox model. They go deeper into the mechanism of cause–effect than the conventional models and may have wider applications. The term *frailty* was first suggested by Vaupel et al. [1] in the context of mortality studies. For further details of frailty models, see Weinke [2]. For technical details and extensions, see Hougaard [3].

According to Vaupel et al. [1], the concept of frailty assumes that all persons have different frailty at birth but it stays same all through life. This can be modified to have variation with age. It surely varies from person to person. Frailty is assumed to have a statistical distribution (generally a gamma distribution) with a mean and a standard deviation different for different age groups. Subjects with frailty more than the mean at any age are those who experience increased hazard. This can happen with mine workers for the age at which they work in mines, or for drivers who are on the road for much of their working life.

Because of the random component, frailty models can give different results than obtained by regular analysis. Zarulli et al. [4] found that the estimate of the effect of education level on mortality after the age of 50 years in Turin is higher by frailty model than by the regular approach. Yewhalaw et al. [5] used frailty models for studying time to malaria in Ethiopia.


1. Vaupel JW, Manton KG, Stellard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demog* 1979;16:439–54. http://www.ressources-actuarielles.net/EXT/ISFA/1226.nsf/0/2d6493f9f0d93b50c125774b0045c00b/$FILE/Vaupel-Demography-16-1979-3.pdf, last accessed January 12, 2016.
2. Wienke A. *Frailty models*. MPIDR Working Paper WP 2003-032, September 2003. http://www.demogr.mpg.de/papers/working/wp-2003-032.pdf, last accessed July 24, 2014.
3. Hougaard P. Frailty models for survival data. *Life Time Data Analysis* 1995;1:255–73. http://link.springer.com/article/10.1007%2FBF00985760#page-1
4. Zarulli V, Marinacci C, Costa G, Caselli G. Mortality by education level at late-adult ages in Turin: A survival analysis using frailty models with period and cohort approaches. *BMJ Open* 2013 Jul 3;3(7). pii: e002841. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3703572/

5. Yewhalaw D, Getachew Y, Tushune K, W Michael K, Kassahun W, Duchateau L, Speybroeck N. The effect of dams and seasons on malaria incidence and anopheles abundance in Ethiopia. *BMC Infect Dis* 2013 Apr 3;13(1):161. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3667047/

## Framingham Heart Study

In 1948, a long-term study into chronic circulatory disease was started in Framingham, Massachusetts, in the United States with the objective of identifying possible **risk factors** and to characterize the natural history of the disease. This has popularly come to be known as the Framingham Heart Study. It is a **longitudinal study** with regular follow-up and is now studying the offspring of those first studied. In fact, the study is on its third generation. For example, see Coviello et al. [1].

The cohort initially recruited for this study is a random sample of 5209 persons of age 30–62 years. Both men and women were recruited. This forms nearly two-thirds of the population of the city of Framingham at that time. Thus, this is an example of an unusually large proportion of the population being sampled. The participants regularly undergo an examination and laboratory investigations almost every 2 years. As of 2014, 31 examinations had been done. The cohort has been expanded to include other people from Framingham and by recruiting the second and third generations. Further details of the study are available at their website [2].

This study is credited with coining the term *risk factor* [3]. Among the biostatistical contributions of this study was the development of new **multivariate** statistical methods to analyze the incidence of multifactorial diseases that help us to estimate the effect of different risk factors. It is also credited with developing a simple coronary prediction model on the basis of measurements such as blood pressure and cholesterol level [3].

1. Coviello AD, Zhuang WV, Lunetta KL, Bhasin S, Ulloor J, Zhang A, Karasik D, Kiel DP, Vasan RS, Murabito JM. Circulating testosterone and SHBG concentrations are heritable in women: The Framingham Heart Study. *J Clin Endocrinol Metab* 2011;96:E1491–5. http://press.endocrine.org/doi/full/10.1210/jc.2011-0050
2. The Framingham Heart Study. https://www.framinghamheartstudy.org/participants/original.php
3. O'Donnell CJ, Elosua R. Cardiovascular risk factors. Insights from Framingham Heart Study. *Rev Esp Cardiol* 2008;61:299–310. http://www.revespcardiol.org/en/cardiovascular-risk-factors-insights-from/articulo/13117552/

## *F*-ratio, *see* *F*-test

## frequencies, *see also* frequency curve/distribution/polygon

Frequency in statistics is the count of occurrence of an event or a value among the subjects of a study or the number of subjects that display a particular **attribute**. You can see that no frequency can be negative. Calculating frequencies requires no advanced statistical knowledge, just basic numerical skills: perhaps the ability to calculate percentages. Frequency tables represent the most initial method for analyzing **categorical data**. Categories can be **nominal**, **ordinal**, or **metric**—that does not matter—but generally, the number of categories does not exceed 12 in such a table. Whereas nominal and ordinal categories are almost always small in number and discrete, metric data can be **discrete** or **continuous**. Examples of *metric* data that are discrete and naturally fall into categories are parity, number of angina episodes, and number of organs affected in an injury. For continuous data, frequency is obtained for values in various specified class intervals. **Class interval**, also called bin, refers to the division of the range of values into smaller parts. For example, for systolic level of blood pressure (BP), the class intervals (in mmHg) can be 100–109, 110–119, etc. Frequencies can be used to review how values in the sample are distributed over different categories. This is used to postulate a statistical **distribution** of the values, which in turn determines the specifics of the statistical method to be used for analysis of the data.

Whereas frequencies reduce to simple counts for discrete variables, they can be time consuming and prone to error for continuous variables when computed manually. Use the help of a computer as much as possible. Most statistical packages have the facility to do this. If you have a small data set and want to do it manually, proceed as follows.

Step 1: Ready the data to be plugged into the frequency table.
Step 2: Decide how many categories would be appropriate and what the class intervals should be. They do not have to be equal, but equal intervals help in proper interpretation regarding percentages in different intervals.
Step 3: Prepare a table format with the first column containing class intervals in ascending order and the second column for tally marks representing the counts. The third column will contain the frequencies as finally obtained. See Table F.7 for systolic BP levels in 33 subjects. The fourth column can have cumulative frequency or percentages as needed.
Step 4: Tally the data into class intervals. Each data value falls exactly into one class interval. Count the tallies and record the result. This is illustrated in Table F.7. Note how the fifth mark is crossed. This makes manual counting easy.

The last column, containing cumulative frequencies, is explained next.

### *Cumulative Frequencies*

Sometimes, the interest is not in the frequencies in different intervals but in the number or percentage of subjects that have not reached a particular threshold. Cumulative frequency at threshold $x$ is the count of subjects with value $\leq x$. In Table F.7, cumulative frequency till level 119 mmHg is 3 and till level 129 mmHg is 14. This means that 14 persons have a systolic level less than or equal to 129 mmHg. This is obtained by sequentially adding the frequencies in the class

**TABLE F.7**
**Manual Construction of a Frequency Table**

| Class Intervals Systolic BP (mmHg) | Tally Marks | Frequencies (*f*) | Cumulative Frequencies |
|---|---|---|---|
| 100–109 | // | 2 | 2 |
| 110–119 | / | 1 | 3 |
| 120–129 | ## ## / | 11 | 14 |
| 130–139 | //// | 4 | 18 |
| 140–159 | ## / | 6 | 24 |
| 160–179 | ## | 5 | 29 |
| 180+ | //// | 4 | 33 |
| **Total** | | **33** | |

intervals. If you want to know what percentage of people have a normal level of systolic BP (i.e., less than 140 mmHg), the answer from Table F.7 is 18 out of 33, or 55%.

Cumulative frequencies are commonly used for durations such as what percentage of people survived for at least 2 years after a particular surgery or what percentage stayed in a hospital for at least 10 days. Statistical use of cumulative frequencies is in the calculation of various **quantiles** (percentiles, deciles, quintiles, quartiles, and tertiles). A graphical method of obtaining quantiles requires that we draw an **ogive**, which is a graphical depiction of the cumulative frequencies.



## frequency curve/distribution/polygon

A frequency curve and a frequency polygon are graphical representations of **frequencies** in different class intervals of a **continuous variable**. These terms are not used for discrete variables. The curve is a smoothened polygon for a sample of subjects and tends to indicate the shape of the statistical **distribution** of the data. We explain these with the help of an example.

Consider the data in Table F.8 on total cholesterol level in a sample of 82 subjects attending a hypertension clinic. The **class intervals** are chosen in accordance with their clinical implications. To start with, note the **histogram** of these data in Figure F.6a. The frequency histogram (Figure F.6a) transforms to a frequency polygon (Figure F.6b) and then to a frequency curve (Figure F.6c). These are drawn in consideration of unequal class intervals, as explained under the topic **histogram**.

### *Frequency Polygon*

A polygon is a shape enclosed by straight lines. A frequency polygon is drawn to depict frequencies of a continuous variable, as in cases where a histogram can be drawn. That is, the continuous variable is plotted on the *x*-axis and the frequencies on the *y*-axis. For a polygon, plot the points corresponding to the frequency (or percentage) on the midpoint of the class intervals, and join them with straight lines. This procedure is the same as joining the midpoints of the tops of the bars in a histogram. For the data in Table F.8, the polygon is shown in Figure F.6b. Note how zero frequency in the interval 320–339 mg/dL is shown. The frequency before the first interval and that after the last interval re each 0, and so lines are also drawn to join zero points. This completes the polygon as an enclosed shape. This is commonly used to study the shape of the distribution. The polygon is closer to the theoretical curve than the histogram, as explained next.

**TABLE F.8**
**Distribution of Subjects Attending a Hypertension Clinic by Total Serum Cholesterol Level**

| Cholesterol Level (mg/dL) | Number of Subjects (*f*) | Percent (%) |
|---|---|---|
| <199 | 3 | 3.7 |
| 200–239 | 13 | 15.9 |
| 240–259 | 16 | 19.5 |
| 260–279 | 17 | 20.7 |
| 280–299 | 24 | 29.3 |
| 300–319 | 6 | 7.3 |
| 320–339 | 0 | 0 |
| 340–399 | 3 | 3.7 |
| **Total** | **82** | **100.0** |

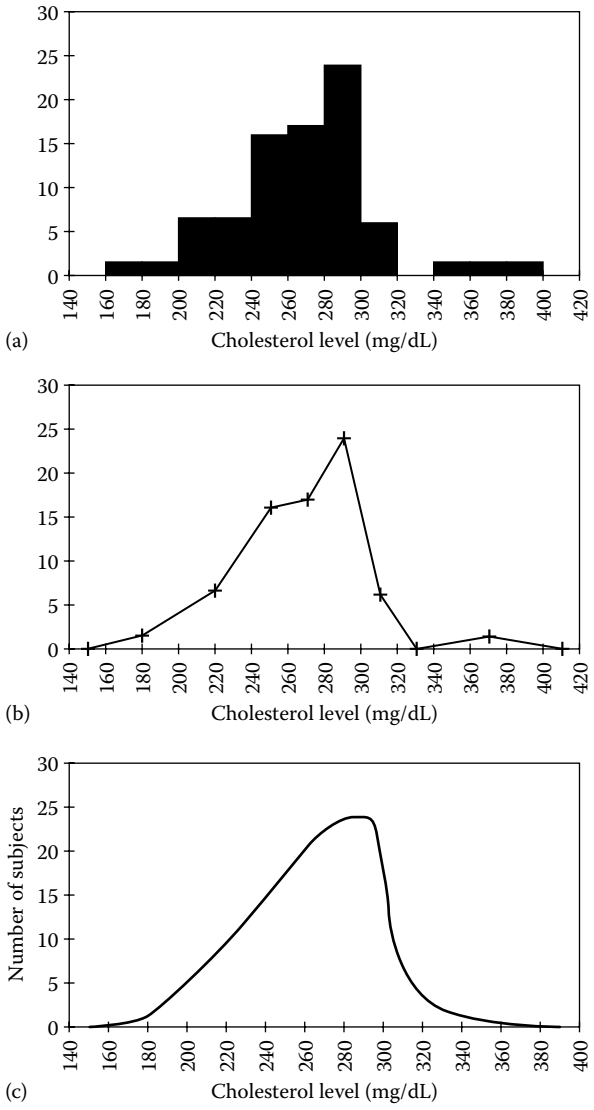



**FIGURE F.6** Different forms or representation of frequencies: (a) Histogram for cholesterol data in Table F.8; (b) frequency polygon for cholesterol data in Table F8; (c) frequency curve for the cholesterol data in Table F.8.

### *Frequency Curve*

Imagine the shape of a frequency polygon when the number of subjects is extremely large and the width of intervals is extremely small. If the data are for 3000 cases and the cholesterol intervals are only 5 mg/dL, the polygon will tend to take the shape of a curve. This can be considered a smooth version of the polygon. Theoretically, a curve would be obtained when the subjects are infinite and the intervals are infinitely small. This is called a frequency curve. A fairly good approximation of this "infinite" curve can be obtained by drawing a smooth curve closest to the polygon. A few points may be far off from the curve so drawn. The curve for the data in Table F.8 is shown in Figure F.6c. Note that a frequency of 0 in the interval 320–339 mg/dL is ignored to achieve a smooth and regular shape of the frequency curve. When an extremely large number of measurements are available, the frequency in the interval 320–339 mg/dL will not be 0.

Quite often, a frequency curve takes a regular shape and is best obtained by means of a mathematical equation. Such an equation seems to work well as a good approximation to the data arising in

practice. This is called a frequency distribution or just a statistical distribution. **Gaussian distribution** is the most common example of a frequency distribution.

Basic information provided by the histogram, polygon, or curve is the nature of the **distribution** of the subjects over various values of the variable, i.e., whether they are evenly distributed or are concentrated around some value, and whether the values are widely scattered or are compact. Thus, these figures are more of an exercise in data exploration than an analysis of data. The method of **statistical analysis** of the data depends on the shape of the distribution.

## frequency (in a cell), *see* cell frequency

## frequentist approach, *see* Bayesian inference

## Friedman test

Developed by Milton Friedman in the years 1937–1940 [1–3], the Friedman test is a **nonparametric test** for **repeated measures** in a **one-way design** for quantitative outcome for the null that the responses in repeated measures have the same distribution. This test is primarily for correlated values such as at different points in time but can be used in a two-way setup when values at different levels of the same factor could be correlated [4]. The Friedman test will test for either factor 1 or for factor 2 considering that the other factor is nuisance and is just a replicate with no intrinsic importance. Thus, the Friedman test is an alternative for an ***F*-test** for a **two-way design** with one observation per cell for the situations where the observed values do not follow any **Gaussian** pattern. The *F*-test is not applicable to non-Gaussian data, particularly when the sample size is small. However, an *F*-test can be used to test for interactions, whereas the Friedman test cannot do this.

We explain the Friedman test with the help of an example of a two-way design with one observation per cell. Suppose one is studying the cholesterol level in different types of hypertension, and suppose the interest is in simultaneously considering obesity as well. Both are not to be considered together, but only one at a time. Let the data be as in Table F.9. As in possibly all nonparametric tests, we need to **rank** the values from the minimum to the maximum. These ranks are given in the parentheses in the table.

**TABLE F.9**
**Total Cholesterol Level (mg/dL) in Persons with Different Levels of Obesity and Types of Hypertension**

| | Obesity (Factor 2) | | |
|---|---|---|---|
| **Hypertension Group (Factor 1)** | **Thin** | **Normal** | **Obese** |
| No hypertension | 248 | 233 | 263 |
| | (2) | (1) | (3) |
| Isolated diastolic hypertension | 247 | 249 | 258 |
| | (1) | (2) | (3) |
| Isolated systolic hypertension | 261 | 275 | 267 |
| | (1) | (3) | (2) |
| Clear hypertension | 225 | 290 | 285 |
| | (1) | (3) | (2) |
| **Total of ranks ($R_{\cdot k}$)** | **(5)** | **(9)** | **(10)** |

*Note:* Ranks are in parentheses.

This is a two-way layout. Factor 1 is hypertension status, and factor 2 is obesity. The former has $J = 4$ levels and the latter has $K = 3$ levels. A separate test for the effect of hypertension status and obesity on cholesterol level can be carried out by the Friedman test. The underlying distribution can be non-Gaussian. The procedure for testing for differences between levels of factor 1 in a two-way layout can be stated as follows:

Step 1. For each $k$th level of factor 2, rank $J$ observations belonging to $J$ levels of factor 1 in order of magnitude from 1 to $J$. Denote the rank of the observation for the $j$th level of factor 1 and $k$th level of factor 2 by $R_{jk}$ ($j = 1, 2, \ldots, J$; $k = 1, 2, \ldots, K$).

Step 2. Calculate $R_{j\cdot} = \Sigma_k R_{jk}$. This is the sum of the ranks received by $K$ subjects, one each in $K$ levels of factor 2. If $H_0$ of no difference between levels of factor 1 is true, $R_{j\cdot}$ would be nearly same for all $j$ ($j = 1, 2, \ldots, J$). This leads to the criterion $S$ due to the Friedman, as specified in the next step.

Step 3. Calculate

$$\text{Friedman test for factor 1: } S_1 = \frac{12}{JK(J+1)}\Sigma_j R_{j\cdot}^2 - 3K(J+1).$$

Step 4. If $S_1$ is large, the probability of the sample values coming from a distribution as specified by $H_0$ is small. Reject $H_0$ if the ***P*-value** is less than the predetermined **significance level**. Computer packages may give the *P*-value right away. Otherwise, use Friedman tables, for example, as given by Lindley and Scott [4]. This provides critical values of $S_1$ for small values of $J$ and $K$ for $\alpha = 0.05$. If $K$ or $J$ is large, $S_1$ has an approximate **chi-square distribution** with $(J - 1)$ degrees of freedom (df's). In that case, use the chi-square distributions to obtain the *P*-value.

The preceding procedure is for testing differences in the levels of factor 1 ignoring factor 2. For testing differences in the levels of the other factor, $J$ and $K$ switch their position.

$$\text{Friedman test for factor 2: } S_2 = \frac{12}{JK(K+1)}\Sigma_k R_{k}^2 - 3J(K+1).$$

Again, for large $J$ or $K$, $S_2$ has an approximate chi-square distribution with $(K - 1)$ df's.

Consider the data in Table F.9 on total plasma cholesterol level (in mg/dL) in 12 subjects belonging to different hypertension groups. In parentheses are ranks within each hypertension category. In this example, $J = 4$ and $K = 3$. For levels of obesity (factor 2),

$$S_2 = \frac{12}{4\times 3(3+1)}(5^2 + 9^2 + 10^2) - 3\times 4(3+1)$$
$$= 3.5.$$

For this value of $S_2$, when $J = 4$ and $K = 3$, $P = 0.273$ from a statistical package. This is more than 0.05. Thus, the evidence is not enough to conclude that obesity affects cholesterol level in these subjects. You may wish to calculate $S_1$ for hypertension groups. This is $S_1 = 3.40$ and corresponds to $P = 0.446$. Thus, there is no sufficient evidence for a difference in cholesterol levels in different hypertension groups either.

The following comment for the Friedman test may be useful. The chi-square approximation is used for fairly small sample sizes. This is quite an approximation. Such an approximation becomes

necessary mostly because the exact distribution of these criteria, particularly for the Friedman test, is very difficult to obtain. In case you have software that gives you exact probabilities, there is no need to use chi-square approximation.

For $K = 2$, the Friedman test is reduced to a two-sided **sign test**. For more details, see Conover [5].


1. Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Stat Assoc* 1937 Dec;32 (200):675–701. http://sci2s.ugr.es/keel/pdf/algorithm/articulo/1937-JSTOR-Friedman.pdf
2. Friedman M. A correction: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Stat Assoc* 1939 Mar;34 (205):109. http://www.jstor.org/discover/10.2307/2279169
3. Friedman M. A comparison of alternative tests of significance for the problem of m rankings. *Annals Math Stat* 1940 Mar;11(1):86–92. http://www.jstor.org/discover/10.2307/2235971
4. Hollander M, Wolfe DA. *Nonparametric Statistical Methods*. Wiley, 1973.
5. Conover WJ. *Practical Nonparametric Statistics*, Third Edition. Wiley, 1999.


## *F*-test

An *F*-test is used to compare variance in one group of subjects with variance in another group of subjects when both groups are independent of one another. The values also should be independent of one another in each group. If they are serial values where a value depends on its previous value, independence is lost, and this test will not be applicable. Independence is also lost when values belong to some kind of affinity group or a cluster, such as family members or those sharing some other environment. The **null hypothesis** under test is that the variances in their respective populations are the same, i.e., $H_0: \sigma_1^2 = \sigma_2^2$.

Equality of variances is an important requirement of many statistical procedures, such as the **Student *t*-test** for two groups. The *F*-test helps to find whether that requirement is violated in the two groups you are studying. For this test, the ratio of the two variances is used as the test criterion, and because of this, it is also called the **variance ratio test** or ***F*-ratio test**. Under the null, this ratio would be equal to 1, and any large deviation from 1 is evidence that the variances are unequal. Since two variances are involved, it has a pair of **degrees of freedom** (df's), one for the numerator and the other for the denominator. The method is applicable to a variety of situations as described next, such as analysis of variance (ANOVA) for comparing means in three or more groups and regression where, also, the test actually is of equality of variances.

The *F*-test is valid only when the underlying distribution of values in each group follows a **Gaussian distribution**, at least nearly so. For other underlying distributions, the **Levene test** can be used for equality of variances under the conditions stated under those topics. Conventionally, though, **transformations** are advised so that the transformed data follow a Gaussian distribution. Generally, logarithm, square root, or inverse transformation does the trick.

Statistically, *F* is the ratio of two independent **chi-square** variables, each divided by its df's. Chi-square with $v$ df's itself is defined as the sum of squares of $v$ independent **standard Gaussian variables**. When the underlying distribution is Gaussian, it can be shown that the sample variance is the sum of squares of $(n - 1)$ independent standard Gaussian variables divided by $(n - 1)$. Thus, the ratio of sample variances in two independent groups is the ratio of two independent chi-squares, hence equal to *F*. The shape of the distribution of *F* is generally highly skewed just as is the shape of chi-square.

### F-*Test for Equality of Two Variances*

The most elementary application of the *F*-test is as just stated—comparison of variances in two independent groups. The criterion for this test is $F = s_1^2/s_2^2$, which is the ratio of the two sample variances. This follows an *F*-distribution provided that the variances are equal. The df's for this test are $(n_1 - 1, n_2 - 1)$, where $n_1$ and $n_2$ are the respective sample sizes. It is customary to consider the group with larger sample variance as group 1 so that the ratio is not less than 1. But this is not necessary. One can use the relationship that $1/F$ also has *F*-distribution with reversed $(n_2 - 1, n_1 - 1)$ df's. While working with statistical software packages, you will not need to do so as the software will take care of this and give you the correct ***P*-value**. Reject the null when the *P*-value is less than the predetermined **level of significance** $\alpha$. Let us reiterate that this test is valid only when the distribution of values in both groups is Gaussian.

### F-*Test for Equality of Means in Three or More Groups (ANOVA)*

The most common application of the *F*-test is in **analysis of variance**, where the objective is to find whether the means of three or more groups are equal or not. (For two groups, the **Student *t*-test** is used, although *F* can be used for two groups also.) Since there are three or more groups, the *F*-test cannot be used for testing equality of variances, although that is an important requirement of the *F*-test. Instead, the Levene test or **Bartlett test** is used for testing equality of variances in this setup. The *F*-test is used in this case to test whether or not the between-groups variance is the same as the within-groups variance in this setup. Note that between-groups variance is calculated as the variance between the group means. Substantially higher between-groups variance compared with within-groups variance is an indication that means across groups are substantially different from what are expected from natural variation between the values. These variances for the sample values are obtained by the concerned **sum of squares** divided by the respective df's. For a brief on this, see the terms **one-way ANOVA** and **two-way ANOVA** in this volume. For example, in the case of one-way ANOVA,

$$F = \frac{n\Sigma_k \left(\bar{y}_k - \bar{\bar{y}}\right)^2 /(K-1)}{\Sigma_k \Sigma_i \left(y_{ik} - \bar{y}_k\right)^2 /[K(n-1)]},$$

where $K$ is the number of groups and $n$ ($i = 1, 2, \ldots, n$) is the size of sample in each group; $\bar{y}_k$ ($k = 1, 2, \ldots, K$) is the mean of the $k$th group; and $\bar{\bar{y}}$ is the overall mean. You can see that the numerator is the between-groups variance in the sample and the denominator is the within-groups variance in the sample. Thus, basically, this is the same as the ratio of two sample variances. This can be extended to two-way and higher-way ANOVA and for testing the statistical significance of the **interactions** also. Another extension is for **repeated measures ANOVA**. As explained under the respective terms in this volume, all these tests use the ratio of two **mean sums of squares**, which are the estimates of the corresponding variances. The null of equality of means is rejected when the value of *F* exceeds a threshold at specified level of significance and the corresponding df's. Equivalently, the value of *F* also gives the *P*-value. This also can be used for decision one way or the other.

### F-*Test in Regression*

The next common application of the *F*-test is in checking the statistical significance in **regression**. For this, the regression variance is estimated by the regression sum of squares divided by its df's. This

can be obtained for any particular regressor or a subset of regressor variables in the regression. This is compared with the **mean square due to error (MSE)**, which is based on the residuals that remain unaccounted for by the regression. This is the sample estimate of the error variance. Again, under the null hypothesis of no effect of regressors on the dependent, these two estimates of variance should be equal and should give $F = 1$ under the condition of homoscedasticity. For the regression to be statistically significant, the $P$-value corresponding to the value of $F$ should be less than the prespecified level of significance $\alpha$. This can also be checked for any particular predictor or any subset of predictors.

## funnel plot

A funnel plot is the **scatter plot** of the estimates of the **effect size** under consideration from different sources against the size of the sample. The premise is that as the sample size increases, the variation in the values of the effect will decline. This happens because the variation of the effect is measured in terms of its **standard error (SE)** and SE has $n$ in the denominator, which decreases as $n$ increases. An example of a typical funnel plot is shown in Figure F.7 for odds ratios (ORs). Because this is a ratio, the plot on the $x$-axis is on a log scale.

A funnel plot is frequently used in **meta-analysis** where the effects reported in many studies are combined to obtain a more reliable estimate. Results based on small sample sizes or with high SE in different studies will obviously spread across a broad range of values. If you plot **odds ratios** in three studies each with a small sample size, they are likely to be further apart from one another compared with ORs in three other studies each with a large sample size. If you are reviewing a large number of studies—some of small size and some of large size—and plot OR on a horizontal axis and sample size on a vertical axis, the plot generally will be as shown in Figure F.7. This is called a funnel plot because of its resemblance with an inverted funnel.

In place of OR, you can have any other effect size such as **relative risk** (RR) and difference in means or proportions. On the vertical axis, you can have the inverse of the SE instead of sample size. An asymmetric shape of the funnel plot raises suspicion over the results

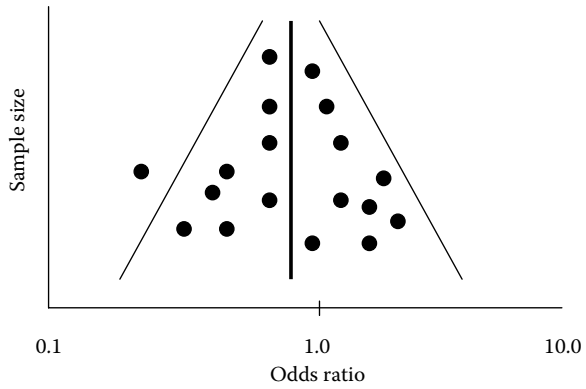



**FIGURE F.7** A typical funnel plot.

of meta-analysis since the selected studies may suffer from publication bias, favoring either higher or lower effect sizes. It also suggests the possibility of a systematic bias in smaller studies. Check whether most smaller studies tend to consistently give a larger (or smaller) effect size compared to larger studies. If so, the bias is evident, and the results of meta-analysis would be invalid. When biased studies are not included in meta-analysis, heterogeneity among results of various studies does not cause much of a problem. Your final confidence interval would depict this heterogeneity. However, funnel plot asymmetry can arise due to a host of other reasons as well, as discussed by Sterne et al. [2]. Besides publication bias, these include location bias, poor methods, inadequate analysis, and perhaps fraud. Asymmetry can also arise just due to chance in the studies included in meta-analysis.

1. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315:629–34. http://www.bmj.com/content/315/7109/629
2. Sterne JAC, Sutton AJ, Loannidis APJ, Terrin N, Jones DR, Lau J, Carpenter J et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011;343:d4002. http://www.bmj.com/content/bmj/343/bmj.d4002.full.pdf9

# G

**GAIC,** *see* **Akaike information criteria (AIC) and general AIC (GAIC)**

## gain from a medical test

The objective of a medical test is to help a clinician to come to a more **valid** decision regarding the presence or absence of disease. Gain from a medical test is the addition it makes to the chance of correct prediction. This is the difference between the pretest and posttest probabilities of the disease.

As explained under the term **predictivities**, these can be understood as the posttest probabilities. They measure the likelihood of the disease or of no disease after the test result becomes known. The pretest probability is the **prevalence** in the group from which the patient has come. Assuming that the patient is a random arrival from the group on which the test is done, the chance that he/she has the disease is the same as the prevalence rate of the disease in that group. A test is useful only if it substantially alters the posttest probability of disease or its absence compared with the pretest probability. This difference is the gain attained by the test. This is easily appreciated as gain when the terms *pretest* and *posttest probabilities* are used instead of *prevalence* and *predictivity*. The gain intimately depends on the prevalence of the disease, or its pretest probability. In any case, this depends on the sensitivity and specificity of the test. The following example illustrates the gain from a medical test.

Consider α-fetoprotein determination as a test for hepatoma in patients with cirrhosis. If this test has a sensitivity of 96% and specificity 88%, what is the gain from the test at different levels of prevalence?

Posttest probabilities for positive and negative outcomes have been obtained in Table G.1 with the following formulas as explained under the term **predictivities**:

$$\text{Positive predictivity: } P(+) = \frac{S(+) * p}{S(+) * p + [1 - S(-)] * (1 - p)}$$

$$\text{Negative predictivity } P(-) = \frac{S(-) * (1 - p)}{S(-) * (1 - p) + [1 - S(+)] * p},$$

where $p$ is the prevalence rate per unit, and $S(+)$ and $S(-)$ are the **sensitivity and specificity**, respectively.

In this example, where the sensitivity is 96% and specificity is 88%, the gain in $P(+)$ is maximum when the prevalence is between 20% and 40%. That is, when a person before the test is estimated to have nearly a 1-in-3 chance of the disease, the test result would be most useful in enhancing confidence in arriving at a decision about the presence of the disease. This substantially enhanced confidence may not be enough, even less than 80%, but that is the best achievable with the help of this test for that pretest likelihood.

The gain from a test has an important clinical implication. If you already believe before the test that the person has an extremely high chance of being positive or an extremely high chance of being negative, then generally, the application of the test could make only a marginal difference in your belief. In our example, though, when the pretest likelihood of the absence of the disease is as low as 10%, a negative test firms up the belief to 71% that the disease is absent. That is a very substantial gain.

These gains would vary from one set of sensitivity—specificity values to another set. You may want to confirm that for $S(+) = 80\%$ and $S(-) = 74\%$, the gains are maximum when the pretest chances are nearly 50–50.

For an example of the real application of the gain, see Wojcinski et al. [1]. They reported that sonoelastography increased the probability of detecting malignancy in BI-RADS-US 3 lesions of the breast from 4.5% pretest to 13.2% posttest. They concluded that sonoelastography yields additional diagnostic information in the evaluation of BI-RADS-US 3 lesions. They have also mentioned sensitivity, specificity, and predictivities of this test.

**TABLE G.1**
**Gain by the Test at Different Levels of Prevalence for Sensitivity 96% and Specificity 88%**

| Probability of Presence of Disease (%) | | | Probability of Absence of Disease (%) | | |
|---|---|---|---|---|---|
| **Pretest (Prevalence)** | **Posttest *P*(+)** | **Gain by the Test** | **Pretest (100 − Prevalence)** | **Posttest *P*(−)** | **Gain by the Test** |
| 10 | 47 | 37 | 90 | 99 | 9 |
| 20 | 67 | 47 | 80 | 99 | 19 |
| 30 | 77 | 47 | 70 | 98 | 28 |
| 40 | 84 | 44 | 60 | 97 | 37 |
| 50 | 89 | 39 | 50 | 96 | 46 |
| 60 | 92 | 32 | 40 | 94 | 54 |
| 70 | 95 | 25 | 30 | 90 | 60 |
| 80 | 97 | 17 | 20 | 85 | 65 |
| 90 | 99 | 9 | 10 | 71 | 61 |

1. Wojcinski S, Boehme E, Farrokh A, Soergel P, Degenhardt F, Hillemanns P. Ultrasound real-time elastography can predict malignancy in BI-RADS®-US 3 lesions. *BMC Cancer* 2013 Mar 27;13:159. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3618252/

## gamma distribution

For those not aware, a statistical **distribution** is the pattern of values over their range—what values are common and what values are rare, etc. Most well known is the **Gaussian distribution**, with a peak in the middle and symmetric decline on either side. Gamma distribution typically is **skewed** to the right and is used for those variables that cannot have negative values. For example, blood sugar level in a population is likely to have a right-skewed distribution since many people will have values more than the **mode** and few people will have less (Figure G.1). Thus, blood sugar level in a population can be modeled to follow a gamma distribution. The name comes from the mathematical gamma functions used as the divisor in this distribution (referred to in the mathematical literature as the Euler integral of the first kind—the second kind being beta) to normalize it so that the area under the curve is 1, as is needed for any statistical distribution.

Just as a Gaussian distribution has two **parameters**, namely, the mean μ and standard deviation (SD) σ, the gamma distribution also has two parameters, generally denoted by α and β. These parameters do not have simple interpretation such as mean and SD. As a medical professional, you may never have to use its mathematical form, but for the information of those interested, the density function of the gamma distribution that defines the pattern of values is given by

$$\text{Gamma distribution: } f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x};\ x > 0;\ \alpha \text{ and } \beta > 0,$$

where Γ(α) is, what is called a *gamma function*. In its simple form, when α is an integer, Γ(α) = (α − 1)!. This is read as (α − 1) factorial and, for a positive integer α, defined as α × (α − 1) × (α − 2) × … × 2 × 1. For example, 5! = 5 × 4 × 3 × 2 × 1 = 120. The values of α and β determine the shape of the distribution. For example, in Figure G.1, α = 2 and β = 0.5. The mean of this distribution is α/β, and the variance is $\alpha/\beta^2$. This gives β = (mean)/(variance) and α = $(\text{mean})^2$/(variance). In case you have data such as blood sugar level of 80 persons, you can find the sample mean and sample variance, and use these values to estimate α and β.

The application of the gamma distribution in health is common for modeling the cost of care, even when calculated per day. Generally, this cost is around a specific value *x*, but in some cases, it is relatively high, and in some others, steeply high—giving rise to the right-skewed pattern of a gamma distribution. For example, Parthan et al. [1] used gamma distribution for modeling the cost of care of nonmetastatic breast cancer patients. The pattern of gamma distribution is also seen for most durations, such as duration of hospitalization or duration of disease. This could be around 3–4 days for most patients but could be 10 or more days in some patients and even 20 days in isolated cases. In neurosciences, interspike interval also follows a gamma distribution. Thus, gamma distribution has some notable applications in health and medicine.

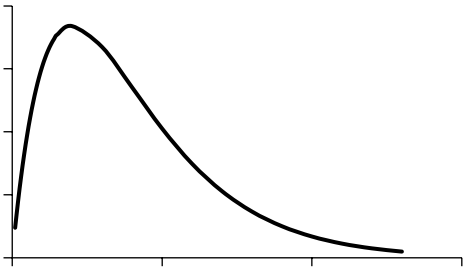


**FIGURE G.1** Gamma distribution with α = 2 and β = 0.5.

For those interested, note that **exponential distribution** is a special case of gamma distribution when α = 1. The popular **chi-square distribution** is also a special case of the gamma distribution where α = ν/2 and β = ½, ν being the degrees of freedom.

1. Parthan A, Santos E, Becker L, Small A, Lalla D, Brammer M, Teitelbaum A. Health care utilization and costs by site of service for nonmetastatic breast cancer patients treated with trastuzumab. *J Manag Care Pharm* 2014 May;20(5):485–93. http://www.amcp.org/WorkArea/DownloadAsset.aspx?id=18024

## Gantt chart

A Gantt chart is a visual depiction of the proposed timeline for carrying out various stages of a project (Figure G.2) and generally is a part of the research **protocol**. Bars one below the other are made for sequential activities, and the length of the bars is proportional to the time proposed to be taken. Figure G.2 shows for the project that identifying a research problem is envisaged to take a period of 30 days, and 15 days down the line, the process of collecting the existing information is expected to start. This activity is planned to take 60 days. And so on. Two or more activities can go together, and this is shown by overlapping bars. The collection of data will take 240 days, and scrutiny of the data will start after the data collection and will go on for 90 days.

A Gantt chart is useful to provide a full view of the stages of the project and the timeline. We have illustrated its simple use in the context of research. For another more complex usage of Gantt chart, see Sakaguchi [1].

1. Sakaguchi Y, Ishida F, Shimizu T, Murata A. Time course of information representation of Macaque AIP neurons in hand manipulation task revealed by information analysis. *J Neurophysiol* 2010;104:3625–43. http://jn.physiology.org/content/104/6/3625.long

## garbage-in garbage-out (GIGO) syndrome

A processing system requires inputs and delivers output. Garbage-in garbage-out syndrome says that when inputs are bad, the outputs cannot be good. This is often abbreviated as GIGO and is mostly used for computer systems. Most computers are programmed to process any input by a specified algorithm and produce some output, mostly without worrying about the applicability or the appropriateness of the process. Statistically, a **regression** model, for example, can be obtained irrespective of the values of dependent and independent variables. If these values or the variables are irrelevant or nonsensical, the model will still be obtained, but that could be equally useless. Thus, no regression model can be taken at face value. This is true for any statistical model. One can calculate mean, median, percentage, etc. for any set of data. Most statistical software packages still have not been given the intelligence to judge the appropriateness of the method. If the values were obtained by a wrong method or without care, the output values will also lead to wrong conclusions. Poor data can only lead to inapplicable results—sometimes even nonsensical.

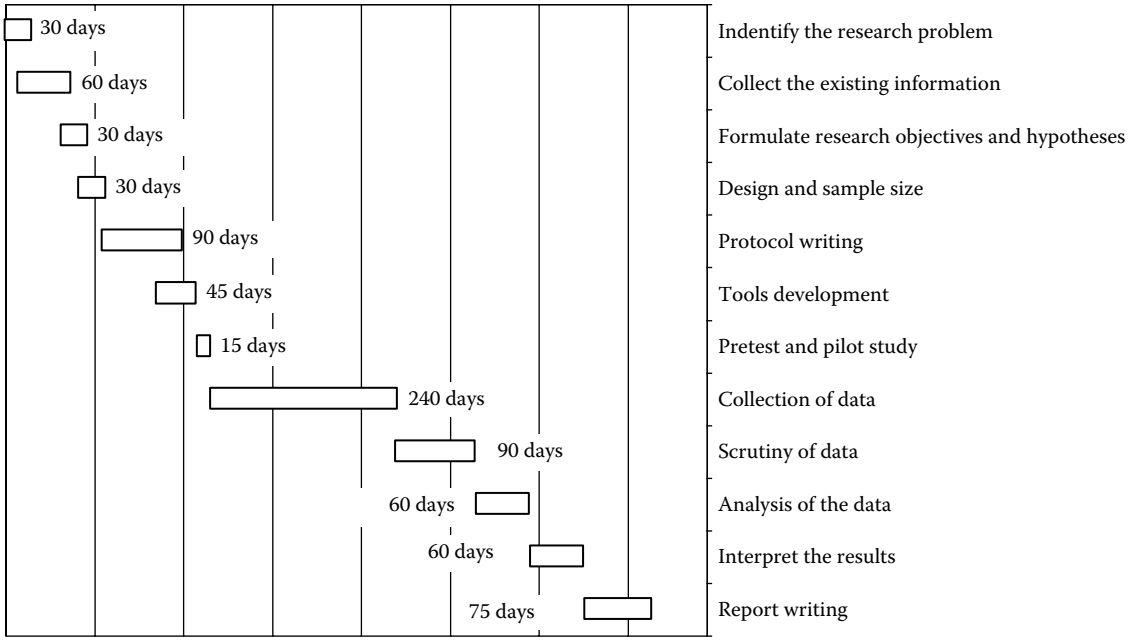



**FIGURE G.2** A Gantt chart showing the time to be taken for completing various stages of a project.

If medical research is not meticulously planned, the results cannot be expected to provide sufficiently valid conclusions. You will get some conclusions, but those conclusions will not apply to the situation you target. Thus, it is necessary that high-quality data are obtained on a well-defined topic and that they are adequately processed to get valid conclusions. Many research studies fail to reach to this expectation. Anthonisen [1] illustrates this without inhibition for research for noncredible results on obstructive sleep apnea.

1. Anthonisen NR. Garbage in, garbage out. *Can Respir J* 2010 Sep–Oct; 17(5):211–2. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2975499/

## Gaussian conditions

Gaussian conditions are those statistical requirements under which many **Gaussian distribution**–based biostatistical methods are applicable to non-Gaussian data. For example, strictly speaking, the **Student *t*-test** is applicable only when the data follow a Gaussian distribution. However, the test is robust and can be used with sufficient confidence even when the distribution is mildly different from Gaussian provided that Gaussian conditions are fulfilled. This is true for analysis of variance (ANOVA), ***F*-test**, and related tests such as the **Tukey test** and **Dunnett test**. **Binomial distribution** for proportions also approaches Gaussian distribution under Gaussian conditions. Primarily, the condition requires that the sample size ($n$) is large and the sample summary under consideration is of the type *mean*. No fast rule is available, but the following are generally accepted as Gaussian conditions:

For means of continuous data: $n \geq 30$
For proportions: $np \geq 8$ and $n(1 - p) \geq 8$, where $p$ is the proportion in the sample

The latter could mean that $n$ should be at least 26,667 when $p = 0.0003$. The smaller the $p$, the larger the requirement of $n$.

Note that these conditions work for (i) a slight departure from a Gaussian pattern and not for violent departures, and (ii) mean-type summaries and not, say, median or percentiles. These conditions arise from what is called the central limit theorem (CLT), which says that mean-type sample summaries that are based on a linear combination of values tend to become Gaussian as the sample size increases. As explained for the term **central limit theorem**, proportion and correlation coefficient, too, are mean-type summary measures and amenable to this manipulation, as are the differences between means, proportions, correlations, etc.—even the odds ratio and relative risk after taking logarithm, as they become linear combinations with this transformation.

When Gaussian conditions exist, there is no need to worry too much about the form of the distribution of the data unless you have extraneous information that the data would be extremely different from Gaussian. This gives considerable freedom to use various statistical tests and is one of the important reasons for statisticians advocating a large sample size.

## Gaussian deviate, *see also* Gaussian distribution

A Gaussian deviate is the difference of a Gaussian variable from its mean, measured in standard deviation (SD) units. If a variable $x$ has a Gaussian distribution with mean $\mu$ and SD $\sigma$, the Gaussian deviate is $z = (x - \mu)/\sigma$. More fully, this is called a *standardized Gaussian deviate*, and sometimes **Z-score**. This will have what is called a **standard normal distribution** or a standard Gaussian distribution. If pulse rate in a population of healthy adults has a mean of 70 per minute and the SD is 4, the Gaussian deviate for a person with pulse rate = 63 per minute is $(63 - 70)/4 = -1.75$, and for a person with pulse rate = 81 is $(81 - 70)/4 = 2.75$. The first person's pulse rate is 1.75*SD away from the mean on the lower side, and that of the second person, 2.75*SD away on the upper side.

Where do we need a Gaussian deviate? This deviate converts a variable with a Gaussian distribution to its standard in the sense that the deviate $z$ has mean = 0 and SD = 1. The primary advantage of such a conversion is achieving comparability. If the first person in our example has a respiration rate of 16 per minute and the population mean is $\mu = 17$ and SD $\sigma = 2.5$, the Gaussian deviate is $(16 - 17)/2.5 =$

−0.4. Now you can see that the pulse rate of this person is 1.75*SD below the mean, whereas respiration rate is only 0.4*SD below the mean. The pulse rate is far below its average compared to the respiration rate. This kind of information may have significance in clinical evaluation of the person. Such a comparison cannot be made without resorting to the deviates—in this case, a Gaussian deviate if the distribution of pulse rate and respiration rate is Gaussian. In addition, well-known properties of the Gaussian distribution can be used to say that the pulse rate is on the border of being unlikely in a healthy person since it is almost 2*SD away from mean, whereas respiration rate is just about the average seen in healthy adults.

As an extension, consider the distribution of sample mean $\bar{x}$. Because of the **central limit theorem (CLT)**, this follows a near-Gaussian distribution for large $n$ with mean $\mu$ and standard error (SE) (the SD of a sample summary is called SE) $\sigma/\sqrt{n}$ irrespective of the distribution of $x$. Thus, the Gaussian deviate is $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$. Similarly, for sample proportion $p, z = \frac{p-\pi}{\sqrt{\pi(1-\pi)/n}}$. These kinds of deviates are the basis for the Gaussian ***z*-test**, as described under this topic.

## Gaussian distribution

If you are not already familiar with **frequency curves** and **distributions (statistical)**, you may want to review those topics. That will help you to understand this section better.

A continuous variable $x$ is said to have a Gaussian distribution when it has the highest chance of being around the value of its mean and the chance tapers off in a specific fashion symmetrically on either side to become extremely unlikely to have values far away from the mean. The shape is typically that of a bell, and the curve is called a *bell-shaped curve*. Mathematically, a Gaussian distribution with mean $\mu$ and SD $\sigma$:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}; \quad -\infty < x, \mu < +\infty; \ \sigma > 0.$$

This looks like a complex form, but it conveys what is already stated in nonprecise terms regarding the highest chance around the mean and its tapering off on either side. The equation specifies the peak and how the chances decline on either side. The shape of the distribution is symmetrical around the mean. The value of the mean defines its location, and the value of the SD defines its scatter (Figure G.3). Nothing else is required to specify a Gaussian distribution.

This kind of curve arises with random errors that concentrate around 0 but occasionally can be far away too. For this reason, this is called the normal curve of error. As a side note, many would be surprised that something as chaotic as random errors follow a nice predictable pattern. This distribution originated with de Moivre in 1733 [1] and was popularized by Karl Gauss when he showed that this distribution holds for random errors.


Karl Gauss

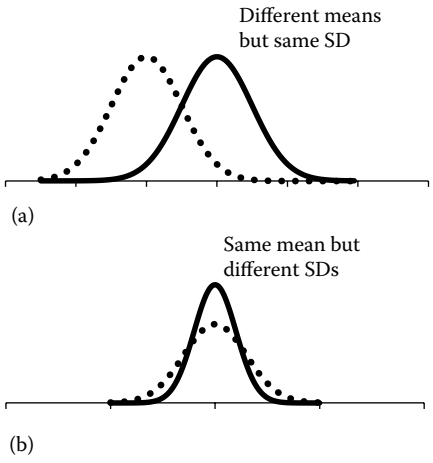



**FIGURE G.3** Gaussian distributions with (a) different means but same SD and (b) same mean but different SDs.

The Gaussian distribution is the most commonly occurring statistical distribution and has been extensively studied. This is commonly called **normal distribution**, but we prefer to call it Gaussian because of other connotations of "normal" in health and medicine. Many medical parameters follow this pattern in healthy subjects. For example, when unhealthy people are excluded, pulse rate, body temperature, kidney functions, liver functions, etc. mostly follow a Gaussian pattern. However, in sick subjects, the values may be **skewed**, sometimes highly skewed. For contrasting with other shapes of statistical distributions, see the topic **distributions**.

### *Properties of a Gaussian Distribution*

A Gaussian distribution has the following properties:

Property 1: The shape is symmetric like a bell.
Property 2: The mean, median, and mode coincide.
Property 3: The limits from (mean − 2SD) to (mean + 2SD) cover the measurements of nearly 95% of subjects. These are referred to as ±2SD limits or sometimes as 2-sigma limits (Figure G.4).

Another often-cited property of a Gaussian distribution is that the limits from (mean − 3SD) to (mean + 3SD) cover almost all subjects (99.7% to be exact). These 3-sigma limits are rarely used in health and medicine except in **quality control**, such as in a laboratory.

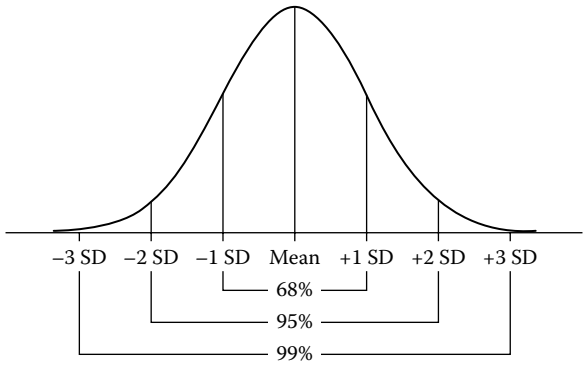



**FIGURE G.4** Mean ± 1, 2, and 3 SD limits in a Gaussian distribution.

Properties 1 to 3 hold true for the population as a whole but can also be used for samples as an approximation. The approximation works well when the sample is representative and the size is large. Sometimes, these properties are used inversely. When these properties hold for a set of data, the distribution is considered Gaussian. This inverse property also works well in most practical situations.

For illustration, consider the serum iron data in Table G.2 for a random sample of 165 healthy subjects. This has mean = 1.207 mg/L, median = 1.199 mg/L, and mode = 1.196 mg/L. These three are nearly equal. The smooth shape of the distribution is shown in Figure G.5. For these data, SD = 0.227 mg/L. Thus, the ±2SD limits are

$$(1.207 - 2 \times 0.227) \text{ to } (1.207 + 2 \times 0.227),$$

or 0.75 to 1.66.

Nearly 94% of healthy subjects mentioned in Table G.2 have serum iron levels within these limits, whereas ±2SD should have 95%. This difference of nearly 1% is due to **sampling fluctuation** and is likely to vanish as the sample size increases.

The utility of ±2SD limits in the practice of medicine is severalfold. Foremost is in determining the **normal range** for healthy subjects so that the value seen in a new subject can be assessed to be within the limits, borderline, or grossly outside. For this, the normal range must be based on the mean and SD of values in healthy subjects. Values far away from these limits are suspected to arise from health aberrations and investigated for factors causing such "abnormal" values. The second utility is in finding the **confidence interval** for the value of a parameter such as population mean μ and probability of an occurrence π. For this, the Gaussian distribution of the sample summary statistics, such as sample mean and sample proportion, is used provided that **Gaussian conditions** hold. Third is in the testing of a hypothesis for finding the chances of a sample coming from a population where the null hypothesis is true. Because of such wide and varied uses, many consider Gaussian distribution as the backbone that holds up most statistical methods.

When you know that a variable has a Gaussian distribution, you can easily find the probability that its value is within a specified range or more than or less than a particular value. For example, if the serum iron level has a Gaussian distribution, you can easily find the chance that it is more than 1.32 mg/L or that it is between 0.75 and 1.12 mg/L for any random person. How to do this is explained under the topic **Gaussian probability (how to obtain)** in this volume. Gaussian distribution can also be **bivariate Gaussian** or **multivariate Gaussian distribution**, as explained under those respective topics.

1. Frequency curve. http://www-groups.dcs.st-and.ac.uk/~history/Curves/Frequency.html, last accessed 23 July 2015

**TABLE G.2**
**Serum Iron Level in 165 Healthy Subjects**

| Serum Iron (mg/L) | Number | Percent |
|---|---|---|
| 0.50–0.69 | 2 | 1.2 |
| 0.70–0.89 | 12 | 7.3 |
| 0.90–1.09 | 33 | 20.0 |
| 1.10–1.29 | 68 | 41.2 |
| 1.30–1.49 | 34 | 20.6 |
| 1.50–1.69 | 13 | 7.9 |
| 1.70–1.89 | 3 | 1.8 |
| **Total** | **165** | **100.0** |

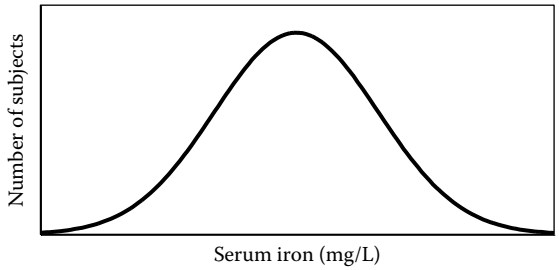



**FIGURE G.5** Distribution of serum iron in healthy subjects (smooth curve).

## Gaussian probability (how to obtain)

Gaussian probability is the chance that the value of a variable with **Gaussian distribution** lies in certain specified range. Statistical packages give this probability easily once you specify its mean and standard deviation (SD), but the meaning and application of this probability is best understood when obtained with the help of what is called a *Gaussian table*. This table is built for the standard **Gaussian** (normal) **distribution**. When a variable has mean μ, the difference (variable − μ) has a mean of 0. The SD, however, remains σ as before. This difference is called a deviate. When a deviate is divided by its SD σ, it is called a relative deviate or a **standardized deviate**. For a Gaussian distribution, this is called a **Gaussian deviate**. For a variable $x$, the standardized deviate is $(x - \mu)/\sigma$ whether the distribution is Gaussian or not. The mean of this deviate is 0, and SD is 1. The standardized deviate for any variable is denoted by $z$.

The standardized Gaussian deviate can be used to find a probability of any type related to a Gaussian distribution with any mean and any SD. For this, convert the variable to its standard form with the help of the just-mentioned equation and then use a Gaussian table. This table generally gives the probability to the right of a given value of $z$, that is, $P(z \geq a)$. This probability is the same as the corresponding area under the curve. See the shaded area in Figure G.6a.

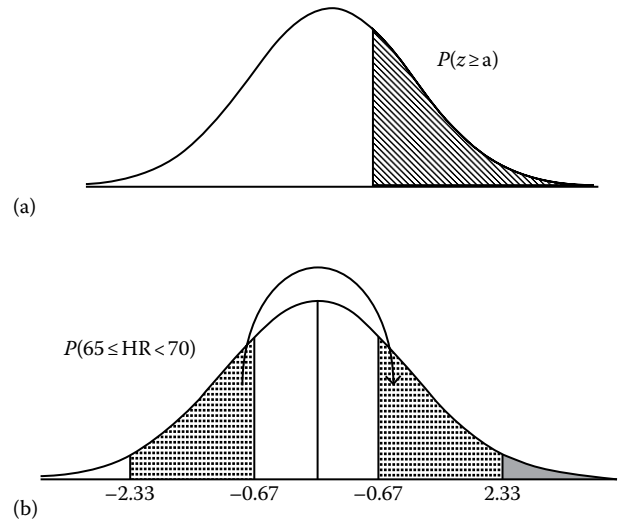



**FIGURE G.6** (a) The shaded area is the probability generally given in Gaussian tables. (b) Use of the symmetric property of the Gaussian distribution.

Figure G.6b illustrates how to use the symmetric property of the Gaussian distribution to get other kinds of probabilities.

The value of $z$ such that the probability to its right is $a$ is denoted by $z_a$, i.e., $P(z \geq z_a) = a$. Statistical software can give this probability, but we illustrate with the help of Gaussian tables available in statistical books that give this probability. For example, from a Gaussian table, $z_{0{\cdot}025} = 1.96$. This notation is different from the one used in other topics but is more convenient here. Its complement, $P(z \leq z_a)$, is called the cumulative probability till $z_a$ and is denoted by $\Phi(z_a)$. For example, $\Phi(1.96) = 0.975$. The following example illustrates the calculation and one use of these probabilities.

You know that heart rate (HR) varies from individual to individual even in healthy subjects. Suppose this follows a Gaussian pattern in a population with mean HR = 72 per minute and SD = 3 per minute. (i) What is the probability that a randomly chosen subject from this population has an HR of 74 or higher? (ii) What percentage of people in this population will have HR between 65 and 70 (both inclusive) per minute?

Since mean = 72 and SD = 3, the standardized Gaussian deviate is

$$z = \frac{\text{HR} - 72}{3}.$$

For HR = 74, $z = (74 - 72)/3 = 0.67$. Thus, $P(\text{HR} \geq 74) = P(z \geq 0.67)$. From the Gaussian table, this probability is 0.2514. In other words, nearly 25% of these healthy subjects are expected to have an HR of 74 or higher.

For (ii), proceed as follows:

$$P(65 \leq \text{HR} \leq 70) = P(\text{HR} \leq 70) - P(\text{HR} < 65)$$

$$= P\left(\frac{\text{HR} - 72}{3} \leq \frac{70 - 72}{3}\right) - P\left(\frac{\text{HR} - 72}{3} < \frac{65 - 72}{3}\right)$$

$$= P(z \leq -0.67) - P(z < -2.33)$$

$$= P(z \geq 0.67) - P(z < 2.33) \text{ because of the symmetric property of the Gaussian distribution (Figure G.6b)}$$

$$= 0.2514 - 0.0099 \text{ from Gaussian distribution}$$

$$= 0.24.$$

Thus, nearly 24% of these subjects are expected to have HR between 65 and 70. However, this answer is far too approximate, as explained next.

### Continuity Correction

The Gaussian distribution is meant for continuous variables. For a really continuous variable, $P(z > 2.33) = P(z \geq 2.33)$, that is, it does not matter whether or not the equality sign is used. This is what was done in the preceding calculation. Consider the following.

Realize that a variable such as HR is measured as **discrete** but can be considered **continuous** because of the large number of its possible values. However, in doing so, it is assumed that the rate 70 is a manifestation of values between 69.5 and 70.5. Strictly speaking, when HR is measured, it is not necessarily exactly 70 per minute. While counting beats, it is possible that there are 70 in 59.7 s and 0.3 s remains. In other words, if these are counted for 10 min, the number may reach 704. Thus, a rate of 70.4 per minute is not impossible. In that sense, it is not wrong to say that the rate 70 really means that it is between 69.5 and 70.5. This is called correction for continuity.

When this is acknowledged, HR between 65 and 70 (both inclusive) is actually HR between 64.5 and 70.5. Thus, the exact probability that HR is between 65 and 70 is

$$P(64.5 \leq \text{HR} < 70.5) = P(\text{HR} < 70.5) - P(\text{HR} < 64.5)$$

$$= P\left(\frac{\text{HR} - 72}{3} < \frac{70.5 - 72}{3}\right) - P\left(\frac{\text{HR} - 72}{3} < \frac{64.5 - 72}{3}\right)$$

$$= P(z < -0.50) - P(z < -2.50); \text{ see the equation for } z$$

$$= P(z > 0.50) - P(z > 2.50) \text{ because of the symmetry}$$

$$= 0.3085 - 0.0062 \text{ from Gaussian distribution}$$

$$= 0.30.$$

Now, with the correction for continuity, nearly 30% of subjects in this healthy population are expected to have an HR between 65 and 70. This answer is more accurate than the 24% reached earlier without the continuity correction. Note how this correction can affect the probability. The probability of HR ≥ 74 calculated earlier will also change accordingly.

Gaussian probability can be computed for any variable that follows a Gaussian distribution. The most common for sample summaries are the sample mean and sample proportion. We illustrate these with the help of an example each.

### Probabilities Relating to the Mean and the Proportion

The same sort of calculations can be done to find various probabilities for sample mean $\bar{x}$ and sample proportion $p$ provided that they follow an approximate Gaussian distribution as per the **Gaussian conditions**. The following examples may fix the ideas.

Suppose a sample of size $n = 16$ is randomly chosen from the same healthy population as in the previous HR example. What is the probability that the mean HR of these 16 subjects is 74 per minute or higher? Since the distribution of HR is given as Gaussian, the sample mean also will be Gaussian despite $n$ not being large. For the sample mean, the mean is the same $\mu$, but the SD of mean, now called **standard error (SE)**, is $\sigma/\sqrt{n}$. Thus,

$$P(\bar{x} \geq 74) = P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \geq \frac{74 - 72}{3/\sqrt{16}}\right)$$

$$= P(z \geq 2.67), \text{ as per the explanation just given for sample mean}$$

$$= 0.0038 \text{ from Gaussian table.}$$

This probability is less than 1%, whereas the probability of individual HR ≥ 74 is nearly 0.25 (see the previous part of this example). This happens because the SE of $\bar{x}$ is $3/\sqrt{16} = 0.75$, which is substantially less than SD = 3. The lower SE indicates that the values of

$\bar{x}$ will be very compact around its mean of 72, and very few $\bar{x}$s will ever exceed 74 per minute if the sample size is $n = 16$.

Realize that the mean, not only of HR but of any variable, in any case can be a fraction. In this example, the mean 74 is actually 74.0. Thus, there is no need for any continuity correction when calculating probabilities for the mean.

Now take an example of qualitative data where the interest is in proportion instead of mean. Consider an undernourished segment of a population in which it is known that 25% of births are preterm (<36 weeks). Thus, $\pi = 0.25$. In a sample of $n = 60$ births in a random month in this population, what is the chance that the number of preterm births would be less than 10?

Since $n\pi = 15$ in this case, which is more than 8, the Gaussian condition is fulfilled, and Gaussian approximation can be safely used. The probability required is

$$P(\text{preterm births} < 10) = P(p < 10/60),$$

where $p$ is the proportion of preterm births in the sample. Since the mean of $p$ is $\pi$ and $\text{SE}(p) = \sqrt{\pi(1-\pi)/n}$, this probability can be obtained as follows:

$$P(\text{preterm births} < 10) = P\left(\frac{p-\pi}{\text{SE}(p)} = \frac{\frac{10}{60} - 0.25}{\sqrt{0.25(1-0.25)/60}}\right)$$

$$= P(z < -1.49), \text{ as per the application for sample proportion}$$

$$= P(z > 1.49) \text{ because of the symmetry of } z$$

$$= 0.0681 \text{ from Gaussian table.}$$

Thus, there is a less than 7% chance that the number of preterm births in this population on a random day would be less than 10 out of 60.

## Gaussianity (how to check)

Many statistical methods are probably overly dependent on **Gaussian distribution**, although this is not a strict requirement for **sampling distributions**, because of the **central limit theorem**. Many times, you would want to know whether the pattern of your data is Gaussian or not—for example, to decide whether the median would be a more appropriate **central value** or the mean would be more appropriate. For sampling distribution also, knowledge about the pattern of the original values is helpful. For example, there is a tendency for small sample $\bar{x}$ to follow the same kind of distribution as the individual $x$'s. The distribution of duration of labor in childbirth is known to be **skewed** to the right. That is, a long duration (relative to the mode) is more common than a short duration. The distribution of mean duration in a small sample of, say, eight women is also likely to follow the same pattern, although in attenuated form. In such cases, if the pattern is not known, it is worthwhile to investigate whether or not it is Gaussian.

Some gross methods for assessing Gaussianity are (i) studying shape of a **histogram** or **stem-and-leaf plot** and checking if it follows the symmetric pattern expected for a Gaussian distribution; (ii) approximating equality among the values of mean, median, and mode; (iii) a **quartile plot** or **box-and-whiskers** plot; and (iv) a **proportion-by-probability (P-P) plot** and **quantile-by-quantile (Q-Q) plot**. All these are presented in this volume under the respective topics. The other alternative is to calculate **standardized deviate** for each observed value, order them from minimum to maximum, and plot them against the corresponding Gaussian probability on a probability paper. This is called **normal probability plot**. If the distribution of observed values is Gaussian, the plot will be nearly a straight line. Another method is to check if mean ±1SD covers nearly two-thirds and mean ±2SD nearly 95% of the values. The range should be nearly 6SD.

More exact methods are based on calculations that check statistical significance from the postulated pattern, such as Gaussian. An overview of these is as follows.

### *Overview of Significance Tests for Assessing Gaussianity*

Although the following tests are discussed in the context of assessing Gaussianity, the methods are general and can be used to assess whether the observed values fall into any specified pattern. Ironically, all these methods require large $n$, in which case the sampling distribution of $\bar{x}$ and of $p$ tends to be Gaussian anyway. The methods would still be useful if the interest is in assessing the distribution of original values rather than of a sample mean. A useful method is the **goodness-of-fit test** based on chi-square. This is based on the proportion of values in various class intervals and is presented under that topic. Other methods are as follows.

Among several statistical tests for Gaussianity, the three most popular are the Shapiro–Wilk test, Anderson–Darling test, and Kolmogorov–Smirnov test. All these are mathematically complex. Statistical software packages generally have a routine for these tests that you can easily apply. However, it is important that you understand the implications.

The **Shapiro–Wilk test** focuses on a lack of symmetry particularly around the mean. This test is not very sensitive to differences toward the tails of the distribution. On the contrary, the **Anderson–Darling test** emphasizes the lack of a Gaussian pattern in the tails of the distribution. This test performs poorly if there are many ties in the data. That is, for this test, the values must be truly continuous. The **Kolmogorov–Smirnov test** works well for relatively larger $n$, and when the mean and SD of the postulated distribution are known a priori and do not have to be estimated from the data. This also tends to be more sensitive near the center of the distribution than at the tails. Details of these tests are mentioned under the respective topics.

The critical value beyond which the hypothesis is rejected in the Anderson–Darling test is different when a Gaussian pattern is being tested compared to when another distribution such as lognormal is being tested. The Shapiro–Wilk critical value also depends on the distribution under test. But the Kolmogorov–Smirnov test is distribution-free, as the critical values do not depend on whether Gaussianity is being tested or some other form. Only the procedure for calculation differs.

It may sound strange to some, but all these statistical tests cannot confirm Gaussianity, although they confirm, with reasonable confidence, the lack of it when not Gaussian. Gaussianity is presumed when its lack is not detected. This is a serious limitation for small sample size since no statistical test is sufficiently powerful to detect deviation if the sample size is small. For reasonable assurance of Gaussianity, an **equivalence test** possibly can be devised, although not directly available in books.

## Gaussian test, *see* *z*-test

## generalized estimating equations (GEEs)

The method of generalized estimating equations (GEEs) is used to estimate the parameters of a model where several correlated response (**dependent**) **variables** are investigated for their relationship with several explanatory (**independent**) **variables**. Thus, this is

an extension of the **generalized linear models** to the setup where the responses are correlated. The correlation can be because the same subjects are measured at several points in time (as in **longitudinal data**), each subject is measured at several sites (such as a particular brain function at several locations in the brain), subjects share a common environment (such as those living together in a family), or any other such setup. In short, such data can be called clustered in the sense that they are more similar within the cluster than outside the cluster. The GEE method provides a framework for analyzing such correlated data and could be useful for cluster randomized trials. The method is comprehensive since the data can be of almost any type—continuous (**Gaussian** or non-Gaussian) or discrete (proportions or counts), rate or ratio. It can be adapted to estimate **fixed effects**, **random effects**, and mixed effects. The difference between general linear models, generalized linear models, and GEEs is shown in Table G.3.



Consider the cholesterol level of 200 members of 70 randomly selected families. Note first that the number of persons in different families would be different and second that $n$ actually is 70 and not 200. You cannot calculate the **standard error (SE)** of the estimates with $n = 200$ as family is a cluster in this example. Because of common heritance and similar diet, the cholesterol values of members of the same family will be correlated. While the families are statistically **independent**, the individuals within the families are not, and thus, the methods of generalized linear models are not applicable. You can also have information on the sex of each person, his/her age, blood group, body mass index, physical exercise, etc., which could be the regressors in this setup. These could be **discrete** or **continuous** or a combination—that does not matter. The objective is to find the extent and form of the relationship between the regressors and the response variable. The response variable is the cholesterol level of the members of the families in this example. When the response variable is continuous, as in this example so far, you would hope that clustered values jointly follow a multivariate Gaussian pattern so that the usual multivariate method of **general linear models** can be used to estimate the **regression coefficients** and to test a **hypothesis** on them for their statistical **significance**. Besides the Gaussian pattern, this will also require an estimate of the **correlations** between the members of the same family and that these correlations are similar across families. The estimates of correlations will come from the sample values in this setup. Unequal clusters and the requirement of multivariate Gaussianity along with the estimate of the correlations and their homogeneity can be a challenge. In the generalized linear model setup, in place of actual cholesterol level, if you only have the information that it is within normal limits or is high, the dependent variable will be **dichotomous**. This will require that a **logistic** kind of relationship be studied. However, because of correlation among the family members, the usual logistic regression is not applicable. If there is no correlation, both these setups (and many others) can indeed be woven into a unifying method of generalized linear models. When correlations are present, these various setups can be studied by the GEE method.

The same setup arises when a characteristic is measured repeatedly over a period of time in a **longitudinal study**. For example, you may want to measure the pain score before and at 1, 2, 5, and 10 min after administering an anesthesia in a patient being prepared for a surgery. Some patients can be measured only one or two times and some others three or four times depending on how they respond to the anesthesia. The number of observations available on each subject may or may not be the same. But these pain scores of the same patients at different points in time will be correlated. The regressors in this setup could be hemoglobin (Hb) level as a marker of nutrition, blood pressure, body mass index, etc. The GEE method is the most commonly used method for analyzing such longitudinal data. Multiple observations on each patient at different points in time form a cluster in this example.

Since the mathematical details are too complex for most medical professionals, we try to explain the GEE method heuristically so that you have at least the elementary knowledge about this method. The strength of the method stems from (i) using only the mean (and to some extent variance) of the values but not any particular **distribution**—this makes it a semiparametric method, (ii) not requiring any joint or multivariate distribution of the clustered values, and (iii) not worrying much about the specific correlation structure. In fact, the correlation structure is considered kind of a nuisance, and it has been shown that the estimates obtained by the GEE method are statistically *consistent* in the sense that as $n$ increases, it tends to become the actual value in the population with certainty, mostly even when the correlation structure is misspecified. This method is able to produce reasonably valid SEs of the estimates of the regression coefficients—thus, believable **confidence intervals** can be obtained when the sample is truly representative. Unlike the usual methods of generalized linear models, the GEE method does not explicitly model between-cluster or within-cluster variation but directly models the mean response. Note that within-cluster variation is a kind of correlation that we have stated as a nuisance under this method.

The GEE method can be implemented by using an appropriate statistical package. But specifying it correctly for commands in the software is difficult. Also, deciphering the output provided by the software can be a challenge. Thus, this method should not be used by nonexperts.

The method requires that a working correlation structure for responses within clusters be specified, although the actual values of the correlations are not required. If you consider this appropriate, you have the option to consider that there is no correlation. This would mean that the values within clusters are independent and reduces the GEE method to the usual generalized linear modeling. The other option is that the correlation between the first and second values within a cluster is the same as between the first and third values, between the second and the fourth values, etc. This is called **exchangeable** or compound symmetry of the correlations. The third option is that the longitudinal values are **autocorrelated** with lag 1—that is, if the correlation between the first and second values is $\rho$, the correlation between the first and third values is $\rho^2$, etc. The fourth option is to consider that different values within

**TABLE G.3**
**Difference between General Linear Models, Generalized Linear Models, and Generalized Estimating Equations**

| Type of Response Variable | Condition on Responses | Method |
|---|---|---|
| Continuous, nearly Gaussian pattern | Uncorrelated | General linear models |
| Continuous (Gaussian or non-Gaussian) or discrete (proportions or counts, rates or ratios) | Uncorrelated | Generalized linear models |
| Continuous (Gaussian or non-Gaussian) or discrete (proportions or counts, rates or ratios) | Correlated | Generalized estimating equations |

*Note:* In all of these, the regressors can be of almost any type—continuous or discrete—and may define fixed or random effects.

the cluster have different correlations, called unstructured. One of these structures is required to solve what are called GEEs. These equations are obtained by maximizing the likelihood without using the Gaussian distribution, called quasi-likelihood. The choice of the correlation structure is not terribly important under the GEE method, but correct specification does help in getting more reliable estimates of the SEs.

Tang et al. [1] used the GEE method to identify the determinants of quality of life during the dying process of terminally ill cancer patients who were longitudinally followed till death, and concluded that optimal quality of life during the dying process may be achieved by interventions designed to adequately manage physical and psychological symptoms, enhance social support, lighten perceived sense of burden to others, and facilitate experiences of posttraumatic growth. Van Rijn et al. [2] studied the effects of single or multiple concordant human papillomavirus (HPV) infections at various anatomical sites (anal canal, penile shaft, and oral cavity) on type-specific HPV seropositivity by using logit link in GEEs. For details of such links, see **link functions**.

The GEE method was developed by Liang and Zeger [3] in 1986. For further details, see Hanley et al. [4]. For technical details, see Agresti [5].


1. Tang ST, Chang WC, Chen JS, Su PJ, Hsieh CH, Chou WC. Trajectory and predictors of quality of life during the dying process: Roles of perceived sense of burden to others and posttraumatic growth. *Support Care Cancer* 2014 May 28. http://link.springer.com/article/10.1007%2Fs00520-014-2288-y
2. van Rijn VM, Mooij SH, Mollers M, Snijders PJ, Speksnijder AG, King AJ, de Vries HJ et al. Anal, penile, and oral high-risk HPV infections and HPV seropositivity in HIV-positive and HIV-negative men who have sex with men. *PLoS One* 2014 Mar 20;9(3):e92208. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3961332/
3. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986 (April);73:13–22. http://www.biostat.jhsph.edu/~fdominic/teaching/bio655/references/extra/liang.bka.1986.pdf
4. Hanley JA, Abdissa Negassa A, deB. Edwardes MD, Forrester JE. Statistical Analysis of correlated data using generalized estimating equations: An orientation. *Am J Epidemiol* 2003;157:364–75. http://aje.oxfordjournals.org/content/157/4/364.full.pdf
5. Agresti A. *Categorical Data Analysis*, Third Edition. Wiley, 2013.


## generalized linear models, *see also* general linear models

The generalized linear model (GLM) is an extension of the general linear model to the setup where the response variable may have a **distribution** far from **Gaussian**. The response can be **continuous** (with Gaussian or non-Gaussian distribution) or **discrete** (proportion or count or rate). Other conditions remain the same as in general linear models. In case you are not familiar with general linear models, we recommend that you familiarize yourself with the general linear models before trying to grasp the essentials of the GLMs. A brief description of these models is in under the topic **general linear models**.

As in the case of general linear models, there is no restriction on the **independent** or regressor variables in the GLMs. These variables could be continuous or discrete, and may pertain to **fixed or random effects** (or may be mixed). When they are mixed, the model would be called a *generalized linear mixed model*. But these regressors must affect the response through linear coefficients, although the variables themselves could be square or log or any such function. If the effect is not linear, the GLM will study only its linear part. Equation-wise, the GLM is

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \varepsilon_i,$$

where $y_i$ is the response of the $i$th person or unit; the regression coefficients β's are linear (i.e., no $\beta^2$, $e^\beta$, etc.) but $x_2$ could be $x_1^2$, etc.; and $\varepsilon_i$ may *not* follow a Gaussian distribution in GLM. For the general linear model, these errors are required to follow a Gaussian pattern, or nearly so if $n$ is large. In the case of GLM, these errors may follow any of the wide spectrum of distributions of what is known as the *exponential family*. This family includes **binomial**, **Poisson**, **exponential**, **Weibull**, **beta**, **gamma**, etc. Notice that these distributions can be highly **skewed** and can belong to **discrete variables**.

You may be aware that the general linear model is a combination of **regression**, **analysis of variance**, and **analysis of covariance**, but the response variable must be continuous in all these setups. **Gaussian distribution** is required in general linear models for building up confidence intervals and for tests of hypotheses on the model parameters, although point estimates can be obtained with the **least squares method** even when the distribution is not Gaussian. These restrictions are dispensed with in GLM by using what is called a link function, which converts the response to a form that has a relatively easily analyzable distribution. For example, proportions that follow a binomial distribution are converted to **logits**—this transforms the probability between 0 and 1 to values that can be positive, negative, or 0. Similarly, counts (e.g., number of patients coming to a clinic) that follow a Poisson distribution are transformed to logarithms to yield to nearly a Gaussian pattern. Logit and log are the respective link functions in these situations. There are other link functions for other setups, as presented under the topic **link functions**. No transformation is required if the response variable is already Gaussian. This is called *identity link*. However, in the GLM, just as in general linear models, various values of the response variable must not be correlated; that is, they must belong to separate persons or units that do not affect each other and do not, for example, belong to the same family, whose members are likely to provide similar values at least to some extent. When the values are correlated, use **generalized estimating equations (GEEs)**.

Estimates of the regression coefficients β's are obtained such that the likelihood of the sample coming from the distribution postulated by the link is maximum. These are popularly called **maximum likelihood estimates (MLEs)**. For Gaussian distributions, these MLEs are well known and can be easily derived, but many other distributions admissible under GLM require an iterative weighted least squares procedure. *Iteration* in effect means that a start is made with some plausible estimates such as the mean, the model is checked whether it fits well with the observed data, and the estimates are revised according to the discrepancies found. This can go on for several iterations till such time that the updated estimates by two successive iterations are nearly the same. (This is called *convergence*—there may be situations where the estimates do not converge, in which case we say that we are not able to obtain the plausible estimates.) Statistical packages are well trained to do these iterations for you, and you would not get wrong estimates if a standard package is used. These packages will give you the estimates of the β's and their standard errors (SEs) and will also test the statistical significance of each regression coefficient. You can then decide which of the regressors are worth retaining and which ones can be discarded.

As in the case of general linear models, **standardization** by subtracting the mean and dividing by the standard deviation (SD) is recommended for explanatory continuous variables so that each gets similar importance. If standardization is not done, the variable with

large values, such as cholesterol level compared with hemoglobin level, sways the estimates and the statistical tests. In statistical terms, the cholesterol level will get disproportionately large weight in calculations relative to the hemoglobin level in the absence of standardization.

The goodness of fit of the model and the statistical significance of the contribution of each variable or a set of explanatory variables is tested by **deviance**.

The GLM method is originally from Nelder and Wedderburn [1]. Further details of the method are available in Dobson and Barnett [2].

1. Nelder JA, Wedderburn RWM. Generalized linear models. *J Royal Statistical Soc, Ser A* 1972;135;370–84. http://biecek.pl/MIMUW/uploads/Nelder_GLM.pdf
2. Dobson AJ, Barnett A. *An Introduction to Generalized Linear Models*, Third Edition. Chapman & Hall/CRC Press, 2008.



## general linear models, *see also* generalized linear models

General linear model is the name given to the method that unifies ordinary **regression**, **analysis of variance**, and **analysis of covariance**. In all these setups, the response variable, also called the **dependent**, should be continuous and is expected to follow a nearly **Gaussian distribution**. The values must be independent and not correlated. This condition in effect means that the observations must belong to separate units or different persons and not belong to one family, one person at different points in time, or different body sites, or any other affinity group whose members have a tendency to be similar to one another.

To fix ideas, denote the response of the *i*th person by $y_i$ and his/her regressors by $x_{1i}$, $x_{2i}$, …, $x_{Ki}$. If you are trying to explain a particular kidney function by the person's weight, age, sex, water intake per day, fiber content in the diet, etc., this notation is like saying that third person ($i = 3$) in our study has kidney function $y_3 = 54$ mg/dL and her weight $x_{13} = 62$ kg, age $x_{23} = 45$ years, sex $x_{33} = 0$ (where 0 is the notation for females and 1 for males), water intake per day $x_{43} = 3.4$ L, and fiber content $x_{53} = 340$ g. Note that $i = 3$ in all these $x$'s. Under these notations, a general linear model is given by

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_K x_{Ki} + \varepsilon_i,$$

where $y_i$ is the response or dependent and $\varepsilon_i$ is the error that has Gaussian distribution with a mean of 0 and any standard deviation (SD) $\sigma$. This is written as $\varepsilon_i \sim N(0, \sigma)$, where $N$ stands for normal distribution, which we like to call Gaussian, since the term *normal* has a different meaning in health and medicine. Note that the SD is the same for each $i$—the condition popularly known as homogeneity of variances or **homoscedasticity**.

The $x$'s are the regressors or the **independent** variables, and the $\beta$is are the **regression coefficients**. These coefficients are the **parameters** of the model. The expression $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_K x_{Ki}$ is the mean of all the persons in the target population whose regressor values are $x_{1i}$, $x_{2i}$, …, $x_{Ki}$. This can be denoted by $\mu_i$. This and the previous explanation regarding $\varepsilon_i$ imply that $y_i \sim N(\mu_i, \sigma)$. In our example, $\mu_i$ is the population mean of those persons whose weight is 62 kg, age is 54 years, etc. If there are 16 persons in the population with exactly same values of all $x$'s, they will most likely have different kidney function, and $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_K x_{Ki}$ is the notation of their mean. Thus, general linear models are for means and not for individual values—a fact that many forget while interpreting a model. You can see that the mean response will change if any $x$ value changes, for example, if age is different or weight is different. Now it should be clear that $\varepsilon_i$ is the deviation of the response of the *i*th person in the population from its mean of similar people; thus, $\varepsilon_i$ has zero mean. This is what is being called error. If the model is a good fit (which really means that the estimates of $\beta$'s are such that the residuals are minimal, the SD will be small. A large SD will indicate that the model is not a good fit. This SD is estimated by the square root of **mean square error (MSE)**, which is the sum of the squares of the residuals divided by its **degrees of freedom**. Residuals are explained later on in this section.

The model is linear so long as the coefficients are linear. That is, there is no $\beta^2$, $e^\beta$, or log$\beta$ type of coefficient. This means that when any $x_k$ increases by one, the response $y$ is supposed to increase by its coefficient $\beta_k$. The regression coefficient $\beta_k$ is generally interpreted as the *net contribution* or net effect of the variable $x_k$ ($k = 1, 2, \ldots, K$), but the term *net* is too strong. For this to be really net contribution, the model must include all possible variables that can affect the response. This is a tall order first because it is not generally feasible to include *all* the variables and second because only those variables that are known or suspected can be included. Many are unknown, and this **epistemic uncertainty** is many times forgotten. An example given later in this section will clarify one aspect of the interpretation of the regression coefficient. Also see the topic **linear regression**. In case the relationship is not linear, a general linear model will limit itself to whatever is the linear part and ignore the rest. In this case, the model will not be a good fit to the data.

However, there is no restriction on the $x$'s. The type of $x$'s is the feature that distinguishes among ordinary regression, analysis of variance, and analysis of covariance. In ordinary regression, all $x$'s have to be quantitative; in analysis of variance, all $x$'s have to be discrete—mostly defining the groups through **indicator variables**; and in analysis of covariance, these are mixed. It is easy to unify all these into one theoretical framework because basically all $x$'s are considered fixed in this setup and the value of $y$ is estimated for *given* values of the $x$'s. The requirement of Gaussian distribution, independence, and homoscedasticity is handy in pursuing what is called the **maximum likelihood estimates (MLEs)**. This method finds those estimates of the parameters that make the observed values most likely. These estimates are denoted by $b$'s. Any standard statistical software will easily obtain these estimates for you when the model is properly specified. Proper specification means that you correctly tell the computer program which variable is to be treated as continuous, which as **categorical**, etc. When these estimates are used, the model can be written as

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_K x_{Ki} + e_i,$$

and the estimated value of $y$ for the *i*th person is $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_K x_{Ki}$. Note that the error term is now denoted by $e_i$ and called **residual** in the context of the sample. This is the difference between the *mean* response in the sample for the persons with values $x_{1i}$, $x_{2i}$, …, $x_{Ki}$ of the regressor variables and the actual observed response for the *i*th person. If there is only one person with these values of the $x$'s, $e_i$ is the difference between the observed $y_i$ and the estimated $\hat{y}_i$.

Under a Gaussian distribution of the response variable, the estimates $b$'s of the regression coefficients $\beta$'s also have Gaussian distribution. This allows us to easily find the **confidence intervals (CIs)** for estimates of $\beta$is and to **test hypotheses** on them by using the property, for example, that the estimate $\pm\, t_{\nu 0.025}$*(estimated SE) is the 95% CI and (estimate − its mean)/(its estimated SE) has **Student *t*** distribution with $\nu$ degrees of freedom (df's). This SE is estimated by the MSE as just explained.

A simple example of a general linear model is the work of Ainslie et al. [1]. They examined how blood flow velocity in the middle cerebral artery (MCAv) in healthy humans is affected by physical activity, body mass index (BMI), blood pressure

(BP), and age. Physical activity was assessed as active or inactive. Thus, this is categorical. The response variable is MCAv, which is quantitative, and it must have nearly the same variance for different ages, different physical activity, etc. for general linear models to be applicable. The authors found that BMI and BP did not have a statistically significant contribution, while age and physical activity were important for MCAv. They have reported a separate model for active and inactive persons, but these combine into the following:

$$\text{MCAv (in cm/s)} = 87.8 - 0.73*\text{age (in years)} + 9.2*\text{activity} - 0.03*\text{activity}*\text{age},$$

where activity = 1 for physically active persons and activity = 0 for inactive persons. When these values of activity are substituted, the models become

$$\text{MCAv (in cm/s)} = 87.8 - 0.73*\text{age (in years) for inactive persons}$$

$$\text{MCAv (in cm/s)} = 87.8 - 0.73*\text{age (in years)} + 9.2\times 1 - 0.03\times 1\times \text{age} = 97.0 - 0.76*\text{age (in years) for active persons.}$$

This model means that MCAv was reduced on average by 0.73 cm/s for each year increase in age in inactive persons but by 0.76 cm/s in active persons, although the baseline for active persons was high (97.0 versus 87.8). Gaussian distribution is not a requirement for getting these equations, because the estimates of the regression coefficients can be obtained by the **least squares method**. But Gaussian distribution is needed to work out the CI. The authors also reported the CI for these regression coefficients. Those who are aware will realize that this model is the same as the analysis of covariance where age is the covariate. Analysis of covariance is the most generalized of the general linear models since it has both continuous and discrete regressors.

The statistical significance of a general linear model is assessed by an ***F*-test**, which is obtained as the mean **sum of squares** due to the model and the mean sum of squares due to error (MSE). The statistical significance of each regression coefficient can also be tested. Appropriate statistical software will do it for you, but the model must be properly specified. Many modifications of the model can be done to test other kinds of **null hypotheses**. For a complete description of the general linear models, their strengths, and their weaknesses, see Vik [2].

When the distribution of the response variable is far from Gaussian, we need to fall back on the **generalized linear models (GLMs)**, and if the responses are correlated, we get help from the **generalized estimating equations (GEEs)**.


1. Ainslie PN, Cotter JD, George KP, Lucas S, Murrell C, Shave R, Thomas KN, Williams MJ, Atkinson G. Elevation in cerebral blood flow velocity with aerobic fitness throughout healthy human ageing. *J Physiol* 2008 Aug 15;586(16):4005–10. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2538930/
2. Vik PW. *Regression, ANOVA, and the General Linear Model: A Statistics Primer.* Sage, 2013.


## generalized Wilcoxon test, *see* Breslow test

## gender inequality index

The gender inequality index (GII) measures the inequality between male and female achievements in three dimensions, namely, health, empowerment, and labor participation. This index was originally proposed by Seth [1] and is now promoted by the United Nations Development Program (UNDP) as a part of the **human development index (HDI)** since gender inequality has been found to be a potential source of adverse human development. This index varies from 0 for areas where men and women fare equally well to 1 where either gender fares as poorly as possible relative to the other [2].

Gender inequality has lately assumed importance around the world, and efforts are made to quantify this inequality so that the trend can be studied and comparison across communities can be made. This certainly has great social implications, but it may affect physical health as well through, for example, smoking, suicides, domestic violence, and life expectancy. It is also strongly related to birth weight, infant mortality, fertility, etc. Varkey et al. [3] have discussed health issues related to gender empowerment in detail, which is a part of GII. In fact, gender empowerment was the earlier index used by UNDP but has now been replaced by GII. **Gender empowerment** was assessed by the proportion of seats held by women in national parliaments, percentage of women in economic decision-making positions, and female share of income. Beside gender empowerment in a modified form, GII includes female disadvantages due to maternal mortality, adolescent fertility, and lack of education.

For calculation of GII, the HDI uses the following indices: (i) **maternal mortality rate (MMR)**; (ii) adolescent fertility rate (AFR), which is the number of births per 1000 girls of age 15–19 per year; (iii) share of parliamentary seats held by each sex (PR) measured in fraction; (iv) attainment at secondary and higher education (SE) levels measured in fraction for females and males separately; and (v) labor market participation rate (LFPR), which is the fraction of working-age women and men in employment. The index uses the **geometric mean** (GM) instead of the usual arithmetic mean of these indices because of their highly skewed distribution. GM is the *n*th root of multiplication of *n* values. If the value of any of these indices is 0, which is not impossible, it is assumed to be 0.1 so that the GM does not become 0.

The first step in the calculation of GII is to obtain

$$G_F = \sqrt[3]{\left(\frac{10}{MMR}\times\frac{1}{AFR}\right)^{1/2}\times(PR_F\times SE_F)^{1/2}\times LFPR_F} \text{ for females.}$$

This basically is the GM of three GMs. The first part is the GM of the two adverse health components, namely, the MMR (whose minimum is assumed to be 10) and the AFR; the second is the GM of the two empowerment components, namely, the parliament seats and the educational attainment; and the third component is labor participation as a stand-alone factor. In the case of males, the adverse health component is 1 since there is no risk of maternal mortality or adolescent fertility. Thus,

$$G_M = \sqrt[3]{1\times(PR_M\times SE_M)^{1/2}\times LFPR_M} \text{ for males.}$$

In the second step, the preceding two indices are aggregated by the **harmonic mean**, which captures the inequality between men and women and adjusts for association between the dimensions. The harmonic mean is the reciprocal of the mean of reciprocals:

$$\text{HARM}(G_F, G_M) = 1/\left(\frac{1/G_F + 1/G_M}{2}\right).$$

This creates an index with equal weight to the two genders, called the equally distributed gender index. This will be used after the third step.

The third step is to aggregate the original indices for males and females with equal weight given to each gender. This is done again by the GM of the "averages" of the three dimensions as follows:

$$G_{F,M} = \sqrt[3]{\overline{Health} \times \overline{Empowerment} \times \overline{LFPR}},$$

where $\overline{Health} = \dfrac{\left(\sqrt{\dfrac{10}{MMR} \times \dfrac{1}{AFR}} + 1\right)}{2}$,

$$\overline{Empowerment} = \frac{\left(\sqrt{PR_F \times SE_F} + \sqrt{PR_M \times SE_M}\right)}{2}, \text{ and}$$

$$\overline{LFPR} = \frac{(LFPR_F + LFPR_M)}{2}.$$

Finally,

$$\text{GII} = 1 - \frac{HARM(G_F, G_M)}{G_{F,M}}.$$

This is the difference between the standard, which is 1, and the equally distributed gender index adjusted for $G_{F,M}$.

For an illustration of the calculations, see the technical notes of the *Human Development Report*, 2013 [2].


1. Seth S.Inequality, interactions and human development. *J Human Development Capabilities* 2009;10:375–96. http://www.tandfonline.com/doi/full/10.1080/.U5X0YRZPy18
2. UNDP. *Human Development Report 2013: Technical Notes*. http://hdr.undp.org/sites/default/files/hdr_2013_en_technotes.pdf
3. Varkey P, Mbbs, Kureshi S, Lesnick T. Empowerment of women and its association with the health of the community. *J Womens Health (Larchmt)* 2010 Jan;19(1):71–6. http://online.liebertpub.com/doi/abs/10.1089/jwh.2009.1444


## general fertility rate, *see* fertility indicators

## geometric mean, *see* mean (arithmetic, geometric, and harmonic)

## geriatric health (epidemiological indicators of)

Geriatric health is gaining increasing importance as the older population is rapidly rising in most countries. Some epidemiological measures of geriatric health are given here—they can be used for individual assessment. Several others are listed in Ref. [1]. Many more are coming up, and you should keep an eye on the new developments.

### *Activities of Daily Living*

In the case of old age or handicaps, the degree of disability can be measured in terms of an activities of daily living (ADL) index. Scores are assigned to the level of independence assessed on several activities of daily living such as walking, bathing, use of toilet, and dressing. The score could range from 0 for complete dependence to, say, 4 for complete independence on each item. The sum of these scores is called the ADL index (see, e.g., Katz and Akpom [2]). A disadvantage of such an index is that it is insensitive to change when the level improves on some items and deteriorates on the others. Another index is derived from 14 questions ranging from difficulty in self-care (e.g., eating or dressing) to higher-level activities (e.g., carrying weights or doing housework), used by the WHO Eleven Countries Study [3], which can be adapted to suit local conditions. No index, including those just mentioned, is widely acceptable.

### *Mental Health of the Elderly*

Physical limitations emerging from degeneration in old age are recognized, but mental agility also deserves attention. Among many instruments available for measuring the mental health of the elderly, one in common use is the mental health component of quality-of-life questionnaires, such as the short form with 36 items (SF-36) [4]. This is no different from the tool used for the general population of adults. SF-36, however, is restricted to functional status, including for mental health.

Insomnia and anxiety are common in old age. For these, the Pittsburgh Sleep Quality Index [5] and *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (DSM-V) [6], respectively, can be used. Both are for the general population and not specific to the elderly. The Beck Depression Inventory is used for assessing the severity of depression symptoms. This is an old instrument but continues to be commonly used, particularly for people of old age.


1. Making geriatric assessment work: Selecting useful measures. *Physical Therapy* June 2001;81(6):1233–52. http://www.physther.net/content/81/6/1233/T1.expansion.html
2. Katz S, Akpom CA. Index of ADL. *Med Care* 1976;14 (Suppl 55):116–8. http://www.ncbi.nlm.nih.gov/pubmed/132585
3. Heikkinen E, Waters WE, Brezinski ZJ (Eds.). *The Elderly In Eleven Countries: A Sociomedical Survey*. Copenhagen: World Health Organization Regional Office for Europe, 1983.
4. SF-36® *Health Survey Scoring Demonstration*. http://www.sf-36.org/demos/SF-36.html
5. Buysse DJ, Reynolds CF III, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: A new instrument for psychiatric practice and research. *Psychiatry Res* 1989;28:193–213. http://api.ning.com/files/BRJZJzqTpgw20OdSsvFFi23FbZ38XOoBHvsBuzu92rQJFINhaAwRB34kvYIyu1tzsNXXuVWM2Umnh6paQxTvfmyiIvE5QN1*/PSQI.pdf
6. American Psychiatric Association, DSM-V. *Diagnostic and Statistical Manual of Mental Disorders*, Fifth Edition. American Psychiatric Association, 2013. http://dsm.psychiatryonline.org/book.aspx?bookid=556

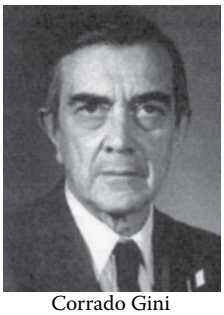

## Gini coefficient, *see also* health inequality, Palma measure of inequality

The Gini coefficient is just about the most popular measure of inequality among various values of a variable. This is used to measure health inequality in a community. For $n$ values, this is given by

$$\text{Gini coefficient: } G = \frac{\Sigma_i \Sigma_j \left| x_i - x_j \right|}{2n^2 \bar{x}}; \; i, j = 1, 2, \ldots, n.$$

The Gini coefficient is applicable to positive values only. If negative values are present and the mean is close to 0, the Gini coefficient may blow up to an unacceptable level. Note also that the Gini coefficient is similar to the **coefficient of variation**. Very different

statistical **distributions** of the variable can give the same value of the Gini coefficient. The value of this coefficient lies between 0 and 1. A value less than 0.2 can be considered tolerable, between 0.2 and 0.4 middling type, and more than 0.4 high inequality. Gini coefficient equal to 1 for an area would mean that just one person in that area is in complete health and all the others have zero health on whatever scale you use, and the Gini coefficient equal to 0 means that all people have exactly similar health—they all could be in poor health or all could be in good health. This coefficient was originally devised by Corrado Gini in 1912 for measuring income disparity [1].


Corrado Gini

Health inequality among different segments of the population has received attention as it doesn't just concern moral issues regarding fairness but also affects the rate of improvement in population health. Inequalities definitely work to the disadvantage of the deprived sections but affect all, as disparities hamper the generation of human capital and inhibit sustainable improvement. While inequalities are inherent in the social fabric, health inequality caused by factors amenable to human intervention is considered unjust. The Gini coefficient remains a favorite index for measuring this inequality, although as described later in this section, there are deficiencies in this coefficient as an index of inequality.

In the case of health, the level in the top and bottom income **quintiles** is a customary measure of health inequality. This is the health of people in the top 20% income bracket and the bottom 20% income bracket. This assumes that the people with the top income will be in the best health and those with the bottom income in the poorest health. The ratio of maximum to minimum can be used, but this ratio disregards the dispersion. The following example illustrates the calculation of the Gini coefficient in the context of **infant mortality rate** (IMR).

In Table G.4 are IMRs in the 20% of the population with the lowest income (lowest quintile), the next 20%, etc., and the highest 20% (top quintile) in a developing country. For these data,

**TABLE G.4**
**Infant Mortality Rates in Income Quintiles in a Developing County**

| Quintile | Infant Mortality Rate |
|---|---|
| Lowest quintile | 73 |
| Second quintile | 68 |
| Third quintile | 51 |
| Fourth quintile | 32 |
| Top quintile | 16 |
| Average ($\bar{x}$) | 48 |

$$\text{Gini coefficient} = \left[|73-68|+|73-51|+\ldots+|73-16|+|68-73| + \ldots+|16-32|\right]/(2\times 5^2\times 48) = 0.25.$$

The low value of 0.25 in this example despite a ratio of more than 1:4 between IMR in the minimum and maximum quintiles shows that the Gini coefficient is not too sensitive to the inequalities. Now you know why an apparently low value such as 0.4 is considered high. You may notice that for other statistical measures between 0 and 1, such as correlation, 0.4 is not considered high.

The problem with the Gini coefficient as a measure of inequality lies in it being nonspecific. We cannot say where the inequality exists. It is oversensitive to the changes in the middle values and insensitive to the changes at the top and bottom [2]. Also, it has no physical interpretation. Another measure, called the **Palma measure of inequality**, has been suggested to remedy these problems. Details are provided under that topic, but briefly, it compares the highest 10% of values with the lowest 40%—these percentages seem to be popping up everywhere in the sense that the remaining **deciles** 5–9 capture just about half the population in almost every setting.

1. Gini, C. Variabilità e Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche, C. Cuppini 1912. Reprinted in *Memorie di metodologia statistica* (Ed. Pizetti E, Salvemini, T).
2. Cobham A, Summer A. Is inequality all about the tails? The Palma measure of inequality. *Significance* Feb 2014;10–13. http://onlinelibrary.wiley.com/store/10.1111/j.1740-9713.2014.00718.x/asset/sign718.pdf?v=1&t=hw9y1qu5&s=c4cfec0d52d21374ffd2547aa319796969a8004b

## GLM, *see* generalized linear models

## global burden of disease (GBD), *see* burden of disease

## gold standard

Gold standard is an oft-repeated term in biostatistics parlance but remains quite nonspecific. Ideally, this refers to something that is infallible, that you can rely on completely, that has 100% **sensitivity** and 100% **specificity**. A gold standard is required for comparing the performance of newly developed regimens or tests. Concepts such as sensitivity, specificity, and predictivities exist only when a gold standard exists. However, due to the omnipresence of uncertainties in health and medicine, there is no gold standard—no treatment is 100% effective; no test is perfect. Thus, *whatever is best available is considered the gold standard*. This, of course, varies from place to place, from time to time, from context to context.

Kratiras et al. [1] have discussed the need to redefine the gold standard for the treatment of advanced prostate cancer in the context of continuous and intermittent androgen deprivation therapy. Rapaso-Amaral et al. [2] has questioned considering bone transplant as the gold standard for the repair of alveolar bone defects. These are just a couple of illustrative examples to emphasize that the gold standard continuously evolves as new knowledge become available. In assessing the new against the gold standard, we tend to forget that the gold standard itself can be in error and, thus, our assessment can

be imperfect. Indices such as positive and negative **predictivities** need to be reevaluated as the gold standard is redefined.

There are other uses of the term *gold standard* in medicine. For example, the forehead flap is the gold standard for nose reconstruction [3], and the randomized placebo-controlled clinical trial [4] is considered a gold standard methodology for assessing the performance of a regimen. The former is the gold standard so long as skin at another site is not found more suitable, and for the latter, see the topic **bias pyramid**, where we mention that randomized controlled trials also do not completely rule out bias.


1. Kratiras Z, Konstantinidis C, Skriapas K. A review of continuous vs intermittent androgen deprivation therapy: Redefining the gold standard in the treatment of advanced prostate cancer. Myths, facts and new data on a "perpetual dispute". *Int Braz J Urol* 2014 Jan–Feb;40(1):3–15; discussion 15. http://brazjurol.com.br/january_february_2014/Kratiras_003_015.htm
2. Raposo-Amaral CE, Bueno DF, Almeida AB, Jorgetti V, Costa CC, Gouveia CH, Vulcano LC, Fanganiello RD, Passos-Bueno MR, AlonsoN. Is bone transplantation the gold standard for repair of alveolar bone defects? *J Tissue Eng* 2014 Jan 16;5:2041731413519352. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3924878/
3. Correa BJ, Weathers WM, Wolfswinkel EM, Thornton JF. The forehead flap: The gold standard of nasal soft tissue reconstruction. *Semin Plast Surg* 2013 May;27(2):96–103. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3743909/
4. De Serres G, Skowronski DM, Wu XW, Ambrose CS. The test-negative design: Validity, accuracy and precision of vaccine efficacy estimates compared to the gold standard of randomised placebo-controlled clinical trials. *Euro Surveill* 2013 Sep 12;18(37). pii:20585. http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=20585


## Goodman–Kruskal gamma, *see* association between ordinal categories (degree of)

## goodness of fit

The term *goodness of fit* is used for a model agreeing with the observed values, or in some cases, to describe whether the observed values follow a specified pattern. You might be aware that the observed values of a variable have a random component yet tend to follow a pattern. Various statistical methods are used to identify this pattern, called a **model**, and then a test of goodness of fit is used to check if the model has really been able to capture the essential features of the data. Mostly, but not always, the **chi-square test** is used for this purpose. There are situations where other tests are used, and these also are listed later in this section.

The most common use of the goodness-of-fit test is in assessing the pattern of categorical data. For example, these categories could be none, mild, moderate, and serious forms of anemia; calendar month of occurrence of sudden infant death syndrome (SIDS) in a year; or cirrhosis, hepatitis, and malignant forms of liver disease. The first two are on an **ordinal scale**, and the last is on a **nominal scale**. The first has 4 categories, the second 12, and the last 3 categories. The interest may be in (i) whether sudden death syndrome occurs twice as often in winter as in summer months; (ii) whether the patterns of none, mild, moderate, and serious hypertension in cases of myocardial infarction (MI) are 10%, 20%, 40%, and 30%, respectively; and (iii) whether the proportions of full, partial, and no recovery within 2 years after surgery in cases of breast cancer are 40%, 50%, and 10%, respectively. Such a problem is known as the problem of goodness of fit because the interest is in finding whether or not the pattern observed in the sample fits the specified pattern well. The procedure is best explained with the help of an example.

**TABLE G.5**
**Blood Group Pattern in a Hypothetical Sample of AIDS Cases**

| Blood Group | O | A | B | AB | Total |
|---|---|---|---|---|---|
| Number of AIDS patients | 57 | 36 | 51 | 6 | **150** |

The blood group of a random sample of 150 patients with acquired immunodeficiency syndrome (AIDS) is investigated to examine the possibility of a preponderance of a particular blood group in AIDS cases. If there were no preponderance, the profile would be the same as in the general population. Suppose this is 6:5:8:1 for blood groups O, A, B, and AB, respectively. The sample observations are shown in Table G.5. If the cases are in the same ratio as in the general population, the number of cases with blood group O should have been $6 \times 150/20 = 45$. The observed number is 57. In view of this difference, does this pattern in AIDS cases really conform to that in the general population? How does one draw a conclusion such that the chance of **Type I error** does not exceed 0.05? A solution is provided under the term **chi-square–overall**. This is valid under certain conditions, particularly, that the sample size is large and almost no cell frequency is less than 5. For small frequencies, a **multinomial test** can be used.

### *Chi-Square Test of Goodness of Fit of a Gaussian Distribution*

A common problem in goodness of fit is to find whether the pattern of quantitative data is different from what is expected under the Gaussian distribution. If the pattern is not different from **Gaussian**, the usual methods of **Student *t***, **analysis of variance (ANOVA)**, etc. can be safely used. If the pattern is very different, you may have to use **transformation** or use **nonparametric methods**. The procedure to test a Gaussian pattern is as follows. This can be used when (i) the data are in grouped form with at least 5 **class intervals** and possibly not more than 12 intervals (fewer intervals make it extremely difficult to assess a Gaussian pattern, and more than 12 intervals tend to negate the advantage of convenient display of grouped data), and (ii) the sample size is large so that number of subjects in any interval is not less than 5 (chi-square is not valid when any cell frequency is less than 5). Otherwise also, the statistical **power** to detect a difference from a Gaussian pattern is low when the sample size is small.

The procedure basically is to calculate the mean and standard deviation of your values. Use these to calculate the **Gaussian probability** in each class interval. Convert this probability to frequency by multiplying by the sample size $n$. These will be the **expected frequencies** in the intervals. Compare the observed frequencies with expected frequencies and calculate chi-square. Use the calculated value of chi-square to find the ***P*-value**. Reject the null of Gaussian pattern when the value of $P$ is less than the prespecified **level of significance** and conclude that the pattern is different from Gaussian. All this is easily done by statistical software.

### *Other Kinds of Goodness of Fit*

There are several other methods to check whether your model is a good fit with the data or not. These depend on the kind of problem

you are investigating. For example, goodness of fit of a linear regression model is assessed by square of the **multiple correlation coefficient** ($R^2$) and tested by an ***F*-test**. This is described under the topic **linear regression**. The goodness of a **logistic model** can be assessed by the classification accuracy and tested by the **Hosmer–Lemeshow test**. The **receiver operating characteristic (ROC) curve** method can also be used for logistic model. For some models, such as **Cox regression**, chi-square-based **deviance** gives a pretty good idea of the goodness of fit. See **log-linear models** for assessing their goodness of fit based on the standardized deviate $z = (O - E)/\sqrt{E}$, where $O$ and $E$ are the observed and expected frequencies, respectively.

## graphs and diagrams, *see also* charts (statistical), maps (statistical)

Tools such as graphs, diagrams, charts, and maps are commonly used for visual display of data. They are generally referred to as *figure* in the literature, although this term also includes non-data-based pictures, such as of a bone or of a lesion. Our description here is restricted to statistical graphs and diagrams.

**Visual display** is considered a powerful medium for communication and for understanding the basic features of a set of data. Salient features are often easily brought forth by an appropriately drawn figure. Also, the impression received from such a figure seems to be more vivid and lasts longer than the impression from numeric data. The main function of a figure is to provide perception and cognition of the basic features of the data, and sometimes, it provides a much deeper understanding of the data. Busy professionals cannot devote time to read the numerical values in a table and make a picture in their mind. Also, numerical tables can be assiduous and can cause fatigue. Some people need a ready-made picture that gives the message regarding the trends and differences. However, for this, an appropriate diagram must be chosen and correctly drawn. If the display requires long time to understand, consider it not sufficiently effective. A large number of methods are available, and the choice is not always easy. All these methods are listed later in this section and described under their respective topics.

Generally speaking, a graph is a figure drawn to a scale. The scale could be on a horizontal axis or on a vertical axis or both. Graphs are used to display the relationship of variables or to display the pattern of their **distribution**. Perhaps the first recorded graph was made by Playfair, published in 1786, which depicted balance of trade between North America and the England during the years 1700–1780 [1].

The term *graph* is disappearing from the medical literature, and the term *diagram* is preferred instead. The **bar diagram**, **line diagram**, and **area diagram** are essentially graphs but are colloquially called diagrams. So are the **frequency curve**, **polygon**, **histogram**, **scatter diagram**, and **radar graph**. Among others are the **pie diagram** and **donut diagram**. Factually, a diagram is a figure depicting data not necessarily to a scale. The schematic depiction of three methods of **observational studies** under that topic is an example of this kind of diagram. There are many others in this book. Graphs and diagrams together are sometimes called **charts**, but we like to use the term chart for text-based figures, as explained under this term. **Bubble charts** are an ingenious way of depicting three- or four-dimensional data.

Some diagrams are more complex than others. Among the complex ones are the **box-and-whiskers plot**, **biplot**, **bihistogram**, **nomogram**, and **Lexis diagram**. This list is restricted to those that are discussed in this volume. There are many others that are either rarely used or far too specialized, and thus not included in this book.

Electrocardiograms (ECGs), electroencephalograms (EEGs), and chymographs are examples of data-based figures commonly used in the practice of medicine. They are examples of far-too-specialized topics to be discussed in this book on biostatistics. But there are other not-so-specialized diagrams such as **growth charts**, **dendrograms**, and **partograms** that are typically used in health and medicine. They are discussed under these respective topics.

A good source for learning about various graphs and diagrams is the celebrated book by Tufte [2]. The topic has evolved enormously lately. Integrated explanations through pictures have given rise to **infographics**. These provide much more information than graphs. Now, we also have interactive graphics where a graph or diagram pops up when selected among the given choices.

### *Choice and Cautions in Visual Display of Data*

It may be helpful to get a clear picture of the situations in which one diagram is more appropriate than others. This is shown in Table G.6 for graphs and diagrams commonly used. As already stated, all these graphs and diagrams are separately explained in this volume under the respective topics.

When the range of values on the $y$-axis is extremely large, examine whether or not a **logarithmic scale** (usually called log scale)

**TABLE G.6**
**The Appropriate Type of Diagram for Various Situations**

| | Scale on Horizontal Axis | | | |
|---|---|---|---|---|
| **Vertical Axis (Must Represent a Quantity)** | **Metric and Continuous—No Categories** | **Metric Categories or Discrete** | **Ordinal** | **Nominal** |
| Frequency or percentage | ×[a] | Histogram[b] | Bar | Bar |
| | | Pie[c] in these three cases when contribution of each category relative to the others is to be shown | | |
| Mean, rate, or ratio | ×[a] | Bar | Bar | Bar |
| | | Line in these two cases for trend | | |
| Values of a variable | Scatter | Scatter | Scatter | Scatter |

[a] Frequency, percentage, mean, rate, or ratio is not possible on $y$-axis when $x$ variable is continuous and has no categories.
[b] Or polygon or frequency curve.
[c] Pie diagram has no horizontal or vertical axis.

would be more appropriate. This scale converts 10 into 1, 100 into 2, 1000 into 3, 10,000 into 4, etc. One such representation appears under the topic **scatter diagram (Figure S.2a)**, where the total bilirubin on the *x*-axis is plotted on a log scale. If the ordinary scale, called linear, were used, the points for low levels of bilirubin would become indistinguishably close. Another log scale is on the *y*-axis of a figure under the topic **line diagram (Figure L.6d)**. One advantage of a log scale is that it can represent change relative to the previous value. Cautious interpretation is needed in this case.

Note the following points regarding various graphs and diagrams:

1. A diagram should not be too complex. A large number of relationships can be shown in one diagram, but it is advisable to restrict one diagram to not more than two relationships or not more than three variables. Although a diagram is considered good if it instantly conveys the substance, in science, diagrams may require some effort on the part of the reader to be able to interpret them properly.
2. In terms of choice of scale, a diagram can be made to show a steeper or flatter relationship by choosing the scale accordingly. Figure G.7a and b shows the same relationship between age and average forced vital capacity (FVC), but FVC is plotted on different scales in the two panels, giving a different impression.
3. The scale of calibration should be clearly indicated. Choose a scale that is suitable and does not exaggerate or understate the values. For comparing two or more groups, use the same scale for both groups.
4. In case secondary data are used to draw a figure, provide the source.
5. Figures displaying all the data points, such as a scatter diagram, are preferable because they allow readers to draw their own conclusions. However, many times, it is not feasible to show the entire set of data. Only summary statistics such as percentages, rates, or averages are shown.
6. All graphs and diagrams must be self-explanatory, containing informative (what, where, when) yet concise titles, legends to identify various components of a diagram, labels for axes, unit of measurement, etc.
7. Use diagrams mostly for exploring the data rather than to draw conclusions. Because the sample size *n* is important for conclusions and small or large *n* does not affect many diagrams, conclusions based on a diagram alone can be very fallacious.
8. As much as possible, directly label the categories, lines, etc., so that reader does not have to refer to legend or text for this purpose. If the audience for a figure is the general population, as in a newspaper, consider inserting text boxes to explain the salient features.
9. Never ever try to depict the obvious. For example, a graph of a proportion *p* on one axis and $(1 - p)$ on the other axis (probability of disease versus probability of its absence) is bound to give a perfect line. Such a graph does not tell us anything except the ignorance of the person who drew this type of graph. The two axes must be meaningful in all graphs and able to communicate something to the viewer.

A graph or a diagram is necessarily an approximate depiction. The values 103.2 and 103.4 can be shown to be distinct in a table but would look the same in a graph. Thus, graphs are good for visual display of a pattern or trend but not for depicting exact values.

1. HBR Reprint F1406Z. *Harvard Business Review* June 2014:32–3. http://hbr.org/product/the-story-of-the-first-charts-in-three-charts/an/F1406Z-PDF-ENG
2. Tufte ER. *The Visual Display of Quantitative Information*, Second Edition. Graphics Press, 2001.

## gross domestic product (GDP)

Gross domestic product (GDP) is a combined measure of the total income of people living in an area. The term actually applies to an established area, such as a country or a province, but is generally used per capita when divided by the population. Grossly speaking, GDP is the monetary value of the total output (goods produced and services provided) of the area during a 1-year period. This obviously severely depends on the population. GDP per capita is an indicator of productivity but is considered the most valid indicator of the standard of living of the people of an area since people spend most of what they produce. Many health parameters have been seen to be affected by GDP per capita. Availability and utilization of health care facilities are directly related, and parameters such as body mass index (BMI) [1], homicides [2], and infectious diseases such as tuberculosis [3] are indirectly related.

GDP per capita serves as the base for calculating some health-related indicators. The most common of these is the expenditure on health. This can be separately calculated for private expenditure by

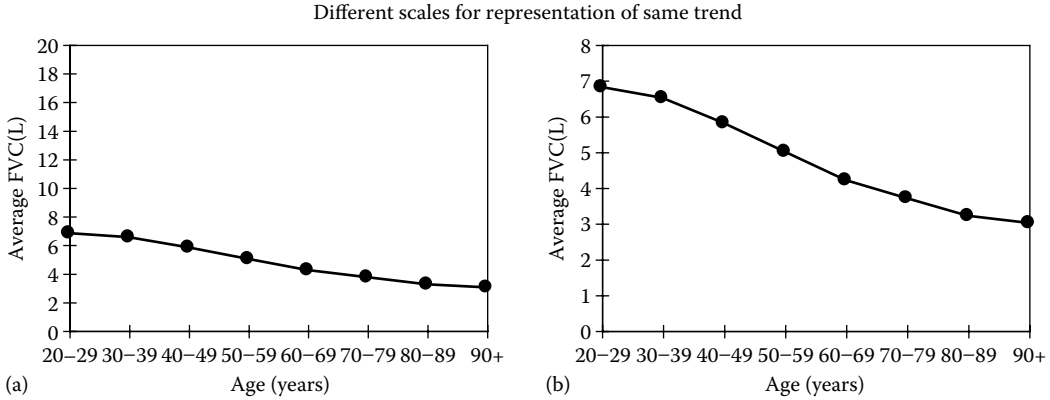



**FIGURE G.7** Effect of choice of scale on the slope or a line: (a) relation of average forced vital capacity with age; (b) the same trend with a different scale on the *y*-axis.

individuals and public expenditure by the government. According to the World Bank estimates for the years 2009–2013, the total (private + public) expenditure on health (excluding water and sanitation) in Brunei Darussalam was only 2.3% of the GDP, whereas it was 10.8% in Belgium [4].


1. Neuman M, Kawachi I, Gortmaker S, Subramanian S. National economic development and disparities in body mass index: A cross-sectional study of data from 38 countries. *PloS One* 2014 Jun 11;9(6):e99327. http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0099327
2. Sousa CA, Silva CM, Souza ER. Determinants of homicides in the state of Bahia, Brazil, in 2009. *Rev Bras Epidemiol* 2014 Mar;17(1):135–46. http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-790X2014000100135&lng=en&nrm=iso&tlng=en
3. Chen M, Kwaku AB, Chen Y, Huang X, Tan H, Wen SW. Gender and regional disparities of tuberculosis in Hunan, China. *Int J Equity Health* 2014 Apr 27;13(1):32. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4013307/
4. The World Bank. Data: *Health Expenditure, Total (% of GDP)*. http://data.worldbank.org/indicator/SH.XPD.TOTL.ZS


## gross enrollment ratio, *see* education indicators

## gross reproduction rate, *see* fertility indicators

## group average method of clustering, *see* centroid method clustering

## grouped data, *see also* class intervals

When the values of the **quantitative variable** under study are tabulated for the number of subjects in specified class intervals, this is called grouped data. The numbers of subjects in the class intervals are called frequencies. Table G.7 has grouped data on the cholesterol level of 200 subjects coming to a hypertension clinic.

Exact measurements on the **metric scale** are indeed statistically preferable to grouped measurements. The irony is that sometimes, circumstances force the grouping of metric data into categories even after exact data are obtained. The weight of a woman may be recorded to the nearest kilogram but may have to be categorized into 5 kg intervals such as (40–44), (45–49), etc. Data reported in this manner are called grouped data. The process is commonly referred to as categorizing a quantitative variable, and the groups are called **class intervals**. The reasons for doing this may be one or more of the following:

- Consider a data set containing systolic blood pressure (BP) levels of 1200 persons. The only effective way to present these in a report is by using groups such as (100–109), (110–119), (120–129), etc., and stating the number of subjects in each such group. This saves space and at the same time makes the data more intelligible. Storage of grouped data may take only about half a page or 1 kB of space, whereas storage of 1200 individual values may take four pages or 8 kB of space. Such grouping also makes the data more sensible, while 1200 ungrouped values may be difficult to comprehend. This certainly is done for the research papers published in journals, although now, some journals in health and medicine back it up with a website where all the original data are stored for anybody to see.

  Groups such as (0–4) and (5–9) for years of age assume that age is noted in terms of *completed* years or age at last birthday. The interval (5–9) actually means 5 to less than 10 years and can also be written as "(5–10) years." It is customary in such statistical grouping that the upper end of the interval is considered to belong to the next interval. Mathematically, these intervals are written as [0–5), [5–10), etc., but this kind of exact notation is seldom used in practice. Wherever the intervals are continuous, the convention of (0–4), (5–9), etc., is generally followed in health and medicine.
- It is well known that the end digit is predominantly 0 or 5 in many data values. This happens either because of approximation done by subjects themselves at the time of inquiry, such as stating one's age as 45 years instead of the more exact 44, or because of the observer's bias, such as in recording a systolic 130 mmHg instead of the exact 132 mmHg. Intervals (105–114), (115–124), (125–134), etc., or (108–112), (113–117), (118–122), etc., would dilute the effect of such digit preference. In another setting, suppose waist and hip sizes are measured without sufficient care and could be in error of up to 5 mm. Grouping of waist–hip ratio in intervals (0.7–0.8), (0.8–0.9), (0.9–1.0), etc. would minimize the effect of such errors, and the purpose of assessing central obesity could still be adequately achieved despite errors, provided they are minor.

The preceding two reasons are valid for grouping at the stage of reporting or analysis, but sometimes, even the recording is done in a grouped form. This is done for the following reasons:

- Eliciting a woman's age or anybody's income is sometimes considered immodest. Some people prefer to keep such information confidential. Stating them in a grouped form may be more acceptable. The exact value remains confidential, yet data are available in a usable form.
- Many clinicians are accustomed to think in terms of anemia present or absent and its degree as mild, moderate, or severe in place of exact hemoglobin (Hb) or hematocrit values. Thus, they sometimes prefer grouped values. Two or more measurements can also be simultaneously considered in this kind of grouping. Categorization of growth of

**TABLE G.7**
**Cholesterol Level in 200 Subjects Coming to a Hypertension Clinic**

| Cholesterol Level Class Interval (mg/dL) | Number of Persons (Frequency) |
|---|---|
| <150 | 5 |
| 150–199 | 76 |
| 200–239 | 39 |
| 240–259 | 35 |
| 260–279 | 24 |
| 280–299 | 12 |
| 300+ | 9 |
| **Total** | **200** |

a child into excessive, normal, retarded, or dismal depends not only on height and weight but also on the age of reaching different milestones of development. Such multifactorial grouping is sometimes more relevant for the practice of medicine.

- In an experiment on a lethal dose of a drug in mice, it is much easier to observe each morning and record the number of dead mice than to keep a continuous watch and note the exact time of death. In this case, the survival time would be available in 24 h categories. Serum glucose level is measured in units of 5 mg/dL because the analyzer in some cases is so calibrated. Thus, 5 mg/dL categories are inadvertently formed. Greater accuracy may be redundant in this case. If better accuracy is needed, cost and efforts may substantially increase.
- For a continuous variable, grouped data are essential to make a histogram or frequency polygon. The shape of these figures gives a good idea of the statistical distribution the values follow. For example, you can guess from these shapes whether the distribution is Gaussian or not.

Whenever data are available in an exact form, statistical analysis should be done using the exact data. Grouping means that the actual values are lost and various calculations are done assuming that all the values in an interval are equal to the midpoint of the interval. To see how this is done, see, for example, the topic **mean**. Midpoint-based calculation obviously is an approximation. Thus, grouping renders analysis less efficient in the sense that some important features of the data may fail to emerge. Statistical inference becomes less efficient because of the assumption that the values in grouped data are flat within each interval. This is against the factual position. Also, discontinuity is imposed by categorization across interval boundaries. An interval 160–169 of systolic BP disregards that 168 mmHg is more than 162 mmHg and the difference is more than that between 168 and 170, which now belong to separate categories. Cut points of intervals are mostly arbitrary, and different cut points can give different results.

In some cases, though, a larger sample size can compensate for this loss. A larger sample size helps to capture a better spectrum of values even when data are grouped. In fact, in cases in which grouped data can be rapidly obtained at substantial savings, more reliable results can be obtained by investigating a larger sample within the same cost. In such a situation, grouped data on metric measurements can be rightly advocated. However, it is important that the number of groups and width of intervals are appropriately chosen so that the essential features of the data are not compromised and the relevance is not lost. It is generally considered desirable that the number of class intervals be between 6 and 12, and class intervals of equal width are preferred. For some measurements, though, unequal categories with medical implications can be made, as for cholesterol level in Table G.7. End intervals may be open, as in our example, but these should be closed as much as possible so that there is no uncertainty about the maximum and minimum values. For family size, each number by itself could be a category, but this would not be called grouped data.

Despite our advice for using the exact values in place of grouped data for statistical treatment, there are examples where exact values lead to kind of a wrong conclusion and grouped data give the right conclusion. One such example is the relation between BMI and, say, disease frequency. Disease is high at both ends (very low and very high BMI), and a regression on exact data will yield a nice U-shaped curve with minimum disease frequency at 23 or 24 kg/m$^2$ and higher at BMI values on either side. However, when BMI is divided into class intervals, the bottom is found to be rather flat, implying that the disease frequency at BMI of 26 or 27 kg/m$^2$ is not higher than at 23 or 24 kg/m$^2$. For details of this example, see Welch et al. [1].

You will need the aforementioned basic concepts to understand medical literature, when trying to analyze data from your own clinic, and when trying to do research.

1. Welch HG, Schwartz LM, Woloshin S. The exaggerated relations between diet, body weight and mortality: The case for a categorical data approach. *CMAJ* March 29, 2005;172(7): 891–5. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC554875/

## group sequential design

In some studies, particularly in **clinical trials**, it may be worthwhile to plan to conduct them in stages by adding a group of subjects after reviewing the results from the subjects already entered. If the results at any stage provide convincing evidence either of the desired efficacy of the regimen or of futility, there is no need to add subjects. This essentially means that the data are analyzed in stages so that the trial can be stopped or continued depending on the results obtained. This can be planned to be done at as many stages as considered necessary. In the extreme case, one can think of analyzing the data sequentially after each pair of observations (one subject for the test group and one for the control group) and decide to continue or stop. In this case, the term *group* is dropped—it is just a **sequential design**. This is difficult and not considered logistically practical for clinical trials. Instead, the data are analyzed at equispaced stages, such as four stages, each after completing one-fourth of the trial. The null hypothesis is either rejected or not rejected at each stage. No further addition is made in case the hypothesis is rejected or continuing the trial is found to be futile. This is called a group sequential design and is accepted by most regulatory agencies as a valid design.

Group sequential designs are initially planned to have a large number of subjects so that they have adequate statistical **power** to detect relatively small gain, although the actual target may not be to detect such a small gain. Commitment of a large sample up front may be scary to some researchers, but that is a feature of group sequential designs. If interim analysis at any stage provides clear evidence of the desired efficacy, the trial is stopped prematurely, and a decision is made with reduced sample size. If evidence emerges that there is hardly any chance of achieving the desired minimum efficacy and it is futile to continue, then also, the trial is stopped. Otherwise, the trial continues. This interim assessment is done in such a manner that the level of significance is not compromised. For this, methods such as **O'Brien–Fleming** and **Lan–deMets** are used. Your statistical software may have the facility to do these tests not just for means but also for efficacy in terms of proportions and for durations that require log-rank tests.

The total number of stages and stopping criteria are specified in advance in the **design** itself. Minimum clinically relevant gain or efficacy must also be specified a priori. **Stopping rules** are devised in such a manner that the **level of significance** and power are least affected. In this approach, the sample size or the design is not modified once committed to in the protocol. This modification is done in an **adaptive design**, although in this also, the protocol contains what types of modifications are planned and when. For a comparison of group sequential designs with adaptive designs, see Kelly et al. [1]. Pocock [2] has discussed group sequential design in detail.

1. Kelly PJ, Sooriyarachchi MR, Stallard N, Todd S. A practical comparison of group-sequential and adaptive designs. *J Biopharm Stat* 2005;15(4):719–38. http://www.tandfonline.com/doi/abs/10.1081/BIP-200062859?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%3dpubmed
2. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977;64(2):191–9. http://biomet.oxfordjournals.org/content/64/2/191.short

## growth charts, *see also* growth indicators of children

Growth charts are figures that show specific **percentiles** of parameters of children at various stages of growth. Among the dimensions of physical growth of children are weight, height, body mass index, and head circumference. Each can be assessed against age, or they can be assessed against each other. The requirement of different growth charts for different body measurements arise as different parts (brain, arms, trunk, etc.) grow at varying rates at different times from birth to adulthood. For the purpose of illustration of the type of diagram they are, see Figure G.8a and b, which are weight-for-age charts for girls, developed by the United States National Center for Health Statistics (NCHS) [1] and the World Health Organization (WHO) [2].

The charts show various percentile curves based on the measurements seen in a large number of *healthy* children. Thus, this technically is a line diagram but is called a chart. The 50th percentile is like the median, which can be used as a reference, and the 3rd and 97th percentile curves define the lower and upper limits for healthy growth. The space between these two percentile curves is considered the *road to health*. Sometimes, ***Z*-scores** are used instead of percentiles.

A growth chart is used for longitudinal monitoring rather than for one-time cross-sectional assessment. The *trend* should follow the same pattern as the reference curve. A flattening or declining trend relative to the reference is an indication of decline in nutritional status, and a steeper upward trend not crossing the upper limit indicates improvement. For further details of how these charts are used for assessing various aspects of growth, see the topic **growth indicators of children**.

Preparing growth charts is an intricate process. This is primarily because (i) the growth measurements such as height and weight in children generally do not follow a **Gaussian** pattern—thus, calculation of various percentiles is difficult, and (ii) it is difficult to identify a smooth pattern that does not miss out on natural periods of steep and shallow growth, even no or negative growth. Most old charts used the **LMS method**, which takes care of the **skewness** of data, but now, the **BCPE method** is generally used, which takes care of deviation in **kurtosis** also. Both these methods are described in this volume, and the details for medical professionals have been succinctly provided by Indrayan [3].

1. CDC. Vital and Health Statistics Series 11, Number 246. *2000 CDC Growth Charts for the United States: Methods and Development*. Department of Health and Human Services 2002. http://www.cdc.gov/growthcharts/2000growthchart-us.pdf
2. WHO. *Child Growth Standards: Weight for Age Charts—Girls*. http://www.who.int/childgrowth/standards/cht_wfa_girls_p_0_5.pdf?ua=1
3. Indrayan A. Demystifying LMS and BCPE methods of centile estimation for growth and other health parameters. *Indian Pediatr* 2014;51:37–43. http://www.indianpediatrics.net/jan2014/jan-37-43.htm

## growth indicators of children, *see also* *Z*-score, *T*-score, velocity of growth

The physical growth of a child is assessed by anthropometric measurements such as weight, height, chest circumference, and head circumference. For this, norms are constructed that help to decide whether a child is on the road to health or not. The norms are generally obtained in the form of percentiles for age intervals of 6 months after 2 years, i.e., age 2 years, $2\frac{1}{2}$ years, 3 years, $3\frac{1}{2}$ years,

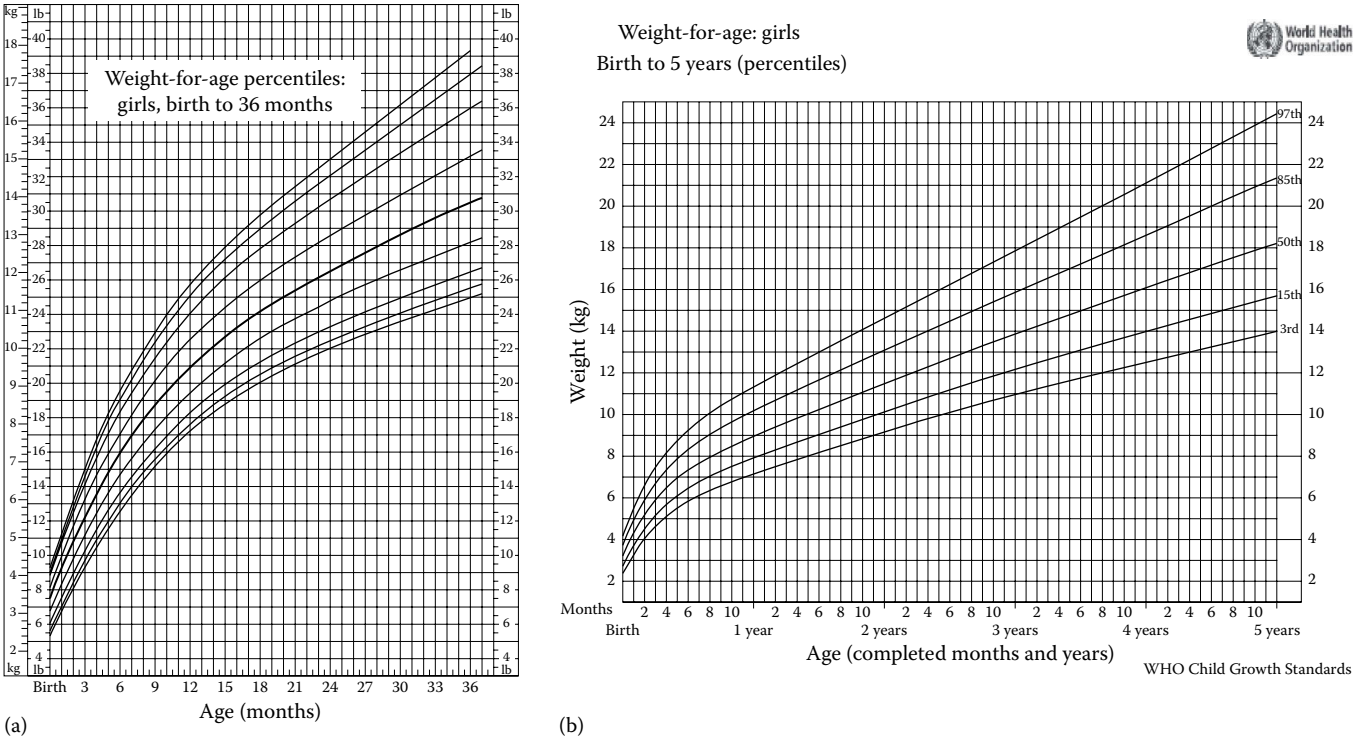



**FIGURE G.8** Weight-for-age charts (girls): (a) NCHS, birth to 36 months; (b) WHO, birth to 5 years.

etc. Smaller age intervals are desirable for age-specific norms for younger children and infants. These norms are plotted and connected by a smooth curve to get **growth charts**. In addition, mathematical models for growth in stature of children are also obtained as discussed by Ledford and Cole [1].

Each of the anthropometric measurements of a child can be independently assessed by either the **percentile** point achieved by the child relative to the healthy children of that age and gender in the same population, or by his/her ***Z*-score**, ***T*-score**, or percent of median. Sometimes, **velocity of growth**, is assessed as discussed in that topic.

Weight for age is the most commonly used indicator of growth of children but provides more effective assessment when the trend over age for the same child is studied. This trend is compared with the trend seen in healthy subjects in that population. The assessment is population specific. The difficulty with weight for age, however, is that it fails to distinguish a thin but tall child from a well-proportioned child.



The weight for height index obviates the need to know age particularly if it is between 1 and 10 years, and can be safely used for children when age is in doubt. This index measures the balance between weight and height (length in the case of children less than 2 years). A weight less than the 3rd percentile point for a particular height indicates wasting (i.e., thinness) associated with failure to gain weight or loss of weight [2]. This is considered an indicator of acute undernourishment. Weight for height fails to detect abnormalities when both height and weight are affected.

Another index for assessing growth is height for age. Low height for age is an indicator of stunting (i.e., shortness) when weight for height is normal. This is frequently associated with chronic undernourishment resulting mostly from poor nutrition or repeated illness, or many times, a combination of these two.

### *Percentiles, Percent of Median, Z-Scores, and* T-*Scores*

In the case of percentiles, the median (50th percentile) is regarded as a reference value, and 3rd and 97th percentiles as thresholds to indicate abnormally low or abnormally high values. Interpretation of the health of children with measurements outside the 3rd and 97th percentiles can be difficult. No matter how healthy the children are who measured to construct the chart, 3% of them will have weight less than the 3rd percentile curve. Thus, even some fully healthy children may show a weight in the low category. This is an acknowledged limitation of a growth chart, but the chart is still useful. Note that low weight in such children is in a relative sense only—relative to the other 97% healthy children. Thus, a low weight does not necessarily indicate poor health in an absolute sense.

The other index used to assess growth is percent of median. If the median weight of healthy children of height 1.10 m is 21.0 kg, and the weight of a child of this height coming to a clinic is 18.5 kg, this is 18.5/21 × 100 = 88% percent of median. A measurement above 80% of the median is regarded as normal, between 71% and 80% as indicating undernutrition of grade I, between 61% and 70% as grade II, and 60% or less as grade III. Each population can evolve its own classification. The parents are advised of suitable corrective steps depending on the grade of undernutrition found in a child. Now, consider the following example.

The interpretation of anthropometric measurements sometimes becomes easier when the **Z-score** is computed. This is the difference of the value from the mean in standard deviation (SD) units. A *Z*-score below –2 is considered low and below –3 exceedingly low. These *Z*-scores can be obtained for almost all anthropometric measurements. This score works well for measurements that follow an approximate **Gaussian distribution**.

Healthy children from well-to-do families were surveyed for their height, weight, age, and gender. The distribution of 450 girls by weight, taken within a week of their seventh birthday, is as follows:

| Weight (kg) | 14–16 | 16–18 | 18–20 | 20–22 | 22–24 | 24–26 | 26–28 |
|---|---|---|---|---|---|---|---|
| No. of girls | 4 | 36 | 85 | 182 | 96 | 40 | 7 |

These give mean = 21.1 kg, median = 21.0 kg, and SD = 2.3 kg. A girl who is nearly 7 years old comes to a clinic from the same area. Her weight is 18.3 kg. Can she be considered underweight?

$$\text{Z-score for the girl} = \frac{18.3 - 21.1}{2.3} = -1.22.$$

$$\text{Weight as percent of median} = \frac{18.3}{21.1} \times 100 = 87.$$

A negative *Z*-score and weight less than 100% of the median both indicate that the girl's weight is less than the average. However, because the *Z*-score is not less than –2 and the percent of the median is not less than 80, the weight can be regarded as within the normal variation, and the girl is not classified as underweight. Thus, there is no cause for alarm. What is important in this case is the longitudinal follow-up to monitor that the pattern remains on normal trajectory.

As explained under the term ***T*-score**, the only difference between a *Z*-score and a *T*-score is that the mean in *Z*-score is replaced by the most desirable or optimum value for that age. Thus, the *T*-score assesses the growth against the optimum.


1. Ledford AW, Cole TJ. Mathematical models of growth in stature throughout childhood. *Ann Hum Biol* 1998;25:101–15. http://www.ncbi.nlm.nih.gov/pubmed/9533510
2. WHO. *Child Growth Standards: Length/Height-for-Age, Weight-for-Age, Weight-far-Length, Weight-for-Height, and Body Mass Index-for-Age—Methods and Development.* World Health Organization, 2006. http://www.who.int/childgrowth/standards/Technical_report.pdf


## growth velocity, *see* velocity of growth in children

## Grubbs test

This test is used to detect outliers in a data set assuming that the distribution of values is Gaussian. **Outliers** are unusual values that do not fit into the general pattern of other values. These affect the values of the mean and standard deviation (SD)—and thus disturb the inference. Data managers are generally careful and ensure that no outliers are present unless there are valid reasons for them to be so. The Grubbs test helps to do this statistically and check that a particularly high value or particularly low value can be regarded as inconsistent with the other values. This test was proposed by Frank Grubbs in 1950 [1].

For any potentially outlier *i*th value, this test is obtained by computing

$$G = \frac{|y_i - \bar{y}|}{s},$$

where $\bar{y}$ is the sample mean and $s$ is the SD. Both are calculated after excluding the potential outliers. Thus, this is just the difference of the potential outlier from the mean of the regular values in SD units. In the first step, $y_i$ is either the maximum or the minimum value. In subsequent stages, you can have the second highest

and second lowest value if these are suspected to be outliers. In any case, the test has to be **one-tailed** since you already know that it is an abnormally high or abnormally low value. Grubbs has provided tables of critical values of $G$ that can be used to decide that the chance ($P$-value) of a value being an outlier is less than the significance level, such as 0.05. Your software may have a provision for this test that will give the $P$-value. If $P < 0.05$, the value can be regarded as an outlier and excluded from analysis so that the results do not become contaminated. This can be done for many values if many are suspected to be outliers. In that case, though, it is advisable to use the **Bonferroni procedure** for apportioning the significance level since the same data set will be used again and again.

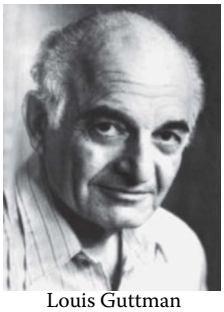
There are a couple of important considerations in determining a value as an outlier or not based on the Grubbs test. First, any outlier so determined should be thoroughly examined for whether this was not a data entry error, recording error, or reporting error. If any error is detected, the correct value should be found and replace the incorrect one. Second, the Grubbs test is based on Gaussian distribution. If the data do not follow this distribution, you may erroneously label a value as an outlier. Third, and the most important, is that the mean and, consequently, the SD themselves can be affected by outlier values, particularly when $n$ is not sufficiently large—thus our advice to calculate these after excluding the potential outliers, although this judgment has to be as objective as possible. This can be done only when the potential outliers are just one or two values. If you suspect many outlying values, examine whether the entire data set is contaminated.

There are some other methods for identifying outliers. A **box-and-whiskers plot** has this provision. A simple **histogram** can also show if one or two values are outliers on either side. A **normal probability plot** would also indicate if any value is an outlier or not.

1. Grubbs FE. Sample criteria for testing outlying observations, *Ann Math Stat* 1950;21(1):27–58. http://projecteuclid.org/euclid.aoms/1177729885

## Guttman scale

A binary (yes/no) set of questions or items are said to be on a Guttman scale, developed by Louis Guttman in 1944 [1], when they follow a hierarchy so that a "yes" to any item implies that the answer is yes to the subsequent (or preceding) items. The characteristic you want to measure is arranged in a continuum from minimum to maximum (or vice versa). For example, visual acuity is assessed in stages, and if a person with a cataractous eye can read font size 24 from a distance of 14 in., this implies that he/she can read font size 48. Thus, a threshold can also be obtained, and one response can predict all other responses in lower hierarchies. Gothwal et al. [2] has done such a study on subjects with a cataract using the Guttman scale. This scale is mostly used to develop short questionnaires that can still discriminate, such as for assessing disease severity and the level of satisfaction with hospital services.


Louis Guttman

Contrast the Guttman scale with, say, a **Likert scale**, where the scores for different items are added. In a Likert scale, generally, all items are supposed to have same difficulty. In Guttman, no addition is needed as the nature of the items is itself cumulative and of increasing difficulty. If a person scores 6, it means that he/she agrees with the first 6 statements.

The scale cited in the preceding example on visual acuity is fully scalable in the sense that there will be hardly any case who is able to read font size 16 but not font size 24. In many practical situations, this hierarchy is not so good. Efforts are sometimes made to frame items from the least to the most difficult in the hope that more able persons will be able to correctly answer more difficult items. However, in this case, a correct answer to item 3 does not necessarily imply a correct answer to items 2 and 1. Such "errors" make the scale really probabilistic.

Sometimes, the study is in reverse mode, where an attempt is made to decide on a Guttman scale of questions from the answers. Such recovery of a good Guttman scale from noisy data is challenging. In this setup, the potential items are placed in a kind of random order and analyzed for difficulty on the basis of the number of correct scores. Then they are placed in hierarchy. This can be understood as the rank ordering of the items. In this case, errors, as mentioned in the previous paragraph, are used to calculate

Coefficient of reproducibility:

$$C_R = 1 - \frac{\text{number of actual errors}}{\text{maximum number of possible errors}}.$$

For this, $K$ hierarchical items on $n$ persons are arranged into $K \times n$ configuration—one column for each person containing Y for correct response and N for wrong response. The number of possible errors is the same as $K \times n$. The items are generally accepted to be on a Guttman scale if $C_R > 0.90$. This means that the errors are less than 10%. Items with large errors can be omitted to improve this coefficient. The *coefficient of reproducibility* does not work if many items are too easy or too difficult for everybody.

Another important criterion is the *coefficient of scalability*. This is the proportion of items that can be correctly identified by the score, i.e., the items that exactly fall into the nice hierarchical pattern. This also is between 0 and 1, the larger the better, but a coefficient exceeding 0.60 is considered acceptable. This means that at least 60% of items must fall into the exact hierarchical pattern. The following example illustrates the procedure.

Consider a scale with $K = 7$ items on motor functions 1 week after stroke in $n = 10$ patients. These are arranged in order of number of Y's in Table G.8. Those who have complete paralysis will have hemiplegia, hemiparesis, etc. They are placed in hierarchy according to the number of Y's. But the hierarchy between hemiparesis and lower extremity, and between balancing and self-care is not apparent as the number of Y's is the same.

The procedure just mentioned can be easily implemented when the number of items is small. When it is large, a **scalogram analysis** is done to select those items that fall best into the Guttman hierarchy of difficulty. A subjective element may creep into the final selection, and it may not be possible to devise a perfectly cumulative scale. An illustration of this is in Table G.9 based on the data in Table G.8. The patient with the most number of Y's is in row 1, and the patient with the least number of Y's is in the last row. This rearrangement makes it easier to spot the errors in the sense that the items do not follow the hierarchy. If they follow the hierarchy stipulated in the Guttman scale, the Y's should form nearly a triangle. An error is an interior N and exterior Y in the scalogram.

**TABLE G.8**
**Motor Functions in 10 Patients 1 Week after Stroke**

| | Motor Function | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Patient No. | Complete Paralysis | Hemiplegia | Lower Extremity | Hemiparesis | Upper Extremity | Difficulty in Self-Care | Difficulty in Balancing | No. of Y's |
| 1 | N | N | Y | N | N | Y | Y | **3** |
| 2 | N | N | Y | Y | Y | Y | Y | **5** |
| 3 | Y | Y | Y | Y | Y | Y | Y | **7** |
| 4 | N | N | Y | Y | Y | Y | Y | **5** |
| 5 | N | N | N | N | Y | Y | N | **2** |
| 6 | N | Y | Y | Y | Y | Y | Y | **6** |
| 7 | N | N | N | Y | Y | Y | Y | **4** |
| 8 | N | N | N | N | N | Y | N | **1** |
| 9 | N | N | N | N | N | N | Y | **1** |
| 10 | N | N | N | N | N | N | Y | **1** |
| **No. of Y's** | **1** | **2** | **5** | **5** | **6** | **8** | **8** | **35** |

**TABLE G.9**
**Table G.8 Reordered for Number of Y's—Scalogram**

| | Motor Function | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Patient No. | Complete Paralysis | Hemiplegia | Lower Extremity | Hemiparesis | Upper Extremity | Difficulty in Self-Care | Difficulty in Balancing | No. of Y's |
| 3 | Y | Y | Y | Y | Y | Y | Y | **7** |
| 6 | N | Y | Y | Y | Y | Y | Y | **6** |
| 2 | N | N | Y | Y | Y | Y | Y | **5** |
| 4 | N | N | Y | Y | Y | Y | Y | **5** |
| 7 | N | N | [N] | Y | Y | Y | Y | **4** |
| 1 | N | N | [Y] | N | [N] | Y | Y | **3** |
| 5 | N | N | N | N | Y | Y | [N] | **2** |
| 8 | N | N | N | N | N | Y | [N] | **1** |
| 9 | N | N | N | N | N | N | Y | **1** |
| 10 | N | N | N | N | N | N | Y | **1** |
| **No. of Y's** | **1** | **2** | **5** | **5** | **6** | **8** | **8** | **35** |

In Table G.9, there is a complete hierarchy for "complete paralysis" and "hemiplegia" since if the response for these is Y, the response to all the lower ones is also Y. This is not so for "lower extremity"—patient nos. 2 and 4 follow the hierarchy, whereas patient no. 1 has Y for this but N for "hemiparesis" and "upper extremity," which are lower down the order. This is where subjectivity creeps in. One researcher may want to put "lower extremity" lower than "hemiplegia" amongst motor functions after stroke. The responses that do not follow the hierarchy are in the box in Table G.9. The number of actual errors is 5, and the maximum number of possible errors is 70. Thus, $C_R = 1 - 5/70 = 0.93$. If we remove difficulty in balancing, $C_R = 1 - 3/60 = 0.95$. Since 5 of the total 7 items are in perfect hierarchy, the coefficient of scalability = 5/7 = 0.71. Both of these measures reveal that the questions in this example are in good agreement with the Guttman scale. This really means that if the response is "yes" to "complete paralysis," there is no need to assess other motor functions. If the answer to this is "no," then you go down to the next assessment.

These are easily illustrated for this small table, but if the number of items and number of subjects are large, as would be the case in most practical situations, you may have to use the help of a software that does this.

For further details of Guttman scale, see Garson [3].

1. Guttman, LA. A basis for scaling qualitative data. *Am Sociol Rev* 1944;9(2):139–50. http://www.jstor.org/discover/10.2307/2086306?uid=3739560&uid=2129&uid=2&uid=70&uid=4&uid=3739256&sid=21104160367727
2. Gothwal VK, Wright TA, Lamoureux EL, Pesudovs K. Guttman scale analysis of the distance vision scale. *Investig Ophthal Visual Sci* 2009;50(9):4496–501. http://www.iovs.org/content/50/9/4496.full.pdf+html

# H

## half-life of medications

When a drug is ingested, it may reach its peak concentration in the system gradually but then starts to decay quickly, or may reach the peak quickly and starts decaying slowly. Half-life is the time taken to go from the peak concentration to half that concentration. Figure H.1 shows the half-life for a drug that reaches its peak gradually but decays quickly. Half-life is utilized to measure the duration that a drug remains in the system and may form the basis for prescribing the periodicity (once a day, twice a day, etc.) of intake. For example, Schulze-Bonhage and Hintz [1] suggested that the antiepileptic drug perampanel can be prescribed once daily because of its long half-life.

If the half-life of a drug is 3 h, it reduces to one-half of its peak concentration in 3 h. It does not mean it will vanish in the next 3 h. After 6 h, the drug concentration will be one-half of what it was after 3 h, i.e., one-fourth of the peak concentration; and after 9 h, it will be one-eighth of its peak concentration. It takes more than four half-lives to reduce the concentration to less than 5% of its peak concentration. After this, the concentration will not vanish but tends to stabilize. A rule of thumb says that it takes about four times the half-life of the drug for the concentration of that drug in the system to reach a steady state irrespective of the half-life. So if you administer a drug with a half-life of 12 h, the steady state will be achieved after 48 h. Drugs with a short half-life reach a steady state relatively quickly compared with those with a long half-life. An example of the statistical use of half-life is in the field of **crossover trials**. For a usual crossover experiment to be successful, there must be no **carryover effect**. The time required for the carryover effect to vanish is called the **washout period**. This should generally be at least four times the half-life of the substance to be confident that there is only a negligible amount of the substance in the system.

The half-life is obtained by fitting a **regression** of concentration on time and locating the time point where concentration is one-half of the peak. In most pharmacologic applications, concentration decline is fast in the beginning and increasingly slow later on, which suggests that a log scale is appropriate. Hence, a linear regression of log(concentration) versus time is generally used. If the slope of this line is $b$,

$$\text{Half-life: } T_{1/2} = \frac{\ln(0.5)}{b}.$$

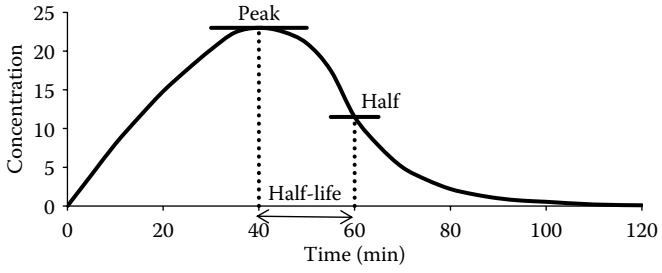



**FIGURE H.1** Half-life of a rapidly vanishing drug after reaching its peak concentration.

Since the concentration is declining over time, $b$ in this case would be negative. The value for ln(0.5) is −0.693 (note that it is also negative). Thus, $T_{1/2}$ would be positive, with a value dependent on $b$. The basic equation is $C_T/C_0 = (1/2)^r$, where $C_T$ is the concentration at time $T$, $r = T/T_{1/2}$ (that is, $r$ is the number of half-lives and $T$ is measured in terms of half-life units), and $C_0$ is the peak concentration. If the half-life of a drug is $T_{1/2}$ = 30 min, the concentration at $T$ = 90 min would be $(1/2)^3$ = one-eighth of what it was at peak concentration since $r$ = 90/30 = 3. The time to reach 10% of concentration in this example is given by $0.10 = (1/2)^r$, which yields $r$ = 3.32, and $T = 3.32 \times 30 = 100$ min.

Feng et al. [2] observed the half-life of Arg-Gly-Asp (RGD)-modified aclacinomycin A (ACM) liposomes to be 1.2 times of that of ACM liposomes and concluded that RGD-modified ACM liposomes have a better antitumor effect than the unmodified ones in lung adenocarcinoma. This illustrates one of the many uses of half-life.


1. Schulze-Bonhage A, Hintz M. Perampanel in the management of partial-onset seizures: A review of safety, efficacy, and patient acceptability. *Patient Prefer Adherence* 2015 Aug 11;9:1143–51. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4542413/
2. Feng C, Li X, Dong C, Zhang X, Zhang X, Gao Y. RGD-modified liposomes enhance efficiency of aclacinomycin A delivery: Evaluation of their effect in lung cancer. *Drug Des Devel Ther* 2015 Aug 11;9:4613–20. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4541546/


## haphazard sampling

Amongst the many sampling techniques, nonrandom methods of sampling can be used for convenience: these fall under the term **purposive samples**. Examples are given here. Going one step further, a combination of two or more methods of purposive sampling can be used, and this is termed *haphazard sampling*. The main types of purposive sampling are as follows.

Subjects who are easily available or who can easily submit to the study form a **convenience sample**. The sampling of volunteers is the most prominent example. Volunteers have a definite place in medical studies, particularly in phase I of a clinical trial, but they may be class apart from the target population for the regimen under trial. Another convenience sample is that from a captive population, such as medical students. They can be easily persuaded by the faculty to be subjects of their study for effects with no potential harm. For example, they may be persuaded to undergo lung function testing before and after an exercise under trial, or can be asked to fill in an anonymous questionnaire on sexual behavior. Medical students may be physically and mentally healthier than other youths, and those agreeing to participate may be of a specific type—either too bright or too dull. Thus, the results cannot be generalized to the population of interest. A third type of convenience sample in medicine is that of referred cases. These are easily available and may also be easily persuaded to participate. Referred cases, however, are almost invariably complicated cases and do not represent the general class of patients. Ironically, they may not even represent complicated cases: sometimes, referral is done of noncomplicated cases for a particular investigation or examination that another center specializes in.

Other purposive sampling techniques include telephone sampling, snowball sampling, Delphi sampling, and quota sampling.

From the description of these nonrandom methods, you can see that they have extremely limited utility. Besides difficulty in identifying the population that these samples represent, nonrandom samples cannot be used for statistical inference, such as tests of significance and confidence intervals, because these statistical methods require some random element. At the same time, however, nonrandomly selected subjects can sometimes provide useful leads for further studies that can be based on random samples. Nevertheless, large-sized samples, even when nonrandom, may still represent the cross-section of the corresponding population and so can be adequate for statistically valid inferences.

For example, consider a relatively rare and not easily identifiable population such as gay men for sexually transmitted diseases. To survey such a population can be expensive and technically difficult [1], and the data available cannot be counted on as being reliable. Yet, researchers often abandon **probability sampling** methods, relying instead on convenience samples such as patrons of gay bars for this segment of the population because that at least helps in identifying the subjects.

The sampling becomes haphazard when some subjects are included as volunteers, some from those easily available, some from those who are referred cases, etc. The use of such haphazard sampling seriously undermines the credibility of the results. Some researchers may use this method of sampling without acknowledging that they are doing so, or without acknowledging its limitations.

1. Blair J. A probability sample of gay urban males: The use of two-phase adaptive sampling. *J Sex Res* 1999;36(1):39–44. http://www.jstor.org/stable/3813629

## hard data and soft data

For statistical analysis, the data are mostly entered in terms of numerics into a computer suite of software. These are hard data. They lend validity to what you are doing or plan to do. Some features of the health spectrum are soft in the sense that they can only be understood in the mind but are difficult to put on paper. This applies particularly to psychological variables such as depression, frustration, opinions, and feelings. These form the soft data. Even if put on paper, soft data may defy coding, particularly if this is to be done before the collection of data. A precoded proforma is considered desirable these days because it makes computer entry so easy, but it should be ensured in this process that the medical sensibility of the information is not lost. The entries are created in a structured manner that many soft data defy; consequently, useful soft data many times remain unexploited.

Dixon-Woods [1] reviews the history of safety measurements through to the modern day and explores how different methods of measurement can yield varying results. Her key message is that not everything we need to consider can be measured as hard data; soft intelligence is also needed. "If you're not measuring, you're not managing. If you're measuring stupidly, you're not managing. If you're only measuring, you're not managing." Despite such intricacies of soft data, they still have a place in medicine in situations where perception matters more than reality.

Despite lacking statistical rigor, soft data serve the purpose of providing the context. They help in adding the emphasis, explaining the importance, and communicating the urgency. Soft data may include verbal communications, awards and honors, appreciation and brickbats, endorsements and repudiations, etc. In the context of health care, these could be concerns of the patients, media reports about the success of a new treatment modality or its failure, success stories, etc. A good report would not ignore soft data to leverage a point emerging from the hard data. Media reports generally start with soft data and then reach the hard data, if any at all. Scientific reports are primarily based on hard data but still use soft data for emphasizing the importance of the findings. Sometimes, soft information provides "meat" to the discussion part of a medical research paper.

It is often helpful to statistically relate soft data with hard data. For example, you can find out how improvement in severity score after a treatment translates into satisfaction or increases self-confidence in dealing with illness. This kind of exercise can provide a complete picture, say, before and after a treatment. A step further would be to examine different domains of severity score to find out which particular aspect contributes more to satisfaction than others. The difficulty would be in measuring the level of satisfaction, although something like a **Likert scale** can be used. Those domains that provide a high level of satisfaction can be exploited to advance the clinical objectives.

1. Dixon-Woods M. *Hard Data and Soft Intelligence: How Can We Know Care is Safe*. http://www.health.org.uk/multimedia/slideshow/hard-data-soft-intelligence/, last accessed November 8, 2015.

## harmonic mean, *see* mean (arithmetic, geometric and harmonic)

## harmonic regression

This kind of regression is used to model periodic oscillations in a time series. Periodic oscillations can occur, for example, due to seasonality of some diseases such as dengue occurring in some areas where mosquitoes breed in stagnant water after the rainy season each year, due to circadian rhythm in some of our body functions, due to menstrual cycle, etc. Generally, the trigonometric functions sine and cosine are used for modeling these oscillations. This is a called *sinusoidal curve*. This can incorporate frequency (peak once a year as in the case of dengue, once a week such as sickness absences on Fridays or Mondays, etc.) as well as amplitude, which depends on whether the peak is sharp or shallow and how high it goes from the secular trend. The model can be modified to include asymmetric oscillations such as rapid rise but gradual fall, and the analysis can be done as in the case of regression. An extension of this is **spectral analysis**, which considers the whole process simultaneously. This analysis includes periodicity as well as secular trend (Figure H.1). Ordinary time series analysis requires a priori fixing of periodicity, while spectral analysis discovers the periodicity and decomposes the series into its systematic components that otherwise may appear random.

Pocock [1] describes the harmonic analysis of a 5-year series of weekly records of sickness absence in a factory. There were $n = 260$ weekly observations ($5 \times 52 = 260$). Plotting "spells commencing per week" against "week no." (Figure H.2) gives an irregular plot on a smooth curve. The former shows a seasonal trend, and the latter is a single sinusoidal curve that can be seen to have a period of 1 year. A comparison of the two demonstrates that the seasonal trend is also nearly sinusoidal. There are, however, clear departures from the simple curve, particularly at annual holiday periods.

Bridges et al. [2] used harmonic regression methods to study seasonality of suicides in the United States—as they were above average all spring and below average all winter. Scavone et al. [3] observed that medications administered into the epidural or intrathecal space for labor analgesia may demonstrate variable effects dependent on the time of day, and this may affect clinical research

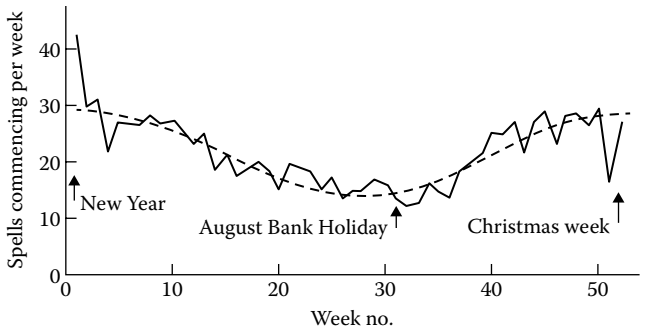



**FIGURE H.2** The estimated seasonal trend in weekly spells of sickness absence, 1960–1964. (From Pocock SJ. *Appl Stat* 1974; 23:103–20. http://www.jstor.org/discover/10.2307/2346992?uid=3738256&uid=2134&uid=2474376657&uid=2&uid=70&uid=3&uid=247437664&uid=60&sid=21104481367491, with permission.)

trials investigating the pharmacology of specific drugs. Their analysis demonstrated a 24 h harmonic cycle for cervical dilation at first analgesia request with maximum values occurring near 17:00 hours and minimum values near 05:00 hours, but the amplitude of the difference was very small. They called it *rhythm analysis*.

As mentioned earlier, the method of harmonic regression involves trigonometric sin and cos terms, and the method of fitting is mathematically complex, which we wish to avoid. Appropriate software may help, but our description may have given you a good idea of what harmonic regression is and where this could be applied.


1. Pocock SJ. Harmonic analysis applied to seasonal variations in sickness absence. *Appl Stat* 1974;23:103–20. http://www.jstor.org/discover/10.2307/2346992?uid=3738256&uid=2134&uid=2474376657&uid=2&uid=70&uid=3&uid=2474376647&uid=60&sid=21104481367491
2. Bridges FS, Yip PS, Yang KCT. Seasonal changes in suicides in the United States, 1971 to 2000. *Perceptual Motor Skills* 2005;100:920–4. http://uwf.edu/fbridges/PMS-June-2005-Part-2_0003.pdf
3. Scavone BM, McCarthy RJ, Wong CA, Sullivan JT. The influence of time of day of administration on duration of opioid labor analgesia. *Anesth Analg* 2010 Oct;111(4):986–91. http://www.ncbi.nlm.nih.gov/pubmed/19897803


## harmonization of data

Harmonization of data refers to the process of standardizing the inputs and outputs in research studies, especially if the goal is to combine the data into a single, integrated data set [1]. The concept of harmonization appears more commonly in comparative research since the objective of the process of harmonization is achieving comparability, particularly when the data are coming from heterogeneous sources. It can be applied at any stage of the study, e.g., questionnaire design, sampling, method of eliciting, data editing, but generally, antecedents in an **analytical study** are separately harmonized from outcomes. Separate harmonization helps in ensuring better comparability and is checked before the data are subjected to statistical analysis. This is also a prerequisite for exchange of information and automation, and helps in delivering improved health care as the evidence base becomes more reliable.

On a practical plane, data harmonization refers to the steps taken to ensure that the data from different sources have the same categorizations (e.g., each center uses the same age categories), same definitions (such as hypertension defined uniformly by all involved as blood pressure > 140/90), same method of measurement (each center using the same kind of instrument), same method of interview (e.g., in home or in clinic), same form for recording (this is generally done), same assessment (e.g., for classifying cases into mild/moderate/serious), etc. The procedure to identify outliers, wrong entries, missing values, etc. is also standardized across various sources. If any inconsistencies remain, **adjustments** as needed are made to make the data comparable before analysis is undertaken.

Harmonization has special application to laboratory results, which have a significant role in diagnosis and monitoring of patients. Such harmonization is required not just across laboratories but also across tests in the same laboratory. Different tests in the same laboratory must not give disparate findings just because one is done in standardized conditions and the other in not-so-standardized conditions, or one follows far a more accurate method than others. Also, if the aliquots of the same sample are analyzed by different laboratories, they must reach the same result. Wherever needed, recalibration is done, and for other situations where this cannot be done, limitations are explicitly stated. Reference material is made available for scrutiny by those who suspect the results. For further details, see AAAC [2].

Abner et al. [3] have discussed harmonization of longitudinal studies on aging, cognition, and dementia in the United States. See Magalhaes et al. [4] for efforts to develop a common methodology that could enhance the opportunity of data harmonization in the etiology of multiple sclerosis across several countries.


1. Granda P, Blasczyk E. *XIII. Data Harmonization*. 2010. http://ccsg.isr.umich.edu/pdf/13DataHarmonizationNov2010.pdf
2. AACC. *Clinical Laboratory Test Harmonization*. http://www.harmonization.net/
3. Abner EL, Schmitt FA, Nelson PT, Lou W, Wan L, Gauriglia R, Dodge HH et al. The statistical modeling of aging and risk of transition project: Data collection and harmonization across 11 longitudinal cohort studies of aging, cognition, and dementia. *Obs Stud* 2015 Mar;1(2015):56–73. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4431579/
4. Magalhaes S, Pugliatti M, Casetta I, Drulovic J, Granieri E, Holmøy T, Kampman MT et al. The EnvIMS study: Design and methodology of an international case-control study of environmental risk factors in multiple sclerosis. *Neuroepidemiology* 2015;44(3):173–81. http://www.karger.com/Article/FullText/381779, last accessed January 3, 2016.


## Hawthorne effect

In the 1920s, researchers at the Western Electric Company's Hawthorne plant in Chicago observed that workers changed their behavior and responses when they knew that they are being observed. This psychological phenomenon is termed the Hawthorne effect. The effect, however, diminishes over time: it is a short-term effect. In the long term, performance reverts to previous levels.

Many years later (in 1939), Roethlisberger [1] documented the results of the Hawthorne studies. At that time, it was generally thought that financial rewards would be the main driver of performance. The Hawthorne effect negated that as it was observed that greater productivity resulted when management made workers feel valued and aware that their concerns were taken seriously. Although the conclusions of the Hawthorne studies have been called into question, the theory persists—probably because most people understand that they usually do perform better when observed than they do otherwise. For example, Stringly et al. [2] found hand hygiene event rates to be three times higher in hallways within eyesight of the auditor compared with when no auditor was visible.

In medical care, patients may show improvement just because they perceive that they are being looked after well. In clinical trials,

the placebo group can show enhanced efficacy because the subjects know that they are in a trial and are being observed. This is in addition to the placebo effect that comes from the false impression that they are being treated. Thus, the Hawthorne effect can cause a considerable unsuspecting bias.

A good review of the Hawthorne effect is provided by McCambridge et al. [3], who called it research participation effect. After a review of several studies, they observed that it does exist in most studies, although not much can be said about their mechanism or magnitude, and concluded that further research is required to understand this effect. Verstappen et al. [4] proposed a block design that allows for control of the Hawthorne effect in randomized controlled trials of test ordering.


1. Roethlisberger FJ. *Management and the Worker: An Account of a Research Program Conducted by the Western Electric Company, Hawthorne Works, Chicago*. Harvard University Press, 1939.
2. Srigley JA, Furness CD, Baker GR, Gardam M. Quantification of the Hawthorne effect in hand hygiene compliance monitoring using an electronic monitoring system: A retrospective cohort study. *BMJ Qual Saf* 2014 Dec;23(12):974–80. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4251174/
3. McCambridge J, Witton J, Elbourne DR. Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *J Clin Epidemiol* 2014 Mar;67(3):267–77. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3969247/
4. Verstappen WH, van der Weijden T, ter Riet G, Grimshaw J, Winkens R, Grol RP. Block design allowed for control of the Hawthorne effect in a randomized controlled trial of test ordering. *J Clin Epidemiol* 2004 Nov;57(11):1119–23. http://www.ncbi.nlm.nih.gov/pubmed/15567627




## hazard functions

Hazard function describes the **hazard rate** per unit of time at different points in time. The concept of hazard function is easily understood in the context of the general population, where the hazard of death is high at the beginning of life (neonatal period), is very low for age 5–59 years, and steeply increases thereafter. Thus, it has nearly a **bathtub** shape (Figure H.3a). The rate is higher at both ends of life in developing countries than in developed countries. In other setups, it can remain constant throughout the time period under observation or may continuously rise (Figure H.3b). An increasing hazard indicates weakening, whereas a declining hazard indicates hardening (i.e., as time passes, it becomes more and more difficult to experience the event). Several other forms can be visualized.

Perhaps, it is easier to understand *cumulative* hazard function in place of the hazard function itself. It sequentially adds the hazard from one time to the next time and describes the probability $P(T < t)$, where $T$ is the variable failure time and $t$ is the actual value of the time. Thus, cumulative hazard function is the probability that a failure has occurred before time $t$. This is like the probability that a person dies before the age of 70 years. The death can occur at age 50, 60, or 69 years—all are included. You may realize that the cumulative hazard function is the complement of the survival function and is thus also called *failure function*. This depends on the **distribution** of the failure time.

The hazard of death is not uniform throughout the life of the general population. Though rare, if it is indeed uniform for some conditions, such as for a disease that strikes the young and the old at the same rate and causes death at the same rate, the hazard function $P(T < t)$ takes the form

$$\text{exponential hazard: } F(t) = 1 - e^{-\lambda t}.$$

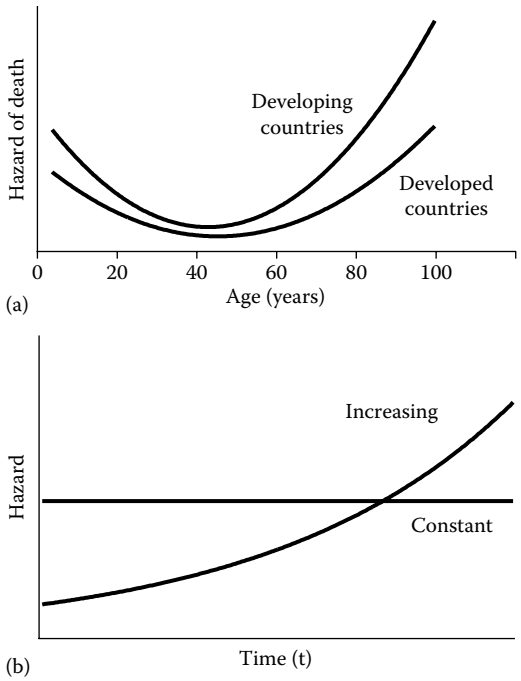



**FIGURE H.3** (a) Hazard of death at different age in general population; (b) constant and increasing hazards.

The function given in this equation is called an **exponential function**. Since hazard is not uniform in most situations, the exponential form of cumulative hazard does not provide an adequate representation. Nonuniform hazard can be understood further by the example of an automobile whose failure rate in the beginning after coming out of the showroom is low and high in, say, the seventh year of running. In this case, the hazards rapidly increase as time passes. In the case of heart transplantation, the initial few days and weeks are critical. After this period, the hazard declines, at least for some time. One such cumulative hazard function that models varying hazards over time is the **Weibull distribution**. The third commonly used cumulative hazard function is **lognormal**. This is appropriate when the duration of survival has a hugely **skewed** distribution to the right—long and very long durations are also present, though mostly, they are not very long. The life span of cells may have this pattern. Data from the Surveillance, Epidemiology, and End Results (SEER) study revealed in a 27-year follow-up that the survival times of cancer patients who died of their disease can be considered to follow a lognormal distribution [1].

All models described in the preceding paragraph are called parametric since they depend on the **parameters** that define the shape of the distribution. Many workers feel that parametric models impose difficult-to-justify restrictions. Also, there are many options, and prior information is rarely available to make a correct choice. You may have to try more than one model to get an adequate fit. On the other hand, a semiparametric model, particularly **Cox regression**, is more flexible and safer. This has become a standard choice for many applications where dependence of the hazard ratio on a set of covariates is studied.

One rather simple nonparametric estimate of cumulative hazard function is

$$\hat{H}(t) = \sum_{T \le t} \left( \frac{d_t}{n_t} \right),$$

where $d_t$ is the number of failures out of $n_t$ at time $t$. Thus, this is evaluated as the sum of the estimated discrete hazards at all the event times up to $t$.

While analyzing survival data, if the hazard function is of interest, choose the one that meets the eligibility requirement. If you are examining a hazard function used by someone else, view it in light of the preceding discussion.

1. Tai P, Yu E, Cserni G, Vlastos G, Royce M, Kunkler I, Vinh-Hung V. Minimum follow-up time required for the estimation of statistical cure of patients: Verification using data from 42 cancer sites in the SEER database. *BMC Cancer* 2005;5:48. http://www.biomedcentral.com/content/pdf/1471-2407-5-48.pdf

## hazard rate/ratio, *see also* proportional hazards

Let us first explain the term *hazard*. This is the same as *risk* and is the probability of occurrence of an event, generally an adverse event such as death. However, hazard is invariably linked to time, whereas risk may not be. Hazard of death at age between 70 and 71 years is the probability of death among those who are alive at age 70 years. This is the conditional probability (see the topic **probability**) and is not the same as the (unconditional) probability of death. Probabilities of death at different ages will add up to 1 because the ages are mutually exclusive and exhaustive, while conditional probabilities do not have this feature. Conditional probability of death between 80 and 84 years may be 0.6 and of death between 85 and 89 years may be 0.7. Their sum is more than 1. This cannot happen with unconditional probabilities. Hazard at any time can be lower or higher than the unconditional probability for that time. Hazard is more relevant in some situations and statistically easier to work with in a semiparametric model, as in **proportional hazards**.

Now we come to the hazard rate. In a disaster such as an earthquake, thousands of people may die in a few hours. The intensity of death or force of mortality in persons exposed to such disasters is extremely high. Three days is less than 1% of a year, and if 800 people die, the rate is 800/(3/365) = 97,333 persons per year! Such a force of mortality at a particular instant is called the hazard rate.

Note that the hazard rate also is very different from the **probability** of death. While the probability of death cannot exceed 1—the sum total of the probabilities from all the different causes has to be 1—hazard rate can exceed 1. Probability is for the occurrence of an event at the end of a period, while the hazard rate is always per unit of time for any specified duration. It measures the speed of occurrence—probability does not. The interest sometimes is not in the probability but, indeed, in the rate of failure in a particular time interval. The hazard rate is the "limit" of this rate as this interval approaches 0. Statisticians use the concept of limit to obtain a mathematical form that works at a particular point in time when time is continuously observed. The hazard rate at time $t$ is the rapidity of failure in the next instant. This is obtained by dividing the conditional probability of failure in the next instant by the length of that instant provided the hazard remains constant over that period. Hazard rates are absolute, and this concept can be applied to all health conditions that change from one state to another in the course of time. Because of its intimate dependence on time, hazard rate is commonly used in **survival analysis**, where the outcome is measured in terms of duration.

While a hazard rate evidently depends on time, it may also depend on several other factors. In a clinical setup, the hazard rate of serious side effects may depend on the covariates, such as the characteristics of the person like age, gender, and nutritional status, as well as on the type of regimen, type of domiciliary care, alertness, competence of the attending physician, etc. When two or more groups are available, such as an experimental and a control group, the groups can be compared with respect to the hazard rate at a particular point in time for a given set of covariates. For such a comparison, since many factors are involved, it is sometimes helpful to obtain the hazard rate as a function of various covariates. One such model is a form of regression, known as **Cox regression**. In this type of regression, the term *hazard rate* is generic and is not restricted to death. It can be used for any other event of interest, such as appearance or reappearance of symptoms, or even for a favorable event, such as discharge from hospital, cessation of smoking, or resumption of daily activities.

Turning now to the hazard ratio, this is obtained as the ratio of hazard rates in the presence of a risk factor and the rate in its absence, such as with a particular disease and without that disease. This can vary with time, and this is what makes it different from odds ratio or relative risk, which have no time element. Presence or absence of a risk factor could be in terms of levels also. For example, we can have the hazard ratio of death of patients with mild disease and with serious disease. In a clinical trial setup, hazard rates could be the "hazard" of recovery, say, per week in patients on the test regimen and on the control regimen.

We all know that the hazard rate of death increases after the age of 60 years as we age further. This is so for healthy persons and more so in persons with some terminal disease. The rate rapidly increases in both groups as we age, but the rate is much higher in persons with terminal disease. In none of these groups is the hazard rate constant as we move from age 60 years to age 90 years. Since the hazard rates increase in both groups, the ratio of hazards at different ages may still be constant. If the hazard rate of death in persons of age 70–74 years is 82 per thousand per year in healthy people, the rate may be 133 per thousand in persons with terminal disease of the same age group. The hazard ratio is 1.5 in this age group. In age group 80–84 years, the rates may be 150 per year and 225 per year, respectively. The ratio is still 1.5 in age group 80–84 years as in the age group 70–74 years despite a more rapid increase in the cases with disease. This is what is called **proportional hazards**. See this topic for further details. Sometimes, hazard rate is reported without reference to the time. For example, it is sometimes calculated as the ratio of median survival duration in the two groups. This can mislead just as all other averages could—hazard ratio could be high during early follow-up and low in the late phase of the follow-up period. For an interesting account of the hazards of hazard ratios, see Hernán [1].

A common misinterpretation occurs when a hazard ratio of 2 is understood to mean that the occurrence of events (such as death) is twice as fast in one group as the other. This is not so. A hazard ratio of 2 in age, say, 70–74 years means that the chance of occurrence of an event (e.g., death) is twice as much in one group as in the other.

1. Hernán MA. The hazards of hazard ratios. *Epidemiol* 2010;21(1):13–5. http://journals.lww.com/epidem/Fulltext/2010/01000/The_Hazards_of_Hazard_Ratios.4.aspx

## health-adjusted life expectancy, *see* life expectancy

## health inequality, *see also* Gini coefficient, Palma measure of inequality

As the name implies, health inequality is the disparity in health of different sections of the society. People with income in the top 10% bracket enjoy disproportionately far superior health than people with the bottom 10% income. This is like the popular perception

that the top 20% of people control 80% of the wealth in almost any free society—called the **champagne glass effect**, as per the details given later in this section. Health also has a similar skewed distribution.

When looking across different segments of the population, health inequality affects the rate of improvement in population health overall. It is true that deprived sections of the population suffer most from health inequalities, but health inequalities affect all sections as disparities hamper the generation of human capital and inhibit sustainable improvement. In addition, all inequalities are unfair and violate social norms.

The most common statistical measure of health inequality is the **Gini coefficient**, as described under this topic along with an illustrative example of how this is calculated. This coefficient is generally between 0.2 and 0.6 and does not capture inequalities as much as it should. Cobham and Sumner [1] make a passionate plea that the Gini coefficient should not be used, although their argument is in the context of income. This more or less applies to health also. An alternative method, called the **Palma measure of inequality**, has been proposed that seems to be more sensitive to inequalities. This measure also is described separately. Asada et al. [2] consider the Gini coefficient as a univariate measure and propose a three-stage approach for measuring health inequality involving cross-disciplinary dialogues.

### *Champagne Glass Effect*

A champagne glass (Figure H.4) is a very effective tool to depict inequality. In this glass, the top 20% of height contains 80% of the champagne. In health also, generally, the top 20% of people use 80% of the health resources either to keep themselves fit and healthy or to enjoy medical care facilities to reduce suffering and to live long. Only 20% resources are left to meet the needs of the remaining 80% of the population. In this depiction, the size of the bowl can be increased if the distribution is even more skewed or can be reduced if it is not so skewed.


1. Cobham A, Sumner A. Is inequality all about the tails? The Palma measure of income inequality. *Significance* 2014;11(1):10–3. http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2014.00718.x/pdf
2. Asada Y, Hurley J, Norheim O, Johri M. A three-stage approach to measuring health inequalities and inequities. *Int J Equity Health* 2014 Nov 1;13(1):98. http://www.equityhealthj.com/content/pdf/s12939-014-0098-y.pdf

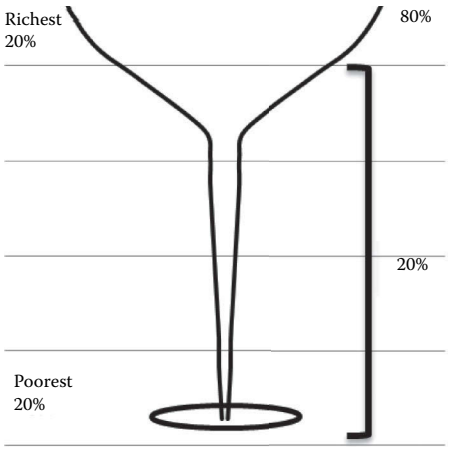




**FIGURE H.4** Champagne glass: top 20% height contains 80% of the total.

## health infrastructure (indicators of)

Availability of hospitals, health centers, dispensaries, and doctors, on one hand, and availability of food, safe drinking water, and sanitation facilities, on the other, together form the health infrastructure. The second component of this infrastructure (food, safe water, etc.) is not important for developed countries because nearly 100% of the population in those countries have these facilities, but it is very important for developing countries.

The following indicators can be used to measure the extent of availability of the health infrastructure:

$$\text{Population served per bed} = \frac{\text{population}}{\text{number of beds}}.$$

This can also be stated as beds per 1000 population and can also be termed density of beds, or **bed–population ratio**. According to the World Bank estimates, the number of beds per 1000 population was 0.8 in Sudan, 3.6 in Sri Lanka, and 6.5 in Belgium in 2012 [1]. These can be divided into intensive care unit (ICU) beds, general care beds, beds for cancer patients, beds for pediatric patients, etc., depending on the focus of the study.

The second health infrastructure indicator is the density of doctors, measured by the number of physicians per 1000 population, also known as **doctor–population ratio**. This can be calculated as

$$\frac{\text{number of physicians}}{\text{per 1000 population}} = \frac{\text{number of physicians}}{\text{population}} \times 1000.$$

The information available with the World Health Organization (WHO) by 2014 shows that nearly 46% of member states have less than 1 physician per 1000 population [2]. The definition can be extended to include doctors of other systems such as homeopathy and Ayurveda. These can also be broken down into doctors of various specialties and superspecialties. The formula remains the same for other personnel such as dentists, nurses, pharmacists, and field workers.

For accessibility, one can think of the percentage of the population with access to an approved health care facility within, say, 5 km. This is in accordance with considering health care accessible in a developing country when it can be reached by the public on foot or by local means of transport in no more than an hour.

A large number of other indicators of health infrastructure can be devised to serve a specific need. For example, you may be interested in medical seats per million population available in educational institutions, laboratories available, ambulance services for emergency cases and their response time, etc. In addition are the following:

$$\text{Per capita availability of food grains per day} = \frac{(\text{total production of food grains}) - (\text{waste}) - (\text{exports}) + (\text{imports}) \text{ in a year}}{\text{population} \times 365}.$$

Water supply is considered safe when it is treated surface water or untreated but uncontaminated water such as from springs, sanitary wells, and protected boreholes. Excreta disposal (sanitation) facilities are considered adequate if they can effectively prevent human, animal, and insect contact with excreta.

The percentage of the population with such access is an important indicator of health infrastructure. According to a WHO/United Nations Children's Fund (UNICEF) report for the year 2014 [3], the world sanitation coverage has increased by 21 percentage points in developing countries since 1990. As of the year 2014, the coverage was 64% in the world, ranging from 36% in the least developed countries to 96% in developed regions. For water coverage, the increase since 1990 in developing countries is 17 percentage points, and it stood at 89% in 2014, ranging from 67% in the least developed countries to 99% in developed regions. Even this basic amenity is not available to a large section of the population in some parts of the world.

Among other indicators that could be considered for infrastructure is expenditure on health. The World Bank computes this in three indicators for each country of the world, namely, health expenditure per capita in US dollars, public health expenditure as percent of total health expenditure, and total health expenditure as percent of gross domestic product (GDP). The country-wise data are available on their associated websites [4].


1. The World Bank. *Hospital Beds (per 1,000 people).* http://data.worldbank.org/indicator/SH.MED.BEDS.ZS, last accessed October 12, 2015.
2. WHO. *Global Health Observatory: Density of Physicians (total number per 1000 population, latest available year).* http://www.who.int/gho/health_workforce/physicians_density/en/, last accessed November 12, 2014.
3. WHO/UNICEF. *A Snapshot of Progress—2014 Update: WHO/UNICEF Joint Monitoring Programme for Water Supply and Sanitation (JMP).* http://www.wssinfo.org/fileadmin/user_upload/documents/Four-page-JMP-2014-Snapshot-standard-on-line-publishing.pdf, last accessed November 14, 2014.
4. The World Bank. *Data: Indicators.* http://data.worldbank.org/indicator/


## health measurement

We describe these separately for individuals and populations.

### *Measures of Individual Health*

Health measurement at the community level is fairly structured and discussed later in this section, but how do you know that you are healthier than your neighbor? This requires that we first understand individual health. One appealing explanation is as follows [1].

The human body has a mechanism to adjust itself to minor variations in the internal as well as external environment. Perspiration in hot weather, shivering in the cold, increased respiration during physical exercise, excretion of redundant nutrients, replacement of lost blood after hemorrhage, and decrease in the diameter of the pupil in bright light are examples of this mechanism. This is a continuous process that goes on all the time in our body and is referred to as homeostasis. Health could be defined as the dynamic balance of body, mind, and soul when homeostasis is going on perfectly well. The greater the capacity to maintain internal equilibrium, the better the health. Perhaps human efficiency is optimal in this condition. Sometimes, infections, injury, nutritional imbalances, stress, etc. become too much for this process to handle, and external help is needed. Medicine can be defined as the intervention that tries to put the system back on track when aberrations occur.

Health being a multidimensional concept, its measurement is intricate and mostly reduces to perceptions. Measurement of physiological and biochemical parameters can give an idea of the level of health of a person, but a person with an amputated leg could be healthier in his/her own right than a person with both legs but not able to walk! This is a normative issue and could be a biostatistical adventure. The measurement of health can be difficult at this stage of our knowledge. Yet, one can think of capacity to work as expected from a healthy person of one's age, both physical and mental, as an acceptable measure of individual health. If you include "a joyful attitude towards life, and a cheerful acceptance of the responsibilities that life puts upon the individual" [2], the measurement becomes even more challenging. Health should also not be equated with well-being [2] as well-being is governed by a host of factors of which health is just one. A definition that can help develop operational indicators of individual health is elusive. Perhaps measures such as **quality of life** or health-related quality of life can be considered as good surrogates. Irrespective of what metric is used, there are issues, such as whether health is to be measured for current status, for the past 1 year, or over a lifetime of experience.

Limiting the discussion to physical health, in Figure H.5, person A has had a healthier life than person B, although both lived the same number of years. Another measure could be the healthy years of life, but that would be more meaningful when considered in old age. This would subtract years spent with any disability, or the degree of disability can be discounted to compute weighted healthy years. This is similar to adjustment by **disability weights** done in computing **disability-adjusted life years (DALYs)** for community health. Note, however, that this kind of metric is mostly restricted to physical health and ignores social and mental health in terms of capacity to do intellectual work or in terms of contributions to the well-being of the family and the society. Another approach could be to divide health into its domains, such as level of mobility, level of self-care, level of

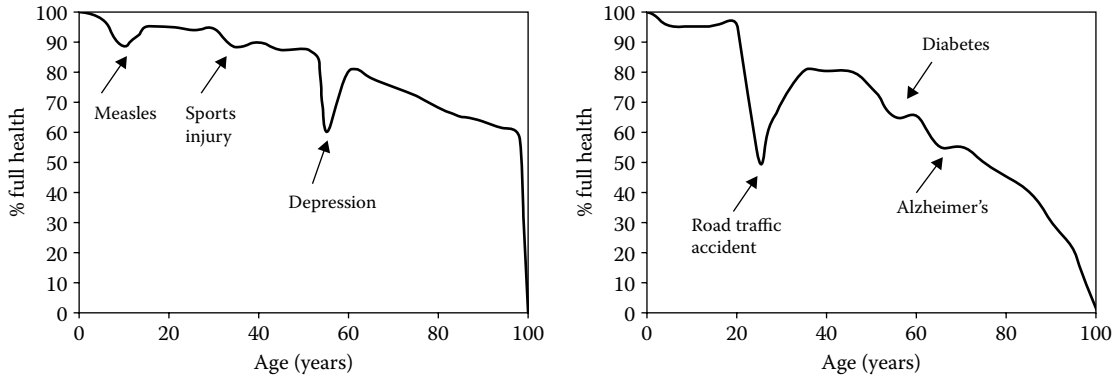



**FIGURE H.5** Lifetime physical health for person A (left panel) and person B (right panel). (From Sigerist HE. *Medicine and Human Welfare*. Yale University Press, 1941, with permission.)

pain, and level of cognition [3], and measure each domain just as we do for quality of life. Also see the topic **positive health**.

### *Measures of Population Health*

Measurement of population health or of a group of people is not so difficult, particularly the physical component. Several measures are available that are well defined and extensively used. Among them are the **incidence** and **prevalence** of various health conditions. These can be calculated for a specific age-sex group or any other group of interest, such as people in a particular occupation or people working in a particular factory. Another is duration of morbidity per year, for which sickness absenteeism could be a surrogate. Third is the severity of disease—a less healthy population will have a higher number of severe cases per 1000 population.

The second common indicator of population health is the **mortality rate**. This is measured in terms of crude death rate, standardized death rate, infant mortality rate, age-specific death rate, etc. Higher age-specific death rate in comparison with another population is a definite sign of poorer health. A good statistical property of death is that this can be ascertained without controversies, unlike morbidity, which has many gray areas.

A comprehensive measure of population health (in fact, a lack of it) is the **disability-adjusted life years (DALYs)** lost. This combines years of life lost due to early deaths and equivalent years lost due to disability from diseases and other adverse health conditions. This computation also requires **disability weights**. Another good measure of population health is the **life expectancy**, particularly the healthy life expectancy. All these indicators are for absence of disease or mortality, and would not be able to compare two populations with, say, the same life expectancy. The difference in health levels may not be substantial in countries with the same life expectancy now but may become important in future when, for example, many countries attain a life expectancy of more than 90 years—a threshold beyond which further improvement would be rare. At that stage, the percentage of the population living healthy at age 90 years could be an effective indicator.

The concept of positive health as conceived by Indrayan [1] for individuals is probably not applicable to communities. Indicators such as average total lung capacity and average hemoglobin level can be considered as possible candidates.


1. Indrayan A. *Medical Biostatistics*, Third Edition. Chapman & Hall/CRC Press, 2012.
2. Sigerist HE. *Medicine and Human Welfare*. Yale University Press, 1941.
3. Chatterji S, Ustün BL, Sadana R, Salomon JA, Mathers CD, Murray CJL. The conceptual basis for measuring and reporting on health. *Global Programme on Evidence for Health Policy Discussion Paper No. 45*. World Health Organization, 2002. http://www.who.int/healthinfo/paper45.pdf


## health programs (evaluation of), *see* evaluation of health programs/systems

## health situation analysis

This is the assessment of the health conditions of a population. Health situation analysis is generally undertaken before devising or launching a health program so that the program can be fine-tuned to meet the specific needs of the population, or for identifying the priority areas of concern.

Health situation analysis would include the quality and type of health services required for a community based on indicators such as the size of the community, its age–gender structure, its socioeconomic profile, and the prevalence of various conditions of health and ill-health in different sections of the community. The health services also depend on culture, traditions, perception, socioeconomic status of the population, existing infrastructure, etc. Expectations and demands are also considered. All these aspects need to be properly assessed to prepare an adequate plan for the provision of health services. The basic aim of this analysis is to provide the baseline information for such a plan. The situation can rapidly change because of either natural growth of the population or interventions, so a time perspective is always kept in view in this analysis.

For launching a health program, generally, the broad problem requiring a plan of action is fairly well known before embarking on the health situation analysis, and only the specifics are to be identified. It seems ideal to talk about both good and bad aspects of health, but in practice, a health plan is drawn up so as to meet the needs as perceived by the population. These needs are obviously related to the adverse aspects of health rather than to the positive aspects. Sometimes a **survey** is required, and sometimes an expert group is set up to identify the specifics of the problem.

The health situation analysis is also expected to provide clues to deal with the problem at hand. The magnitude of the problem in different segments of the population can provide an epidemiological clue based simply on excess occurrence in specific groups. Determinants or risk factors for the health condition under review are also assessed. These could also be available from the literature or from the knowledge and experience of experts in that type of situation.

Assessment of the functionally available **health infrastructure** is also an integral part of the health situation analysis. This includes facilities such as hospitals, beds, and health centers; staff in terms of doctors, nurses, technicians, etc.; supplies such as equipment, drugs, chemicals, and vehicles; and most of all, their timely availability at the functional level so that they can be effectively used.

## health statistics

Used as a plural, *health statistics* are data relating to various aspects of health. These include, but are not limited to, data on births, growth of children, nutrition, morbidity, health behavior, availability and utilization of health services, health costs, and deaths. **Vital statistics** that relate to fertility and mortality are part of health statistics. Some workers include **demographic** data, such as on age-sex structure, growth of population, migration, education, and occupation, in the ambit of health statistics since these factors affect health. Some would include even environmental data, such as on pollution, occupational hazards, and traffic accidents. In its broad sense, **medical biostatistics** relating to incidence and prevalence of diseases, hospital data, causes of death, **case fatality**, etc. can also be considered part of health statistics. All these can be studied for different segments of the population as needed, such as urban/rural, male/female, child/adult, etc.

Health statistics can be best understood by reference to the National Center for Health Statistics (NCHS) under the remit of the Centers for Disease Control and Prevention [1]. Using data open to the public, NCHS generates statistical tables, charts, and graphs illustrating various health statistics for the US population. For example, in 2014, NCHS published a special issue of health statistics from the Federal Interagency Forum on Child and Family Statistics entitled *America's Young Adults* [2]. Among others, this report concludes that there has been a steady downward trend in smoking since 1983 in 18- to 24-year olds regardless of race or gender. Also, from 1990, there has been a

steady trend downward in death rates among 18- to 24-year olds, with the most common cause of death being unintentional injuries.

You can see that the health statistics can be those that are directly collected from persons, households, health facilities, or even published reports. But these can also be derived from those data after necessary calculations or other processing. The data are further analyzed to derive useful health statistics that can be operationally utilized for taking action, or to convey meaningful conclusions. The example cited in the preceding paragraph illustrates this kind of derived statistic. The calculation of **disability-adjusted life years (DALYs)** lost based on morbidity and mortality is a fine example of how routine health statistics can be converted to information with far-reaching implications.

As always, the derived health statistics would be only as good as the original data themselves. All efforts must be made at the primary source level to ensure that the information is correctly obtained based on appropriate instrumentation and proper assessment. The definitions must be fully standardized, and the staff should be fully trained to use uniform methods. The data should be both **valid** and **reliable**. Exercise special caution while combining secondary data with primary data—the two may have followed very different procedures and may differ in unanticipated vital aspects. Some adjustment for known deficiencies in the data can be made at the time of analysis, but these adjustments may be available for highly restricted conditions.

1. CDC. *National Center for Health Statistics*. http://www.cdc.gov/nchs/
2. Snyder T. *America's Young Adults: Special Issue, 2014*. National Center for Education Statistics. http://www.census.gov/content/dam/Census/newsroom/c-span/2014/20140829_cspan_young_adults.pdf

## healthy life expectancy, *see* life expectancy (types of)

## hierarchical clustering, *see also* cluster analysis

Hierarchical clustering is a method of successive grouping of given units into an unknown number of homogeneous groups such that the values of the units within each group are alike but the units between groups are different. Technically, these groups are called clusters, which are sought to be internally homogeneous and externally isolated as much as possible. The method starts with identifying two units that are most similar and putting them into one group. Call this *entity*. Then similarity between this entity and all the other units is examined, and again, the most similar ones are merged into groups. This could mean that either the number of units in the previously formed entity becomes three or another entity of two units is formed. This procedure is repeated, merging one unit at each stage. That is why it is called hierarchical. The full name of this procedure is *hierarchical agglomerative clustering* because of sequential merging. The process can go on till such time that all units merge into one big group, but the purpose is to stop as soon as we find that subsequent merging is destroying the internal homogeneity of the clusters. We explain this in a short while in this section. A **dendrogram** is a graphical display of the merging taking place at each stage, and this can help in deciding when to stop.

A similar procedure can be adopted by beginning with all the units together as one big cluster. Divide this into two clusters such that these are very different from one another. Then divide them into three clusters, and so on. This process is called a *hierarchical divisive algorithm*. There are not many examples of the use of this method. Most clustering applications use an agglomerative algorithm.

Statistically, the problem is of dividing a fixed number of units, $n$, into a small but unknown number of homogeneous clusters, $K$, with respect to $J$ measurements on each unit. Now these clusters will have $n_1, n_2, \ldots, n_K$ units, respectively, such that $n_1 + n_2 + \ldots + n_K = n$. For example, we might divide various hemoglobinopathies into clusters that are internally similar but distinct from others, as measured by, say, anion high-performance liquid chromatography (HPLC). Clusters so formed would include hemoglobin (Hb) A2, Hb variants, HbF, even Hb, and red blood cell (RBC), as measured by mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), etc. The number of clusters is determined by calculating a measure of internal homogeneity at each stage of the clustering process. When the homogeneity is significantly disturbed, the process is stopped.

The process obviously requires a measure of affinity or similarity. For quantitative data, generally, **Euclidean distance** is used, which is a measure of differences instead of similarity. Units with the least distance are considered most similar. This can be obtained not just for univariate values but also for multivariate values. For qualitative binary data, the Jaccard coefficient or any other measure described under the topic **association between dichotomous characteristics (degree of)** can be used. For ordinal data, measures such as Somer $d$ are available (see **association between ordinal characteristics**), and for polytomous data, measures such as the contingency coefficient (see **association between polytomous characteristics**) are available. However, there are not many examples of the use of the clustering method for qualitative data.

The primary problem in hierarchical clustering is measuring the distance between two entities, where one entity has, say, $n_1$ units and the other entity has $n_2$ units. It is for this purpose that one of the methods among **complete linkage**, **simple linkage**, **Ward**, **median**, etc. is used. Generally, the method of complete linkage is preferred as it was found best at not discovering false clusters of random data in a simulation study [1].

The most difficult decision in the hierarchical clustering process is regarding the number of clusters naturally present in the data. The decision is made with the help of criteria such as pseudo-$r$ or the **cubic clustering criterion** [2]. These values should be high compared with the adjacent stages of the clustering process. Another criterion could be the distance between the two units or entities that are being merged in different stages. If this shows a sudden jump, it is indicative of a very dissimilar unit joining the entity. Thus, the stage where the entities are optimal in terms of internal homogeneity and external isolation can be identified. The entities at this stage are the required natural clusters.

See Everitt et al. [3] for further details of the hierarchical clustering process.

1. Jain NC, Indrayan A, Goel LR. Monte Carlo comparison of six hierarchical clustering methods on random data. *Pattern Recognition* 1986;19:95–9. http://www.sciencedirect.com/science/article/pii/0031320386900385
2. Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 1985;50:159–79. http://link.springer.com/article/10.1007%2FBF02294245#page-1
3. Everitt BS, Landau S. Leese M, Stahl D. *Cluster Analysis*, Fifth Edition. Wiley, 2011.

## hierarchical designs

Hierarchical design is a synonym for **nested design**, where one factor works within another factor. Consider a study on 600 stroke patients treated in 20 hospitals located in five countries. The recovery of these

patients would depend, among others, on (i) patient characteristics such as disease severity and the nutrition level, (ii) hospital practices such as nursing care and competency of the doctors, and (iii) country characteristics such as work culture and attitudes. The final outcome is the joint effect of these factors. Note that country-level characteristics can be measured only for the five countries participating in the study, and hospital practices can be assessed for 20 participating hospitals, whereas patient data would be available for 600 patients.

The analysis of data from hierarchical designs will consider, for example, within-hospital variation (across patients) as well as between-hospital variation. Within-country variation will be between hospitals as well as between patients. There will also be a between-country variation. Usual analysis does not consider this hierarchy, and **multilevel models** are required to analyze such data. In these models, the **analysis of variance (ANOVA)** is modified to include hierarchical structure. Whereas patients can be considered random samples from each hospital, if hospitals and countries are also random samples representing a bigger population, these also will be modeled as **random effects** in ANOVA. Setting up such an ANOVA with statistical software will require skill, and we would suggest getting the help of a qualified statistician to do such analysis for you. The analysis will have to take this hierarchy into consideration.

The details of analysis for two hierarchical factors are provided by Ellison and Barwick [1]. For another real-life example, see Kong et al. [2]. They have discussed methicillin-resistant *Staphylococcus aureus* (MRSA) transmission by patient proximity in hospitals with a nested structure where beds are positioned within cubicles and cubicles are positioned within wards. Any study on MRSA in this kind of hospital will follow a hierarchical design.

1. Ellison SLR, Barwick VJ. *Practical Statistics for the Analytical Scientist: A Bench Guide*, Second Edition. Royal Society of Chemistry, 2009.
2. Kong F, Paterson DL, Whitby M, Coory M, Clements AC. A hierarchical spatial modelling approach to investigate MRSA transmission in a tertiary hospital. *BMC Infect Dis* 2013 Sep 30;13:449. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3854069/

## hierarchical models/regression

Inclusion of **regressors** of different levels in a **regression analysis** gives rise to hierarchical regression. For example, for studying risk factors of coronary diseases, you can have some variables at an individual level, such as cholesterol level and body mass index, and some variables at a family level, such as housing and interpersonal support. In this situation, you cannot run the usual regression because (i) the number of families will be different from the number of individuals and their information will be on different variables, and (ii) the individuals within a family will tend to provide correlated information. This nesting and resulting clustering effect violates the requirement for validity of the usual regression. In our example, cholesterol level and body mass index among members of the same family are likely to be correlated because of a common diet, environment, and lineage. The same situation occurs while studying ambulation time after bariatric surgery done by different surgeons. Each surgeon uses his/her own precautions before, during, and after surgery, which may have a profound effect on the ambulation time besides, of course, the patient characteristics, such as extent of obesity and age. Models that describe hierarchical regression are called hierarchical models. These are also called **multilevel models/regression.** The other name is *nested models.*

The reasons that hierarchical models cannot be analyzed by the regular methods are twofold. First, some information will be available for each individual (level 1), some for subgroups (level 2), and some for groups (level 3). A model is built that incorporates this structure. Second, ordinary **regression** works well when the **regressors** are considered fixed—a condition that is rarely met in a hierarchical structure. Consider the duration of stay of patients in critical care in small, medium, and large hospitals. You might want to know whether or not this duration of stay varies according to the size of hospital, which in turn determines the quality of the facilities, the care available, and the confidence of the patients. If so, the chosen small, medium, and large hospitals are considered random samples of such hospitals in the city or state. In that case, their effect is not fixed but is random (see **fixed and random effects**). In addition, of course, the duration of stay will depend on the condition of the patient at the time of admission, which can be assessed by the **APACHE score**. Let the size of the hospital be defined by the number of beds: small if the number of beds is less than 100, medium if the number of beds is between 100 and 399, and large if the number of beds is 400 or more. The exact number of beds is not under consideration. A sample of $n_1 = 15$ patients from a small hospital, $n_2 = 30$ patients from a medium hospital, and $n_3 = 50$ from a large hospital are chosen from those undergoing critical care. In this example, the hospital effect is fixed, but the effect of the APACHE score is random as the sample of patients is random. The linear regression model in this situation will be of the following type:

$$y_{ij} = a_i + b_i x_{1i} + c_i x_{2ij} + d_i x_{1i} * x_{2ij},$$

where

- $y_{ij}$ = duration of hospital stay of $j$th patient of the $i$th-sized hospital ($j = 1, 2, \ldots, n_i$; $i = 1, 2, 3$)
- $x_{1i}$ = size of the $i$th hospital ($i = 1$ is small, $i = 2$ is medium, $i = 3$ is large), assuming that the size of hospital has a linear effect on the duration of hospital stay
- $x_{2ij}$ = APACHE score of $j$th patient of the $i$th hospital ($j = 1, 2, \ldots, n_i$; $i = 1, 2, 3$)
- $a_i$, $b_i$, $c_i$, and $d_i$ together define the intercepts and slopes for the patients, APACHE score, and the hospitals of different sizes

The basic premise in hierarchical regression in the context of this example is that the duration of hospital stay tends to follow different patterns depending on the size of the hospital, and that both hospital size and the APACHE score should be considered together as the factors affecting the duration of hospital stay. Depending on the data, the regressions may appear as shown in Figure H.6. The lines have different intercepts and different slopes. The usual line obtained after ignoring the hospital size, which assumes that all durations and APACHE scores are from the same target population, is also given. Note how this composite line hides the differences present in the lines for different hospital sizes.

The analysis basically is in terms of fitting two models: first for individuals using the usual regression analysis and second for considering the estimates of this regression as dependent on the second-level factors. See Garson [1] for details.

This example is for two levels (size within hospital and subjects within size) and can be extended for more levels. For example, one extension could be to include doctors as the factor influencing the hospital stay through their knowledge and attitude. The levels can be fixed effects if both levels are fixed (hospitals and wards—neither is a sample, and the conclusions are valid for these only), both can be random, or one can be fixed and the other random.

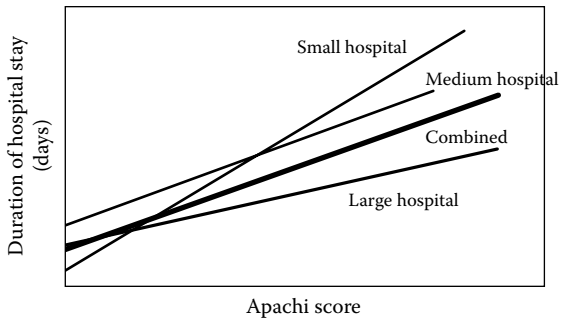



**FIGURE H.6** Hierarchical (multilevel) regressions.

This kind of regression can also be used in the analysis of **longitudinal** data. Suppose that after surgery, each patient is recorded for pulse pressure every 5 min till such time that the patient stabilizes. The first patient becomes stable in 40 min so that he is recorded for $n_1 = 8$ time points, the second in 30 min so that $n_2 = 6$, etc. If the patients are considered to be fixed (i.e., not a random sample), the situation is the same as mentioned previously. If they are a random sample, a slight modification of the model will be required.

1. Garson GD. *Hierarchical Linear Modeling: Guide and Applications*. Sage, 2012.

## histogram

A histogram is a set of contiguously drawn bars showing a **frequency distribution**. The bars are drawn for each group (or interval) of values such that the area so enclosed is proportional to the frequency in that group. Generally, the variable values are plotted on the horizontal ($x$) axis, and the frequencies are plotted on the vertical ($y$) axis. The vertical axis can represent percentage instead of frequency. This will affect the scale but not the shape of a histogram.

A histogram for the data in Table H.1 is shown in Figure H.7a. Note the following for these data. The first is an open interval because the lower end point is not mentioned. For the purpose of drawing the histogram, this interval is assumed to be of the same width as the next interval, i.e., 160–199. Thus, the first and the second intervals have width of 40 mg/dL each, whereas all others except the last have width of 20 mg/dL. For the area of the bar to represent frequency, the height of the bar for the first two intervals should half of the frequency in these intervals. This suitably adjusts for the double width of these intervals. The interval 320–339 has no subject, and this is represented by a blank space in the histogram. The last interval has width 60, which is three times the width of most of the other intervals. To adjust the area for this width, the height of the bar for this interval is one-third of the frequency in that interval.

**TABLE H.1**
**Distribution of Subjects Attending a Hypertension Clinic by the Total Serum Cholesterol Level**

| Cholesterol Level (mg/dL) | Number of Subjects ($f$) | Percent (%) |
|---|---|---|
| <199 | 3 | 3.7 |
| 200–239 | 13 | 15.9 |
| 240–259 | 16 | 19.5 |
| 260–279 | 17 | 20.7 |
| 280–299 | 24 | 29.3 |
| 300–319 | 6 | 7.3 |
| 320–339 | 0 | 0 |
| 340–399 | 3 | 3.7 |
| **Total** | **82** | **100.0** |

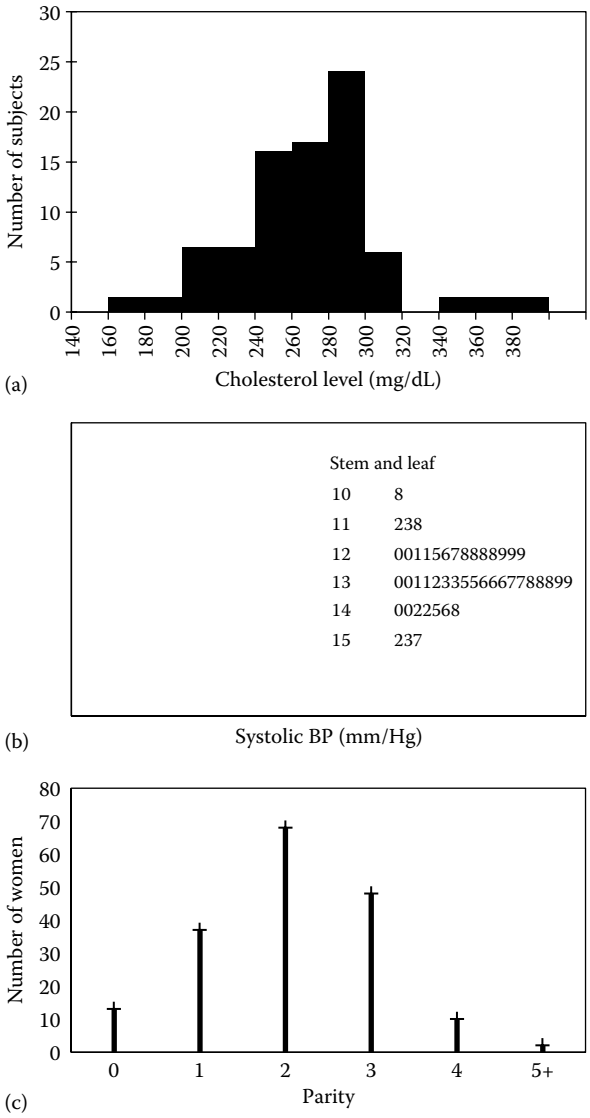



**FIGURE H.7** (a) Histogram for the data in Table H.1; (b) stem-and-leaf plot; (c) line histogram for discrete data.

The following two conditions are prerequisites for a histogram to be a valid representation: (i) Characteristics to be represented on the horizontal axis must be on a metric scale, preferably a continuous variable, such as cholesterol level in Table H.1. The categories must be numeric intervals and must be mutually exclusive and exhaustive. Ordinal and nominal scales are not represented by a histogram, nor are **multiple responses**. (ii) The data to be represented on the vertical axis must be either percentages that add up to 100 or frequencies that add up to total $n$. They cannot be other values such as means, rates, or ratios. These two conditions are fairly severe and restrict the use of histograms to a very specific kind of data.

Figure H.7b and c presents two variants of the histogram. The first is a **stem-and-leaf plot** in Figure H.7b, which seems like a horizontal version of a histogram. Our plot shows systolic blood pressure

of a group of people where the first two digits of the values are considered the stem and the third digit the leaf. Since actual values are depicted, recurrence of the same values becomes more evident in this representation, but intervals that are not multiples of 10 or that are unequal are difficult to display in a stem-and-leaf plot. This kind of plot is attributed to John Tukey [1].

When the variable is really discrete with a small number of values (such as parity), the frequency can be represented by vertical lines in place of bars, as shown in Figure H.7c. This is the second variant of the histogram.

1. Ramseyer GC. *John W Tukey Hall*. http://my.ilstu.edu/~gcramsey/TukeyHall.html, last accessed November 26, 2015.

## historical cohort, *see* prospective studies

H

## historical controls, *see* controls

## homogeneity of variances, *see* homoscedasticity

## homoscedasticity

This is the equality of variances of values in different groups. Suppose you measure cholesterol level in different obesity groups: thin, normal, overweight, and obese. If the variances of cholesterol level in these groups are the same, we say that the groups are homoscedastic with respect to cholesterol levels.

If there are $K$ groups in a study, they are homoscedastic if $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_K^2$. You can see that this is applicable to quantitative data only, and it is a requirement for testing the null hypothesis of equality of means across groups. In most practical situations, this can be depicted as in Figure H.8, where only means differ but the variances are the same.

Popular statistical tests such as **Student *t*** (for two groups) and **ANOVA *F*** (for three or more groups) are valid only when the condition of homoscedasticity is fulfilled. They also require a Gaussian pattern of the sample means and independence of values, but that is not an issue here, and we have not shown a Gaussian pattern in Figure H.8. Homoscedasticity is not as easily met as it seems. In our example, generally, the cholesterol level will be higher with increased obesity, and the variances will also increase. As the values increase, the variability also increases, violating the homoscedasticity. It is always advisable to check that this is not the case before using these statistical tests.

There are many ways that homoscedasticity can be checked. Check this graphically by a **box-and-whiskers plot** for different groups. Varying height of the boxes would indicate different variances. Statistically, the variation must be substantial in sample values for violation of homoscedasticity. Generally, the largest variance should be no more than four times the smallest. The conventional statistical test for checking homoscedasticity is the **Bartlett test**.

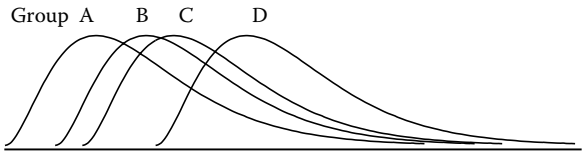



**FIGURE H.8** Distribution in four groups differing only in mean but not variances.

This is heavily dependent on a Gaussian pattern for the distribution of the values. Many software packages now use the **Levene test**. This is based on the median and so is more robust to departure from Gaussianity. **Transformations** of the data, such as logarithm (ln$y$), square ($y^2$), and square root $\left(\sqrt{y}\right)$, are tried in cases where violation occurs. Experience suggests that a transformation can be found that converts a grossly non-Gaussian distribution to an approximately Gaussian pattern and, at the same time, also stabilizes the variance across groups.

If the distribution pattern is already Gaussian and the test reveals that the variances in different groups are significantly different, then the $F$-test should not be used. In fact, there may not be much reason for doing the test for equality of means when the variances are found to be different. It is rare to find that the variance of a variable is different from group to group but the mean is the same. The nature of the variable and measuring techniques should ensure roughly comparable variances. A difference in variances is itself evidence that the populations are different. However, in the rare case in which the interest persists in equality of means despite different variances, try transformation of the data as suggested before. But this should not disturb the Gaussian pattern too much.

## Hosmer–Lemeshow test

This is one of the tests used to check the adequacy of a model (or the lack of it), particularly for logistic regression. The other tests for this purpose are likelihood-based **deviance**, *Nagelkerke* $R^2$, **receiver operating characteristic (ROC) curve**, and bootstrap, in addition to the usual chi-square. Simply stated, the Hosmer–Lemeshow test compares the observed frequencies in arbitrary but rational-looking strata of regressor values ($x$'s) with the expected frequencies in those strata as per the model using a slight modification of the usual chi-square. Because of several regressors in most logistic regressions, these strata are sometimes referred to as profiles. Corresponding to these strata of $x$'s, find the expected frequencies based on the fitted logistic model and compute the following:

$$\text{Hosmer} - \text{Lemeshow test: } \chi^2_{HL} = \sum_{g=1}^{G} \frac{(O_g - E_g)^2}{E_g(1-p_g)} (g = 1,2,\ldots,G),$$

where $O_g$ and $E_g$ are the observed and expected frequencies, respectively, in the $g$th stratum, and $p_g$ is the proportion of subjects in the $g$th stratum. We have stated its simplified version so that its basic features are clear. This chi-square will have 2 degrees of freedom (df's) less than the number of strata. If the $P$-value is less than the prefixed level of significance, such as 0.05, conclude that the model is not consistent with the data. Remember that $P \geq 0.05$ does not mean that the model is good but only that the evidence against it is not sufficient with this amount of data.

The default procedure in many statistical software packages is to divide predicted probabilities from the fitted logistic regression into, say, 10 **decile** groups and compare the observed number of responses in these groups with the predicted number of responses using cells defined by the groups. For 10 groups, this chi-square will have 8 df's. Your software will do all of this for you. The validity condition of chi-square applies, which says that not many expected cell frequencies should be less than 5 and that none should be close to 0. This implies that $n$ should be really large.

Consider a study to explore the predictors of malnutrition in children from a deprived section of society. One hundred malnourished and 100 normal children of age between 1 and 5 years—matched for age within ±6 months and sex—were considered for the study.

**TABLE H.2**
**Contingency table for Hosmer–Lemeshow test**

| | Group = Normal | | Group = Malnutrition | | |
|---|---|---|---|---|---|
| Strata No. | Observed Frequency | Expected Frequency | Observed Frequency | Expected Frequency | Total Observed |
| 1 | 19 | 17.244 | 0 | 1.756 | **19** |
| 2 | 21 | 20.765 | 3 | 3.235 | **24** |
| 3 | 15 | 15.451 | 5 | 4.549 | **20** |
| 4 | 11 | 12.959 | 9 | 7.041 | **20** |
| 5 | 11 | 10.897 | 9 | 9.103 | **20** |
| 6 | 10 | 9.019 | 12 | 12.981 | **22** |
| 7 | 6 | 6.558 | 14 | 13.442 | **20** |
| 8 | 2 | 4.095 | 17 | 14.905 | **19** |
| 9 | 4 | 2.521 | 17 | 18.479 | **21** |
| 10 | 1 | .492 | 14 | 14.508 | **15** |

The potential predictors were decided either from the previous knowledge or significant predictors from univariate analysis. The dependent variable is group (malnutrition = 1, normal = 0), and the predictors are maternal education (illiterate = 1, literate = 0), daily income of parents (low = 1, not low = 0), immunization (no = 1, yes = 0), colostrum (not given = 1, given = 0), breast-feeding till 6 months (no = 1, yes = 0), and mode of feeding (bottle = 1, others = 0). Consider only these six predictors for this exercise.

A statistical software package formed 10 strata of values of the predictors and produced Table H.2 for the observed and expected frequencies in normal and malnourished groups in these strata after running a logistic regression. These start with nearly 20 subjects in each stratum since the total number of subjects is 200. These frequencies give Hosmer–Lemeshow $\chi^2 = 6.001$ with 8 df's and $P = 0.647$. Thus, there is no evidence that the model-expected frequencies differ from the observed frequencies, and the model can be considered to be adequate. Since some frequencies are less than 5, that raises questions about the validity of this test for these data.

The Hosmer–Lemeshow test has been criticized lately for arbitrariness that creeps in while defining the groups and for being insensitive to deviations in individual cases (see, for example, Hosmer et al. [1]). Also, for large samples, this test magnifies the discrepancy between the observed and expected, resulting in an unnecessarily small *P*-value.

For more in-depth exposition, see Hosmer et al. [2] and Allison [3].


1. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med* 1997;16:965–80. http://www2.stat.duke.edu/~zo2/dropbox/goflogistic.pdf
2. Hosmer DW Jr, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*, Third Edition. Wiley, 2013.
3. Allison P. *Logistic Regression Analysis Using SAS: Theory and Applications*, Second Edition. SAS Institute, 2012.


## hospital statistics

Because of digitization almost all around the world, hospitals generate a large volume of data on what kind of and how many patients are served, and with what outcomes. These are the hospital statistics and can be used to serve a variety of purposes.

### *Utilization of Hospital Services*

Hospital statistics are conventionally used to find ways and means to maximize utilization of hospital services. Basic information desired from these data is the following:

$$\textit{Bed - occupancy rate}\ (\%): \frac{\text{bed-days of occupancy}}{\text{total bed-days}} \times 100.$$

This can be calculated for any week, any month, or any year. If a hospital has 300 beds, the total bed-days in a year are 300 × 365 = 109,500, and if the occupancy is 66,364 bed-days, the occupancy is 66,364 × 100/109,500 = 60.6%. Bed-days of occupancy is itself evident from the load of inpatients. Some hospitals with fast turnover may have 90% occupancy, whereas others, particularly those at the periphery, may have just about 30%. Some busy pediatric hospitals in developing countries, particularly in the public sector, may have more than 100% occupancy if one bed is occupied by two patients at the same time for some days and hardly any bed is ever vacant. This is a reality in some hospitals in developing countries.

*Average length of stay* per patient (days):

$$\frac{\text{patient-days in hospital}}{\text{total number of admitted patients}}$$

Separation of a patient can be due to discharges or death. This kind of indicator may be extremely useful for the casualty department of a hospital, where patients are supposed to move fast to regular care. Such an indicator should be calculated separately for each department since in some departments, such as those dealing with cancer cases, the length of stay could be substantially longer, and for others (say, hernia surgery), it may be short.

### *Quality of Care*

Quality should be assessed against a prefixed standard, but in the case of hospitals the objective is to achieve 100% excellence. This is generally negatively assessed in terms of errors:

$$\text{Average medication errors per patient-day: } \frac{\textit{medication errors}}{\text{patient-days}}.$$

This can be calculated for any month, any year, etc. Similar indicators can be developed for transfusion infections and surgical site infections, maybe even incidence of bedsores, ventilator-associated pneumonias, and needlestick injuries. Correct assessment on the basis of such indicators depends mostly on honest reporting, and that is suspected in many situations. They require more sincerity on the part of the hospital administration or proper enforcement by regulatory agencies such as accreditation boards.

Most valid indicators of quality of care in hospitals are based on mortality. These indicators can be classified into those related to inpatient procedures, such as esophageal and pancreatic resections, craniotomy, and hip replacement, and those based on inpatient conditions, such as acute myocardial infarctions, acute stroke, and gastrointestinal hemorrhage. These rates are generally calculated per 1000 separations (discharges and deaths) of patients undergoing these procedures or suffering from these conditions. Although these rates would depend on the severity of cases that a hospital gets (some hospitals tend to receive more patients in critical conditions because of their good reputation), they may be used for comparing areas, hospitals, and time, to get an idea of how good a hospital is relative to its previous performance. If the case mix is not substantially

different, these indicators can also be used for assessing how good hospitals are relative to the hospitals in other areas. These indicators, however, do not consider the cost aspect at all.

The Agency for Healthcare and Quality works intensively in this area, and they provide a tool kit [1] for assessing quality of services in hospitals.

### *Research Based on Hospital Statistics*

Two kinds of research can be done on the basis of routinely generated data in hospitals. First is the analytical research that tries to link the outcome with the condition of the patient at admission as mediated by hospital care and patient cooperation. This requires that the records be immaculately maintained and complete for each patient. This kind of research also includes an assessment of how the final diagnosis matches with the initial diagnosis or how it matches with the investigation results. This can be done for each consultant, each department, etc. to build up a profile. Second is developing norms based on the measurements of those patients who do not have any abnormality. Many hospitals run preventive health checkup programs for corporate employees or for insurance companies, and investigate thousands of subjects each year. Many of these people are absolutely healthy. The levels of parameters such as biochemical measurements, heart measurements as revealed by echocardiograms, and bone mineral density can be used to establish local norms in place of depending on the international norms.

1. AHRQ. *Quality Indicator Toolkit for Hospitals: Fact Sheet*. http://www.ahrq.gov/research/findings/factsheets/quality/qifactsheet/index.html

## Hotelling $T^2$

This is the **multivariate** analogue of the **Student *t*-test** and can be used to test whether the means of a particular *set* of variables have predefined values or, more generally, whether the means in two groups for a set of variables are the same. The test was first proposed by Harold Hotelling in 1931 [1].


Harold Hotelling

Consider kidney functions (such as creatinine clearance, urea clearance, and diotrast or *p*-aminohippurate clearance) in healthy females of age 50–59 years, where the interest is in finding whether the means seen in vegetarians are the same as in nonvegetarians. The only difference between this setup and that of the Student *t* is that now we want to consider three or four kidney functions together in a multivariate setup in place of considering them one at a time. A multivariate setup allows consideration of the correlations among the variables that a univariate setup ignores.

As in the case of the Student *t*, Hotelling $T^2$ is valid under certain conditions. These are as follows: (i) either the sample size is large or the distribution of the values is nearly Gaussian—in this case, **multivariate Gaussian**; (ii) the **dispersion matrices** of the variables in the two groups are nearly equal (this corresponds to equal variances in the univariate setup)—this condition is especially important if the sample sizes in the two groups are unequal; and (iii) the values across individuals are independent—that is, the values in one person does not affect the values in the other person (for example, the individuals do not belong to the same family). Equality of dispersion matrices is tested by the **Box *M*** test. This test gives correct results only when the underlying distribution is multivariate Gaussian. Alternatively, the multivariate analog of the **Leven test** can be used, which is robust to non-Gaussianity.

Realize that Hotelling $T^2$ is a special case of the multivariate analysis of variance (MANOVA) test. In other words, MANOVA for two groups gives the same result as Hotelling $T^2$, just as ANOVA gives the same result as Student *t* for two groups. The underlying formula of Hotelling $T^2$ requires an understanding of matrices and determinants, which we are avoiding in this book for medical professionals. Most statistical packages have a provision to do MANOVA and, hence, Hotelling $T^2$. If you are interested in its mathematics, see Anderson [2].

1. Hotelling H. The generalization of Student's ratio. *Ann Math Stat* 1931;2(3):360–78. http://projecteuclid.org/download/pdf_1/euclid.aoms/1177732979
2. Anderson TW. *An Introduction to Multivariate Statistical Analysis*, Third Edition. Wiley, 2003.

## human development index

Human development is considered to be an important aspect of well-being and thus of health. The United Nation's Development Programme (UNDP) each year computes a human development index (HDI) for each country by using a common base so that it can be realistically compared across countries, sometimes even across times. This is calculated by combining three components of development, viz., education, income, and life expectancy. The methodology is periodically reviewed and revised, and the one available in their 2014 report [1] requires the following:

- Per capita income in real terms by converting it to **purchasing power parity** (PPP) dollars. Complex calculations are done for this conversion. This is needed to restore parity in income in different countries with different price levels.
- **Life expectancy** at birth (ELB).
- Mean years of schooling.
- Expected years of schooling. This is the number of years of schooling that a child of school entrance age can expect to receive if prevailing patterns of age-specific enrolment rates were to stay the same throughout the child's life.

Each of these four components is converted to an index relative to the observed minimum and the observed maximum value. These are known as goal posts. For per capita income, the minimum estimated by UNDP is PPP $100, the maximum is PPP $75,000, and the logarithm is taken. ELB is used as a surrogate for health. For this, these goal posts are a minimum of 20 years and a maximum of 85 years. For mean years of schooling and expected years of schooling, these are from 0 to15 years and from 0 to 18 years, respectively. Education index is the average of the indices for mean years of schooling and expected years of schooling. The HDI is the **geometric mean** of the

income index, life expectancy (health) index, and education index. The following example is taken from the Human Development Report 2014, and illustrates the calculations for HDI of Costa Rica for the year 2011.

Indicator values for Costa Rica:

ELB (years), 79.93; mean years of schooling, 8.37; expected years of schooling 13.50; gross national income per capita (PPP 2011 dollars), 13,011.7

Health index = $(79.93 - 20)/(85 - 20) = 0.922$

Mean years of schooling index = $(8.37 - 0)/(15 - 0) = 0.558$

Expected years of schooling index = $13.50/18 = 0.750$

Education index = $(0.558 + 0.750)/2 = 0.654$

Income index = $[\ln(13{,}011.7) - \ln(100)]/[\ln(75{,}000) - \ln(100)] = 0.735$

HDI = $(0.922 \times 0.654 \times 0.735)^{1/3} = 0.763$

Such values are calculated for each country each year, and the countries are ranked from maximum to minimum. In absolute value, an index of 0.5 is sometimes considered the minimum for reasonable development.

Indrayan et al. [2] proposed that the index of components of HDI henceforth be computed on the basis of the percentage of the population that reached a minimum threshold. For income, this threshold could be the minimum required for adequate food, housing, and clothing (say PPP $10,000 per capita per year); for ELB, the percentage of persons of age 60 years and above; and for education, the percentage of adults (18 years and above) with at least high school education. This modification will make the index and its components readily interpretable as it would indicate how well the country has been able to meet the minimum level.

1. UNDP. *Human Development Report 2014—Sustaining Human Progress: Reducing Vulnerabilities and Building Resilience*. http://hdr.undp.org/en/content/human-development-report-2014
2. Indrayan A, Wysocki MJ, Chawla A, Kumar R, Singh N. 3-decade trend in human development index in India and its major states. *Social Indicators Res* 1999;96:91–120. http://link.springer.com/article/10.1023%2FA%3A1006875829698#page-1

## Huynh–Feldt correction, *see also* repeated measures ANOVA

This is used for correcting both the degrees of freedom (df's) of the *F*-test for univariate **repeated measures ANOVA** in the case where the condition of **sphericity** of the dispersion matrices at different points in time is violated [1]. The sphericity condition is that the differences between all pairs of times have the same variance. Violation of sphericity is detected by the **Mauchly test**. Remember that the ability of all statistical tests to show significance depends on the sample size—thus, nonsignificance can arise due to inadequate sample size. This can happen with the Mauchly test as well. If this test is statistically significant, the Huynh–Feldt correction is applied to the df's of the *F*-test, although the value of *F* remains the same. This correction will reduce the df's by a factor called epsilon, making it difficult to reject the null hypothesis. The greater the departure from sphericity, the lower the epsilon. If the df's for your *F*-test are (5, 25) and epsilon = 0.4, then the corrected df's are (2.0, 10.0). Note that lower value of epsilon means a bigger correction to the df's. Rarely, the Huynh–Feldt correction can exceed the value of 1, but if it does, 1 is used, and no correction to df's is done.

Many statistical packages provide results with the Huynh–Feldt correction and also the corresponding *P*-value for repeated measures ANOVA. This is also known simply as Huynh correction. You may find other types of epsilon also (such as Greenhouse–Geisser) in your software output, but the Huynh–Feldt epsilon is widely accepted.

For a more in-depth exposition of the Huynh–Feldt correction, see Tamhane [2].

1. Huynh H, Feldt LS. Estimation of Box correction for degrees of freedom from sample data in randomized block and split plot designs. *J Educ Stat* 1976;1(1):69–82. http://www.jstor.org/discover/10.2307/1164736?uid=3738256&uid=2&uid=4&sid=21105254571903
2. Tamhane AC. *Statistical Analysis of Designed Experiments: Theory and Applications*. Wiley, 2012.

## hyperpopulation



This term is used in two senses in biostatistics: first for an imaginary population from which a sample can be assumed to have come and second for the population generated through **simulations** or by **resampling**.

All statistical methods of inference such as **confidence interval** and **test of significance** require a random sample of subjects. Consider a study of homosexual men attending clinics in two cities. Neither group can be considered to be truly random. Yet confidence intervals and ***P*-values** for statistical significance of the difference between the two groups are calculated, such as done by Koblin et al. [1]. These *P*-values are valid only when the cohorts are considered a random sample from what is called a hyperpopulation of homosexuals in the two cities. This hyperpopulation exists in concept but not in reality, and may possibly include the past and, more importantly, cases arising in the immediate future.

Patients coming to a clinic during a particular period can be considered a random sample from the population of patients that are currently coming to that clinic. But this limited definition of population is sometimes forgotten, and generalized conclusions are drawn on the basis of *P*-values. This practice is quite frequent in medical literature and is mostly accepted without question with the underlying assumption that this generalization is for the hyperpopulation.

The second use of the term is either when you do simulations to generate a population of values of the type needed for inference, or when repeated sampling is done with replacement from the available sample. In the first case, with high-speed computers at your disposal, you can generate millions of values that follow, for example, a Gaussian distribution with a specified mean and standard deviation. These values are not real—thus the term *hyperpopulation*. In the second case, if a sample of reasonable size is available, you can generate innumerable samples of any size by replacement. This is called **resampling** and has strategies such as **bootstrap** and **jackknife**. This technique allows generation of a hyperpopulation of samples and has been found to be effective in drawing valid conclusions in cases where the **sampling distributions** are intractable.

1. Koblin BA, Hessol NA, Zauber AG, Taylor PE, Buchbinder SP, Katz MH, Stevens CE. Increased incidence of cancer among homosexual men, New York City and San Francisco, 1978–1990. *Am J Epidemiol* 1996;144:916–23. http://aje.oxfordjournals.org/content/144/10/916.full.pdf

## hypothesis (null and alternative), *see* null and alternative hypotheses

## hypothesis (research)

A research hypothesis is a precise expression of the expected results regarding the state of a phenomenon in the target population. Realize that research is about replacing existing "hypotheses" with new ones that are more plausible. In medical research, a hypothesis could purport to explain the etiology of diseases, preventive strategies, screening and diagnostic modalities, distribution of occurrence in different segments of the population, the strategies to treat or mange a disease, methods to prevent recurrence or adverse sequel, etc. Consider which type of hypothesis can be investigated by the proposed research.

Hypotheses are not guesses but reflect the depth of knowledge of the topic of research. They must be stated in a manner that can be tested by collecting evidence. For example, the hypothesis that dietary pattern affects the occurrence of cancer is not testable unless the specifics of diet and the type of cancer are specified. Antecedent and outcome variables, or other relevant variables, should be exactly specified in a hypothesis. Generally, a separate hypothesis for each major expected relationship is generated.

The hypotheses must correspond to the general and specific objectives of the study. Thus, carefully examine each objective, and assess which of these generate a new hypothesis. Whereas objectives define the key variables of interest, hypotheses are a guide to the strategies to analyze the data.

Beware that, just like objectives, hypotheses, too, are governed by the current knowledge. For example, nobody wants to know how music aptitude affects cancer risk, although a future endeavor may reveal an association between the two. Researchers are already examining the hypothesis that infections may trigger some cardiac ailments. The hypothesis that childhood nutrition, perhaps even during pregnancy, can manifest in chronic diseases 40–50 years later surprised many when it was initially proposed, but it is now accepted as not only plausible but also likely and real.

A research hypothesis can be very different from statistical null and alternative hypotheses. Statistical hypotheses are grounded to real values as seen in empirical research, whereas a research hypothesis can be philosophical or abstract.

## hypothesis testing (statistical), *see also* null and alternative hypotheses

Hypothesis testing in statistics is a general term for assessing whether sample data are consistent (or otherwise) with statements made about the **population** of interest at the time of planning the study. Hypothesis testing is an approach used for choosing between two statistical **hypotheses**, namely, **null and alternative**, regarding the population. These are regarding the values of the parameters of a distribution of the variable of interest or regarding the shape of its distribution.

Over the years, statistical testing of hypotheses has become important in the development of ideas emanating from empirical research. It hails from the ideas put forward by Neyman and Pearson [1] and further developed by Lehmann and Romano [2].

Let the competing hypotheses be $H_0$ (null hypothesis) and $H_1$ (alternative hypothesis), and let a summary of the data (or statistic) denoted by $T$ be considered appropriate to test one hypothesis against the other. Chi-square, Student $t$, and ANOVA $F$ are examples of such summaries. The value of $T$ is calculated on the basis of the sample values under the assumption that the null hypothesis is true. The statistical **distribution** of $T$ should also be known, particularly under the null hypothesis, so that we can find how likely or unlikely the value obtained with these calculations is. This would indicate whether or not the sample values are in conformity with the null. To do hypothesis testing, the conventional procedure is to first define a critical region (see **acceptance and rejection regions**) in advance of collecting data and then reject $H_0$ in favor of $H_1$ depending on whether the calculated value of $T$ falls within the critical region or not. The procedure now used is not based on the critical region but just to find the probability that the value of $T$ is as much as obtained or more extreme. This is what is called the ***P*-value**. This is the crucial value in the testing of hypotheses. The lower the $P$-value, the greater the inconsistency of the sample data with the null. A sufficiently low value of $P$ is an indication that we reject the null in favor of the alternative. Although this procedure can be understood as providing graded evidence in terms of exact $P$-value, a cut-off is needed to decide when to reject the null. Generally, a $P$-value less than 0.05 is considered sufficiently small to reject the null. This threshold is called the **level of significance**.

As explained under the topic **tests of hypothesis (philosophy of)**, statistical testing never allows acceptance of a null hypothesis. We either reject or do not reject the null. The following example may help you to understand the steps and the philosophy behind the statistical tests of hypotheses.

Suppose an argument erupts concerning the percentage of births in a community that are premature. Based on everyday experience, a practitioner asserts that 10% of births are premature, neither less nor more. To test this assertion, a random sample of 60 births is systematically observed, and 8 of them are found to be premature. This is 13.3%. Can we conclude that the percentage of premature births in the population is not 10%? Consider the following steps:

1. Assertion of 10% premature births in this case is the null hypothesis. This is what is to be tested (and possibly to be refuted). Thus, $H_0$: $\pi = 0.10$, where $\pi$ is the probability of premature birth in this segment of the population. Since there is no assertion about a particular direction of difference, it could be negative or positive. That is, $H_1$: $\pi \neq 0.10$. Let the level of significance be fixed at $\alpha = 0.05$. That is, the chance of Type I error should be less than 5%.
2. The quantity of consequence in this case naturally is the proportion, $p$, actually observed in the sample. For this sample of 60 births, $p = 8/60 = 0.1333$. Since $n$ is 60 and $np \geq 8$, we can invoke the **central limit theorem** and assume that $p$ will approximately have a Gaussian distribution. The standard Gaussian deviate in this case is

$$z = \frac{p - \pi}{\mathrm{SE}(p)}.$$

   This can be used as a test criterion. When $H_0$: $\pi = 0.10$ is true, then from the usual formula of the standard error (SE) is $\mathrm{SE}(p) = \sqrt{0.10 \times 0.90/60} = 0.0387$. Thus, for this example, under $H_0$, $z = (0.1333 - 0.10)/0.0387 = 0.86$.
3. The $P$-value is the probability of obtaining this value of $z$ or a value more extreme toward $H_1$. Since $H_1$ is two-sided,

$$P\text{-value} = P(Z \leq -0.86) + P(Z \geq 0.86) = 0.1949 + 0.1949 = 0.39$$

   from the Gaussian distribution. This $P$-value is certainly very high in comparison with the conventional threshold

0.05. If $H_0$ is rejected, the chances of its being wrongly rejected are as much as 0.39 (or 39%). This is too high an error. Since $P \geq 0.05$, $H_0$ is plausible, and the assumption $\pi = 0.10$ cannot be rejected. The sample does not provide sufficient evidence against $H_0$.

The difficulty is that if the null is $\pi = 0.11$, that also will not be rejected by this set of data, and the same is true for, say, $\pi = 0.08$. This is true for many other values of $\pi$. If we accept, which value of $\pi$ do we accept? For this reason, statistical tests never provide evidence in favor of the null—just enough evidence against the null or nothing at all. When a null is not rejected, not much of significance is added to the current knowledge except that the null is not implausible.

The values in this example can provide sufficient evidence to reject some other values of $\pi$. For the purpose of illustration, change $H_0$ to $\pi = 0.25$. Under this $H_0$,

$$z = \frac{0.1333 - 0.25}{\sqrt{0.25 \times 0.75/60}} = -2.08.$$

Now $P = P(Z \leq -2.08) + P(Z \geq 2.08) = 0.0188 + 0.0188 = 0.038$. This is less than 0.05. Thus, this $H_0$ is not plausible and is rejected. It is concluded that the percentage of premature births is not 25. The percentage (13.33%) observed in the sample is sufficiently different from 25 in a statistical sense but not sufficiently different from 10. Thus, the $H_0$ that the premature births are 25% can be rejected but not the $H_0$ that they are 10%.

The size of the *P*-value comes from two things: the size of the estimated treatment difference and its estimated intersample variability (which derives partially from the sample size). Thus, the *P*-value partly reflects the size of the study, which has no biological importance, and partly the size of the effect and sampling variability. An extremely low *P*-value can arise even when the effect size is extremely small, perhaps negligible, in situations where the sample size is inordinately large. Thus, many consider the **confidence interval** a much better procedure than the testing of hypothesis for valid inference.

We also need to consider the types of error that can arise using hypothesis testing. **Type I error** occurs if there is no treatment effect or difference but our data happen to wrongly conclude that there is. **Type II error** occurs when the data fail to detect a treatment effect or difference that is actually present. Its complimentary is the **power** of the test that measures the chance of declaring a treatment effect or difference of a given size to be statistically significantly different from the value under the null hypothesis.

1. Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. *Phil Trans Roy Soc A*.1933;231:289–337. http://www.stats.org.uk/statistical-inference/NeymanPearson1933.pdf
2. Lehmann EL, Romano JP. *Testing Statistical Hypotheses*, Third Edition. Springer, 2008.

# I

## $I^2$ (index of heterogeneity) in meta-analysis, *see also* meta-analysis

$I^2$ is an index of heterogeneity of the **effect sizes** in different studies compiled for **meta-analysis**. Effect size could be the difference in means or in proportions, odds ratio, relative risk, or any other measurement. Meta-analysis is the term used for the method of combining evidence from various studies on the same effect. One of the requisites for studies compiled for this purpose is that they should not have widely different results from one another. Some variation is natural, and that is why we need meta-analysis, but that should be within the range generally perceived as natural across studies in different areas and in different populations with different sample sizes, and possibly with some minor variation in methodology. It is preferred that the studies chosen for meta-analysis have similar methodology, but that might be difficult to ensure.

$I^2$ is the percentage of total variance attributed to the between-study variation. Generally, this kind of contribution is obtained by the method of analysis of variance (ANOVA), but that can be done only when the original individual values are available. In the case of meta-analysis, only the effect size is available for the study groups as a whole, and raw data are rarely available. Thus, we need another measure that takes the form of $I^2$ as follows.

The first step in this is to calculate **Cochran** $Q = \Sigma_k w_k (T_k - \overline{T})^2$, where $T_k$ is the effect size found in the $k$th ($k$ = 1, 2, …, $K$) study and $\overline{T}$ is the weighted average of these effect sizes. This is given by $\overline{T} = \Sigma_k w_k T_k / \Sigma_k w_k$. $w_k$ is the weighting factor of the $k$th effect size: this could be any measure you choose, but the most common choice is the inverse of the variance of the $k$th effect size. The actual variance would almost never be known and is estimated by the square of the estimated standard error (SE) of the effect size that most studies report. Cochran $Q$ follows a **chi-square distribution** with ($K$ − 1) degrees of freedom (df's) provided $T_k$s have approximately **Gaussian** (Normal) **distribution**. $Q$ itself can be used as an index of heterogeneity—the higher the value of $Q$, the higher the heterogeneity. The null hypothesis of equality of effect sizes can be rejected if the chi-square-based ***P*-value** for the value of $Q$ is less than the predetermined **level of significance**, such as 0.05. But the difficulty with $Q$ is that this can become large just because the number of studies is large. Thus, it would unnecessarily reject the null even if there is homogeneity in effect sizes in a large number of studies. More importantly, $Q$ depends on the metric of the effect sizes.

The next step is to make $Q$ independent of $K$ and the metric of effect size as much as possible. For this, Higgins and Thompson in 2002 [1] proposed several measures, but the following has come to be widely accepted as the index of heterogeneity:

$$\text{index of heterogeneity: } I^2 = \left(1 - \frac{K-1}{Q}\right) * 100\%.$$

This requires $Q > (K - 1)$; luckily, this generally is the case. If that is not so, the value of $Q$ is considered equal to 0—it cannot be negative. A value 25% of this index is generally considered low, 50% moderate, and 75% high. If this index is high for the effect sizes in the studies selected for meta-analysis, the studies that are causing this high value are excluded from the analysis if not important. However, just one or two studies with very large $n$ can also inflate the value of $I^2$ because they get a large weight due to their small SE, and these should not be excluded.

For more details, including the confidence interval for $I^2$, see Huedo-Medina et al. [2]. Many statistical software packages that have a capability for meta-analysis have incorporated this index in their output. For a real-life example, see the work of Viguiliouk et al. [3], who found $I^2$ = 37% in 12 studies they included in a meta-analysis studying the effect of tree nuts on glycemic control in diabetes and concluded that the difference between a nuts diet and nonnuts diet was not statistically significant, though the direction was in favor of tree nuts.


1. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002 Jun 15;21(11):1539–58. http://www.ncbi.nlm.nih.gov/pubmed/12111919
2. Tania Huedo-Medina T, Sanchez-Meca J, Marin-Martinez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *CHIP Documents* University of Connecticut Paper 19, 2006. http://digitalcommons.uconn.edu/cgi/viewcontent.cgi?article=1019&context=chip_docs, last accessed August 2, 2015.
3. Viguiliouk E, Kendall CWC, Blanco Mejia S et al. Effect of tree nuts on glycemic control in diabetes: A systematic review and meta-analysis of randomized controlled dietary trials. *PLoS ONE* 2014;9(7): e103376. http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0103376


## ICD, *see* International Classification of Diseases

## icicle plots

Icicle plots depict the process of **hierarchical clustering** of values, where the most similar ones are clustered together in the first stage, the next most similar in the second stage, and so on, till all units are plotted to form one big cluster. What we have stated is the **agglomerative algorithm**, but we can have a divisive algorithm also, which starts by considering all units together as one big cluster and divided on the basis of some metric of distance. Icicle plots can be used for divisive algorithms also, although practical use is commonly for agglomerative algorithms.

For an example, see Figure I.1a, which is an icicle plot for hierarchical agglomerative clustering of nine subjects with respect to their values of urea, uric acid, and creatinine levels. This resembles a row of icicles hanging from a windowsill. The plot shows, for example, that case nos. 2 and 7 had the most similar values and were merged in the first stage. This merging is shown in the bottom row of the figure. One row up shows the merging of case nos. 4 and 6, and so on. Thus, the entire agglomerative process is shown. In the last stage, all merge together in the top row. The method we followed for this clustering is **average linkage**, but any other method can be used

For contrast, we have also shown the corresponding **dendrogram** in Figure I.1b. This is not as good in showing the process but is

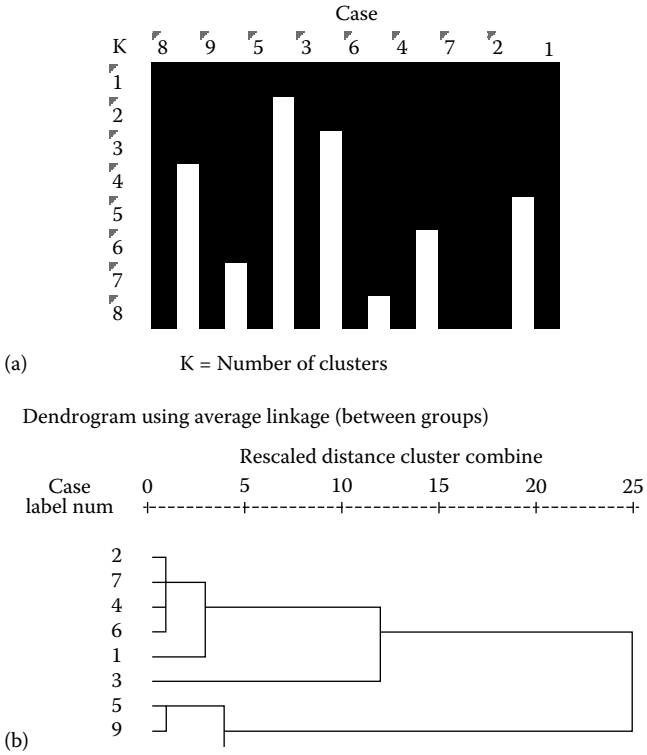



**FIGURE I.1** (a) Icicle plot and (b) dendrogram for hierarchical clustering of nine cases with respect to kidney functions.

good in showing the distances between entities merged at different stages. This can be used to decide the optimal number of clusters. The icicle plot does not have this property. In our example, the distances (depicted by the horizontal length of the lines) are not big for three clusters, but when we reach the two-cluster stage, the distance shoots up, and it becomes huge for one cluster, indicating that three clusters are appropriate for these data.

## imprecise probability, *see* probability

## imputation for missing values, *see also* nonresponse

Imputation is the process by which the missing values in a set of data are sought to be replaced by plausible values. Because of various problems as discussed for **missing data**, these are quite common in medical setups. Excluding them altogether from analysis may yield a biased picture since experience suggests that missing values generally are not random but are special in the sense of being either too high or too low or having any such aberration. Perhaps no imputation is needed when the sample is large and the missing values are less than 5% because in this case, the available 95% of values can lead you to the right conclusion, or at least one as good as with imputations. If there are more missing values, they can be imputed by various statistical methods, as described later in this section, provided the their structure is known or can be modeled.

If a higher attrition is anticipated, the usual practice is to inflate the sample size accordingly so that enough data are available for analysis. Despite enough data available at the end, missing data can still cause bias in the sense that specific types of values are missing and the ones left for analysis fail to reflect the full spectrum of values.

Imputation helps in (i) providing the comfort of working with complete data; (ii) using predetermined statistical methods of analysis that otherwise may require substantial alteration to take care of the **unbalanced** data due to missing values; and (iii) computationally recovering the complete record of each subject—otherwise, the cost of collecting data can be prohibitive in some situations. But it is not without problems since imputed values are artificial and not real—thus, they may lead to concocted results. Because of this, imputation should not be done if, say, more than 10% of values are missing, and the results from the available data should be acknowledged as possibly biased. When imputations are limited, they are not likely to impact the results much.

In an experiment or a trial, missing values are imputed to be the most adverse to the regimen under trial so that a positive conclusion about its efficacy and safety does not breach the conventional 5% probability of Type I error. For example, the cases that could not be evaluated for the outcome are counted among the failures. For this, **intention-to-treat** analysis is advised. However, if 10% of values are missing, this failure rate itself may be too high in some setups, and imputation may become necessary if all missing data cannot be considered as failures.

For data missing apparently at random, the imputation methods are as follows. The first is the **regression** method, which requires that the available values be used to find a regression equation that can predict the missing values. Not only the available values of the variable in question are used, but other variables such as age and sex for which relatively complete data are available are also incorporated in this equation. Baseline markers of disease and other pretreatment and posttreatment measures that can affect missingness could also be useful. One can also think of treatment compliance and tolerability of the regimen as predictors of missingness. A prerequisite for this is that the regression must be statistically adequate with, say, a **coefficient of determination** of more than 90%. When this is not feasible, the second method is to replace the missing value with the average of the available values. This may not work well, as it takes away the chance component and unnecessarily tends to reduce the variance and produce false statistical **significance**. The third is the closest match method, which requires examination of the data of all subjects to find the one that is closest in all other reports and substitute this subject's value for the missing value. Among other methods are (i) replacement of a missing value by generating a random value from the distribution of the variable when known and (ii) adding a random residual to the respective mean values. Both these methods can work only when the missing values are entirely random. Yet another method is to determine classes of similar subjects and replace missing values with a random value from the same class.

Another popular method is to replace the missing value by carrying forward the last value. This is done in **longitudinal studies** on the assumption that no change occurred during the interregnum. This may work well in a **clinical trial** setup, where it is known that a subsequent value will be the same or will have improved because of the regimen, and will not be more adverse. This is in accordance with the principle that missing values should be replaced by pessimistic values. If you are concerned about missing values in clinical trials, consult the article by Powney et al. [1] that has discussed various methods used in a sample of 100 trials.

It is easy to imagine that imputing a missing value with its estimate by any method fails to account for uncertainty about the missing value. Analysis based on such single imputations, as they are called, treats this imputed value as though this is the actual observed value. To account for uncertainty about the actual missing value, sometimes, one missing value is replaced by several plausible values. It is like a random sample of missing values and is called **multiple**

**imputations**. Such multiple-imputed data sets are then analyzed as usual. The process mostly results in relatively more valid inferences. See Carpenter and Kenward [2] for details.

Adjustments done by imputation or otherwise certainly reduce the validity of the data, but that is the best one can do when the missed values cannot be ignored. Remember that imputation is an exercise in salvaging the damage, which helps but does not rectify. The only better course is to do everything possible to obtain complete data.

1. Powney M, Williamson P, Kirkham J, Kolamunnage-Dona R. A review of the handling of missing longitudinal outcome data in clinical trials. *Trials* 2014 Jun 19;15(1):237. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4087243/
2. Carpenter J, Kenward M. *Multiple Imputation and its Application*. Wiley, 2013.

## inception cohort, *see* cohort studies

## incidence

Incidence is the number of conditions or the number of persons having an onset of a condition in a specified period. A condition is considered to have had its onset when it is first noticed. This could be the time when the person first felt sick or injured, or it could be the time when the person (or the family) was first told of a condition that was previously unknown. Thus, a person who has had diabetes for a long time is considered to have onset at the time when diabetes was first detected. Sometimes, the time of onset depends on whether or not the induction period (from exposure to subclinical) and latent period (from subclinical to clinical) are considered.

The incidence rate is an important indicator of burden of disease in a community. It can be calculated per cent, per thousand, per million, etc., as convenient, of exposed subjects, and it measures the risk of developing morbidity in a randomly chosen subject from that group. Incidence rate may increase without a real rise in new cases when the reporting or case detection improves. A real rise indicates that the existing strategy to control the disease has not succeeded and alternative or improved strategy is needed. Analysis of differences in incidence in various socioeconomic, biological, and geographical groups may provide useful tips to devise a better strategy for control of the disease. Also, note the following:

- Incidence rate is necessarily associated with duration. If the period is 1 year, the rate obtained will also be for that year, and 1-year incidence would be different from 1-month incidence, although not necessarily one-twelfth of the annual incidence. Notice that incidence can be obtained only through a **follow-up study**.
- Incidence rate can also be calculated for spells or episodes instead of the affected persons. This would count a person two or more times if the same person has repeated spells within the reference period, such as for diarrhea and angina. The term *incidence density* is preferred in this case.
- Incidence reflects causal factors and is better interpreted as **risk**. This is used to formulate and test hypotheses on the etiology of the disease.
- *Incidence rate ratio* can also be obtained for two exclusive groups. If the incidence of liver diseases in alcohol users is 5 per 1000 per year and in nonusers 0.5 per 1000 per year, the incidence rate ratio is 10. However, this has a better and more popular name, **relative risk**.

Just like **prevalence**, precise incidence is central to the epidemiology of any disease, but it is very difficult to obtain. Full enumeration is typically too expensive, very labor intensive, and almost impossible to achieve for the monitoring of disease frequency. Thus, an estimate is obtained by studying a sample. A critical component in disease monitoring is the degree of undercount. One way to improve the estimate of the rates is by correcting them for the level of ascertainment. Comparison of different populations, communities, or countries is accurate when the incidence estimates from a surveillance system or some other source are adjusted for the degree of underascertainment.

Prevalence is greatly affected by the duration of disease—the higher the duration, the higher the prevalence, since the cases tend to accumulate. This is not so for incidence, since incidence is absolute and depends on the risk factors, susceptibility, demographic factors such as age and sex, etc., but not on the prevalence or duration of disease. For this reason, many times, incidence is considered a better indicator of morbidity than prevalence. For example, the definition of an epidemic is based on incidence and not prevalence. An increase in incidence raises an alarm for initiating new control measures, whereas prevalence indicates how much medical infrastructure is needed to handle the cases.

A uniform follow-up is not necessary to calculate incidence rate, although it is required in any case. In case some subjects are followed up for, say, 3 years, some for 4 years, some for 2 years, etc., we can use the concept of **person-time** for calculating the incidence. For person-years, this would be given by

$$\text{incidence rate per 100 person-years} = \frac{\text{new cases occuring in the observed period}}{\text{person-years observed}} * 100.$$

However, the incidence must be evenly distributed over different periods for this to be valid. This does not apply for diseases arising from something like smoking, since the initial years of smoking may be more harmful than later years, when the damage is already done. The multiplier in the formula given in this equation is not necessarily 100, and it can be chosen as per convenience. Also, any other time unit such as person-weeks can also be used.

### *Incidence from Prevalence*

A follow-up study is relatively more expensive than a **cross-sectional study**. Thus, it is easier to obtain the prevalence than the incidence. The duration of sickness can also be generally easily guessed from experience. Because prevalence depends on the incidence and duration, the relationship can be exploited to find the incidence based on prevalence and duration. In a steady state, *if there are no intervening factors*,

$$\text{incidence} = \frac{\text{prevalence}}{\text{average duration of sickness}},$$

where incidence and duration are in the same time unit. If the annual incidence is to be calculated, then the duration too should be measured in terms of years. If the average duration of sickness for a particular disease is 15 days, then this is 15/365 = 0.041 years. If the prevalence rate is 1.2 per 1000 and the average duration is 15 days, the incidence rate is 1.2/0.041 = 29 per 1000 per year under a steady state. If the prevalence rate is per 1000 persons, then incidence too is a rate per 1000 persons. Incidence can be calculated for specific groups, by age, gender, occupation, and region, by inserting the prevalence and duration for the chosen group.

The relationship between incidence and prevalence can be visualized as a reservoir where water coming in by way of incidence builds up a level (prevalence) that remains constant if the outflow due to deaths and cures together is the same as the inflow (see Figure P.12 under the topic **prevalence and prevalence rates**). If deaths are reduced by health measures or when a new treatment becomes available that cures faster, the prevalence will decline provided that the incidence remains the same. Interestingly, if the treatment is such that it prevents death but does not cure, as happened with antiretroviral therapy for HIV, the prevalence would increase rather than decrease. The net effect of such a treatment is prolonged disease (increase in duration of disease). If incidence increases and cure rate and deaths together remains the same, the level will rise.

The concept of duration of sickness is generally applicable to all acute conditions but not so much to chronic conditions. Some not-so-severe conditions such as hypertension, varicose veins, and lower vision are rarely fully reversed in a manner that would allow leading a normal life without ongoing treatment. For such conditions, the duration of disease is anybody's guess. If the condition is more concentrated in the elderly, such as cataracts, mortality affects the prevalence. Mortality itself may be higher in the group with disease than in the nondiseased group. Such conditions interfere with the parameters of this equation, and more elaborate calculations may be required to estimate incidence from the prevalence of such conditions.

## inclusion and exclusion criteria

The concept of inclusion and exclusion criteria is generally restricted to **clinical trials** for deciding the eligibility of subjects to be in the trial. Actual participation depends on whether (i) the eligible subject provides consent and (ii) the subject is in the randomly selected sample from the eligible subjects.

Inclusion and exclusion criteria are important components of the study **protocol** since they are part of the case definition that delineates the **target population**. This is decided before the study starts with the purpose that only appropriate subjects participate and that they remain safe. The inclusion and exclusion criteria also help in reaching a focused and reproducible conclusion. The likelihood of exploiting the subjects is minimized, and the possibility of harming them is ruled out upfront. Each of these criteria must be fully justified in the protocol because reviewers can always question them, and they must also meet the requirement of regulatory agencies.

Inclusion criteria are those characteristics that are necessary for subjects to be considered as eligible for inclusion keeping in view the objectives of the study. Besides the disease under consideration, these characteristics could be age of at least 20 years, any sex, positive electrocardiogram, complaint of acute chest pain, etc. These establish the baseline about the type of cases on which the trial would be done and also make the subjects relatively homogeneous. In case healthy controls are needed, these also should be fully specified; for example, specify that merely the absence of that particular disease or absence of any disease would be considered for eligibility of these subjects.

Exclusion criteria are those characteristics the presence of which can affect the correct assessment of the outcome in the subjects. If a person is already under treatment for cardiovascular problems, he/she may become ineligible for inclusion in the study. A person can be excluded also because he/she (i) is too old or too weak to pass through the rigors of a clinical trial, (ii) has a serious form of disease that requires immediate attention, (iii) has a condition for which the test regimen or the control regimen can be harmful, (iv) is not likely to complete the follow-up, or (v) has comorbidity that could contaminate the response. Also, conditions such as obesity, anemia, subclinical diseases, and psychological conditions can compromise the results. These are just examples—in fact, inclusion and exclusion criteria are trial and regimen specific, and cannot be generalized.

Sometimes, research requires that only a very specific type of cases are included so that unadulterated results are obtained that prove a point, just as is mostly done for clinical trials with efficacy as the end point. But then the generalizability suffers. You will want your results to apply to the patients seeking medical help but not to be applicable to such general class of patients. Nevertheless, investigation of such a restricted class of cases fulfills the basic objective of coming to a conclusion about the efficacy and safety of the regimen.

The terms *inclusion criteria* and *exclusion criteria* are also used in the context of selection of studies for **systematic review** and **meta-analysis**. These criteria specify parameters such as minimum sample size, type of subjects included, proper methodology, specific measure of effect, being published in indexed journals, etc. This is done to include only the relevant studies and exclude frivolous ones. These criteria enhance the credibility of results and specify the population for which the ultimate conclusion will apply.

## incomplete tables, *see* contingency tables

## independence (statistical), *see also* dependence and independence (statistical)

An event A is considered to be statistically independent of event B when the chance of occurrence of one is not affected by occurrence or nonoccurrence of the other. The sex of an unborn child is not affected by the sex of another children in the same family. They are independent. My being hypertensive depends, to an extent, on my mother being hypertensive—thus, my hypertension status is not independent of the status of my mother. Liver disorders and injury in an accident of the same person are independent, but diabetes and hypertension are not. In 100 patients of kidney disease enrolled for a trial, the treatment response of one person will not be affected by the response of another person in the trial. They are independent. Persons with some **affinity** or belonging to a **cluster** are not independent, as they can give similar responses. Intraocular pressure (IOP) in the right eye is not independent of the IOP in the left eye. **Repeated measures** of a subject, say on pain score after a surgery, are not independent since pain score at time point 2 depends on what was the score at time point 1.

Strictly speaking, blood pressure of a male of age 34 years living in Nigeria will not be entirely independent of the blood pressure of a male of age 72 years living in Japan, because both are males and sex is one of the determinants of blood pressure. But this kind of far-fetched dependence is ignored in statistics since this is negligible in a probability sense.

Statistically,

necessary and sufficient condition for independence:

$$P(\mathrm{AB}) = P(\mathrm{A})*P(\mathrm{B}),$$

where $P$ stands for probability, $P(\mathrm{AB})$ is the joint probability of the two events occurring together, and $P(\mathrm{A})$ and $P(\mathrm{B})$ are individual probabilities. *Necessary and sufficient condition* means that independence implies this, and if this holds, independence is assured. This is known as the multiplication rule of probability and can be extended to three or more events. In practice, this equation may hold only approximately because of **sampling fluctuations**. Another probabilistic formulation of independence is that $P(\mathrm{A|B}) = P(\mathrm{A})$ and

$P(B|A) = P(B)$, where what comes after the vertical slash is what is already known to have occurred. These two mean that knowledge of occurrence of an event does not alter the probability of occurrence of the other. This is intuitive also—if they are independent, one should not affect the other.

Most statistical methods require that the values be independent. For example, this is a requirement for the **analysis of variance (ANOVA)** and **regression** analysis. When a sample of *n* subjects is randomly selected from a population for a study, these are independent since the values in one subject will not be affected by values in the other subjects. It is because of this independence that the joint probability of the sample values, called **likelihood**, can be obtained by simply by multiplying the individual probabilities. Many statistical procedures have been devised using this likelihood. Repeated measures require a modification of these methods because of dependence, as do serial values such as in a time series.

We have described independence in the context of events and values. *Independent variable* has a different meaning. For this, see **dependent and independent variables**.

## independent and paired samples

Two or more samples are considered independent when the values in one sample are not affected by the values in the other group. In a sample of 40 males and 50 females for studying kidney function, the two groups are independent since the values in males will have no effect on the values in females. On the other hand, if you are studying one group of subjects before a treatment and after the treatment, the values obviously are not independent since *after* values will depend on what the values were before the treatment. These are called paired samples. However, while dealing with quantitative measurement in a paired setup, the difference of *after* values from *before* values in one subject will be independent of the difference in another subject in the sample.

Beside *before* and *after* measurements, there are other examples of paired values in medicine. Blood pressure measurements in the two arms or in one arm and one leg of the same person are paired. Values in twins are a popular example of paired measurements. Measurements in husband and wife, living together, say, for 10 years, would also be paired because of a shared environment. One-to-one matching, as in some case–control studies, also provides paired values.

Independent samples require a different statistical method compared to a paired sample. You may be already aware about the **Student *t*-test**, which is different for independent samples compared to for paired samples. Similarly, the **chi-square test** for independent samples is different from chi-square for paired samples, called the **McNemar test**.

A paired setup can be extended to a multiple setup where three or more measurements are clustered. At the time level, this takes us to the setup of **repeated measurements** or serial measurements on the same group of people. At the family level, measurements of members of the family would be dependent on each other in most situations. When you have four observers, the observer-sensitive measurements, such as assessment of attitude, taken by each observer will have some dependence because of his/her method of assessment. Measurements taken at six sites of the brain will also be dependent. If you are measuring lung functions by forced vital capacity, forced expiratory flow in 1 s, peak expiratory flow rate, and total lung capacity in a sample 80 subjects, these four measurements are not independent of one another, although in this case, the measurements are different. Separate statistical methods, mostly **multivariate methods**, are required for analysis of such data.

## independent variables, *see* dependent and independent variables

## indexes, *see also* indicators

Index is a combination of two or more **indicators** that provides a relatively more comprehensive picture of the status. In health and medicine, indicators measure a specific aspect of health, whereas an index combines them to give better context. An index, when properly constructed, can enhance the utility of indicators and can sometimes generate even new information.

The most popular example in medicine is **body mass index** (BMI), which combines height and weight. Popular among others are the **APACHE score**, used for assessment of the severity condition of critical patients, which is an index but is called a score, and shock index, used for patients of ST-elevation myocardial infarction. Other examples are the ankle-brachial (pressure) index, bispectral index, glycemic index, and craniofacial index. At the community level, we have the **human development index**, **physical quality of life index**, and **disability-adjusted life years**. All these are a combination of two or more measurements.

An index is quantitative and therefore involves calculation that could be a nemesis for some clinicians. Some indexing instruments come ready with software to perform the calculations and directly provide the results. The bispectral index is automatically calculated by software. High-pressure liquid chromatography (HPLC) automatically calculates the peak area of intensity of signals corresponding to concentrates of drug-evoked potentials. Thus, for some indexes, calculations are not much of a problem. Perhaps a greater problem is their **validity** and **reliability**. A large number of indexes are available, and many are being devised every year, but studies that provide evidence of their reliability and validity for different segments of population are rare. The most widely used index for obesity is the BMI, but its utility, too, is sometimes questioned in comparison with the waist–hip ratio (WHR), which is sometimes seen as a better correlate of coronary events. The utility of WHR in cancer and lung disease has not been fully evaluated, while BMI has been extensively investigated. Thus choice of an index can be an issue in situations where two or more indexes are available for assessing the same aspect of health, and you need to be judicious in making the choice.

## indicators, *see also* indexes

An indicator is a tool to measure the quantitative level of a characteristic. This definition works well when the characteristic is graded and has a level. This is not always true as, for example, the site of cancer and the sex of a person are characteristics that cannot be assigned a grade at the individual level. For groups, though, the percentage of patients with cancer of different sites and sex ratio in a group of arthritis cases are quantities. These are indicators at the group level and not at the individual level.

The term *indicator* is generally used when the focus is on a *specific* aspect of a characteristic. Severity of disease in a patient can be assessed by the severity of pain, inability to perform essential functions of life, prognostic implications such as the chance of death within a week, etc. Each of these is an indicator since each is concerned with a particular aspect of severity of disease. An indicator is a **univariate** assessment and perhaps the most direct measurement of particular aspects of characteristics of interest. Indicators provide the exactitude that one would like to have in the assessment, or may want to have in the resource material consulted to update one's knowledge.

Critics may be concerned with the limited utility of indicators because they focus on one particular aspect of health and ignore the other aspects, howsoever related. For example, **case-fatality rate** is an indicator of the chance of death but is oblivious to the pain and suffering it brings to the patient and the family. The weight of a person is an indicator that, by itself, loses importance unless related to height. Glomerular filtration rate (GFR) is better interpreted in the context of creatinine excretion. At the same time, though, the focus of an indicator on a particular aspect is its greatest strength. Case fatality does measure the most important aspect of prognosis that is important for both the patient and the doctor, measuring weight does help when monitored in the same person over a period, and GFR does give an indication about kidney function. They are good standalone indicators, although putting them in the context of other related parameters does help in providing a better interpretation.

### *Choice of Indicators*

Quite often, multiple indicators are available for apparently the same characteristic, and one may have to make a choice. Triglyceride, high-density lipoprotein (HDL), and low-density lipoprotein (LDL) are indicators of different lipids. Many times, these are considered together, but if one has to choose, the selection would depend on the relevance for the condition of the patient and the specific aspect of the lipid of interest. For example, triglyceride level can be a good indicator of lipoprotein, called natural fats, and LDL level can be a good indicator of risk of developing atherosclerosis. Conversely, a researcher may want to find which of the three is a better marker of coronary events, and why the others are not as good.

The nutrition level of a child can be assessed anthropometrically by weight, height, and skinfold thickness. Even though these are assessed in relation to age, they are univariate and thus are indicators in a relaxed sense of the term. The choice among weight, height, and skinfold thickness as anthropometric indicators of nutrition level in children would depend on whether one is looking at short- or long-term nutrition, or at the adiposity. Against this, weight for height, for example, requires two measurements; hence, it is bivariate and technically transgresses to the realm of **indexes**.

Blood hemoglobin (Hb) concentration and hematocrit (Hct) values assist in evaluating plasma dilution, blood viscosity, and oxygen-carrying capacity. Hb concentration, in general, is easily affected dietary supplements, whereas Hct is a long-term measure. A related measure is red blood cell (RBC) count. Each of these is an indicator used in specific contexts, and you should be able to decide which ones would serve your purpose.

At the community level, events such as child mortality can be measured by neonatal mortality, postneonatal mortality, and infant mortality. The neonatal period can be divided into early (<7 days) and late (7 to <28 days) periods. Although all of them could be used together to provide a holistic picture, a professional may want to concentrate on one indicator that best measures the specific aspect of interest. If the focus is on antenatal care, maternal nutrition, and skilled attendance at birth, perhaps early neonatal mortality rate is the best indicator. If the focus is on breastfeeding, infections, and child nutrition, postneonatal mortality may be better. If the interest is in studying the inability to thrive in the face of repeated infections and midterm sequel of low birth weight, mortality between 1 and 4 years may be better. In summary, different indicators may seem similar, but each indicator has a specific application. The focus of application helps to decide which particular indicator to use.

## indicator variables, *see* variables

## indirect standardization, *see* standardized death rates

## Indrayan smoking index, *see also* smoking index

Developed by Abhaya Indrayan, the Indrayan smoking index [1] is just about the most comprehensive measure of the burden of smoking in an individual. Besides the usual duration and number of cigarettes, this index incorporates the effect of age at the start of smoking, intermittent smoking, passive smoking, and time elapsed since cessation in the case of past smokers. This can be adapted to include all forms of smoking—cigarette with and without filter, bidi, pipe, cigar, etc. This index is given by

$$\text{Indrayan smoking index: } S = (3 - a/15)\left(\frac{1}{2}\sqrt{\Sigma p_k n_k x_k} - y\right),$$

where

- $a$ is the age (years) at start of smoking (take $a = 30$ for $a > 30$ assuming that starting after the age of 30 years has the same effect as start at age 30 years),
- $p_k$ is the intensity of smoking for $n_k$ years ($p_k = 1$ for regular cigarettes and bidi, 0.67 for filter cigarettes, 5.0 for cigar, 2.5 for pipe, and 0.15 for passive smoking—these are suggested values and can be changed as desired),
- $x_k$ is the number of cigarettes/cigars/etc. smoked per day for $n_k$ years, and
- $y$ is the number of years elapsed since quitting in the case of past smokers (this must be less than $\Sigma p_k n_k x_k$)—$y = 0$ for current smokers.

Abhaya Indrayan

A value less than 0 of this index is interpreted as 0 since smoking cannot have any beneficial effect. This is a comprehensive index of the present burden of smoking as it incorporates (i) the duration of smoking, (ii) the quantity of smoking, (iii) smoking of filter cigarettes and other forms of smoking that can be factored to cigarette smoking, (iv) progressively higher burden from smoking more pack-years in life (but the cumulative effect tends to flatten), (v) benefit of the time elapsed since quitting, and (vi) deleterious effect of starting smoking early in life. The index has a built-in feature to consider current smokers and ex-smokers and obviates the need to divide ever smokers into such a dichotomy. The index models the entire history of smoking into a single metric. The modification of cigarette-years by this index in some typical conditions is shown in Figure I.2. This index does incorporate a large number of aspects of smoking but fails to capture occasional smoking or the beneficial effect of interruption.

To see how this index works, consider the smoking history of the following three persons:

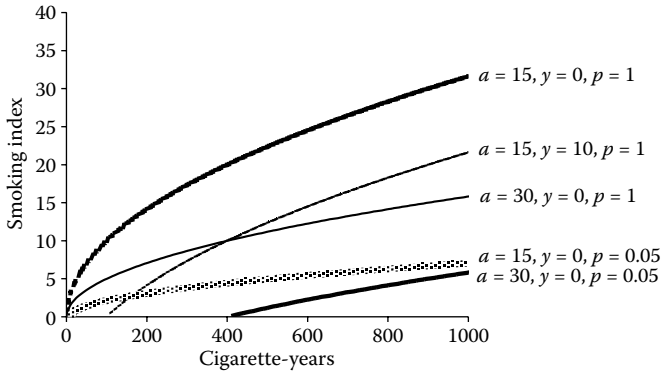



**FIGURE I.2** Indrayan smoking index for some typical values (see text).

Person A: Started smoking at age 12 years. Initially smoked 10 regular cigarettes for $3\frac{1}{2}$ years. Since then has been smoking 20 filter cigarettes a day for the last $17\frac{1}{2}$ years.

Person B: Started smoking at age 21 years. Smoked 12 cigarettes a day for 1 year, 15 cigarettes a day for $2\frac{1}{2}$ years, 20 cigarettes a day for $1\frac{1}{2}$ years, no smoking for 6 months, and 2 cigars a day for 1 year. He has not smoked for the past 4 years.

Person C: Never smoked but spouse smoked. When both were together, an average of 5 cigarettes a day were smoked. This started at age 27 years and went on for 6 years. There has been no exposure to cigarette smoke for the past 3 years.

The smoking index is as follows:

$$\text{Person A:}\quad S=(3-12/15)\left(\frac{1}{2}\sqrt{10\times3.5+0.67\times20\times17.50}\right)$$

$$=2.2\times8.21=18.1.$$

An index of 8.21 increases to 18.1 because of the deleterious effect of starting smoking at an early age of 12 years.

$$\text{Person B:}\quad S=(3-21/15)\left(\frac{1}{2}\sqrt{12\times1+15\times2.5+20\times1.5+2\times1}-4\right)$$

$$=1.64\times(4.73-4)=1.2$$

The present burden of smoking is small because of quitting 4 years ago.

$$\text{Person C:}\quad S=(3-27/15)\left(\frac{1}{2}\sqrt{0.15\times5\times6}-3\right)$$

$$=1.2\times(-1.9)=0\ \text{(negative value is to be taken as 0)}$$

The burden is small because of passive smoking, and that, too, vanished because of no exposure for the past 3 years.

The Indrayan smoking index is a vast improvement over the usual index of pack-years because of incorporation of so many facets. Indrayan et al. [2] have tried to validate this on a data set on smokers, but the index is yet to be adopted on a large scale because of the subjective consideration of the effect of an early start, benefit of quitting, and equivalence of other forms of smoking with cigarette smoking. All these, however, are flexible in this index, and it is a pointer on how a comprehensive index can be devised.

All smoking indexes, including this one, are heavily dependent on age. The higher the age, the greater the chance of a higher index because then the person may have smoked for a longer duration. The burden of smoking naturally accumulates as age advances. Thus, this should be calculated separately for each age group, particularly for comparing one population with the other.

1. Indrayan A. *Medical Biostatistics*, Third Edition, Chapman & Hall/CRC Press, 2012.
2. Indrayan A, Kumar R, Dwivedi S. A simple index of smoking. *COBRA Preprint Series (Berkeley Electronic Press)*, Article 41, 2008. http://biostats.bepress.com/cobra/ps/art40

## infant mortality rate, *see* mortality rates

## infection rate

Infection rate is the number of new infections in a specified duration per 1000 persons at risk. The duration could be per week, per month, per year, etc., depending on how quick the spread of infection is. The population can also be 100 to make it percent, per million, or any other number so that the infection rate becomes a convenient number. In terms of formula,

$$\text{infection rate}=\frac{\text{number of new infections in a unit of time}}{\text{population at risk at midpoint of the time}}\times K,$$

where $K$ is the constant of your choice depending upon whether you want it to be per 100, per 1000, per million, etc. This rate measures the risk of infection per unit of time in a given setting.

Surveillance or any such system would be required to keep track of infections. Many countries have this in place for infections such as HIV. For example, according to one report [1], HIV infection rate in the United States dropped from 24.1 per 100,000 people in 2002 to 16.1 in 2011. This gives some indication of how the epidemic is relenting in that country. When broken up into groups, the infection rate actually increased in the gay and bisexual communities during this period.

A common application of infection rate is in the hospital setting, where infections occurring in the patients admitted in a hospital are a serious concern for administrators. Most good hospitals keep an immaculate record of hospital infections. Only those infections that occur after admission are counted; thus, these can be considered hospital acquired and are called *hospital infection rate* or *nosocomial infection rate*. The denominator can be changed to hospital-days (see **person-time**) of stay in the hospital in place of the number of patients admitted. If the rate is per month and a patient gets two infections in a month, that patient would be counted twice under this scheme. When calculated ward-wise or department-wise, this rate can provide vital information on where the problem needs more attention. However, for comparison across wards or across hospitals, the rates may have to be **standardized** for the age of the patients, severity of disease, etc. It can also be calculated for each surgery site, catheter related, endoscope related, etc.

1. BBC News. HIV infection rate in the US falls by a third in a decade. http://www.bbc.com/news/world-us-canada-28389275, last accessed August 8, 2014.

## infectious disease models, *see also* epidemic models

Infectious disease models are biostatistical equations that try to predict the number or proportion of infectives in a population on the basis of susceptibility of the population, **infectivity** of the disease, latency of the disease, and other such features. These models help in better understanding of the exact role of various factors in the trend of **incidence** and **prevalence** of infectious diseases. They are applied to express the dynamics of infectious diseases and thereby devise strategies for their control. They can be used to evaluate the impact of the programs that are designed to reduce infections or increase immunity (such as vaccinations). If the models are really appropriate, they can help identify the optimal control strategies for minimizing infections. Infectious disease models have been most commonly tried for HIV-AIDS, malaria, and tuberculosis [1].

**Models**, by definition, are simplified versions of a complex process. Thus, they can never be 100% correct, and if they yield correct results in a large percentage of cases, they are considered good. The same is true for infectious disease models. A simple model that works reasonably well under certain regularity conditions is the so-called *SIR model*. Under this model, $S$ is the fraction of susceptible individuals in the population, $I$ is the fraction of infected persons, and $R$ is the fraction of resistant individuals. Resistant individuals are those that are almost permanently immune (either because previous infection has provided this kind of immunity or due to immunization). In a closed population, this would mean $S + I + R = 1$ in the long run. If β is the per capita contact rate, the rate of transmission of infection = $\beta * I * S$, provided, and this is a big *if*, that there is homogeneous mixing in the population. This rate is generally high for dense populations and is associated with greater duration of infectiousness and greater infectiousness of the disease. If you wish to consider those who are exposed but not infected yet (latent cases) also, denote them by $E$, and this becomes the *SIER model*. The rate of transmission is also called the infectiousness of the disease. Such a model can be used to predict the impact of changes in any parameter of the system. Eames et al. [2] discussed these models for studying the impact of school closure on the swine flu epidemic in the United Kingdom in 2009.

The model described in the preceding paragraph is deterministic as there is no probability element. When extended to incorporate the chance element or the distribution (such as distribution of contact rate with the provision that some segments of the population have high contact rate and some have low contact rate), the model becomes stochastic. This helps to prevent ascribing variations in the spread of disease to infectiousness when it is explainable by chance alone. For an example, see the work of Brown et al. [3], who proposed a biosurveillance network for detecting emerging zoonotic outbreaks using a stochastic SIER model.

For further details of infectious disease models, see Vynnycky and White [4].


1. Casals M, Guzmán K, Caylà JA. Mathematical models used in the study of infectious diseases [Spanish]. *Rev Esp Salud Publica* 2009 Sep–Oct;83(5):689–95. http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1135-57272009000500010&lng=en&nrm=iso&tlng=en
2. Eames KT, Tilston NL, White PJ, Adams E, Edmunds WJ. The impact of illness and the impact of school closure on social contact patterns. *Health Technol Assess* 2010 Jul;14(34):267–312. http://www.ncbi.nlm.nih.gov/pubmed/20630125
3. Brown M, Moore L, McMahon B, Powell D, LaBute M, Hyman JM, Rivas A et al. Constructing rigorous and broad biosurveillance networks for detecting emerging zoonotic outbreaks. *PLoS One* 2015 May 6;10(5):e0124037. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4422680/
4. Vynnycky E, White RG. *An Introduction to Infectious Disease Modelling*. Oxford University Press, 2010.


## infectivity, pathogenicity and virulence, *see* disease spectrum

## inference (statistical)

Statistical inference is the scientific induction of results from a sample to the population. This is an extrapolation based on the evidence provided by the sample. This works well under commonsense condition, namely, that the sample is an adequate representative of the population. Representativeness can be almost assured by two statistical tenets: the sample is large enough to reflect the full spectrum of the subjects, and it is randomly drawn. Statistical texts are full of references to these features of the sample, and indeed, these can hardly be overemphasized. They are so much ingrained in the statistical thought process that they are many times taken for granted. *Random* samples allow the use of the probability rules, which are essential for latching onto the mathematical basis for providing rationale and also help in achieving representativeness. Nonrandom samples can also be representative, and they also can be used for inference, perhaps for the right inference, but the inference so arrived at will lack the scientific rationale. On the other hand, a random sample can be nonrepresentative, particularly when small, which can lead to wrong inference but may still escape criticism because of the randomness property. Large samples increase our confidence—thus, reliable inference can be drawn. The best assurance for reliable and valid conclusion inference is a large sample and random selection.

Statistical inference is basically of two types: **estimation** and **testing of hypothesis**. Both generally require specifying the unknown **parameters** that are to be estimated or tested. Exceptions are when the hypothesis of randomness or of a specified form of **distribution** is tested where the focus is not on any specific parameter but is on other features of interest. In routine situations, the parameters targeted for inference are the mean of the population, proportion or probability of occurrence of an event, correlation coefficient and regression among the variables, odds ratio for an event, etc.

Statistical estimation is of two types: the point estimate and the interval estimate. Estimating a parameter by one value is called the point estimate, whereas constructing a range of values outside which the value of the parameter is extremely unlikely is called the interval estimate. The latter is better known as the **confidence interval** (CI) when the range excludes extremely unlikely small values as well as extremely unlikely large values, and called **confidence bound** when only one-sided values are excluded. If you find in a trial that the efficacy of a regimen is 83%, this is the point estimate of the efficacy that you expect in all cases of that type. The CI will say something like "the efficacy is not likely to be less than 77% or more than 89%." If we say that the efficacy is not likely to be more than 93%, this is the upper confidence bound, and if we say that that the efficacy is not likely to be less than 65%, this is the lower confidence bound. The interval estimate heavily depends on the **confidence level** you want to put into the range. Generally, a confidence level of 95% is chosen so that the probability that a random value is outside is 5%. This is the convention generally followed to determine whether an event is extremely unlikely. The procedure for obtaining the CI for a large number of different types of parameters is presented in this book under the respective topics.

Inference regarding the point estimate can be drawn in practically all situations without worrying about the underlying **distribution** of the values. The point estimate almost invariably is the corresponding value in the sample; for example, the point estimate of the population median is the sample median, and for population relative risk (RR), it is the relative risk found in the sample subjects. However, if you want to find the reliability of the point estimate in the sense of how likely it is to hold in repeated samples, you would immediately need to know the distribution of the underlying values. Distribution is also required for finding the CI and the bound. **Nonparametric methods** are available for inference in case a definite distribution cannot be postulated. Another inference method now under considerable discussion is the **bootstrap**, which also is a nonparametric method.

The second type of statistical inference is **testing of hypothesis**. The thrust in this inference is to find whether the data are consistent with the hypothesized value of the parameters. A **null hypothesis** is set up, and the chance of the consistency of the data with this null is evaluated through a test criterion with known distribution. This type of inference also is more efficient when the distribution of the underlying values is known, at least approximately, since that helps to construct a suitable test criterion. Several procedures for testing a hypothesis on specific parameters are discussed in this volume under the respective topics. Nonparametric methods, including bootstrap, are available that do not depend on the underlying distribution, but these have limitations. First, they are not as efficient as the usual parametric methods for known distributions, and second, they are not as developed yet.

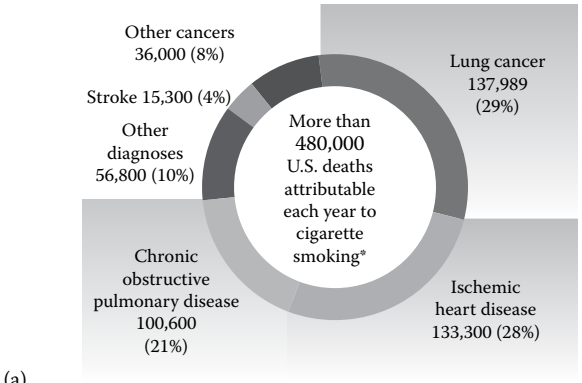
All inferences based on tests of hypotheses involve errors, as discussed for Type I error and Type II error. Statistical inferences are never absolute and can be wrong in individual cases, though they have high likelihood of being right in the long run when based on appropriate methods.

## infographics

Infographics contain some textual information in addition to graphs, as in Figure I.3a on deaths attributable to smoking in the United States, and may even contain many graphs with textual information, as in Figure I.3b on maternal mortality. The purpose is to present a complete picture so that the reader does not have to refer to the text. An infographic is supposed to be complete in itself.

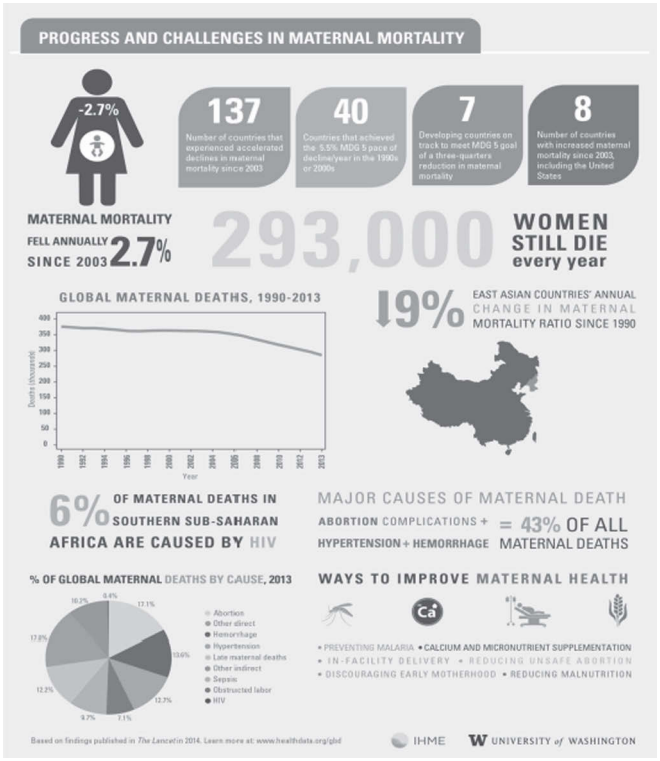
Figure I.3b contains a lot of information: the number of countries with fast and slow progress in controlling maternal mortality; the global trend in maternal mortality rate since 1990 and the decline since 2003; the number of women still dying; the percentage of maternal deaths due to HIV in southern sub-Saharan Africa; the rate of decline in maternal mortality in East Asian countries; the major causes of maternal deaths; and the distribution of maternal deaths in the world by cause. Perhaps the cognitive part of the graphics is lost when so much of the information is depicted in one infographic. The other infographic in Figure I.3a is simple and effective in conveying the number of smoking-attributable deaths occurring due to various causes in the United States during 2005–2009.

1. CDC. *Smoking and Tobacco Use*. http://www.cdc.gov/tobacco/data_statistics/tables/health/infographics/index.htm
2. IHME. *Millennium Development Goal 5: Progress and challenges in maternal mortality*. http://www.healthdata.org/infographic/millennium-development-goal-5-progress-and-challenges-maternal-mortality




(a)

*Average annual number of deaths 2005–2009




(b)

**FIGURE I.3** Infographics: (a) annual deaths attributable to smoking in the United States in the years 2005–2009. (From CDC. *Smoking and Tobacco Use*. http://www.cdc.gov/tobacco/data_statistics/tables/health/infographics/index.htm.) (b) Maternal mortality status in the world, 2013. (From IHME. *Millennium Development Goal 5: Progress and challenges in maternal mortality*. http://www.healthdata.org/infographic/millennium-development-goal-5-progress-and-challenges-maternal-mortality.)

## informatics (medical), *see also* expert systems

Medical informatics is the science concerned with the processing and engineering of the systems of information related to health and medicine. Thus, it has two distinct parts: first is information processing, and second is devising and maintaining information systems. The latter is mostly the domain of computer science experts, but the former has a lot to do with biostatistics, although these days, this part also is heavily dependent on computers.

The MEDLARS (Medical Literature Analysis and Retrieval System) project of the National Library of Medicine of the United States is just about the most well-known product of medical informatics. This started in 1964 and became available online in 1971 as MEDLINE. This system now has more than 22 million citations from more than 5000 medical journals across the world in different languages and is the most comprehensive database of medical literature. Back citations from the year 1951 have also been added. PubMed is the search portal that includes MEDLINE and online books. Its wide and free availability has made it look like an ordinary database, but consider the enormous efforts that went and are going into collecting, retrieving, collating, and organizing the information from such diverse sources.

Another major component of medical informatics is **expert systems**. These collate the medical information regarding diseases and other health conditions, particularly on diagnosis and treatment. With more than 2000 known disease entities, their stages, severity levels, variations across populations, and varied modes of treatment (including many unknowns), it is an uphill task to systematically organize all the information in one place. Not much success has been achieved yet on this front, not even when a single disease is considered. Expert systems require statistical considerations of probabilities of the presence or absence of disease, sensitivity–specificity/predictivities of medical tests, etc.

Beside the two mentioned in the preceding two paragraphs, the third is **data mining**. With almost every hospital now maintaining electronic records of at least the admitted patients, it is possible to dig through these data sets for signals that can identify the modes of diagnosis and of therapy that have yielded successful results. In the course of time, as public pressure for transparency grows, these databases may be available online. One can then pool several databases and come to more reliable results. This activity can gain substantial momentum when the records of patients in different clinics follow a uniform pattern so that the analysis and inference is straightforward from the pooled database.

The fourth is the ongoing effort to computerize and link the medical activities at the most peripheral levels, such as primary health centers in villages. Their linkage with the specialists at, say, district hospitals can substantially improve outcomes. *Telemedicine* is gradually but surely making headway into the medical systems with this objective.

The fifth is about computer applications to health gadgets that can communicate with care professionals for help. The 911 service in the United States is an example that may have saved millions of lives by instant tracing of the location of the distress calls and providing help within minutes. Another is health applications on mobile phones [1] and perceived-health monitors [2]. Your investigation and examination reports reach your phone or computer, which you can use subsequently for consultation with any doctor. Personalized medicine is also moving forward on the strength of information technology (see, for example, the work by Binefa et al. [3] on colorectal cancer).

For more information on medical informatics, see Ong [4].


1. Sama PR, Eapen ZJ, Weinfurt KP, Shah BR, Schulman KA. An evaluation of mobile health application tools. *JMIR Mhealth Uhealth* 2014 May 1;2(2):e19. http://mhealth.jmir.org/2014/2/e19/
2. Indrayan A. A health monitor in your pocket? *CSI Communications* 1996;20(1):8–9.
3. Binefa G, Rodríguez-Moranta F, Teule A, Medina-Hayas M. Colorectal cancer: From prevention to personalized medicine. *World J Gastroenterol* 2014 Jun 14;20(22):6786–808. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4051918/
4. Ong KE. *Medical Informatics: An Executive Primer*, Second Edition. Health Information Management and Systems Society (HIMSS), 2011.


## informed consent, *see* ethics of clinical trials and medical research

## instrumental variables, *see* variables

## intention-to-treat analysis

The intention-to-treat (ITT) strategy analyzes all participants in a **clinical trial** according to their original group assignment regardless of what occurred subsequently. This reduces bias due to nonrandom loss or shift of participants that can occur during the course of the trial. The efficacy of a regimen may be overestimated if ITT is not done. ITT analysis is conservative and provides more confidence. ITT is only for aligning groups—the statistical method of analysis, such as Student *t*-test and ANOVA, remains the same as in the usual setup.

ITT analysis is an ingenious strategy to partially circumvent the limitation imposed by a specific kind of distortion in the data and is particularly advocated for **randomized controlled trials**. Consider a typical trial for a new drug for vertigo in which randomized patients are given a tablet and told to take one if and when an attack occurs. Some subjects will have vertigo and take the tablet, and some will not have an attack and will not take the tablet. Should the analysis be based on those who had the attack or on all those who were randomized? There might be other situations where patients randomized to receive the intended treatment shift loyalty in the middle of the study and go for another treatment. Sometimes, the assigned regimen is not given in full, due to side effects, or one or more subjects receive the incorrect regimen in error. Sometimes, there is noncompliance, and some patients do not follow the complete regimen. Deviation from protocol can occur for a variety of other reasons also. These are examples of some situations where ITT analysis can be helpful. Since this can occur in practice as well, this simulates **pragmatic trials** and can be tried as an additional analysis to the regular per-protocol analysis that excludes the missing or distorted data. *Per-protocol analysis* is done on those who have taken a full course of medication, who took no prohibited drugs, and whose outcome is available as per the protocol. If the results do not materially differ in these two analyses, the confidence in the results naturally strengthens. For a complete ITT analysis, the outcome must be known for all patients, including those who shifted to other treatments. There is no consensus on how to handle nonresponse in ITT analysis, but the most acceptable method is to consider all nonresponses as adverse to the treatment. ITT analysis is recommended especially for superiority, **equivalence**, and noninferiority trials.

An appropriate situation for ITT analysis is given in Table I.1, where medical and surgical treatments are the two treatments under comparison for stable angina pectoris [1]. The outcome measure is 2-year mortality. Some patients switched from medical to surgical and some from surgical to medical under compelling circumstances.

A total of 373 patients were allocated to receive medical treatment (groups A and B) as first-line treatment, but 50 of these ended up getting surgery (second-line treatment; group B). Similarly, out of 394 patients allocated for surgery as first-line treatment (groups C and D), 26 got medical treatment (second-line treatment; group D). The ITT analysis would be (A + B) versus (C + D). Two-year mortality in these groups was 7.8% and 5.3%, respectively. On the contrary, if the protocol is to be strictly followed, the comparison would be of group A with group C, and the deviant groups B and D would be ignored. If the patients actually receiving medical (A + D) and surgical (B + C) treatments are compared, the 2-year mortality was 9.5% and 4.1%, respectively.

**TABLE I.1**
**ITT and Other Analyses to Compare Medical and Surgical Treatments for Stable Angina Pectoris**

| | Treatment—Allocated/Actual | | | |
|---|---|---|---|---|
| | Medical/Medical | Medical/Surgical | Surgical/Surgical | Surgical/Medical |
| | A | B | C | D |
| Number of patients | 323 | 50 | 368 | 26 |
| Number of deaths | 27 | 2 | 15 | 6 |
| Mortality | 8.4% | 4.0% | 4.1% | 23.1% |
| ITT analysis (A + B versus C + D) | 29/373 = 7.8% | | 21/394 = 5.3% | |
| As per protocol (A versus C) | 27/323 = 8.4% | – | 15/368 = 4.1% | – |
| As per actual treatment (A + D versus B + C) | 33/349 = 9.5% (A + D) | 17/418 = 4.1% (B + C) | | |

*Source:* Adapted from European Coronary Surgery Study Group. *Lancet* 1979;i:889–93. http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(79)91372-2/abstract.

Note how different approaches can provide different results. ITT analysis provides a pragmatic estimate of the effect or of the difference in actual situations, and not the potential difference in ideal situations. Thus, this analysis mirrors clinical decisions and avoids bias in most situations. However, there are exceptions, such as safety of a regimen. ITT also fails to take care of fallacies, such as a participant tossing the drug or placebo in a toilet or taking four doses one day after missing them on the previous three occasions. The latter might come out from the interview of the subjects; the former may never surface.

Whereas efficacy of a regimen can be evaluated for clinical realities where some patients end up in an unintended group, common sense dictates that safety analysis should be on as "as-treated" basis. It would be unfair to ascribe a serious side effect to placebo where the patient supposed to receive placebo actually received the test drug. For not-so-serious side effects such as headache and nausea, which can occur due to nonpharmacological reasons and can occur in a placebo group, ITT analysis can still be adopted. In the case of efficacy, the major concern is with avoidance of **Type I error**, while for safety, avoidance of **Type II error** matters more—as in equivalence studies.

1. European Coronary Surgery Study Group. Coronary-artery bypass surgery in stable angina pectoris: Survival at two years. *Lancet* 1979;i:889–93. http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(79)91372-2/abstract

## interaction, *see also* main effect and interaction effect in ANOVA

Biological interaction is generally understood to occur when two or more factors are simultaneously needed to produce or enhance (or retard) an effect. Statistical interaction in the context of **designs** is a departure from the **additive** model. This means that the simultaneous effect of two factors is different from the sum total of their individual effects. *Interaction* is a term that belongs to a *K*-way ($K \geq 2$) experiment but is not restricted to experiments on biological samples and animal experiments. In fact, it is better explained by examples of human experimentation, better known as clinical trials.

The interaction explained in the next paragraph is not a pharmacological interaction between drugs but is a statistical interaction between the *effects* of a regimen when administered in conjunction with another regimen. It is also different from other usages, such as host–environment interaction.

Some factors work more effectively when other conducive factors are also present. Iron supplementation is more effective in increasing hemoglobin level when folic acid is also given. Their combined presence is much more effective than the sum total of their individual effects. This is a positive interaction and is called **synergism**. Since aspirin can reduce the beneficial effect of angiotensin-converting enzyme (ACE) inhibitors in patients with heart failure, they possibly have a negative interaction. This is called **antagonism**. Most interactions cannot be classified into either of these two categories. Osteoporosis is more severe in older women than older men (Figure I.4a). Thus, age and gender interact for severity of osteoporosis. Perhaps age and gender have no interaction for total lung capacity (TLC) as the decline in TLC with age runs almost parallel in men and women (Figure I.4b). A similar pattern of response for various levels of factors, except for a nearly constant difference, indicates absence of interaction. Then they are called factors with an **additive effect**. If the responses are not parallel, interaction is said to be present. Epidemiologically, interaction is called **effect modification**.

In the sense just described, the term *interaction* is used for the product of two or more **antecedent** factors and not for the relationship between an antecedent and an outcome. The differential effect of various levels of an antecedent on outcome is not called an interaction.

Even when interaction is absent and the factors' effects are additive, analysis of a two-way design gives different results from the analysis of two one-way designs—one for each factor. In any case, at least a two-way design is needed to explore whether interaction is present between two factors. Also note that when interaction is present, the average effect (**main effect**) of the factors gives a wrong picture. The interaction could be such that the factor has negative

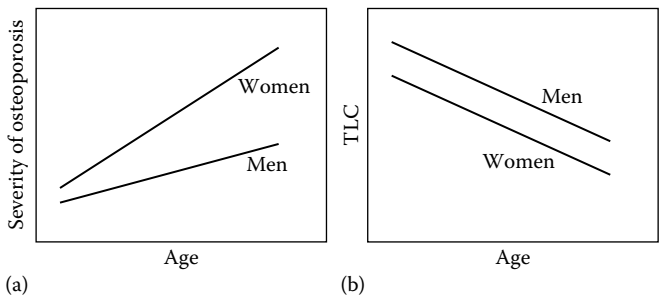



**FIGURE I.4** (a) Age and sex interact to produce differential effect on severity of osteoporosis; (b) age and sex do not interact for effect on TLC.

effect when it is $<a$ and positive effect when it is $\geq a$, which could produce an average close to 0. Thus, main effects have to be carefully interpreted in case interaction is present. The **analysis of variance (ANOVA)** procedure is fully equipped to test for statistical significance of an **interaction effect** through calculation of interaction **sum of squares**. If more than two factors are present, say A, B, and C, you can test for interaction between A and B, between A and C, and between B and C. You can also test for higher-order interaction, in this case, among factors A, B, and C combined. This will be difficult to interpret.

Although methods exist that can evaluate interaction in factors even when only one subject receives each combination of factor levels, generally, it is evaluated when more than one experimental unit is assigned to each factor level combination. More than one subject for each combination of factor levels is essentially the same as **replication**.

We have so far restricted our discussion to interaction between factors in an ANOVA setup, but this can happen in a **regression** setup also. If you have two quantitative regressors such as age and body mass index (BMI), these can interact in the sense that higher BMI in older age can be more harmful than the sum total of these two effects individually. If this is suspected, the product term (e.g., $x_1{*}x_2$) is included as one of the regressors. If the **regression coefficient** of this term is statistically significant, you can conclude that their interaction has a significant effect on the outcome.

## intercept (in a regression), *see* simple linear regression

## interim analysis

As the name suggests, interim analysis is the analysis of data while the data collection process is still going on. This means that the initial part of the data are analyzed with the purpose to look for signals that can indicate either that the objectives are extremely unlikely to be fulfilled or that they are already fulfilled, or that the project is on track and needs no modification, or that some aspects of the study need to be modified.

Interim analysis is not done on an ad hoc basis but is done after careful planning. This must be explicitly stated in the **protocol** with details of when the interim analysis will take place, what the objectives of such analysis are, and what action will be taken on the basis of the interim results. The stage of the project for interim analysis can be specified as after two-thirds of the sample is complete, or after 1 year of enrolment, after 6 months of follow-up, etc. In a phase II trial on hair loss–related quality of life in whole-brain radiotherapy, De Puysseleyr et al. [1] planned interim analysis after 10 patients to assess the futility of continuing. Futility was defined as a mean score of hair loss exceeding 56.7. Nüßlein et al. [2] presented interim analysis after 6 months of a prospective study (it was not a trial) for an otherwise a 2-year study on the effectiveness of abatacept for rheumatoid arthritis. Thus, different stages can be specified as per the requirement.

Interim analysis is commonly done in the case of **clinical trials**. Since these trials are becoming extremely expensive, it would be nice if we can save some effort by analyzing the interim data. Thus, the data are analyzed in two or more stages. The objective of interim analysis in the case of clinical trials is to find if there is sufficient evidence for stopping the trial due to either efficacy or futility. If the interim analysis provides sufficient evidence of the desired efficacy of the regimen at the specified level of significance and good statistical power, the trial is stopped for efficacy. If the interim analysis gives sufficient indication that there is hardly any chance of reaching the desired efficacy, the trial is stopped for futility. As discussed for **stopping rules**, all these are fully specified in the protocol after considerable thinking so that the probability of **Type I error** and of **Type II error** are not adversely affected. If the results of the interim analysis are not going to be revealed to anyone involved with trial conduct and the only effect that the interim analysis may have on the trial is to cause it be curtailed by reason of "futility," then this does not result in an increase in the risk of Type I error. In this case, the investigators know that an interim analysis is going to be run and when. In a reverse argument, if the randomized controlled trial (RCT) is blind and everybody keeps quiet as they should, and the trial continues, the investigators still guess quite rightly that the interim results were positive for continuing. Note that it is rightly guessed even if nobody says anything and everybody is blind. If the investigators thus believe that the drug is working, and even talk about it with the patients, a bias can occur. If the end points are more subjective to assess or psychologically influenced, such as depression, this can introduce a potential bias in the assessor and maybe also in the patient. The bias does not occur if nobody knows that an interim analysis is being run. This builds a case for not including interim analysis in the protocol, so that the investigators are not aware that an interim analysis is being run. You may have to weigh both sides of the argument and reach a conclusion regarding including or not including interim analysis in the protocol.

1. De Puysseleyr A, Van De Velde J, Speleers B, Vercauteren T, Goedgebeur A, Van Hoof T, Boterberg T, De Neve W, De Wagter C, Ost P. Hair-sparing whole brain radiotherapy with volumetric arc therapy in patients treated for brain metastases: Dosimetric and clinical results of a phase II trial. *Radiat Oncol* 2014 Jul 29;9(1):170. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4118657/
2. Nüßlein HG, Alten R, Galeazzi M, Lorenz HM, Boumpas D, Nurmohamed MT, Bensen WG et al. Real-world effectiveness of abatacept for rheumatoid arthritis treatment in European and Canadian populations: A 6-month interim analysis of the 2-year, observational, prospective ACTION study. *BMC Musculoskelet Disord* 2014 Jan 11;15:14. http://www.biomedcentral.com/1471-2474/15/14

## internal attributes, *see* factor analysis

## internal consistency, *see also* consistency

Among various meanings of internal consistency, the one most relevant for us is the psychometric property of the multi-item instrument (**questionnaire/schedule**) used in obtaining and recording the responses from the people surveyed. This property says that the items should be framed in a manner such that the responses for different items of the instrument are consistent with one another. Internal consistency is the property of the instrument and not of the responses, but it is through the responses that the internal consistency of an instrument is assessed.

It is important to make a distinction between reliability of an instrument and its internal consistency. **Reliability** in the context of an instrument refers to consistency of the whole instrument, whereas internal consistency refers to the items within a test. As the number of items increases in an instrument, it tends to become increasingly reliable as its repeated application in similar situations tends to give the same result. A lengthy instrument has an averaging effect. As the number of items increases, the ability of the instrument to get a similar *total* response from people with similar feature increases. This happens because any low response reported under one item is likely to even out a higher response

to another item. This is not so with items. The internal consistency of a 12-item instrument could be the same as of an 18-item instrument—this would not be so with reliability. In the literature, you may find internal consistency being considered as part of the reliability.

Consider an instrument containing several questions or items on, say, a scale of 0 to 5. An example is the assessment of ability to perform activities of daily living (ADL) by a geriatric population. Items such as ability to dress, to bathe, to walk around, and to eat can be scored from 0 for complete inability (full dependence) to 5 for complete independence requiring no assistance. It is expected that the ability score on one item would correspond to the score on other items. If there were a difference, it would generally persist across subjects. In fact, all items are different facets of the same entity—the ADL in this case. That the underlying construct is uniform across all items of a test is an important prerequisite for measuring internal consistency. If that is so, the responses will be consistent with one another provided that the items are properly framed and the questions are appropriately asked.

One way to measure internal consistency is to split the test into ostensible halves where feasible. This is generally possible for a questionnaire by randomly dividing questions into two parts. The other method is to put odd-numbered questions in one half and even-numbered questions in the other half. The product–moment **correlation coefficient** between total scores in the two parts across several subjects is a measure of internal consistency. This is called **split-half consistency**. Note that there is no one-to-one correspondence between one item in one half of the test and an item in the other half of the test. Thus, **intraclass correlation**, which otherwise is a measure of consistency, cannot be used in this case. The usual product–moment correlation coefficient between total scores in the two parts is used to assess split-half consistency.

The most acceptable measure of internal consistency of a multi-item instrument is **Cronbach alpha**. This uses the average of correlations between each pair of items and can be obtained only when the items have a quantitative response. If the responses are binary yes/no, the **Kuder–Richardson coefficient** can be used. For details, see Carmines and Zeller [1].

1. Carmines EG, Zeller RA. *Reliability and Validity Assessment*. Sage, 1979.

## International Classification of Diseases (ICD)

The International Classification of Diseases (ICD) is an internationally acceptable system to classify various diseases and other health conditions in a uniform format so that each disease is understood in the same manner all over the world. The full name is International Statistical Classification of Diseases and Related Health Problems. It helps in achieving uniformity not just in classification but also in collection, processing, and reporting of disease data. Each disease is given an alphanumeric code, and its subdiseases, variations, and complications are given distinct subcodes. Thus, for example, migraine without aura has code G43.0, migraine with aura has code G43.1, hemiplegic migraine has code G43.4, and abdominal migraine has code G43.D. There are 12 subcategories of migraine alone. There is code G43.9 for unspecified also, so that nothing is left out. The codes we have mentioned are four alphanumeric characters, but they can have a fifth character to specify that the condition is not intractable (fifth character 0) or intractable (fifth character 1) and a sixth character that identifies the migraine as occurring with migrainosus (additional code 1) or without migrainosus (additional code 9). Each one of these conditions is specified to avoid any confusion. You can see that the ICD transforms textual descriptions of health conditions to alphanumeric codes, and the hierarchy is the same as seen in these conditions. These are **mutually exclusive and exhaustive** groups, and nothing is left out since a code is provided for "others" also. Such detailed coding substantially reduces chances of any misunderstanding regarding the type of disease under consideration. Now, in place of long twined definitions that are liable to be interpreted differently by different workers, a code is enough to describe the disease. If such detailed specification of any disease is not available, a broad code such as G43 only for all migraines can be used. Code G is for diseases of the nervous system, and the subsequent characters serve to further specify the disease.

Because of the standardization introduced by ICD, the incidence and prevalence of diseases and other health problems can be monitored without fearing much about changing definitions, varying interpretations, and different nomenclatures in different countries and by different professionals at different points in time. Cause of death recorded in death certificates can also be interpreted uniformly around the world.

ICD is piloted by the World Health Organization (WHO) and is periodically revised as new diseases emerge and new knowledge about existing diseases is generated. The present classification is called revision 10 (named ICD-10), and the 11th revision is expected to be available in 2018. The codes given for migraine in one of the preceding paragraphs are as per ICD-10. ICD has several adaptations, such as for oncology, injuries, and primary care.

While detailed coding is a boon for complete specification of the exact disease, it also is a bane for many workers. Medically qualified persons are needed to specify the code, but more important is that many medical professionals too do not have adequate training to adopt these codes. Thus, many diseases get misclassified, and it is extremely difficult in a finished report to find the extent of misclassification. Because of this limitation, ICD is still not used widely, particularly in developing countries, where medical professionals are in short supply.

Full details of ICD are available at the WHO website [1].

1. WHO. *International Classification of Diseases (ICD)*. http://www.who.int/classifications/icd/en/

## interpolation

Interpolation is the process of estimating the unavailable intermediary values on the basis of the available values. For this, the trend of values is identified, and this is used to generate the estimate of the intermediary values. This is nearly the same as imputation, but imputation is done for missing values, and interpolation is done for the values not observed or not reported. Imputation is done by using a statistical method that takes care of **sampling fluctuations**; interpolation is pure mathematics with no consideration of sampling.

Consider a simple example of a child's height of 84 cm at age 2 years and 94 cm at age 3 years. What is your best guess of height at age 2½ years? The most obvious answer is (84 + 94)/2 = 89 cm since 2½ is exactly in the middle of 2 and 3 years. This is the *linear* interpolation in the sense that it is based on the assumption that height grew at a uniform rate between 2 and 3 years. This is nearly so between these two ages—thus, this interpolation will not be wrong by too much. If a provision is made for slightly higher growth in the first 6 months than in the last 6 months of age between 2 and 3 years, the trend would need to undergo the revised interpolation. This might be, say, 89.3 cm. *Linear* interpolation of height at age

27 months would be 84 + 3 × (94 − 84)/12 = 86.5 cm, which again uses the assumption that the height increase of 10 cm during this period of 12 months is uniform—thus, 2.5 cm in 3 months. If that were not so, the interpolation would not be linear.

Interpolation can be used for creating 3-D anatomical models on the basis of images at different cross-sections, say, of ablation of an unresectable liver tumor [1]. Fu et al. [2] used interpolation to generate data on smoking and drinking at smaller geographical levels in India from survey data at a larger level. In both these studies, the algorithm is not as simple as the linear one we used in our illustration, but they illustrate how interpolation can be used in health and medicine.

1. Spinczyk D. Preparing the anatomical model for ablation of unresectable liver tumor. *Wideochir Inne Tech Malo Inwazyjne* 2014 Jun; 9(2):246–51. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4105683/
2. Fu SH, Jha P, Gupta PC, Kumar R, Dikshit R, Sinha D. Geospatial analysis on the distributions of tobacco smoking and alcohol drinking in India. *PLoS One* 2014 Jul 15;9(7):e102416. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4099149/



## interquartile range, *see* variation (measures of)

## interrater reliability

Interrater reliability is the near-uniform assessment of the variable in question by multiple observers when done in identical conditions. *Rater* is a general term and can be used for observers, assessors, clinicians, laboratories, instruments, or any such entity that is used to measure the characteristics of the subjects. Thus, this has many names—interobserver reliability, interlaboratory reliability, interinstrument reliability, etc. However, human beings are particularly prone to errors and subjectivity, and interrater reliability has special significance when the observers are human beings. The terms **agreement** and **concordance** are also used for reliability depending on the context.

The score assigned by different examiners on papers submitted by students is a very convincing example of interrater reliability. When the same answers are scored by different examiners, they should ideally come up with the same scores for each student, and the differences across examiners should be minimal, if any. Similarly, when the same patient is assessed by three clinicians, blind to the assessment done by the others, they are reliable if all three reach the same conclusion. When aliquots of a blood sample are sent to four different laboratories, they should come up with the same values when using the same method or same instrument.

Assessment of agreement between two observers, two methods, etc. for quantitative measurements is discussed under the topic **agreement**. The term *interrater* is generally used when there are more than two observers. Much research uses multiple observers, and it is desirable that they are reliable so that the differential results cannot be ascribed to interrater disagreements. A separate study may be needed to assess reliability, in which all the observers are asked to measure the same subjects. If they are not reliable, focused training on this issue may help. Once their reliability is established, they can be used in research with confidence.

Interrater reliability for qualitative measurements is measured by **Cohen kappa** and for quantitative measurements by **intraclass correlation coefficient (ICC)**. Both of these are described in this volume under the respective terms along with how these values should be interpreted. The conditions under which they provide valid assessment are also stated. For qualitative measurements, **Bangdiwala *B*** can also be used in the case of **ordinal categories**. An account of various measures of interrater reliability has been provided by Hallgren [1].

High interrater reliability does not assure **validity**—high reliability can occur when all observers provide a better rating or a poorer rating than deserved, but it certainly increases confidence in the results. Low reliability indicates large measurement errors, whose effect can be low statistical **power** in a testing-of-hypothesis setup. Low reliability can occur due to poor scale, poor training, pliant characteristics of the subjects, variable response, etc. You can think of deleting the variable with low interrater reliability from your research, or replacing it with the more reliable ones.

1. Hallgren KA. Computing inter-rater reliability for observational data: An overview and tutorial. *Tutor Quant Methods Psychol* 2012;8(1):23–34. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/#!po=50.0000

## interval estimate, *see* confidence intervals (the concept of)

## interval scale, *see* scales (statistical)

## intervention studies, *see also* clinical trials, experimental studies

An intervention study is the one where the interest is in finding the effect of an intentional intervention on some specified outcome. Only human or man-made interventions are considered in this kind of study, and natural interventions, such as disasters and pollution, do not come into this fold. Contrast an intervention study with an **observational study** where naturally occurring events are studied.

The effect of an intervention could be adequately estimated only when the study is carried out in two groups—one with intervention and the other without intervention. When more than one intervention or grades of intervention are under study, accordingly more groups would be needed. These are called **arms** of the study. The average difference in the outcome in the two arms provides an estimate of the effect of intervention in one compared with the other. Efforts are made in an intervention study to control all other factors that can influence the outcome so that whatever effect is observed can be legitimately ascribed to the intervention. This is done by adopting a design that includes equivalent subjects in different arms. When controlling all factors is not feasible, as would happen in most situations, some are controlled by design as much as possible, and the effect of others is extricated by statistical methods such as regression.

Intervention studies in health and medicine are basically of two types. First are laboratory experiments such as on animals and on biological specimens, and the second are **trials** on patients and healthy subjects. Both these can be of several types. For example, trials can be for preventive and promotive intervention and can be done in communities, but most challenging are **clinical trials** done in clinics for assessing the efficacy and safety of newly devised regimens. For details, see the respective topics in this volume.

## interview data/survey, *see also* interviewing techniques

Interview is one of the three major methods of collecting medical data. The other two are examination and (laboratory and radiological)

investigations. Among minor methods are observations and records. All these methods complement each other in presenting a unified picture. However, generally, only interview is used in mass surveys. An example of an exception is the Health Interview and Examination Surveys carried out periodically in the United States and some other countries on a mass scale, where examination is also used.

Interview seems like the easiest and most inexpensive method of collecting health information. It is invasive of the time of the respondent but otherwise is a harmless procedure. The same cannot be said about examination and investigations. However, the information obtained by interview is the least reliable also. Interviewers are seldom trained to elicit exact information, and the respondents tend to vary their response depending on their mood, the context, and their reaction to the interviewer's approach. People tend to hide sensitive information such as on sexual abuses and injuries with legal implications. Even for information as benign as height and weight, there are instances when height is overstated and weight understated so that the body mass index on the basis of interview turns out substantially low compared with that obtained by actual measurements [1]. Interview-based measurements should not be used unless compelled by the circumstances. Results based on interview data can rarely be believed to be as good as those obtained by objective measurements. However, for aspects such as opinion, aptitude, and history, interview is just about the only alternative. Interview surveys carefully planned and carried out can yield useful information, and repeat surveys can point toward the trends (see, e.g., Ref. [2]).

Many large-scale surveys use interview as the sole method of collecting information. Demographic and Health Surveys carried out in many developing countries from time to time are an example of this methodology. These surveys have been very effective in delineating such diverse aspects of health as neonatal mortality [3] and knowledge and awareness about sexually transmitted diseases [4]. Recently, the Global Adult Tobacco Survey has been carried out in several countries to find the extent of use of various tobacco products in different segments of the population. This also is entirely interview based.


1. Acevedo P, López-Ejeda N, Alférez-García I et al. Body mass index through self-reported data and body image perception in Spanish adults attending dietary consultation. *Nutrition* 2014 Jun;30(6):679–84. http://www.nutritionjrnl.com/article/S0899-9007(13)00508-X/abstract
2. Hodge A, Firth S, Marthias T, Jimenez-Soto E. Location matters: Trends in inequalities in child mortality in Indonesia. Evidence from repeated cross-sectional survey. *PLoS One* 2014 Jul 25;9(7):e103597. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4111602/
3. Neupane S, Doku DT. Neonatal mortality in Nepal: A multilevel analysis of a nationally representative. *J Epidemiol Glob Health* 2014 Sep;4(3):213–22. http://linkinghub.elsevier.com/retrieve/pii/S2210-6006(14)00019-7
4. Hossain M, Mani KK, Sidik SM, Shahar HK, Islam R. Knowledge and awareness about STDs among women in Bangladesh. *BMC Public Health* 2014 Jul 31;14(1):775. http://www.biomedcentral.com/content/pdf/1471-2458-14-775.pdf


## interviewer bias, *see* bias in medical studies and their minimization

## interviewing techniques

Interviewing techniques comprise the ways to elicit a response from the subjects of interview. A good technique will ensure that correct and complete responses are obtained from all the subjects. This includes the method of approaching the subject, enlisting his/her confidence, establishing rapport, motivating him/her to devote time, convincing him/her of the utility of the interview, etc. The technique becomes important for health surveys, where sometimes, a person is required to answer a long series of questions. In a clinic, too, a patient is invariably asked about complaints, past history, health behavior, willingness to undergo certain medical or surgical procedures, etc. Unless correct and complete information is obtained, the data analysis, howsoever immaculate, is not going to yield valid results.

A cardinal property of a good interview technique is that the differences in the answers by different subjects are genuine and not ascribable to the language or content of the question, how you put the question, understanding of the subject, the length of the interview, inability to recall, unwillingness to answer, bias of either the interviewer or of the subject, etc. This is easy to accomplish, but errors can be minimized by adhering to the following simple rules.

Make sure that the instrument (**questionnaire/schedule**) is fully tested and standardized for the kind of subjects proposed to be interviewed. Sensitive questions such as on sexual encounters and involvement in crime may have to be reworded so that they are not embarrassing. Complex questions may have to be broken down into parts for easy understanding and easy questioning. These are not part of the interview technique but are stated here because of their importance in getting the right response. The interviewer should undergo training on techniques that can ensure correct response. This would include (i) conducting the interview in a conducive environment where the respondent feels comfortable; (ii) using the language and mannerisms of the subject proposed to be interview and establishing rapport; (iii) explaining the purpose of the interview and convincing him/her of how is this going to help; (iv) remaining neutral during the entire course of the interview; (v) not tweaking the questions meant to be stated verbatim; (vi) restricting the probing, if needed, to "what do you mean," "tell me more," etc. and not suggesting an answer; (vii) recording pertinent answers as in the case of *open-ended questions* and checking off the answer to *close-ended question* only after the respondent has an opportunity to listen to all possible options; and (viii) not passing judgment. These look like simple rules but could be difficult to follow in practice. Try to follow them as much as possible to get the correct response. Remember that you may never have another opportunity to get the right answers, and whatever answers are recorded will be analyzed as they are. If the data are wrong, do not expect right answers from the analysis of survey data.

For more details, see Miller et al. [1] and Fowler and Mangione [2].


1. Miller K, Chepp V, Willson S, Padilla JL. *Cognitive Interviewing Methodology*. Wiley, 2014.
2. Fowler FJ, Mangione TW. *Standardized Survey Interviewing*. Sage, 1990.


## intraclass correlation, *see also* interrater reliability, agreement assessment (overall)

The intraclass correlation coefficient (ICC) is the measure of the degree of consistency or conformity between quantitative elements belonging to the same subgroup. This is especially suited to studying (linear) correlation between twins or other multiple births or between measurements of two eyes or two limbs, or other parts of the body of the same individual. One useful application of ICC is in assessing agreement between different observers and different methods when used on the same set of subjects where the observers and methods can be more than two. This application is discussed

under the topic **agreement** for a pair of measurements. ICC is also frequently used in assessing **interrater reliability**.

The meaning of intraclass correlation should be clear from the dot diagram in Figure I.5. Here, there are five observers (five dots for each subject) and six subjects. Each observer measures each subject (called *fully crossed*). You can see that the values are close to one another in each subject except subject no. 4. Because of similarity of values between observers for most subjects, the ICC would be high. This measures the degree of interrater reliability or agreement or concordance. The degree will further increase in our example if values for the fourth subject are also close to one another, and would be low if the values reported by different observers for most subjects are very different.

The computation of the ICC is slightly different from that of the product–moment **correlation coefficient**. Several variants are available that work in different situations (for example, observers are a random sample or not from a big pool). We are presenting the classical definition, which also seems to be the most acceptable and is applicable when the observers are fixed. The formula is relatively easy for a pair of values compared to multiple values.



intraclass correlation coefficient (a pair of readings):

$$r_I = \frac{2\Sigma_i(x_{i1}-\bar{x})(x_{i2}-\bar{x})}{\Sigma_i(x_{i1}-\bar{x})^2+\Sigma_i(x_{i2}-\bar{x})^2},$$

where

$x_{i1}$ is the measurement on the $i$th subject ($i$ = 1, 2, …, $n$) when obtained by the first method or the first observer,

$x_{i2}$ is the measurement on the same subject by the second method or the second observer, and

$\bar{x}$ is the overall mean of all $2n$ observations.

The denominator in intraclass correlation is different compared with the formula of **product–moment correlation**. Also, the mean used in the numerator is the overall mean in place of the subject mean.

intraclass correlation coefficient (several readings):

$$r_I = \frac{\Sigma_i\Sigma_{j\neq k}(x_{ij}-\bar{x})(x_{ik}-\bar{x})}{(K-1)\Sigma_i\Sigma_j(x_{ij}-\bar{x})^2}; \quad i=1,2,\ldots,n; \quad j,k=1,2,\ldots,K;$$

where $n$ is the number of subjects and $K$ is the number of observers or the number of methods to be compared. The mean $\bar{x}$ is calculated on the basis of all $Kn$ observations. Both these formulas are, in fact, sample estimates of the corresponding correlation in the population. A correlation of more than 0.75 is generally considered enough to conclude good agreement. A value less than 0.40 generally indicates poor agreement. A negative value of ICC indicates systematic disagreement.

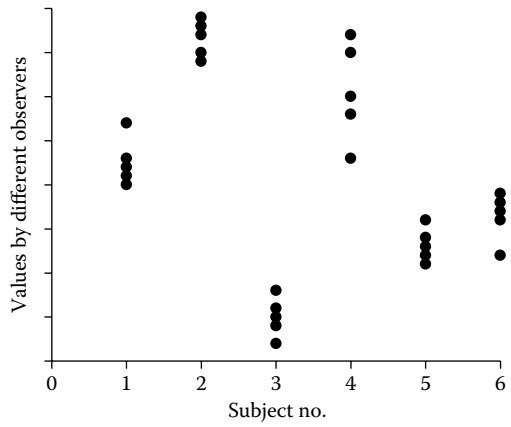



**FIGURE I.5** Illustration of intraclass correlation—values close to one another in each subject except subject no. 4.

The aforementioned formulas are for a fully crossed design. That is, the same set of observers is used for all the subjects. This is also called a two-way setup, as opposed to a one-way setup, in which each subject is assessed by a separate group of randomly drawn observers from a big pool. This is like having a pool of 200 nurses and randomly selecting 3 nurses to score subject 1, another 3 to score subject 2, etc. In the two-way setup also, one possibility is that you select $K$ observers randomly from a big pool, but the same $K$ observers are used throughout. In this case, observers would have a **random effect** because of random selection, and these formulas will not apply. If the observers are fixed and not randomly selected, these formulas are right, but in this case, you will not be able to generalize to any bigger group of observers.

### *ANOVA Formulation and Testing Statistical Significance of ICC*

For testing of significance, we use another formulation of ICC that is based on **analysis of variance** (ANOVA). This is on the premise that similar values across observers for the same subject would yield small within-subjects variance. Sounds reasonable! Between-subjects variance will remain whatever it is. Denote within-subjects variance by $\sigma_w^2$ and between-subjects variance by $\sigma_b^2$. A high value of between-subjects variance relative to the total variance is an indication that within-subjects variance is small, and the values within subjects are similar. This would mean high ICC. Under this formulation, thus,

$$\text{ICC: } \rho_I = \frac{\sigma_b^2}{\sigma_b^2+\sigma_w^2}.$$

This formulation is valid only when the observers are considered a random sample from the "population" of observers. This value of ICC cannot be negative, as opposed to the earlier formula for fixed observers, which can have a negative value. This value of ICC is interpreted as the proportion of the total variance accounted for by the between-subjects variation. Note again that a high value of this ratio would mean small within-subjects variance and, consequently, high intraclass correlation. The results of **one-way ANOVA** give the estimate of the ICC after some algebra as follows:

$$\text{estimated ICC: } \frac{\text{MSB}-\text{MSE}}{\text{MSB}+(K-1)\text{MSE}},$$

where MSB is the between-subjects mean square, MSE is the error mean square (this is the estimate of the within-subjects variance), and $K$ is the number of observers. To understand these **mean squares**, see that topic in this volume. If the observers disagree too much, MSE would be high, and ICC would be low.

This formulation highlights a limitation of ICC. This could be population specific. The between-subjects variance (MSB) could be high in the general population and low in the patients attending a particular clinic as they tend to be similar. Thus, an instrument could be judged reliable in one setup but unreliable in another setup. But there is a positive feature also. If you have five observers in one setup and three in another setup, the ICCs would still be comparable so long as the MSB is the same.

As an example of the use of ICC, consider a nationwide study by Nagler et al. [1] on variability between laboratories performing coagulation tests with identical platforms; prothrombin time, fibrinogen level, and other parameters were measured in a sample of 20 healthy subjects in eight selected laboratories. This variance across laboratories was substantial, and the ICC was low. This led to the conclusion that the standardization from one laboratory to the other is lacking, and efforts are needed to raise the level of standardization of structures and procedures involved in the quantification of coagulation factors.

In case of ICC, the interest would rarely be in the **null hypothesis** that ICC = 0. It would be whether it exceeds a given threshold or not. Thus, mostly, $H_0$: $\rho = \rho_0$ and the alternative $H_1$: $\rho > \rho_0$ (one-sided), and this can be easily tested by

criterion for testing ICC:

$$F = \frac{1+(K-1)\rho_0}{1-\rho_0} \times \frac{\text{MSB}}{\text{MSE}} \quad \text{with} \quad [(n-1), n(K-1)] \text{ df's.}$$

Use ***F*-distribution** to get the ***P*-value**. If it is less than the predetermined level of significance (say, less than 0.05), reject the null in favor of the alternative hypothesis; otherwise, do not.

For more details of ICC, see Shrout and Fleiss [2].

1. Nagler M, Bachmann LM, Alberio L, Angelillo-Scherrer A, Asmis LM, Korte W, Mendez A et al. Variability between laboratories performing coagulation tests with identical platforms: A nationwide evaluation study. *Thrombosis J* 2013;11:6. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3599351/
2. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull* 1979;86(2):420–8. http://www.na-mic.org/Wiki/images/4/4b/Shrout_and_fleiss_ICC.pdf

## intraobserver consistency/reliability, *see also* intraclass correlation

Intraobserver consistency is near equality of values in repeated measurements by one observer of the same characteristic of the same subject at the same point in time. This is also called intraobserver reliability. Repeated measurements can seldom be taken simultaneously at the same point in time in the same subject, but the idea here is measurements within a small period so that the value does not actually change. If an observer is asked to make three readings of diastolic blood pressure in the same subject in the same, say, sitting position after desired rest, the observer will be called consistent when the readings obtained are nearly the same every time. If an observer assesses the extent of burns of the same patient repeatedly within a period of, say 15 min, the answer should be the same. In the case of interpretation of an image, the same image can be given again and again to the same observer without even revealing that this is the same image. A good observer will give the same assessment every time even when, in this case, the time gap is large. The concern here is with the quality of the observer, but this can be extended to the quality of the tools or instruments or quality of laboratories. For example, when four aliquots of the same blood sample are blindly sent to the same laboratory for, say, finding the total cholesterol level, the results of all four should match. If the differences are wider than expected due to minor random variation, the quality of the laboratory is under suspicion.

Statistically, consistency means low variability, and **variance** is the natural criterion for assessing consistency in quantitative values. However, this measure heavily depends on the unit of measurement and the number of subjects. This also is easily distorted by a few outliers. A measure that is relatively independent of these ills is the correlation coefficient among the values reported by the observer on the same subject. This is described under the topic **intraclass correlation coefficient (ICC)**. In Figure I.5, substitute "observers" for "subjects" on the horizontal axis and values for different observers on the vertical axis, and you get an intraobserver consistency setup. Basically, these two setups are the same. Kraeutler et al. [1] used ICC for assessing intraobserver reliability of the radiographic diagnosis and treatment of acromioclavicular joint separations.

For qualitative measurements, intraobserver consistency is measured in terms of **Cohen kappa**. If you want to see an example of its actual use, see the work of Nepple et al. [2], who used Cohen kappa for assessing the intraobserver reliability of arthroscopic classification of acetabular rim labrochondral disease.

1. Kraeutler MJ, Williams GR Jr, Cohen SB, Ciccotti MG, Tucker BS, Dines JS, Altchek DW, Dodson CC. Inter- and intraobserver reliability of the radiographic diagnosis and treatment of acromioclavicular joint separations. *Orthopedics* 2012 Oct;35(10):e1483–7. http://www.healio.com/orthopedics/journals/ortho/2012-10-35-10/%7B53a697e2-1144-4aba-845e-44c099a278c6%7D/inter-and-intraobserver-reliability-of-the-radiographic-diagnosis-and-treatment-of-acromioclavicular-joint-separations
2. Nepple JJ, Larson CM, Smith MV, Kim YJ, Zaltz I, Sierra RJ, Clohisy JC. The reliability of arthroscopic classification of acetabular rim labrochondral disease. *Am J Sports Med* 2012 Oct;40(10):2224–9. http://ajs.sagepub.com/content/40/10/2224.long

## inverse association

The term inverse association is used in several senses. (i) The first is when an increase in one is accompanied by a decrease in the other. Age and lung functions in adults are inversely associated. Since both are quantitative, the term generally used is inversely *correlated* in place of *associated*. This is the same as negative **correlation**. (ii) The second is regarding the association between the chances and not the values. Alzheimer's disease and cancer have been found inversely associated in the sense that Alzheimer's disease decreases the chance of cancer in a person and cancer decreases the chance of Alzheimer's disease [1]. They are much less likely to occur together. (iii) If one percentage rises out of the total, the other is bound to decline. If the percentage of deaths by communicable diseases decreases, as is occurring in most parts of the world, it is natural that the percentage of deaths by chronic diseases would increase. This stems from the fact that the sum total of probabilities of death by different causes is 1, and a fall in one will correspondingly cause a rise in the others. This is kind of obvious to statisticians but not so much to medical professionals. Under this paradigm, with increasing expectation of life almost everywhere, a rise in the percentage of deaths due to chronic diseases could be considered a welcome sign as less people are dying early due to infections, many of which inflict children.

Mathematically, though, there is a difference between negative and inverse. The negative of 3 is –3, whereas the inverse of 3 is 1/3. This kind of inverse association holds true for **odds ratio (OR)**. For a fourfold table with cell frequencies *a*, *b*, *c*, *d*, OR = *ad*/*bc*. If first row–first column is for presence of the **antecedent** and presence of disease, this is the OR for the presence of antecedent in diseased cases. Its mathematical inverse, *bc*/*ad*, is the OR for the absence of the antecedent in diseased cases. If the OR for positive electrocardiogram (ECG) in myocardial infarction (MI) is 4, the OR of negative ECG in MI cases is 1/4. The association between

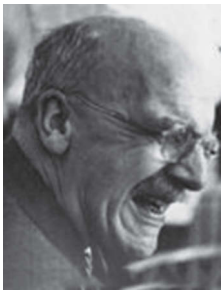
MI and negative ECG is the inverse of the association between MI and positive ECG. This inverse is not in terms of the minus sign but in terms of the mathematical inverse. Similarly, if the **relative risk** (RR) of disease in exposed to nonexposed subjects is *r*, the RR in nonexposed to exposed subjects is 1/*r*. If the RR of oral cancer in regular users of smokeless tobacco to nonusers is 5.8, the RR of oral cancer in nonusers to users is 1/5.8. That is, nonuse of smokeless tobacco is inversely associated with the risk of oral cancer relative to its use.

1. Driver JA, Beiser A, Au R, Kreger BE, Splansky GL, Kurth T, Kiel DP, Lu KP, Seshadri S, Wolf PA. Inverse association between cancer and Alzheimer's disease: Results from the Framingham Heart Study. *BMJ* 2012 Mar 12;344:e1442. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3647385/

## inverse probability, *see* Bayes rule

## inverse sampling

Formally introduced by JBS Haldane in 1945 [1], inverse sampling is the process under which sampling continues till such time that a prefixed number of the outcomes of interest are available. This kind of sampling is advocated for rare outcomes since a fixed sample in this situation may not yield sufficient outcomes of interest for a reliable conclusion.



JBS Haldane

Suppose the interest is in adult cases of epilepsy. This is a rare condition, and even if a sample of few thousand is taken from the general population, or even from those coming to a hospital, the number of epilepsy cases may still be inadequate for any reliable conclusion. The sampling scheme in this situation can be to include any case of epilepsy seen in a clinic, and to continue to do so till, say, 30 cases of epilepsy are available. The number of outcomes of interest is fixed, but the sample size is not fixed in this case and is a **random variable** that depends on chance. You can see that this method of sampling counts the numbers and thus is applicable for events or attributes but not to quantitative outcomes. We cannot say that we will continue to sample till such time that the average direct bilirubin level is between 0.15 and 0.17 mg/dL. But we can say that we will continue sampling till such time that we have 50 cases with a high (say, >0.20 mg/dL) level of direct bilirubin. Such a cutoff converts the quantitative outcome to a qualitative outcome.

Inverse sampling is sometimes advocated for matched **case–control studies**, particularly when there are $K$ (>1) controls per case. In the usual case–control setup, there is a great likelihood that matching on selected predefined characteristics incidentally also matches the exposure under investigation. When matching exposure is found, the pair becomes redundant as it does not contribute to the discrimination under study. In this situation, the sampling may continue till such time that a control subject is found with different exposure. Keogh [2] has discussed this phenomenon and concluded that this inverse sampling offers improved statistical efficiency relative to a comparable study with a fixed number of controls per case.

The usual formulas for **standard errors (SEs)** and other estimates are based on **simple random sampling**. Those will not be applicable to inverse sampling. To work these out for a population proportion $\pi$, we need to consider what is called **a negative binomial distribution**.

Tian et al. [3] have worked out the SE and the confidence interval (CI) of relative risk under inverse sampling. Aggarwal and Pandey [4] compared the estimates of prevalence of leprosy in a population in India obtained by inverse sampling and cluster sampling. They found that the estimates are similar but the SE of the estimate obtained by inverse sampling was higher than obtained by cluster sampling. Thus, this scheme is not as efficient but can still be recommended in a setup where the disease of interest is rare.

1. Haldane JBS. On a method of estimating frequencies. *Biometrika* 1945;33:222–5. http://www.jstor.org/discover/10.2307/2332299?uid=3738256&uid=2&uid=4&sid=21104535588667
2. Keogh RH. Inverse sampling of controls in a matched case–control study. *Biostatistics* 2008 Jan;9(1):152–8. http://biostatistics.oxfordjournals.org/content/9/1/152.long
3. Tian M, Tang ML, Ng HK, Chan PS. Confidence intervals for the risk ratio under inverse sampling. *Stat Med* 2008 Jul 30;27(17):3301–24. http://onlinelibrary.wiley.com/doi/10.1002/sim.3158/abstract
4. Aggarwal A, Pandey A. Inverse sampling to study disease burden of leprosy. *Indian J Med Res* 2010 Oct;132:438–41. http://icmr.nic.in/ijmr/2010/october/14.pdf

## Ishikawa diagram

The Ishikawa diagram is a special type of diagram where potential causes of an outcome or steps of a process are organized in a *fishbone* structure for better understanding. This is also called a *cause–effect diagram*, although its application goes beyond this setup. It helps in sorting the causes or steps in useful categories so that there is no repetition, and the stipulation is that nothing is missed. It can also be used to structure a brainstorming session.

Kaoru Ishikawa

The diagram was devised by Kaoru Ishikawa in 1968 [2]. In Figure I.6 is such a diagram showing the process parameters that affect the critical quality attributes of water precipitation. Note its structure resembling bones of a fish. A new "bone" can be added as and when any new idea comes to mind that deserves to be considered. The length of the "spine" can also be increased when needed. This type of diagram is primarily a quality improvement tool and can be used in health and medicine also. Wong [3] has

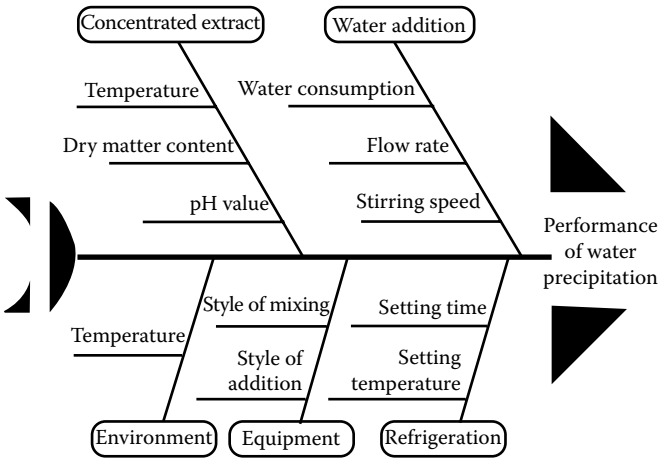



**FIGURE I.6** Ishikawa diagram of water precipitation process. (From Gong X, Chen H, Chen T, Qu H. *PLoS One* 2014 Aug 7;9(8):e104493. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4125280/.)

discussed how Ishikawa diagrams can be used to assist the memory of clinicians while dealing with a case or to improve teaching and training. For example, it can be structured in a manner such that patient issues are on one side of the spine and care-provider issues on the other side. That can help in providing the required focus on different types of issues. One can use boxes or other markers to distinguish main causes from subcauses, as done in Figure I.6.

1. Gong X, Chen H, Chen T, Qu H. Unit operation optimization for the manufacturing of botanical injections using a design space approach: A case study of water precipitation. *PLoS One* 2014 Aug 7;9(8):e104493. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4125280/
2. Ishikawa K. *Guide to Quality Control*. Union of Japanese Scientists and Engineers, 1968.
3. Wong KC. Using an Ishikawa diagram as a tool to assist memory and retrieval of relevant medical cases from the medical literature. *J Med Case Rep* 2011 Mar 29;5:120. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3076252/

## item analysis

Item analysis is a set of statistical techniques used for assessing the efficiency of the items of a test to properly discriminate between the good and not-so-good outcomes. The results of item analysis rate the quality of the test for fair assessment. The analysis can also identify the items that are problematic in fair assessment. This analysis is generally used for developing and assessing the quality of educational tests in classrooms that could distinguish good students from not-so-good students, but the technique has found applications to multi-item medical tests also, for example, to grade disease, to grade quality of life after a surgery, to grade the knowledge regarding a particular health condition, etc. Such tests can be better understood as scales or instruments. For an application, see the work of Compton et al. [1], who have reported item analysis of multiple-choice questions on knowledge regarding mental illnesses.

Severity of disease in a patient can be assessed either by the total score in a test or externally by experts. If this is based on the total score, think of using **tertiles** to divide the subjects into three groups, namely, serious, moderate, and mild. Those who know tertiles will realize that this would classify the top one-third of patients as serious, the middle one-third as moderate, and the bottom one-third as mild. This may or may not actually be so but would serve the purpose of item analysis in most situations. If a large number of patients are available, these can be divided into five groups by **quintiles**. The following methods are described for three groups.

The first step in item analysis is to obtain the "difficulty index" of each item. In educational testing, this would mean the percentage of items answered wrongly. An item wrongly answered by less than 20% of students or more than 80% students is not considered a good item to discriminate between the good student and the poor student. This can be done only if you already know what the correct answer is and can be easily done for multiple-choice questions. Deletion of such items helps in reducing the length of the test, which also is a concern in test development. In medical testing, this translates into questions for, say, health assessment.

Next is the "discrimination index." This is the difference between the proportion of correct responses by serious cases and the proportion of correct responses by mild cases. The closer the value of this index is to 1.0, the more is its discrimination power. For example, if a question on previous history is answered "yes" as often by patients with mild disease as by patients with serious disease, the question has a low discrimination index for severity and is a candidate for deletion from the test.

Third is the **point-biserial correlation** between a patient's response (if dichotomous) on an item and his/her total score. If there are $n$ subjects, you will have $n$ pairs of values. If the item has a quantitative response such as pain score, this would be assessed by the regular product–moment **correlation coefficient**. The higher the correlation, the better the item in the sense that it is scored high for those who have a high score in total. Fourth is the distraction score. This is obtained for multiple-choice questions where some options are used as distracters. More low-score patients should select distracters as their answer than high-score patients.

There are other methods for assessing the quality of a test. These include the **reliability** of the test and its **validity**. They are for the test as a whole and not for individual items. More details of item analysis are provided by Koch [2].

1. Compton MT, Hankerson-Dyson D, Broussard B. Development, item analysis, and initial reliability and validity of a multiple-choice knowledge of mental illnesses test for lay samples. *Psychiatry Res* 2011 Aug 30;189(1):141–8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3156930/
2. Koch JA. *Item Analysis*. BiblioScholar, 2013.

## iterative procedure

An iterative procedure in biostatistics is finding the solution of an equation by trial and error, which requires that improved values are tried every time based on previous trials so that we can get closer and closer to the right solution. Obviously, this procedure is needed for those equations that do not have an explicit solution.

An iterative procedure is commonly followed to find the solution of **maximum likelihood** equations. These equations are derived to find the best estimates that fit the data. For example, sample mean $\bar{x}$ is a maximum likelihood estimate of the population mean μ of a Gaussian distribution. This is an explicit solution of the maximum likelihood equation and needs no iteration. But for the maximum likelihood estimates of the coefficients in **logistic regression**, there is no explicit solution, and we need to iterate to get a solution. In this

case, start with a plausible-looking value, plug it into the equation, find the error, revise the values, plug the revised value, and so on, till such time that the error becomes negligible. Various mathematical algorithms are available that guide us about how to find the error and how to revise the value for the next iteration. Successive iterations in most situations provide a closer solution, but sometimes, the iterations fail to narrow down the error, and it persists. In this case, we say that the iterations fail to *converge*, and no good estimate can be found.

Iteration is a fairly general procedure and has several other applications in the field of health and medicine. For example, iteration is used to reconstruct computed tomography images [1] and for other reconstruction techniques. The same sort of procedure is used in biostatistics, albeit for solving complex equations that do not have direct solutions.

1. Chen M, Mi D, He P, Deng L, Wei B. A CT reconstruction algorithm based on L1/2 regularization. *Comput Math Methods Med* 2014;2014: 862910. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4009238/

# J

## Jaccard dichotomy coefficient, *see* association between dichotomous characteristics (degree of)

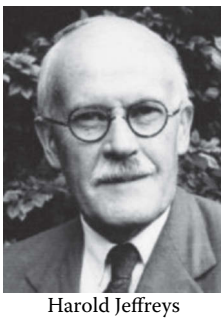
## jackknife resampling, *see* resampling

## Jeffreys interval for binomial π, *see also* Clopper–Pearson bound/interval

The Jeffreys interval is a slight modification of the **Clopper–Pearson interval** for the **binomial** probability π. This is sometimes preferred because this gives an interval with confidence closer to the desired level and has equal tails, whereas the Clopper–Pearson interval has unequal tails. The Jeffreys interval is best expressed in terms of the inverse of the **beta distribution** as

$$\text{Idf.Beta}(\alpha/2, x + ½, n - x + ½), \text{Idf.Beta}(1 - \alpha/2, x + ½, n - x + ½)$$

for confidence level $100*(1 - \alpha)\%$, where $x$ is the number of binomial "successes" out of $n$. This interval is symmetric for $x$ and $n - x$ in the sense that ½ is added to both. This is what tends to make this equal-tailed—each tail with probability $\alpha/2$. On the contrary, in the Clopper–Pearson interval, 1 is added to $n - x$ for the lower limit and to $x$ for the upper limit. Though exact, this kind of continuity correction in the Clopper–Pearson interval tends to yield an interval larger than needed, and this is not symmetric. The Jeffreys interval emerges from the work of Harold Jeffreys that culminated in his publication in 1946 [1] on estimation with invariant form for the prior probability.



Harold Jeffreys

The Jeffreys interval has Bayesian overtones with prior distribution being beta with parameters (½, ½). This is called the *Jeffreys prior.* After observing $x$ successes in $n$ trials, the distribution changes to beta $(x + ½, n - x + ½)$. When $x = 0$, the upper limit remains the same, but the lower limit is set to 0, and when $x = n$, the lower limit remains, but the upper limit is set to 1. This is the practice usually followed for confidence interval (CI) on binomial π because this is a probability that has natural bounds at 0 and 1.

See also **Wilson interval** for another kind of interval for binomial π.

1. Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proc Royal Soc London. Series A* 1946;186(1007):453–61. http://www.jstor.org/stable/97883

## jitter graph/plot

A jitter graph is a plot where overlapping values are shown to be distinct by adding a small amount of noise to the data. In a **scatter diagram**, the values of $y$ on the vertical axis are plotted against $x$ on the horizontal axis. It is possible in your data that $y$ has same value for, say, fours person for the same value of $x$. In this case, these four points will be plotted one over the other, and it would look that this is just one point. Jitter helps to make it clear that these are four points without really disturbing the depiction of the relationship.

Overlapping could happen particularly when the value of $y$ is continuous but rounded off to a convenient number. Consider the birth weight of 1000 babies plotted against their birth order. Birth weight naturally would be recorded to, say, two decimal places. It is possible that 6 out of 1000 babies have birth order 1 and birth weight of 3.59 kg. In the scatterplot, these would be indistinguishable, and all six points will be at the same spot. To make them distinct in the plot, you can consider that these have birth weights (in kilograms) of 3.587, 3.358, 3.589, 3.590, 3.591, and 3.592. Then the overlap will be clear, as they will be slightly visible separately. Note that this jitter is not that much unjustified, because the value 3.59 in fact is a manifestation of values 3.586 to 3.595, as the weight is recorded to the nearest two decimals. We have systematically added 0.001 to the weight, but any random small quantity can be added to achieve jitter.

Consider the data in Table J.1 on blood creatinine level in kidney disease patients of different severity. These are values for a total of 43 patients, but some values are the same within each group. A scatterplot of these values is shown in Figure J.1a, where only 28 distinct values are seen. The other 15 values are overlaps and cannot be seen. After jitter, all 43 values can be seen (Figure J.1b). In this figure, the jitter is applied to the severity of disease since, in this case, slight left or right of mild is still mild, and the case is similar for moderate and serious disease groups. When any one variable is discrete as in this case, it is easy to apply jitter without affecting the meaning of the display.

**TABLE J.1**
**Blood Creatinine Levels in Male Adult Patients with Kidney Disease of Different Severity**

| Severity of Disease | Blood Creatinine Levels (mg/dL) |
|---|---|
| Mild | 1.3, 2.5, 1.8, 1.7, 1.4, 1.2, 1.8, 2.1, 1.7, 1.9, 2.2, 1.5, 1.7, 1.4, 2.0, 2.5, 1.8, 1.6, 2.0 |
| Moderate | 2.8, 3.2, 2.1, 2.7, 2.2, 2.9, 3.0, 3.4, 3.4, 2.9 |
| Serious | 3.9, 3.2, 3.7, 4.1, 4.5, 4.0, 4.3, 4.4, 4.2, 4.1, 3.7, 3.9, 3.2, 3.7, 4.0, 4.2, 3.8, 3.6, 4.0 |

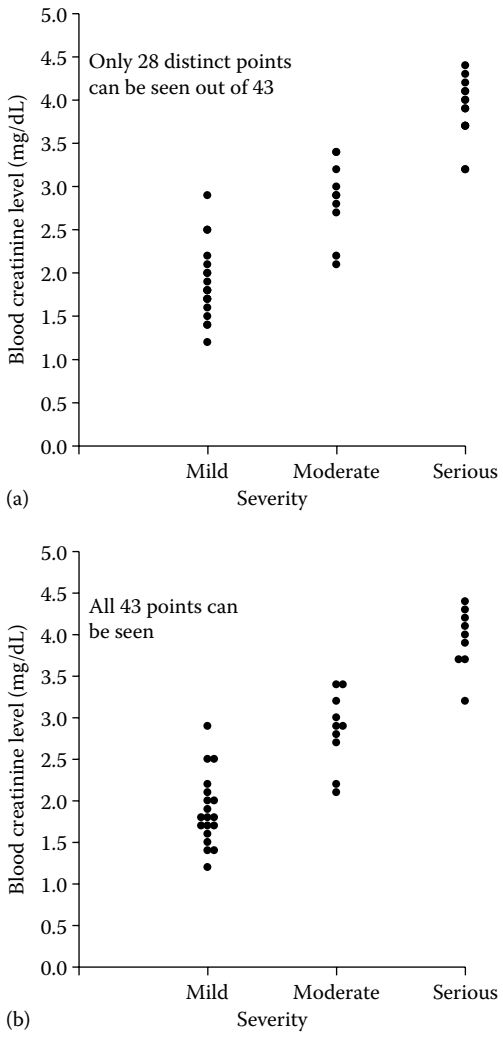



**FIGURE J.1** Plot of the data in Table J.1: (a) without jitter; (b) with jitter.

Jitter may not be needed if both the variables are continuous with a sufficient number of decimals. The need for this kind of adjustment arises generally when at least one variable is discrete. Jitter is only for display and not for analysis of data—the analysis continues to be on the actual values.

## joint distribution, *see* bivariate distributions, multivariate distributions

## J-shaped curve/distribution

A J-shaped curve arises when the values of one variable $y$ gradually decrease for increasing initial values of $x$ and then show a steep increase as the value of $x$ increases after a limit. A popular example

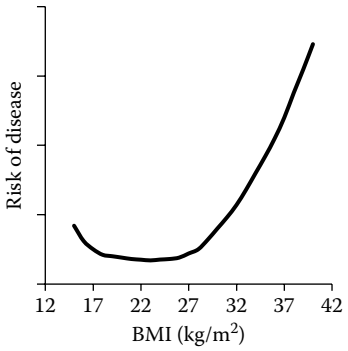



**FIGURE J.2** J-shaped curve between risk of disease and BMI.

is the risk of disease ($y$) and body mass index (BMI) ($x$). The risk of disease is relatively high when the BMI is low, say less than 18 kg/m$^2$; falls as the BMI increases; reaches a bottom when BMI is around 22–23 kg/m$^2$; and then increases as the BMI increases. (Although there is evidence that the risk remains low for BMI between 22 and 27 kg/m$^2$, we ignore this for our example.) This curve is shown in Figure J.2. The relationship between alcohol intake and risk of stroke also follows this pattern in many populations, where risk is slightly higher when no alcohol is taken and low for moderate drinking but steeply increases for heavy drinking. Nove et al. [1] examined maternal mortality in different age groups in 144 countries and noted a J-shaped curve because of high mortality among adolescents, again high at age 40 years or more.

In many applications, a higher value at initial values of $x$ is not seen, but the curve is still called J-shaped, although this is better considered an **exponential curve**. For example, bacterial growth over time, when unhindered, follows this kind of pattern. If the rise is slow with increasing $x$, it can still be possibly represented by an exponential curve by suitably choosing the parameters.

Figure J.2 is a curve and not a distribution, but it becomes a statistical distribution when the $y$-axis is the probability or the frequency or percentage. Thus, if the $y$-axis is the number of subjects and the shape is the same as in Figure J.2, this will be called a J-shaped distribution. This will not happen with BMI, because the number of people with high BMI does not increase as the BMI increases from 30 kg/m$^2$ onward, but can happen with some other variables.

You may occasionally find reference to a *reverse* J-shaped curve or distribution when low values carry much more risk than high values. This has been seen for baseline ankle brachial index and the hazard ratio of death in both men and women in the United States [2]. The authors called it a J-shaped distribution, but it actually is a reverse J-shaped curve.

1. Nove A, Matthews Z, Neal S, Camacho AV. Maternal mortality in adolescents compared with women of other ages: Evidence from 144 countries. *Lancet Glob Health* 2014 Mar;2(3):e155–64. http://www.thelancet.com/journals/langlo/article/PIIS2214-109X(13)70179-7/fulltext
2. Ankle Brachial Index Collaboration et al. Ankle brachial index combined with Framingham Risk Score to predict cardiovascular events and mortality: A meta-analysis. *JAMA* 2008 Jul 9;300(2):197–208. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2932628/

# K

## Kaiser–Meyer–Olkin (KMO) measure

This is an index that compares the magnitude of observed correlations with the magnitudes of **partial correlations** in a multivariate setup, and is used to assess whether or not the data are adequate for **factor analysis**. Partial correlations are the correlations between two variables after removing the linear effect of the other variables. If the variables really share common factors, almost all the partial correlations should be small.

If the partial correlations are nearly equal to the (total) correlations, the value of the KMO index is nearly 0.5. Thus, the value of KMO should be more than 0.5, preferably more than 0.8, to indicate that most partial correlations are sufficiently small. Then, you can expect the existence of common factors as required for a successful factor analysis. KMO is calculated after a **Bartlett test** for sphericity rejects the null hypothesis of no correlations among the variables and assures that some correlations are present among the variables as required for trying factor analysis. If the Bartlett test reveals that significant correlations are not present, there is little justification to proceed with factor analysis.

Also called a measure of sampling adequacy, the KMO was initially proposed by Kaiser [1] in 1970 and modified by Kaiser and Rice [2] in 1974. Gonzáles et al. [3] used this measure for validation of an index of erectile function in Brazil, and Borg et al. [4] used this for a questionnaire to assess treatment outcomes of acetabular fractures in Sweden.


1. Kaiser HF. A second generation Little Jiffy. *Psychometrika* 1970;35:401–15. http://link.springer.com/article/10.1007%2FBF02291817#page-1
2. Kaiser HF, Rice J. Little Jiffy, Mark IV. *Educ Psychol Measurement 1974;34:*111–7. http://epm.sagepub.com/content/34/1/111.extract
3. Gonzáles AI, Sties SW, Wittkopf PG, Mara LS, Ulbrich AZ, Cardoso FL, Carvalho Td. Validation of the International Index of Erectile Function (IIFE) for use in Brazil. *Arq Bras Cardiol* 2013 Aug;101(2):176–82. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3998151/
4. Borg T, Carlsson M, Larsson S. Questionnaire to assess treatment outcomes of acetabular fractures. *J Orthop Surg (Hong Kong)* 2012 Apr;20(1):55–60. http://www.josonline.org/pdf/v20i1p55.pdf


## Kaplan–Meier method, *see also* survival curve/function

The Kaplan–Meier method is a method to estimate the survival pattern of a group of subjects where the duration of survival is exactly measured on a continuous scale. On the other hand, the **life table method** is used for the analysis of survival when the survival is available in grouped data form. *Survival duration* is a generic term used for any duration, also popularly called time to event. These durations need special methods because (i) durations generally have highly skewed distribution and (ii) for some subjects, the complete duration is not available because of dropout, which can happen for various reasons, called **censored** data or incomplete segments. The method was developed by Kaplan and Meier in 1958 [1].

To understand this method, consider an example of 100 cervical cancer patients of whom 1 died after 1 month of enrollment, 4 died at 3 months, and 2 died at 7 months. Since 1 died out of 100 alive at 1 month, 4 died out of 99 alive at 3 months, and 2 died out of 95 alive at 7 months, the estimated survival probabilities are 99/100, 95/99, and 93/95, respectively at 1, 3, and 7 months. The denominator keeps on changing as mortality occurs as these are the conditional probabilities based on the persons surviving at the previous time points. The overall survival after 7 months is 93/100, which also can be obtained as 99/100 × 95/99 × 93/95. This product of the conditional probabilities is the basis for the Kaplan–Meier (K-M) estimate of survival probability, where the censored values are ignored after patients are last seen alive.

When the exact duration of survival is recorded, there is no need for an adjustment of the type used in the life table method. Instead, the censored observations are considered only till the time the subjects were last seen alive, and after that, they are ignored. The probability of survival is computed for each observed duration in place of each interval. Survival rate at time $t$ in this case is the proportion surviving longer than $t$. To estimate this, arrange the subjects according to the known duration of survival, including the censored duration. For $T$ distinct time points for which the duration of survival has been recorded, first obtain

$$p_t = \frac{n_t - d_t}{n_t}, \quad t = 1,\ 2,\ldots,\ T,$$

where

- $p_t$ is the proportion surviving the $t$th time point among those who survived the $(t-1)$st time point;
- $n_t$ is the number of subjects at risk of death at the $t$th time point, i.e., those who are still being followed up (this excludes those with incomplete segments, i.e., $n_{t+1} = n_t - c_t - d_t$)
- $c_t$ is the number of subjects with incomplete segments at time point $t$
- $d_t$ is the number who died at the $t$th time point

Then the estimated proportion surviving the $t$th time point is

Kaplan–Meier survival function at time $t$:

$$s_t = p_t p_{t-1} \ldots p_2 p_1;\ t = 1, 2, \ldots, T.$$

This is the estimated survival function in this case. It is computed for each unique time point. If a time point is applicable to two or more subjects, it is counted only once. This method requires the calculation of as many survival rates by the product rule given in the equation just mentioned as there are events. Hence, the K-M method is also called the **product limit** method. The larger the $T$, the smoother the survival curve. The K-M method for estimating survival probability at the last time point is usually very unreliable because of heavy censoring toward the end of the trial. The following example illustrates the method.

Consider the following data on survival time of 15 patients following radical mastectomy for breast cancer. The study started in January

**TABLE K.1**
**Illustration of Kaplan–Meier Method for Calculation of Survivors**

| Patient No. | Months of Survival | Serial Number of the Time Point $t$ | No. of Patients at Risk (at the Beginning of Time Point $t$) $n_t$ | Deaths at the $i$th Time Point $d_t$ | Number of Survivors $n_t - d_t$ | Proportion Survived $t$th Time Point $p_t$[a] | Proportion Survived since the Beginning $s_t$[b] | Number with Incomplete Segments $c_t$ | Number of Known Survivors $n_t - c_t - d_t = n_{t+1}$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 6 | 1 | 15 | 1 | 14 | 14/15 = 0.933 | 0.93 | 0 | 14 |
| 2 | 8 | 2 | 14 | 1 | 13 | 13/14 = 0.929 | 0.87 | 0 | 13 |
| 3, 4, 5 | 20+ | 3 | 13 | 2 | 11 | 11/13 = 0.846 | 0.73 | 1 | 10 |
| 6 | 24+ | 4 | 10 | 0 | 10 | 10/10 = 1 | 0.73 | 1 | 9 |
| 7 | 25+ | 5 | 9 | 0 | 9 | 9/9 = 1 | 0.73 | 1 | 8 |
| 8 | 30+ | 6 | 8 | 0 | 8 | 8/8 = 1 | 0.73 | 1 | 7 |
| 9 | 35+ | 7 | 7 | 0 | 7 | 7/7 = 1 | 0.73 | 1 | 6 |
| 10, 11 | 37 | 8 | 6 | 2 | 4 | 4/6 = 0.667 | 0.49 | 0 | 4 |
| 12 | 38+ | 9 | 4 | 0 | 4 | 4/4 = 1 | 0.49 | 1 | 3 |
| 13 | 40+ | 10 | 3 | 0 | 3 | 3/3 = 1 | 0.49 | 1 | 2 |
| 14 | 42 | 11 | 2 | 1 | 1 | 1/2 = 0.50 | 0.24 | 0 | 1 |
| 15 | 45+ | 12 | 1 | 0 | 1 | 1/1 = 1 | 0.24 | 1 | 0 |

*Note:* "+" indicates that the duration of survival is at least this much (incomplete segments) for at least one patient.

[a] $p_t$ as in the formula given in the text.

[b] $s_t$ as in the formula in the equation in the text (for example, $s_3 = (11/13) \times (13/14) \times (14/15) = 0.73$).

2012 and continued till December 2015. Thus, the maximum follow-up was 4 years. However, many patients joined the study after January 2012 as and when radical mastectomy was performed.

Survival time (months):

6 8 20 20 20+ 24+ 25+ 30+ 35+ 37 37 38+ 40+ 42 45+

where "+" means that the patients are lost to follow-up or not followed up after this period. They are the incomplete segments. Their exact survival time is not known, but it is at least as many months as shown. The patients are deliberately ordered by survival time. This order is not important but makes the presentation simple. The sample size is small, and there are many incomplete segments. Both are undesirable for **survival analysis**, but the data are still adequate to illustrate the method.

The patients with incomplete segments do not contribute to the calculation of $s_t$ after they are lost. Thus, the rows such as those corresponding to duration 24–35 months in this case can be deleted, although Table K.1 has these rows for completeness. Note how $p_t$ is calculated for survival period of 20 months where there are two deaths and one incomplete segment. At this point, of the 13 patients, 11 survived and 2 died, and the patients with incomplete segments are not counted in this calculation. Thus, $p_3 = 11/13$. The other calculations are similar. The survival curve is shown in Figure K.1. The curve obtained by the life table method is also shown for comparison.

Since time is continuously observed in this setup, many deaths at one time point are shown by a stepladder. The median duration corresponding to 50% survival (0.5 on the $y$-axis) is nearly 37 months by this method. This is very different from the 43 months arrived at if the life table method is used.

Both the life table and K-M methods use all the censored durations, but the life table method assumes half intervals for censored observations. Thus, of the two, the K–M method is considered a better method. In any case, as you can easily appreciate, observation of values in intervals rather than exact values implies loss of information.

The validity conditions of K-M are that (i) censoring is unrelated to prognosis or survival in comparison with the other subjects in the group, (ii) the chance of survival of subjects enrolled early in the study is the same as that of those recruited late, and (iii) duration is exactly recorded and not in intervals.

If you want to see the K-M method in action, there are a large number of articles in the literature. One among them is by Takeda et al. [2], who used the K-M method to study the duration of continuation with a dorzolamide/timolol fixed-combination ophthalmic agent regimen in glaucoma patients who had been treated earlier with monotherapy.

A modification of the K-M method is when the data before reaching a certain landmark are ignored, for example, the events occurring between diagnosis and treatment. This is called **landmark analysis** and is discussed in detail by Dafni [3].


1. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Amer Stat Assoc* 1958;53(282):457–81. http://www.jstor.org/discover/10.2307/2281868?uid=3738256&uid=2&uid=4&sid=21104684031027
2. Takeda S, Mimura T, Matsubara M. Effect of 3 years of treatment with a dorzolamide/timolol (1%/0.5%) combination on intraocular pressure. *Clin Ophthalmol* 2014 Sep 9;8:1773–82. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4164289/
3. Dafni U. Primer on statistical interpretation and methods: Landmark analysis at the 25-year landmark point. *Circulation: Cardiovas Qual Outcomes* 2011;4:363–71. http://circoutcomes.ahajournals.org/content/4/3/363.full.pdf+html


## kappa (statistic), *see* Cohen kappa

## Kendall tau, *see* association between ordinal characteristics (degree of)

## Kolmogorov–Smirnov test

The Kolmogorov–Smirnov (K-S) test is used to find whether observed values fall into a specified **distribution** pattern (one-sample test), or whether the values in two groups follow the same pattern (two-sample test). This test is based on the maximum difference between the cumulative distributions (Figure K.2). This is a nonparametric test that can be used for any distribution, but the variable must be on a continuous scale.

In the case of one sample, this test computes the distances between the observed cumulative relative frequency and the cumulative probability expected under the specified distribution at each observed value. The specified distribution is the **null hypothesis** in this case that we would like to reject. This should be fully specified. If our sample of size $n$ has values ($x_1$, $x_2$, …, $x_n$), the cumulative relative frequency at $x_i$ is the number of values less than or equal to $x_i$ divided by $n$. This would be compared with the corresponding cumulative probability at $x_i$ of the specified distribution, and the difference is obtained. The K-S criterion is the maximum of these differences in absolute value. If the value of the criterion exceeds the value under the null hypothesis at a 5% level of significance, the null is rejected at this level. The values under the null are available in books and statistical software. This is valid for large $n$, but the

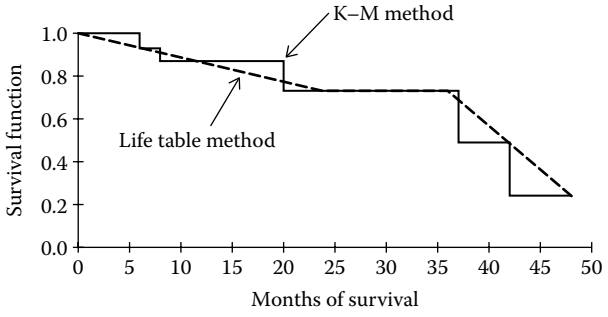



**FIGURE K.1** Survival function of breast cancer cases in our example: life table and K–M methods.

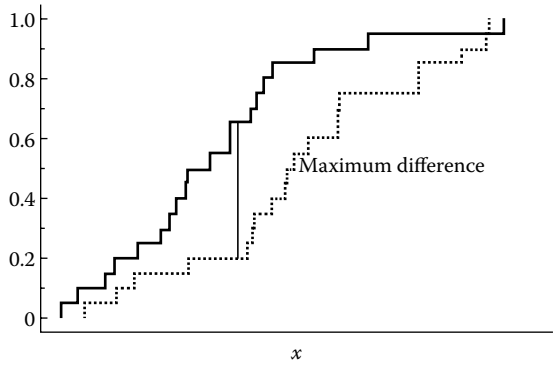



**FIGURE K.2** Maximum difference between cumulative distributions.

exact form is available for certain special cases. In fact, very large $n$ is generally required for a valid K-S test. The criterion was first proposed by Andrey Kolmogorov in 1933 [1], and the values of the criterion under the null were first tabulated by Smirnov in 1939 [2]. These tables became popular only after their publication in 1948 [3].

One practical problem with the one-sample K-S test is that the distribution under the null hypothesis must be fully specified including the values of its parameters. For example, if the specified distribution is Gaussian, its parameters are population mean and population variance, and both must be known. If not known, the usual criterion values under the null cannot be used; instead, they have to be worked out by simulations. When it is known that the distribution under the null is Gaussian, gamma, exponential, or any other, some other parametric test such as the **Anderson–Darling test** may be more powerful in the sense of its ability to detect a difference when really present.

The two-sample analogue of the K-S tests the hypothesis that the two samples come from the same distribution. The groups must be independent for this test. The test computes the differences in the cumulative relative frequencies in the two observed distributions at each observed value in either distribution. The sample sizes can be unequal, but these must be large for K-S test to be valid. The criterion in this case again is the maximum of the differences in the cumulative relative frequencies, as shown in Figure K.2. If the value of this criterion exceeds the value expected under the null (as tabulated by Smirnov), the null is rejected; otherwise, it is not.

The K-S test is conventionally used for uncensored exact continuous (ungrouped) values, but its variations for censored, grouped, or discrete data setups have also been derived [4].

The K-S test is commonly used to find whether the data are violating Gaussianity. This is required for many statistical procedures. For example, Altun et al. [5] used this on serum hepcidin levels and troponin levels in patients with non-ST-elevation myocardial infarction. If the values violate the Gaussian pattern, then nonparametric tests are used. However, just as for almost any statistical test, remember that small samples may not be able to detect a non-Gaussian pattern even if present.


1. Kolmogorov A. Sulla determinazione empirica di una legge di distribuzione. *G Ist Ital Attuari* 1933;4:83–91. URL not available
2. Smirnov N. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull Mathematwue de l'Universite de Moscou* 1939;2:2. URL not available
3. Smirnov N. Table for estimating the goodness of fit of empirical distributions. *Ann Math Stat* 1948;19:279–81. http://projecteuclid.org/download/pdf_1/euclid.aoms/1177730256
4. Pettitt AN, Stephens MA. The Kolmogorov–Smirnov goodness-of-fit statistic with discrete and grouped data. *Technometrics* 1977 May;19(2):205–10. http://www.jstor.org/stable/1268631
5. Altun B, Altun M, Acar G, Kilinç M, Taşolar H, Küçük A, Temiz A, Gazi E, Kirilmaz B. Assessment of serum hepcidin levels in patients with non-ST elevation myocardial infarction. *Anadolu Kardiyol Derg* 2014 Sep;14(6):515–8. http://www.anakarder.com/eng/makale/2845/105/Full-Text


## Kruskal–Wallis test

The Kruskal–Wallis (K-W) test is the nonparametric counterpart of the **one-way ANOVA *F*-test**. This is used to test whether the **central values** such as median in three or more independent groups are equal or not, with the condition that the shape of the distribution is nearly the same in all the groups. The $F$-test is valid only for data in a Gaussian (normal) pattern, but there is no such restriction for the K-W test. Minor violations of Gaussianity are permissible in the case of the $F$-test if the sample size is large, but if the violations are major and the sample size too is small, the K-W test is indicated in place of the $F$-test.

The K-W test was developed by William Kruskal and Allen Wallis in 1952 [1]. This basically is the same as the $F$-test for one-way layout but uses ranks of values instead of the exact values. The use of ranks makes it a nonparametric test and helps to dispense with the condition of Gaussianity.

Consider an example of the cholesterol level in isolated diastolic hypertensives, isolated systolic hypertensives, clear hypertensives, and controls. All subjects are adult females of medium build, the groups are matched for age, and they all belong to the same socio-ethnic group. Thus, the factors that may affect cholesterol level are controlled to a large extent. It is expected that the *pattern* of distribution of cholesterol level in different hypertension groups would be the same but not Gaussian. It is suspected that one or more groups may have measurements higher or lower than the others. The objective is to find whether or not the location (measured by the central values) differences between groups are statistically significant. Because the underlying distribution is not Gaussian, and if, in addition, the number of subjects in different groups is small, the conventional ANOVA cannot be used. The nonparametric K-W test is the right method for such a setup.

Denote the number of groups by $J$, each containing $n$ subjects. For simplicity, let the number of subjects be the same in each group, but that is not a prerequisite. Rank all $nJ$ observations jointly from smallest to largest. Denote the rank of the $i$th subject ($i = 1, 2, \ldots, n$) in the $j$th group ($j = 1, 2, \ldots, J$) by $R_{ij}$. Let the sum of the ranks of the observations in the $j$th group be denoted by $R_{\bullet j}$, i.e., $R_{\bullet j} = \Sigma_i R_{ij}$, If there is no difference in the location of the groups, then $R_{\bullet 1}$, $R_{\bullet 2}$, …, $R_{\bullet J}$ should be nearly equal and their variance nearly equal to 0. The following criterion exploits this premise:

$$\text{Kruskal–Wallis (K–W) test: } H = \frac{12}{nJ(nJ+1)} \frac{1}{n} \Sigma_j R_{\bullet j}^2 - 3(nJ+1).$$

When ties occur (two or more values are the same), the observations are assigned average ranks. In this case, the criterion changes slightly. For details, see Hollander and Wolfe [2]. Standard statistical packages automatically take care of such contingencies.

If the null hypothesis of equality of groups were true, the value of $H$ would be small. The distribution of the criterion $H$ under $H_0$ is known, and the $P$-value corresponding to the calculated value of $H$ for a set of data can be obtained. Again, it is better to leave it to a statistical software package to give the $P$-value. But it has limitations; for example, for three groups, the differences between the groups cannot be statistically significant by the K-W test unless $(n_1 + n_2 + n_3) \geq 7$ and at least two groups have two or more subjects. Software may give exact $P$-values for higher $J$ and higher $n$. If not, for four or more groups or any $n_j \geq 6$, the criterion $H$ in the preceding equation can be approximated by **chi-square** with $(J - 1)$ degrees of freedom (df's). In that case, reject $H_0$ if the calculated value of $H$ exceeds the critical value of chi-square at the desired significance level.

Pursuing the example cited earlier, consider cholesterol level in females with different types of hypertension and controls, as given in Table K.2, where the definition of various types of hypertension is also given. Isolated systolic and isolated diastolic hypertension are no longer considered benign conditions—they are recognized as cardiovascular risk factors. Also given in parentheses in this table are the joint ranks. The last column is the sum of the ranks ($R_{\bullet j}$) for the group. Note the split ranks at the tied values.

**TABLE K.2**
**Cholesterol Level in Women with Different Types of Hypertension**

| Hypertension Group | Total Plasma Cholesterol Level (mg/dL)[a] | | | | | Sum of Ranks |
|---|---|---|---|---|---|---|
| No hypertension—control | 221 | 207 | 248 | 195 | 219 | |
| (DBP < 90, SBP < 140)[b] | (8) | (2.5) | (16) | (1) | (7) | (34.5) |
| Isolated diastolic hypertension | 217 | 258 | 225 | 215 | 228 | |
| (DPB ≥ 95, SBP < 140) | (5) | (17) | (9) | (4) | (11) | (46) |
| Isolated systolic hypertension | 262 | 227 | 207 | 245 | 230 | |
| (DBP < 90, SBP ≥ 150) | (18) | (10) | (2.5) | (15) | (12) | (57.5) |
| Clear hypertension | 218 | 238 | 265 | 269 | 240 | |
| (DBP ≥ 90, SBP ≥ 140) | (6) | (13) | (19) | (20) | (14) | (72) |

*Note:* The hypertension categories are not exhaustive: For example, a subject with DBP = 92 and SBP = 138 mmHg would not be in this study.

[a] Rank of the value in the parentheses.

[b] DBP, diastolic blood pressure (mmHg); SBP, systolic blood pressure (mmHg).

In this case, $J = 4$ and $n = 5$. Using the sum of the ranks in the groups (last column), from the criterion given earlier:

$$H = \frac{12}{5\times 4(5\times 4+1)}\frac{1}{5}(34.5^2+46^2+57.5^2+72^2) - 3(5\times 4+1)$$

$$= 4.4.$$

Since the number of groups is four, we can use the chi-square approximation. A computer package gives $P = 0.2203$. This is not sufficiently small, and therefore, the null hypothesis of equality of locations of the groups cannot be rejected. Note that this conclusion is reached despite major differences in the value of $R_{\bullet j}$ in the four groups. The K-W test reveals that those differences in this example may have arisen from a sampling fluctuation when the groups actually have the same location.


1. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *J Amer Stat Assoc* 1952;47(262):583–621. http://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483441#.VCi_wvmSwWk
2. Hollander M, Wolfe DA. *Nonparametric Statistical Methods*, Second Edition. Wiley-Interscience, 1999.


## Kuder–Richardson coefficient

This is used to test the **reliability** of a set of **binary** responses to items in an instrument such as a questionnaire. The Kuder–Richardson (K-R) coefficient measures the consistency of responses from item to item and determines the degree to which all items measure a common characteristic of the respondent. Thus, this is a measure of the **internal consistency** of an instrument. Reliability of the instrument is important for good inferences from the data. If a correct answer is scored as 1 and an incorrect answer is scored as 0, this is given by

$$\text{Kuder–Richardson coefficient: } r_{\text{KR}} = \frac{K}{K-1}\left(1 - \frac{\Sigma_k p_k(1-p_k)}{s^2}\right),$$

where $K$ = number of questions in the instrument, $p_k$ = proportion of subjects in the sample who answered the $k$th ($k$ = 1, 2, …, $K$) question correctly, and $s^2$ = variance of the *total* scores of all the people taking the test. The value of this coefficient ranges from 0 to 1. A high value indicates reliability, while too high a value (in excess of 0.90) indicates a homogeneous test that may not serve the purpose. This coefficient was developed in 1937 [1] and is known to be affected by the difficulty of the test, the variation in scores, and the length of the instrument.

If a 20-item questionnaire ($K$ = 20) is administered to $n$ = 80 subjects, and if 32 of these 80 answer the 7th question correctly, $p_7 = 32/80 = 0.4$. This can be obtained for each question. Suppose the total score of the first subject is 61, the second 78, the third 49, etc., and the variance of these scores is $s^2 = 37.21$ ($s = 6.1$). If the proportion of correct answers is as in Table K.3, $r_{\text{KR}} = \frac{20}{19}\left(1 - \frac{3.0534}{37.21}\right) = 0.97$. This test is highly homogeneous, and possibly not as good in discrimination. The smaller the variance of total scores, the larger the K-R coefficient.

Hiranyatheb et al. [2] used the K-R coefficient for assessing the reliability of the Thai version of the symptom checklist of the Yale–Brown Obsessive Compulsive Scale, and Kesselheim et al. [3] used it for assessing the reliability of a new test of residents' ethics knowledge for pediatrics in the United States.

The popular **Cronbach alpha** is an extension of the K-R coefficient for quantitative responses in place of binary responses.


1. Kuder GF, Richardson MW. The theory of the estimation of test reliability. *Psychometrika*, 1937;2(3):151–60. http://link.springer.com/article/10.1007%2FBF02288391#page-1
2. Hiranyatheb T, Saipanish R, Lotrakul M. Reliability and validity of the Thai version of the Yale–Brown Obsessive Compulsive Scale—Second Edition in clinical samples. *Neuropsychiatr Dis Treat* 2014 Mar 13;10:471–7. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3958520/
3. Kesselheim JC, McMahon GT, Joffe S. Development of a test of residents' ethics knowledge for pediatrics (TREK-P). *J Grad Med Educ* 2012 Jun;4(2):242–5. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3399620/


## kurtosis

Kurtosis in statistics is the peakedness of a unimodal statistical **distribution** of a continuous variable that also affects the tails. Most

**TABLE K.3**
**Illustration of Calculation of K-R Coefficient**

| Question | Proportion of Correct Answers ($p$, Out of $n$ = 80 Subjects) | $p(1-p)$ |
|---|---|---|
| 1 | 0.2000 | 0.1600 |
| 2 | 0.5625 | 0.2461 |
| 3 | 0.4250 | 0.2444 |
| 4 | 1.0000 | 0.0000 |
| 5 | 0.7000 | 0.2100 |
| 6 | 0.8375 | 0.1361 |
| 7 | 0.6875 | 0.2148 |
| 8 | 0.9750 | 0.0244 |
| 9 | 0.5375 | 0.2486 |
| 10 | 0.7375 | 0.1936 |
| 11 | 0.8250 | 0.1444 |
| 12 | 0.9250 | 0.0694 |
| 13 | 0.4250 | 0.2444 |
| 14 | 0.8875 | 0.0998 |
| 15 | 0.8375 | 0.1361 |
| 16 | 0.9625 | 0.0361 |
| 17 | 0.9000 | 0.0900 |
| 18 | 0.4625 | 0.2486 |
| 19 | 0.6125 | 0.2373 |
| 20 | 0.9250 | 0.0694 |
| **Total** | | **3.0534** |



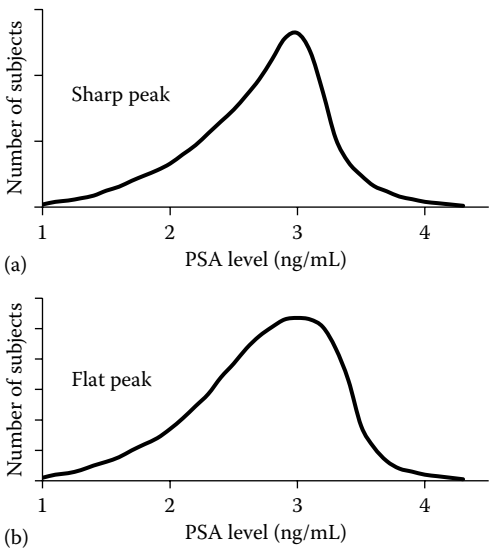



**FIGURE K.3** Distribution of PSA level in two groups: (a) sharp peak; (b) flat peak.

continuous medical measurements have a tendency to concentrate around a particular value, which means that this value is the most common. For example, prostate-specific antigen (PSA) in healthy males is mostly around 3 ng/mL. Consequently, when frequency distribution is plotted, the peak will occur around 3 ng/mL. If there are a large number of people with values around 3 ng/mL relative to the adjacent values, the peak will be steep and sharp, but if there are not as many, the peak will be relatively flat. This peakedness is what is called kurtosis.

Figure K.3 presents a contrast between distributions with a sharp peak and a relatively flat peak. Both the distributions have nearly the same variance and the same **skewness**, but the kurtosis is very different. Statistically, a quantity called the fourth moment of a distribution measures the peakedness. Statistical packages will easily compute this for any distribution when you specify either the values or their **frequency distribution**. This, however, will be reliable for a large sample only.

For a Gaussian (normal) distribution, the value of kurtosis is 3. Generally, 3 is subtracted from the value of kurtosis for any distribution so that a comparison with zero value can be made. After this subtraction, the distribution with near-zero kurtosis is called *mesokurtic*, a flat-peaked distribution will have negative kurtosis (called *platykurtic* distribution), and a sharp-peaked distribution will have positive kurtosis (called *leptokurtic* distribution). Platykurtic will generally have light tails and leptokurtic heavy tails. These are assessed in comparison with the peak in the Gaussian distribution, but that is not required physically. Kurtosis can be calculated numerically, as stated in the preceding paragraph. For example, scores on a **Likert scale** from 0 to 9 will have a platykurtic distribution if a large number of responses are of middling type—between 4 and 6—as is the general tendency with this scale. All **uniform distributions** are platykurtic.

It is possible that the two symmetric distributions apparently look **Gaussian** but have very different kurtosis (Figure K.4). When kurtosis is very different from normal, the requirement of Gaussian distribution as for **Student *t*** and **ANOVA *F*** is not met, and the results would not be valid. Thus, it is desirable that kurtosis is also checked along with skewness, particularly for values such as scores on a Likert scale. Practically, this is seldom done since most distributions tend to have nearly normal kurtosis.

Among the many applications of kurtosis in health and medicine, one is in constructing growth curves for children. Many growth parameters (height, weight, head circumference, etc.) have a leptokurtic distribution at each age—thus, an adjustment is needed. The **Box–Cox power exponential (BCPE)** method has been especially devised to take care of this problem in kurtosis.

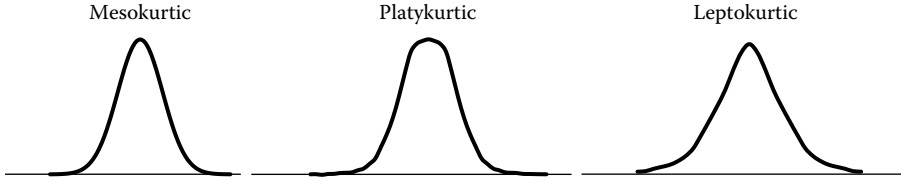



**FIGURE K.4** Symmetric distributions with different kurtosis.

# L

## lag, *see* autocorrelation

## Lan–deMets procedure, *see also* O'Brien–Fleming procedure

The Lan–deMets procedure arises when considering **stopping rules** for clinical trials. An ongoing trial can be stopped for many reasons, two of which are statistical. The first is that the evidence accumulated so far confirms that the regimen has adequate efficacy. The other reason is that the efficacy is found to be too low to go any further. The latter is said to be stopping for futility. Such appraisal has special appeal for phase II trials where the primary aim is to provide proof of concept in the sense that the regimen has minimum efficacy. Stopping for futility will save the researchers from going through the expensive phase III trial. On the other hand, stopping for efficacy can be applied to both phase II and phase III trials but has special appeal to phase III as it can save substantial time and resources.

The procedure for stopping for futility becomes intricate because this is done in a manner that the prefixed **level of significance** and statistical **power** remain unaltered. Software-based intricate calculations are required to ensure this. Stopping for efficacy requires that the level of significance $\alpha$ is judiciously apportioned at each appraisal such that the total **Type I error** does not exceed $\alpha$. A simple but inadequate approach is to spend $\alpha$ in $K$ equal parts if $K$ appraisals are planned, including the final analysis of all the data. This means that the hypothesis at first appraisal is tested at an $\alpha/K$ level of significance, second at a $2\alpha/K$ level of significance, and the last at an $\alpha$ level of significance. Statistically, equal-$\alpha$ spending procedure as just stated is too liberal than required to control Type I error and would relatively easily reject the null. It sounds reasonable to have a procedure that is even more stringent at initial stages. This is done in the group sequential approach to interim statistical testing, but the difficulty is that it includes the number of scheduled analyses. This number must be determined before the onset of the trial, and there should be equal spacing between scheduled analyses with respect to patient accrual. The *alpha-spending function* approach was developed to overcome these problems. One such popular procedure is as follows for a two-tailed test:

$$\alpha\text{-Spending Function: } \alpha(\tau_k) = 2\,[1 - \Phi(z_{\alpha/2}/\sqrt{\tau_k})],$$

where $\Phi$ is the cumulative Gaussian probability at the value specified within the parentheses and $\tau_k$ is the proportion of the information available at the $k$th appraisal. This may appear to be a complex expression but that really is not the case. For example, one-half of the subjects completing one-half of the trial corresponds to $\tau_k = 0.25$, and since $z_{\alpha/2} = 1.96$ for a two-sided Gaussian test at the 0.05 significance level, $z_{\alpha/2}/\sqrt{\tau_k} = 1.96/\sqrt{0.25} = 3.92$ at this $\tau_k$. At this value, $(1 - \Phi) < 0.00005$ from the Gaussian distribution and $\alpha(\tau_k) < 0.0001$. Thus, the critical value of 3.92 for rejecting the null corresponds to a nominal $P$-value of less than 0.0001 at this stage. This is the critical value with this procedure at one-fourth of the trial and assumes that the trend of results later in the trial would be on the same pattern as accrued so far.

**TABLE L.1**
**Lan–deMets α-Spending Function for Unequally Spaced $K = 3$ Appraisals**

| Appraisal No: $k$ | $\tau_k$ | Critical Value | Nominal $P$-Value from Gaussian Table (One-Tailed Test) | Adjusted $P$-Value |
|---|---|---|---|---|
| 1 | 20% | 2.9626 | 0.0015 | 0.0015 |
| 2 | 50% | 2.2682 | 0.0117 | 0.0107 |
| 3 | 100% | 2.0302 | 0.0212 | 0.0128 |
| | | | **Total** | **0.0250** |

A more acceptable procedure is attributed to Lan–deMets [1,2], which is flexible and accommodates unequally spaced appraisals. This can be used for equally spaced sequential designs as well. This preserves the overall Type I error regardless of timing of the appraisals but makes it difficult to stop the trial early unless there is a strong evidence of the desired efficacy. This also imposes a small penalty at the end for interim looks. A suitable software package is needed to find the Lan–deMets critical values for specified appraisals. For example, if there are three appraisals at 20%, 50%, and 100%, the critical values for one-tailed $\alpha = 0.025$ are given in Table L.1. The total of adjusted $P$-values is 0.025, and the critical value at last appraisal at completion of the trial is 2.0302 in place of the usual 1.96, indicating a slight penalty for conducting the previous two appraisals.

No matter which method is used, there is a possibility of bias in early stopping. If you plan $n = 400$ subjects in a trial and stop it after analysis of data on 150 subjects because the data tell you that strict significance has been reached, the question arises as to whether these 150 subjects are as representative of the population as 400 would have been. If the first 150 subjects are not random, further bias is apparent. For details, see Chow et al. [3].


1. DeMets DL, Lan KK. Interim analysis: The alpha spending function approach. *Stat Med* 1994;13:1341–52. http://eclass.uoa.gr/modules/document/file.php/MATH301/PracticalSession3/LanDeMets.pdf
2. DeMets DL, Lan KK. The alpha spending function approach to interim analysis, in: *Recent Advances in Clinical Trial Designs and Analysis* (Ed. Thall PF). Kluwer, 1995. ftp://maia-2.biostat.wisc.edu/pub/chappell/641/papers/paper35.pdf
3. Chow S-C, Wang H, Shao J. *Sample Size Calculations in Clinical Research*, Second Edition. Chapman & Hall/CRC Press, 2007.


## landmark analysis

This analysis ignores data accruing either before or after a landmark event when full data are anticipated to be highly unreliable because of wavered responses. Thus, this is a conditional analysis. Landmark

analysis is generally applied to durations such as for **Kaplan-Meier** analysis that models any duration from occurrence of an event to occurrence of another event (including termination of follow-up). This duration is generally referred to as duration of survival, but it can be any duration such as duration of hospital stay, time elapsed since the last infarction, or simply the duration of survival after onset of a terminal disease. The model can give unreliable, sometimes erroneous, results if the behavior of patients is too erratic for a part of the duration. For example, patients may show highly variable results initially for some days after a particular surgery, but stable results thereafter. In such cases, it is prudent to model the duration after the response stabilizes. This is called the landmark analysis—the time point of stabilized response being the landmark. Note that stabilizing time point may vary from patient to patient. The duration under landmark modeling could be for something like post-remission survival, relapse after hospital discharge, and the time since the treatment compliance stabilizes after some initial wavering. In some situations, this wavering may mean patient switching group membership such as responders becoming nonresponders, as is quite common in cancer treatment. In the landmark analysis, the data accruing only from a landmark are considered and the data preceding the landmark are ignored. This landmark could be at the end point as well, in which case the data accruing after the landmark are omitted. The method was introduced by Anderson et al. [1] in 1983.

For landmark analysis to provide unbiased results, it is imperative that the landmark is decided before data collection on the basis of clinical considerations since data-driven landmark may produce erroneous results. If the landmark is too early, unreliable observations may still creep in, and if it is too late, a significant proportion of events may be omitted, causing loss of statistical power. Some events occurring early in some patients may be important markers of survival but would not be considered in landmark analysis. If the patients were randomized to begin with, they may not remain randomized after some events are deleted, and follow-up time available for analysis also decreases with this method. In addition, there is always a risk of loss of generalizability of results with this analysis on truncated data since extrapolation of the survival pattern will ignore the omitted time period. On the whole, we can say that this method has promise but is yet to gain maturity, although if the landmark time is early enough, the analysis may provide useful results. For further details and intricacies of the method, see Dafni [2].

The following two examples illustrate situations where landmark analysis could be useful. In comparison of the long-term mortality of acute ST-segment elevation myocardial infarction (STEMI) and non–ST-segment elevation acute coronary syndrome (NSTE-ACS) patients after percutaneous coronary intervention, Ren et al. [3] found no difference in all-cause mortality for both STEMI and NSTE-ACS between 6 months and 4 years of follow-up. This conclusion was based on landmark analysis. With landmark analysis at 12 months, Kumar et al. [4] observed that progression-free survival and overall survival continued to remain superior for patients attaining complete response after autologous stem cell transplant in multiple myeloma.


1. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. *J Clin Oncol* 1983;1:710–9. http://jco.ascopubs.org/content/1/11/710.abstract
2. Dafni U. Landmark analysis at the 25-year landmark point. *Circulation: Cardiovascular Quality and Outcomes* 2011;4:363–71. http://circoutcomes.ahajournals.org/content/4/3/363.full
3. Ren L, Ye H, Wang P, Cui Y, Cao S, Lv S. Comparison of long-term mortality of acute ST-segment elevation myocardial infarction and non-ST-segment elevation acute coronary syndrome patients after percutaneous coronary intervention. *Int J Clin Exp Med* 2014 Dec 15;7(12):5588–92. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4307524/
4. Kumar L, Iqbal N, Mookerjee A, Verma RK, Sharma OD, Batra A, Pramanik R, Gupta R. Complete response after autologous stem cell transplant in multiple myeloma. *Cancer Med* 2014 Aug;3(4):939–46. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4303161/


## LASSO, *see* least absolute shrinkage and selection operator (LASSO)

## latent variables

In contrast to a manifest variable, which can be measured directly, a latent variable cannot be measured directly, such as happiness, satisfaction, racial prejudice, and social class. Depending on the context, latent variables are also referred to as *underlying variables* and sometimes *frailties*. Other kinds of latent variables include hypothetical **constructs** that are tied to less concrete aspects of what is being studied. Sometimes, **random effects** are also considered latent.

In some medical studies, latent variables are an important component that needs to be understood and delineated. Since they are hidden, inference on latent variables has to be indirect, requiring statistical models connecting latent variables with the observed variables. This gives rise to **latent variable models** such as **factor analysis** and **path analysis**.

A latent variable may not be specified because of practical considerations. Since they are difficult to measure, these are sometimes perceived as redundant, or may be considered constants that do not need to be observed. However, in some situations, these variables are essential to draw any inference worthy of execution. They permit us to describe relations among a class of events or variables that share something in common, rather than making concrete statements regarding the relation between specific variables [1].

In a study on adulthood personality correlates of childhood adversity, Carver et al. [2] reduced personality-related scales to four latent variables and termed them anger/aggression, extrinsic focus, agreeableness, and engagement. Note that these are unobservable variables but are useful for this kind of study. Zhao et al. [3] investigated the ability of topic modeling to reduce high dimensionality to a small number of latent variables, which makes it suitable for the clustering of large medical data sets. In this example, latent variables serve as a means to reduce dimensionality without giving them specific names.

A good reference for exploring latent variables is the work of Babones [4].


1. Bollen KA. Latent variables in psychology and the social sciences. *Ann Rev Psychol* 2002;53:605–34. http://www.unt.edu/rss/LatentVariablesBollen.pdf, last accessed January 11, 2015.
2. Carver CS, Johnson SL, McCullough ME, Forster DE, Joormann J. Adulthood personality correlates of childhood adversity. *Front Psychol* 2014 Nov 21;5:1357. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4240049/
3. Zhao W, Zou W, Chen JJ. Topic modeling for cluster analysis of large biological and medical datasets. *BMC Bioinformatics* 2014 Oct 21;15 Suppl 11:S11. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4251039/
4. Babones SJ. *Latent Variables and Factor Analysis*. Sage, 2015.

## latent variables models, *see also* latent variables

These models are developed on observed or manifest variables, relating them to the actual variables of interest that may not be directly observable—referred to as **latent variables**. For example, people are secretive about their income, and housing, cars, amenities, etc., are sometimes observed instead as surrogates to guess their income. Health is also an unobservable characteristic, and indicators such as absence of disease, capacity to lift weight, and lung capacity can be used as indirect measures. Latent variable models can also be used for hardcore statistical issues such as measurement errors, unobserved heterogeneity, and missing data. Methods such as **factor analysis**, **path analysis**, and **structural equation models** are examples of latent variable models that link observable data in the real world to symbolic data in the modeled world. These models seem to have roots in the 1968 book by Lazarsfeld and Henry [1].

Specifics of latent variable models depend on whether the observed and latent variables are **quantitative** (generally continuous) or **qualitative**, and whether there are any **covariates** to be included. For example, the conventional factor analysis is used when both the observed and latent variables are quantitative and there is no covariate. When both the observed and latent variables are categorical, the method called *latent class model* is used. The mathematics underlying this kind of modeling can get rather involved, and best avoided here.

In one of the unusual applications, Ghassemi et al. [2] used latent variable models to decompose free-text hospital notes into meaningful features, and the predictive power of these features for patient mortality was examined. In this example, meaningful features that can possibly predict mortality are the latent variables and the free-text hospital notes are the observed variables. Jackson et al. [3] proposed latent models for ordinal variables and their application in the study of newly licensed teenage drivers. In this study, risky driving behavior during the first 18 months of their licensure was used to predict the occurrence of a crash or a near crash event. You can see in these examples that the predicted variables are latent and not actually observed.

For a comprehensive and unified approach to latent variable modeling from a statistical perspective, see Bartholomew et al. [4]. A word of caution may be in order. Many researchers view latent variable models with suspicion because of unverifiable underlying assumptions and naïve inferences regarding causality. For some, this is a strength as this analysis can be viewed as a type of sensitivity analysis for the robustness of the model.


1. Lazarsfeld PF, Henry NW. *Latent Structure Analysis*. Houghton Mill, 1968.
2. Ghassemi M, Naumann T, Doshi-Velez F, Brimmer N, Joshi R, Rumshisky A, Szolovits P. Unfolding physiological state: Mortality modelling in intensive care units. *KDD* 2014 Aug 24;2014:75–84. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4185189/
3. Jackson JC, Albert PS, Zhang Z, Morton BS. Ordinal latent variable models and their application in the study of newly licensed teenage drivers. *J R Stat Soc Ser C Appl Stat* 2013 May;62(3):435–50. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4183151/
4. Bartholomew DJ, Knott M, Moustaki I. *Latent Variable Models and Factor Analysis: A Unified Approach*, Third Edition. Wiley, 2011.


## laws of probability (addition and multiplication), *see also* probability

Among the most elementary laws of **probability** are the law of multiplication, which helps calculate the probability of *joint* occurrence of two or more events, and the law of addition, which helps calculate the probability of one *or* the other event.

### *Law of Multiplication*

Coronary artery disease (CAD) is more common in diabetic patients than in nondiabetic patients. We say that CAD and diabetes are associated. These are termed statistically dependent diseases although such dependence does not imply any cause–effect type of relationship. On the other hand, blindness and deafness in a person are independent events in the sense that occurrence of one does not increase or decrease the chance of occurrence of the other. For such independent events, the joint probability of the two occurring together in a person can be easily computed as the product of the individual probabilities. Thus,

$$P(\text{blindness AND deafness}) = P(\text{blindness})*P(\text{deafness}),$$

where $P$ is the notation for probability.

Symbolically, "AND" is denoted by $\cap$ and is called an intersection. Thus, for two independent events A and B,

$$P(A \cap B) = P(A)*P(B).$$

This is called the law of multiplication or the product rule of probabilities. A useful feature of this relationship is its inversibility in the sense that if the above equation holds, then events A and B are independent; otherwise, not. Computation of the joint probability $P$(blindness and deafness) directly, without using the above equation, requires knowledge of the percentage of subjects in which these two occur together. In many situations, this might be cumbersome to obtain compared with the two individual probabilities on the right side of the above equation. Note that the individual probability in our example is simply the prevalence rate, which would be easily available. If these are 1 per 1000 and 2 per 10,000, respectively, then the prevalence of blindness and deafness occurring together is $0.001 \times 0.0002 = 0.0000002$ or 2 per 10 million. The law of multiplication is useful for obtaining joint probability for such independent events.

Table L.2 is constructed from the table on the correspondence between body mass index and subscapular-to-triceps skinfold ratio in middle-aged Cretan men given by Aravanis et al. [1].

Cutoff points chosen for each category of BMI are **tertiles**. These divide the subjects into three groups of equal size. Thus, nearly equal numbers in the row totals are not surprising: 109, 111, 111. In this case, $P$(low BMI) = 109/331 = 0.329, and $P$(low skinfold ratio) = 110/331 = 0.332. Thus,

$$P(\text{low BMI}) \times P(\text{low skinfold ratio}) = 0.329 \times 0.332 = 0.109.$$

On the other hand, the following is the joint probability on the basis of the cell frequency: $P$(low BMI and low skinfold ratio) = 56/331 = 0.169. Since these two probabilities are very different, low BMI

**TABLE L.2**
**Correspondence between Skinfold and Body Mass Index (BMI) in Cretan Men**

| | Subscapular-to-Triceps Skinfold Ratio | | | |
|---|---|---|---|---|
| **Body Mass Index** | **Low (<1.77)** | **Medium (1.77–2.33)** | **High (>2.33)** | **Total** |
| Low (<25.3 kg/m$^2$) | 56 | 43 | 10 | **109** |
| Medium (25.3–28.7 kg/m$^2$) | 36 | 40 | 35 | **111** |
| High (>28.7 kg/m$^2$) | 18 | 31 | 62 | **111** |
| **Total** | **110** | **114** | **107** | **331** |

and low skinfold ratio are not independent. The above calculations show that if one is low, there is a greater chance that the other is also low. The same is true for "high." Because 62/331 (i.e., 0.187) is much more than (111/331) × (107/331) (i.e., 0.108), there is a greater chance of one being high when the other is high. However, there is no evidence that middle BMI occurs more frequently with middle skinfold ratio. When BMI is in the middle category, the joint probabilities of low, middle, and high skinfold ratios are 36/331 = 0.109, 40/331 = 0.121, and 35/331 = 0.106, respectively. These are not very different from the product of individual probabilities (111/331) × (110/331) = 0.111, (111/331) × (114/331) = 0.115, and (111/331) × (107/331) = 0.108. Thus, when BMI is in the middle category, skinfold ratio is independent and falls in any of the three categories with almost equal frequency.

Theoretically, even a mild difference in the joint probability from the product of individual probabilities suggests association. Practically, the difference should be *substantial* to conclude **association** because of two considerations. First, $n = 331$ is a sample in the above example and there is a need to be cautious about sampling fluctuation that can easily produce minor deviations. Second, a very mild association may not be medically relevant in many situations.

### Law of Addition

After corrective surgery for residual deformity in multiple injuries, the recovery may be full, partial, or none. Because a patient at any particular instance can belong to only one of these categories, these are called **mutually exclusive** categories. In the case of such categories, the probability of belonging to one or the other is computed by the law of addition. That is,

$$P(\text{full or partial recovery}) = P(\text{full recovery}) + P(\text{partial recovery}).$$

If the probability of full recovery is 0.30 and that of partial recovery is 0.40, then the probability of at least some recovery is 0.70. In notation, the symbol ∪, called union, is used for "or." Thus, for mutually exclusive events,

$$P(\text{A} \cup \text{B}) = P(\text{A}) + P(\text{B}).$$

Note that mutually exclusive events cannot be independent and vice versa. There is no bar on independent events occurring together, whereas mutually exclusive events cannot occur together. Because full, partial, and no recovery are the only possibilities, out of which one has to happen, these are called exhaustive categories but that is not a requirement for using the addition rule. However, if the categories are exhaustive, then the probability of any one of them occurring is 1.

1. Aravanis C, Mensink RP, Corcondilas A, Ioanidis P, Feskens EJM, Katan MB. Risk factors for coronary heart disease in middle-aged men in Crete in 1982. *Int J Epidemiol* 1988;17:779–83. http://ije.oxfordjournals.org/content/17/4/779.short

## least absolute shrinkage and selection operator (LASSO)

When there are a large number of predictors under consideration in a **regression** modeling, it is always considered advisable to find a model that uses a small set of predictors so that a parsimonious model can be built. This helps in providing a better handle for control of the outcome and in better understanding the functionality of the system under the model. The LASSO is a method that can help in selecting predictors of a target variable $y$ from a larger set of possible predictors. Developed in 1996 by Tibshirani [1], the LASSO uses the least squares method as usual for estimation of the regression coefficients but puts a restriction that the sum of the absolute values of regression coefficients does not exceed a predefined threshold λ. This threshold is called the tuning parameter. In doing so, the LASSO can drive the coefficients of irrelevant variables down to zero, thus performing automatic variable selection. The method requires that all the predictors are **standardized** since otherwise the regression coefficients depend on the unit of measurement and one variable can outdo others just because the unit of measurement is small.

Incidentally, lasso is also the rope with a loop at one end that is thrown and tightened around the target. This fits well into the nature of the LASSO method that is thrown around the estimates of the regression coefficients to tighten their values. The method is applicable to nonlinear models as well, but we use a linear model in the explanation next for simplicity.

In terms of notations, the linear regression model is

$$\hat{y} = b_1 x_1 + b_2 x_2 + \ldots + b_K x_K,$$

where $(x_1, x_2, \ldots, x_K)$ are the standardized predictors and $(b_1, b_2, \ldots, b_K)$ are the regression coefficients. The usual least squares method chooses regression coefficients $b$'s such that $\Sigma(\hat{y} - y)^2$ is minimum with no other restriction. However, the LASSO method places the restriction that the sum of the absolute values of regression coefficients $\Sigma|b_k| \le \lambda$. When λ is large, the usual regression is obtained but a small λ forces some of the $b$'s to shrink so much so that some even become nearly zero. The predictors' nearly zero coefficient can be excluded from the model. This shrinkage is like penalizing the regression coefficients. This method has been shown to improve the quality of prediction besides increasing the parsimony. For a necessary and sufficient condition for application of LASSO, see Zhao and Yu [2].

Yeh et al. [3] used the LASSO method to identify specific policies of medical schools that are associated with students' reports of interaction with pharmaceutical and medical device industry, and concluded that policy makers should pay greater attention to less research-intensive institutions. Huang et al. [4] developed a computational model for breast cancer prognosis by combining the Pathway Deregulation Score–based pathifier algorithm, using, among others, the LASSO method. This score can help in the more personalized care of breast cancer patients. These examples may give an idea of how and where the LASSO method can be used.

1. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Statist Soc B* 1996;58(1):267–88. http://statweb.stanford.edu/~tibs/lasso/lasso.pdf
2. Zhao P, Yu B. On model selection consistency of LASSO. *J Machine Learning Res* 2006 Nov;7:2541–63. http://www.jmlr.org/papers/volume7/zhao06a/zhao06a.pdf
3. Yeh JS, Austad KE, Franklin JM, Chimonas S, Campbell EG, Avorn J, Kesselheim AS. Association of medical students' reports of interactions with the pharmaceutical and medical device industries and medical school policies and characteristics: A cross-sectional study. *PLoS Med* 2014 Oct 14;11(10):e1001743. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4196737/
4. Huang S, Yee C, Ching T, Yu H, Garmire LX. A novel model to combine clinical and pathway-based transcriptomic information for the prognosis prediction of breast cancer. *PLoS Comput Biol* 2014 Sep 18;10(9):e1003851. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4168973/

## least significant difference (LSD), *see also* multiple comparisons

The LSD is one of the many methods used for **multiple comparisons** of means in different groups in an **analysis of variance (ANOVA)** setup. Other multiple comparison tests are **Bonferroni**, **Tukey**, **Duncan**, and **Dunnett**, each of which is used under specific conditions. All multiple comparison tests compare a set of means in a way that the experimentwise (Type I) error rate remains under control such as within 5% or at any specified α level. (To understand "experimentwise error rate," see the topic **multiple comparisons**.) The hypothesis that the means are equal is first tested by an α-level ***F*-test** in an ANOVA setup. Should this test turns out not significant, the procedure generally terminates without making any detailed inferences on groupwise differences across the means since this implies that there is no significant difference in means in different groups. When $F$ is significant, the group differences are tested by one of the multiple comparison tests depending on the type of comparisons required for your setup. As just mentioned, these tests adjust the Type I error so that this remains under control. The difference between this and, for example, **Student *t*** for two groups is that LSD uses the **mean square error (MSE)** from ANOVA as the pooled variance in place of the pooled estimate from variances of the two groups under comparison used in the Student $t$-test. The MSE is a better estimate as this is based on a larger sample and considers variation across groups.

$$\text{LSD} = t_{\alpha/2,\nu}\sqrt{\text{MSE}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)},$$

where $t_{\alpha/2,\nu}$ is the critical value from Student $t$ distribution corresponding to an α/2 level of significance and $\nu$ df (these are the df for the MSE) and $n_1$, $n_2$ are the sample sizes of the two groups under comparison. As the name implies, this is the smallest difference between means in the two groups under comparison for it to be considered significant based on Student $t$-test.

The procedure is to calculate one LSD for all the pairwise comparisons and check if the difference in means in any two groups is more than this. The groups with a bigger difference in means are considered statistically significantly different at α level.

Since LSD does not adjust α for multiple comparisons, this test has fallen out of practice. However, this has historical importance as this is the first multiple comparison test ever developed. This was proposed by Ronald Fisher in 1935.

## least squares method

This method is used for fitting a **regression** in a manner that the sum of the squared distances between the observed and fitted values is at a minimum. In terms of notations, this sum is expressed as $\Sigma(y-\hat{y})^2$, where $\hat{y}$ is the notation for the fitted value. The difference between $y$ and $\hat{y}$ is also statistically called the error, although this is not a mistake, just an error in fitting. The better name for samples is **residual**. Some residuals are negative and some are positive, and the method ensures that the sum of these residuals is zero. Because of the square in this method, bigger residuals get a higher weight in minimization compared with the small ones, which means that large residuals, in a way, are tolerated less than small residuals. Two such residuals between the observed and the fitted values are shown in Figure L.1. These are the vertical distances of the observed values of $y$ with the fitted values of $y$. The fitted values of $y$ are represented by the regression curve in this figure.

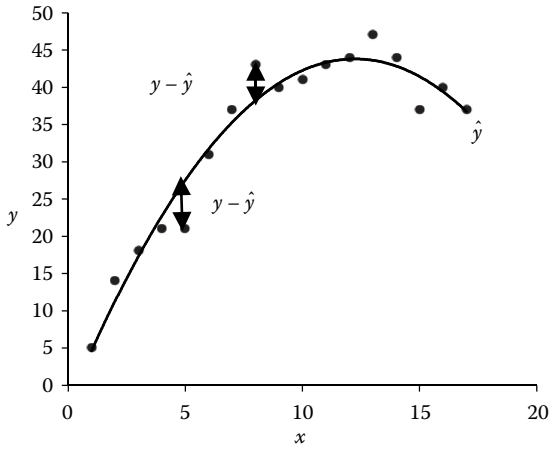



**FIGURE L.1** Differences between the observed and fitted values of $y$.

As explained under the topic **regression**, the data required to fit a regression are the values of the dependent variable $y$ corresponding to different values of the regressor variables $x$'s. The dependent variable must be quantitative. For one set of values of $x$'s, one or many values of $y$ can be observed. The values of $y$ should be available for several distinct values (at least four for the regression to be meaningful) of $x$. The values of the $x$ variables can be *deliberately* chosen to serve the purpose.

The objective of the least squares method is to obtain the **regression coefficients** in a manner that the fitted regression is closest to the observed values. However, the form of the regression—linear, curvilinear, or nonlinear—has to be specified. Square of the differences helps us consider the positive residuals the same way as the negative residuals. The minimization process requires mathematical differential equations that we avoid in this book and are considered not important for our audience. Almost all statistical packages have the facility to provide the estimates of the regression coefficients by the least squares method. The precision of these estimates is best for central values of the regressors and imprecise for larger or smaller values. The method of least squares works well when the variance of the residuals is the same for different values of the regressors. When this does not hold, a **weighted least squares method** is used. When the regression is linear and the $y$ values are independent of one another, the inverse of the variances is chosen as the weights. This may require an iterative process.

The other method generally used for obtaining the estimates of the regression coefficients is the **maximum likelihood**. When the regression equation is $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_Kx_K$, this method finds those values of ($b_0$, $b_1$, …, $b_K$) that make our observed sample most likely to happen. In case the residuals follow the **Gaussian** pattern, this method gives the same estimates as the method of least squares. The method of least squares is more popular as it does not require any particular pattern of the data.

## level of significance, *see also* *P*-values

Level of significance is the prefixed maximum threshold of the chance of **Type I error** in statistical testing of hypothesis. Type I error is explained in the next paragraph. In brief, this is the probability of wrongly rejecting a **null hypothesis** ($H_0$).When performing a hypothesis test, the values observed in the sample serve as evidence, and they are used either to reject or not to reject an $H_0$. However, these values are subject to sampling fluctuation and may or may not lead to a correct conclusion. Errors do occur, which are classified as

Type I and Type II errors. To understand why we need to put a cap on Type I error and give it a new name as level of significance, a detailed account of Type I error is necessary in this section. **Type II error** is explained separately.

### *Type I Error*

This error is best explained with the help of court analogy. It is quite common in court that a real offender is acquitted because of weak evidence. This is not considered a serious error, but the other error also occurs. An innocent is pronounced guilty because there is strong circumstantial evidence. This is serious, and is an example of Type I error.

In a testing of hypothesis setup, if there is no real difference between the test and control groups in a clinical trial, but the data strongly disagree, the true null hypothesis of equality has to be undesirably rejected. A false-positive conclusion is reached. This is serious, and is thus called Type I error, or alpha error.

The seriousness of this error can also be understood from the setup of a trial on a new drug. This type of error occurs when an ineffective drug is declared effective. This gives false assurance, and the ineffective drug is unnecessarily marketed, prescribed, and ingested, and side effects are tolerated. Imagine the cost and inconvenience caused by this error. Statistical procedure requires that the probability of this type of error be kept low, generally within 5%. The actual probability of Type I error in the observed sample is called the ***P*-value**. It can also be understood as the probability that a true null hypothesis is wrongly discarded because the sample values so indicate. In the case of trials, this could occur if the patients included in the trial by chance happen to exhibit a difference between the groups when actually none exists.

The seriousness of a Type I error requires that a threshold of *P*-value is fixed *in advance*, beyond which it would not be tolerated. This is called the significance level and is denoted by α. This is also called the alpha level. This specifies the criterion of doubt, beyond which a null hypothesis is rejected. Statisticians—almost all empirical scientists—generally use a threshold of 5% chance of error. Five percent is an internationally accepted norm for significance level that is rarely breached in medical literature. In situations where 5% chance of error can translate into serious consequences such as death, a lower level can be fixed.

A null hypothesis is generally a statement of the status quo, and the **alternative hypothesis** signifies a change. If it were very costly to make changes from the status quo, you would want to be very sure that the change would be beneficial. The threshold of risk of Type I error is then kept very low, say less than 1%. This may be needed when, for example, a drug presently in extensive use or an existing lifesaving drug is sought to be replaced. In most medical situations, $\alpha = 0.05$ is considered adequate.

To theorists, a 5% chance of error may look high but that is not so as humans and animals vary much more than chemical reactions or electron motion in physics. Medical theories are hard to evolve, and all medical professionals are expected to be well informed about the pros and cons of various errors so that they can be judicious in their decisions despite a 5% chance of Type I error.

When the *P*-value in any study is less than the prefixed level of significance, the presence of difference or relationship is concluded. When $P$ is more than or equal to, say, 0.05, sampling fluctuation cannot be excluded as a likely explanation of the observed difference. There is a convention to call a result statistically significant when $P < 0.05$ and to call a result highly significant when $P < 0.01$. A probability of less than 0.01 is considered exceedingly small in most medical contexts.

It is important to distinguish between a *P*-value and an α level. Both measure the probability of Type I error. The first is obtained for the data set in hand whereas the second is fixed in advance. You may say for a problem that not more than 3% chance of Type I error can be tolerated by setting $\alpha = 0.03$. The calculations later on may reveal that the *P*-value for your data set is only 0.012, 0.362, or any other number. A decision regarding rejecting or not rejecting a null hypothesis is reached according to the *P*-value being higher or lower than the predetermined level of significance.

## Levene test

This test is used to check whether the **variances** in two or more groups are the same or not. Equality, also called homogeneity, of variances across groups is an important prerequisite for many statistical procedures such as the Student *t*-test and analysis of variance (ANOVA) and for residuals in regression. The Levene test helps in finding violations of this requirement. Since procedures such as ANOVA are for testing equality of means, intuitively, this equality has clear meaning when the variances are equal and loses relevance when variances greatly differ from group to group.

The conventional test for equality of variances in two groups is $F = s_1^2/s_2^2$, where $s_1$ and $s_2$ are the sample standard deviations. This is referred to as *F*-distribution for finding the *P*-value. For more than two groups, the **Bartlett test** is conventionally used. Both these tests are heavily dependent on the Gaussian distribution because the sample variances themselves (with numerator involving $(y - \bar{y})^2$) are distorted by outliers, and these outliers can substantially change the mean, thus affecting the deviations from mean, and the use of the square, amplifying the effect. The Levene test, on the other hand, is relatively independent of this requirement. This test was developed by Howard Levene in 1960 [1] using means as explained next but was later extended by Brown and Forsythe [2] using medians. This extension made the Levene test independent of the requirement of Gaussian distribution of the data.


Howard Levene

The Levene test criterion is

$$W = \frac{(n-K)\Sigma_{k=1}^{K} n_k (\bar{y}_k - \bar{y})^2}{(K-1)\Sigma_{k=1}^{K}\Sigma_{i=1}^{n_k} (y_{ik} - \bar{y}_k)^2},$$

where $K$ is the number of groups and the sample size of group $k$ is $n_k$ ($k = 1, 2, \ldots, K$), and $y_{ik} = \left|x_{ik} - \bar{x}_k\right|$. With the Brown and Forsythe extension, the mean in the last expression can be replaced by median. In this case, the means appearing in the formula for $W$ can also be replaced by medians and the criterion is then called the **Brown–Forsythe test**. The test is essentially based on the absolute values of the deviations $y_{ik}$'s. The criterion $W$ approximately follows an *F*-distribution with ($K - 1$, $n - K$) df, which would allow one to obtain the *P*-value.

The Levene test is just about the most popular test for checking the homogeneity of variances, and it is available in almost all

statistical software packages. When the Levene test indicates that the variances across groups are not significantly different, procedures such as ANOVA *F*-test can be confidently used provided other conditions are met. However, when the Levene test is significant, the variances are not equal and the subsequent test for equality of means is the **Welch test** instead of the regular *F*-test.


1. Levene H. Robust test for equality of variances, in: *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, (Ed. Olkin et al.). Stanford University Press, 1960:pp. 278–92.
2. Brown MB, Forsythe AB. Robust tests for equality of variances. *J Am Stat Assoc* 1974;69:364–7. http://www.jstor.org/discover/10.2307/2285659?sid=21105091689171&uid=4&uid=2


## Lexis diagram

A Lexis diagram is a two-dimensional representation of three-dimensional data; it is used more commonly in demographics for age at death, year of death, and year of birth for a cohort of population. An initial version of this diagram was first proposed by Zeuner in 1869, to which Wilhelm Lexis just added a network of parallel lines, but the name "Lexis diagram" has imposed itself in a seemingly invincible way [1]. The diagram could be a useful tool in **age–period–cohort analysis** as explained by Carstensen [2].

A very typical Lexis diagram that shows age-specific mortality in different birth cohorts is a contour-like diagram as shown in Figure L.2a. This figure shows the lung cancer mortality in Australian males in 1950–1985 by age at death and year of death [3]. Note how the dashed lines connect the age, year of birth, and year of death. The authors of the article called it a **synoptic chart**. Displayed along the contours is the number of deaths. The data displayed in Figure L.2a can also be shown in three dimensions by a **surface chart** (Figure L.2b). Perhaps easier to understand is the surface in Figure L.2b. It can be easily seen, for example, that the mortality rate at age around 70 years steeply increased during the period 1970–1980 and showed a slight decline thereafter.

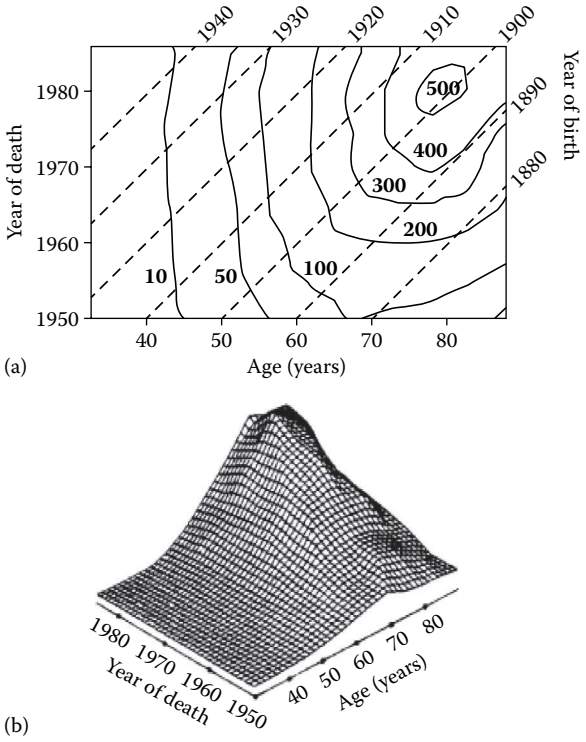



**FIGURE L.2** Mortality from lung cancer in Australian males during 1950–1985: (a) Lexis diagram; (b) surface chart. (From Jolley D, Giles GG. *Int J Epidemiol* 1992; 21:178–82. http://ije.oxfordjournals.org/content/21/1/178.full.pdf. With permission.)

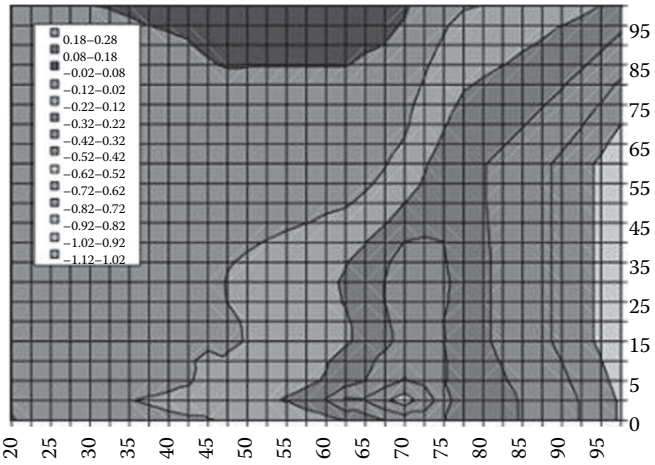



**FIGURE L.3** Log of ratios of death rates between adjacent levels of life expectancy. (From Gerland P, Li N. Modifying the Lee–Carter method to project mortality changes up to 2100. Paper presented at the *Population Association of America Annual General Meeting, 2011.* http://esa.un.org/unpd/wpp/publications/Files/Li_2011_Modifying%20the%20Lee-Carter%20method%20to%20project%20mortality%20changes%20up%20to%202100.pdf.)

Gerland and Li [4] have presented several Lexis diagrams on projected changes in mortality up to 2100 in some parts of the world. One of these is in Figure L.3, where log of ratios of death rates between adjacent levels of life expectancy are shown. Brinks et al. [5] used a Lexis diagram to track different timescales for birth rates, incidence, and mortality rates of chronic diseases.

Vandeschrick [1] has explained each step of the construction of a Lexis diagram in case you are interested to draw one yourself.


1. Vandeschrick C. The Lexis diagram, a misnomer. *Demog Res* 2001;4:97–124. http://www.demographic-research.org/Volumes/Vol4/3/4-3.pdf, last accessed January 15, 2015.
2. Carstensen B. Age–period–cohort models for the Lexis diagram. *Stat Med* 2007 Jul 10;26(15):3018–45. http://bendixcarstensen.com/APC/Carstensen.2007a.pdf, last accessed January 15, 2015.
3. Jolley D, Giles GG. Visualizing age–period–cohort trend surfaces: A synoptic approach. *Int J Epidemiol* 1992; 21:178–82. http://ije.oxfordjournals.org/content/21/1/178.full.pdf
4. Gerland P, Li N. Modifying the Lee–Carter method to project mortality changes up to 2100. Paper presented at the *Population Association of America Annual General Meeting, 2011.* http://esa.un.org/unpd/wpp/publications/Files/Li_2011_Modifying%20the%20Lee-Carter%20method%20to%20project%20mortality%20changes%20up%20to%202100.pdf
5. Brinks R, Landwehr S, Fischer-Betz R, Schneider M, Giani G. Lexis diagram and illness-death model: Simulating populations in chronic disease epidemiology. *PLoS One* 2014 Sep 12;9(9):e106043. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4162544/

## life expectancy (types of), *see also* expectation of life and life table method

The term *life expectancy* is used for **expectation of life**—a term separately explained. However, there are many variants of life expectancy, namely, active life expectancy, disability-free life expectancy, and health-adjusted life expectancy or healthy life expectancy, and we explain all these in this section. Broadly, these terms are used for that part of life expectancy that is expected to be lived in good health as per the current trend. Minor differences in them are as follows.

It was realized a long time ago that increased longevity beyond a point does not generally translate into overall physical well-being of individuals. The advanced technology now available can delay death and prolong life in many cases. However, many of these added years can be more painful or of increasingly restricted activity. Years are added to life, but life not added to years. Thus, the concept of *healthy life expectancy* was evolved. This is the average number of years that a person is expected to live in good health. Good health is subjective and defined variously by different workers, which has caused usage of different terms.

In the context of old age, this can be *active life expectancy*, that is, the period when a person can independently carry out the daily chores of life—walking, bathing, toileting, dressing, and so on. Active life expectancy would be obviously lower in old age than the usual life expectancy. This would include periods of those sicknesses that do not affect the daily activities of life but would exclude the dependent period. One variation is *disability-free life expectancy* (DFLE) that excludes the period of disease and infirmity. This period can arise from acute diseases that occasionally occur for a short duration such as cough and cold, diarrhea, cholera, typhoid, and malaria, as well as from chronic diseases such as ischemias, malignancies, tuberculosis, and ulcers that cause long-term sickness. Disabilities such as of vision, hearing, and movements are also counted. After such exclusion, the procedure to calculate DFLE is basically the same as for calculating life expectancy. However, estimating the duration of disease and infirmity in a lifetime is a big challenge.

A more popular version is the *health-adjusted life expectancy* (HALE), where the period of ill health is adjusted according to the severity of condition, and equivalent years in good health lost are computed in a way similar to the years of disability calculated for **disability-adjusted life years** (DALYs) lost. The process though is somewhat reversed in the sense that our interest now is in equivalent healthy years lived in place of healthy years lost. This is also popularly known as *healthy life expectancy*. According to World Health Organization (WHO) estimates, global HALE at birth in 2012 for males and females combined was 62 years in the world. This is 8 years lower than total life expectancy at birth. These are the equivalent years lost due to poor health. HALE at birth ranged from a low of 49 years for African males to almost 70 years for females in the WHO Western Pacific Region. Approximately 10% to 15% life is lost due to ill health in different parts of the world [1]. For complete methodology for calculating HALE, see the WHO document [2] on this topic.


1. WHO. *Healthy Life Expectancy (HALE) at Birth*. http://www.who.int/gho/mortality_burden_disease/life_tables/hale_text/en/, last accessed January 18, 2015.
2. WHO. *WHO Methods for Life Expectancy and Healthy Life Expectancy*. World Health Organization, 2014. http://www.who.int/healthinfo/statistics/LT_method.pdf?ua=1&ua=1, last accessed January 18, 2015.


## life table, *see* expectation of life and the life table

## life table method, *see* survival curve/function

## likelihood, log-likelihood, and maximum likelihood estimates

We first explain what do we mean by likelihood and then describe log-likelihood and maximum likelihood estimates.

### *Likelihood*

Likelihood is the chance of occurrence, but it looks at an event that has already occurred. If you have three randomly selected persons in the sample and the first two are obese while the third is not, what is the chance of this happening when the chance of one being obese is 0.2? If the subjects are independent, this chance is $0.2 \times 0.2 \times 0.8 = 0.032$. This is the likelihood of this sample and depends on the value of the parameter—in this case, the probability 0.2 of occurrence in a single subject. Likelihood refers to the past events with known outcomes, whereas probability refers to the occurrence of future events.

In statistics, likelihood is a function of the **parameters** of the **distribution** from which the sample is drawn in the sense that the likelihood depends on what the value of each parameter is. For example, if the sample is drawn from a population where the values follow a Gaussian (Normal) distribution, the likelihood would depend on the values of the population mean $\mu$ and population standard deviation $\sigma$. Thus, the full name is *likelihood function* of the sample that signifies that it depends on the values of the parameters. The sample values $(x_1, x_2, \ldots, x_n)$ are considered known as just explained and the likelihood is obtained after these values are obtained. The values of the parameters are considered unknown as almost always is the case. Thus, likelihood is a function of the parameters with a fixed sample. For mathematical derivation, these sample values can continue to be denoted by $(x_1, x_2, \ldots, x_n)$, but these are no longer variables in a likelihood function; instead, they are fixed values. When the sample values are independent of one another, the likelihood function is the product of the individual probabilities of the values in the sample in accordance with the product rule of probabilities. In short, this is the probability of occurrence of the values obtained in the sample and is, thus, calculated after the sample values are available. If you have $n$ independent subjects in your sample with values $x_1, x_2, \ldots, x_n$, the likelihood is $L = P(x_1) \times P(x_2) \times \ldots \times P(x_n)$, where $P$ stands for probability.

The primary use of likelihood function in statistics is in finding those estimates of the parameters of the distribution that make the sample values most likely to happen. These are called the maximum likelihood estimates as explained in a short while. The second use of likelihoods is in the **likelihood ratio test** for which the following may be useful.

### *Log-Likelihood*

The product of probabilities in likelihood may have made it clear why taking the logarithm is useful. Since $L = P(x_1) \times P(x_2) \times \ldots \times P(x_n)$, $\ln L = \Sigma_i \ln P(x_i)$, which is now additive. Logarithm converts multiplications to additions. On top of this, many probabilities based on statistical distributions involve exponents, and taking log converts them to algebraically simple expressions. For example,

the likelihood when 3 deaths occur out of 9 cases is (by **binomial distribution**)

$$L = \frac{9!}{3!6!}\pi^3(1-\pi)^6,$$

where $\pi$ is the probability of death of a case. This gives $\ln L$ = constant + $3\ln\pi + 6\ln(1-\pi)$, which is linear and much easier to handle. However $L$, being the probability, is almost always less than 1 and $\ln L$ is negative. Thus, instead of $\ln L$, $-2\ln L$ is used, which not only makes it positive but also allows us to use the chi-square criterion for large $n$. This property is extensively used in statistical tests such as in the **likelihood ratio test**.

### *Maximum Likelihood Estimates*

Maximum likelihood can be easily explained by considering an example of a **binary variable**. Suppose 3 die out of 9 patients with the Ebola virus. The probability of death is unknown: let us denote this by $\pi$. As before, the **binomial distribution** tells us that the likelihood of 3 deaths out of 9 cases (if independent) is

$$L = \frac{9!}{3!6!}\pi^3(1-\pi)^6.$$

At what value of $\pi$ is this likelihood at a maximum? At $\pi = 0.1$, $L = 0.0446$; at $\pi = 0.2$, $L = 0.1762$; at $\pi = 0.3$, $L = 0.2668$; at $\pi = 0.33$, $L = 0.2731$, at $\pi = 0.35$, $L = 0.2716$; at $\pi = 0.4$, $L = 0.2508$. You can see that the value of $L$ is maximum when $\pi = 0.33$ and all other values of $\pi$ give a lower value of $L$. Thus, the maximum likelihood estimate of $\pi$ is 0.33. This incidentally is the same as the proportion of deaths observed in the sample (3 out of 9).

The method of maximum likelihood provides those estimates of the parameters of a model that make the sample values most likely. Since we cannot change the data, we change the estimate to see what estimate makes the likelihood maximum. For computational convenience, log-likelihood is maximized in place of the likelihood itself. Taking the logarithm tends to linearize the likelihood function as explained earlier. Instead of log-likelihood, $-2\ln L$ is used, which not only makes it positive but also is easy to handle as the distribution of $-2\ln L$ is known (chi-square) for many situations as also explained earlier. Because of the minus sign in this, those estimates are chosen in a model that minimizes $-2\ln L$, which is the same as maximizing $L$.

Maximizing involves differential calculus: maximum likelihood estimates are easily and exclusively obtained for distributions such as Gaussian; this is not so easy for some other distributions. Logistic regression is an example where the estimates of parameters that maximize the likelihood of such distributions are obtained by iteration. This is a trial-and-error method starting with some plausible estimates.

## likelihood ratio of a diagnostic test

Likelihood ratio of a diagnostic test is a measure of its utility in increasing or decreasing our confidence in the presence or absence of disease in a suspected case. Realize first that medical science is based on observations and a diagnosis can seldom be established with 100% confidence. Some uncertainty remains. Suppose the clinical picture in a patient gives you a confidence of 60% that a particular disease X is present in that patient. This is the *pretest probability* of that disease in that patient. A diagnostic test such as prostatic specific antigen is ordered for confirming or ruling out cancer of the prostate. If the result of this test is positive, your confidence in the presence of disease increases, although the confidence level still does not reach 100%, and if the result is negative, the confidence decreases. Likelihood ratio tells how much the confidence will increase or decrease after the result of the test is available.

The likelihood ratio is the ratio of the probability of a test result in those with disease and the probability in those without disease (this is not the same as used in a statistical likelihood ratio test). Since the test can be negative or positive, the likelihood ratio is also of two types—positive likelihood ratio denoted by LR+ and negative likelihood ratio denoted by LR–. These are defined as follows:

$$\text{LR}+ = \frac{\text{probability of positive test in persons with disease}}{\text{probability of positive test in those without disease}}$$

$$= \frac{\text{sensitivity}}{1-\text{specificity}};$$

$$\text{LR}- = \frac{\text{probability of negative test in persons with disease}}{\text{probability of negative test in those without disease}}$$

$$= \frac{1-\text{sensitivity}}{\text{specificity}}.$$

Sensitivity and specificity should be clear from these equations; otherwise, they are explained under the topic **sensitivity and specificity**. If sensitivity is 90% and specificity is 80%, then LR+ = 0.90/(1 − 0.80) = 4.5 and LR– = (1 − 0.90)/0.80 = 0.125. Note that LR+ is more than 1 in this example and LR– is less than 1. This would be the case in most situations. The interpretation of these ratios is something like this: A value of 10 or more of LR+ and 0.1 or less of LR– indicates that the test is extremely good in those whose disease status is known, a value of 5–10 of LR+ and 0.1–0.2 of LR– indicates moderate utility of the test, and all other values are regarded as not helpful. A value of 1 means that the test is not helpful at all, and a test with LRs around 1 should not be ordered as it does not help in enhancing your confidence in the presence or absence of disease in a suspected patient.

Since sensitivity and specificity are indicators of the inherent validity of the test, likelihood ratios are indicators of the inherent validity of the test as well. Despite this, the interpretation is mostly in terms of the likelihood of presence or absence of disease. LR is multiplicative for odds of disease but not for the probability of disease. For distinction between odds and probability, see the topic **odds** on this volume. LR+ measures the increased factor of odds of disease when the test is positive, and LR– measures the decreased factor of odds of no disease when the test is negative. In other words,

$$\text{posttest odds} = \text{pretest odds}*\text{LR}.$$

If the pretest probability of disease is 10%, the odds are 1:9 or 1/9, and if LR+ = 6, posttest odds = 6/9 by the equation just mentioned, or the probability is 6/(6 + 9) = 0.4. That is, the chances of disease being present after testing positive increase from 10% to 40%. Thus, LRs provide further insight into the inherent quality of the test. The procedure is to first covert the pretest probability to odds, multiply it with LR to get posttest odds, and finally convert the posttest odds back to probability.

## likelihood ratio test, *see also* likelihood, log-likelihood, and maximum likelihood estimates

Let us first understand likelihood ratio (LR) as used in this test. Consider a simple example of 3 deaths out of 9 cases of Ebola virus. Let the unknown probability of death be denoted by $\pi$, so that the **likelihood** of 3 deaths out of 9 by **binomial distribution** is

$$L = \frac{9!}{3!6!}\pi^3(1-\pi)^6.$$

If $\pi = 0.1$, this likelihood is 0.0446, and for $\pi = 0.2$, this is 0.1762, giving an LR of 0.1762/0.0446 = 3.95. Thus, the likelihood of $\pi = 0.2$ is just about four times as much as of $\pi = 0.1$ when the number of deaths is 3 out of 9 cases. You can see how LR tells us which value of the parameter is more plausible out of those that we have under comparison.

The LR test uses a similar argument but in a different context. This test is mostly used to check the statistical significance of the loss when some parameters of the model are deleted to achieve a relatively simple model. Since the parameters of a model are seldom known, we must estimate them, and the estimation of less number of parameters is easier. The estimates will be more reliable also because of the smaller number of parameters, and the model becomes simple that it could be easily expressed and easily used.

Although the LR test can be used in a variety of situations, it should be clear from the preceding explanation that it is best used for nested models; that is, the simpler model is part of the more complex model—the only difference being that the complex model has additional parameters. Removing one or more predictor variables will always reduce the likelihood (model will fit less well), but whether or not the loss is statistically significant is tested by the LR test. It is customary to call the more complex model as the full model and the simpler model as the reduced model. What is not clear from our explanation is that the LR test is an approximate test and valid only for large $n$ because of the use of chi-square as explained in a short while.

Suppose the full model has $M$ parameters and the reduced model has $K$ parameters ($K < M$). Thus, the null hypothesis under test is that the remaining $M - K$ parameters, which are examined for deletion, are equal to zero. If our data fail to reject this null, the conclusion is that the reduced model is not significantly different from the full model. That is, the loss in likelihood by deleting $M - K$ parameters is not statistically significant. In that case, we can go ahead with the reduced model. In a special case, $K$ could be zero, in which case the null is that all $M$ parameters in the full model are equal to zero. Now proceed as follows.

Replace the values of the parameters by their **maximum likelihood estimates** and obtain the value of the likelihoods for the full model as well as for the reduced model. Denote these likelihoods respectively by $L_1$ and $L_0$. Since the full model has more parameters, $L_1$ will be more than $L_0$. Calculate the ratio $\lambda = L_0/L_1$. This ratio will always be between 0 and 1. The smaller the value of $\lambda$, the higher is the loss in likelihood due to fewer parameters in the reduced model, and the reduced model becomes less plausible. It may look amazing, but it has been theoretically established that $-2\ln\lambda$ follows approximately a chi-square distribution with $(M - K)$ degrees of freedom for large $n$. Since $\lambda$ is less than 1, $\ln\lambda$ will be negative and $-2\ln\lambda$, called **deviance**, will be positive. The smaller the value of the LR, the larger will be the value of $-2\ln\lambda$. Thus, a large value of $-2\ln\lambda$ will reject the null hypothesis and will indicate that the reduced model is causing significant loss in likelihood—thus, the concerned parameters should not be deleted. Just as a reminder,

$$\text{Deviance: } -2\ln\lambda = -2\ln(L_0/L_1) = -2(\ln L_0 - \ln L_1) \sim \chi^2 \text{ with } (M - K) \text{ df.}$$

Also, recollect that the likelihoods are obtained by multiplication of the probabilities of individual values, the logarithm of which will be a sum. Thus, this transformation helps not only in achieving chi-square distribution but also in converting the ratio to a linear function. The ***P*-value** will be obtained by the chi-square distribution that will determine the statistical significance.

Consider a **logistic model** that predicts the probability of death of patients admitted in the intensive care unit (ICU) on the basis of age, sex, and severity score. This model will be of the type

$$\text{logit}(\pi) = \beta_0 + \beta_1\text{*age} + \beta_2\text{*Sex} + \beta_3\text{*(severity score)} + \varepsilon,$$

where $\pi$ is the probability of death of an ICU patient. If the interest is in finding whether sex has any significant effect on the probability of death, the null hypothesis is $H_0$: $\beta_2 = 0$. In this case, the full model is as stated above and the reduced model will be $\text{logit}(\pi) = \beta_0 + \beta_1\text{*age} + \beta_3\text{*(severity score)} + \varepsilon$. The $\beta$'s and their estimates in the reduced model will be different from the estimates in the full model although we are using the same notation. The full model has $M = 4$ parameters (including $\beta_0$), and the reduced model has $K = 3$ parameters. To test this null, obtain the likelihood $L_1$ of the full model and the likelihood $L_0$ of the reduced model, calculate $-2\ln\lambda$ as per the procedure just mentioned, and compare using chi-square with $(M - K) = 1$ df for finding the $P$-value. If this is less than the preset level of significance such as 0.05, reject the null; otherwise, not. When this null is rejected, the conclusion would be that sex does have a significant influence on the probability of death or, in other words, probability of death is different in males and females even when age and severity score are the same.

## Likert scale

This is a specific kind of tool for psychometric assessment of opinion, belief, and attitude. For a particular construct such as satisfaction from hospital services, a set of items is developed so that all aspects of that construct can be assessed. For hospital services, these items could be on competence of doctors, humanitarian nursing services, proper diagnostic facilities, and so on. A declarative statement such as "I am happy with the nursing services in this hospital" is made in each item and the respondent is asked to specify his level of agreement on generally a 5-point scale such as strongly disagree, disagree, indifferent (neutral), agree, and strongly agree. This is called a Likert scale.

You can have as many items in your questionnaire as is necessary, but a 10-item assessment looks enough to cover different aspects of a construct without burdening the respondent. Sometimes, many more items are devised for discussion by the group of stakeholders who selects "good" items. If the survey is repetitive, **item-analysis** of the responses of a previous survey is done to delete the redundant items and, possibly, to replace them with new ones. Internal consistency can be assessed by a measure such as **Cronbach alpha**. The construct validity of the items is assessed by **factor analysis**.

Response for each item is graded on the same scale—not necessarily 5-point—it could be 7-point, 9-point, or any other scale. Attempts should be made to ensure that the options from strongly disagree to strongly agree are equally spaced so that they can be legitimately assigned scores 0 to 4, 0 to 6, or 0 to 8 depending on the number of options. Instead, these scores can be scaled from −2 to +2, −3 to +3, −4 to +4, and so on. You may like to prefer 1 to 5, 1 to 7, 1 to 9, and so on. Various scoring patterns can give different results and the choice is yours as long as you can justify it.

Options for some items are reversed so that "strongly agree" is the first option and "strongly disagree" is the last option. Some items are negatively framed. This helps in eliciting a well-thought response instead of consistently selecting the same response such as "agree" for most items. At the time of analysis, such reversed items and corresponding scores are rearranged in the proper order.

Add the scores of all the items and get a total score for each respondent. This total is called the Likert score of that respondent for the construct under study. For a 10-item questionnaire and each item on a 5-point scale (0 to 4), the total score for each respondent will range from 0 to 40, which can be analyzed just as any other numerical measurement. Second, the analysis can be focused on each item. If item 3 is on satisfaction with diagnostic facilities, and these are $n = 50$ responses on a 0-to-4 scale for this item, you can calculate the median score to get an assessment of satisfaction with diagnostic facilities, or to compare it with, say, nursing services. Thus, an assessment regarding which component of the construct is adequate in the hospital and which needs strengthening can be made.

Rodriguez et al. [1] used a 1-to-5 Likert scale for assessing satisfaction of the patients, primary care providers, and specialty physicians with electronic consultation in a veteran setting in the United States. Wong et al. [2] used the same scale for patients' beliefs about generic medicines in Malaysia. These examples may give you an idea how Likert scale is used in practice.


1. Rodriguez KL, Burkitt KH, Bayliss NK, Skoko JE, Switzer GE, Zickmund SL, Fine MJ, Macpherson DS. Veteran, primary care provider, and specialist satisfaction with electronic consultation. *JMIR Med Inform* 2015 Jan 14;3(1):e5. http://medinform.jmir.org/2015/1/e5/
2. Wong ZY, Hassali MA, Alrasheedy AA, Saleem F, Yahaya AH, Aljadhey H. Patients' beliefs about generic medicines in Malaysia. *Pharm Pract (Granada)* 2014 Oct;12(4):474. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4282766/


## limits of agreement, *see* Bland-Altman method of agreement

## linear effect, *see also* linearity (test for)

First, the meaning of linear should be clear. Statisticians know it well, but some medical professionals may not be very clear. The term *linear* comes from line. If a relationship between $x$ and $y$ is linear, it really implies that as $x$ increases from $x$ to $x + 1$, the value of $y$ increases by the same quantity, no matter what the value of $x$ is within the values observed. The mathematical equation of a line is

$$\text{Line: } y = a + b*x.$$

In this equation, as $x$ becomes $x + 1$, $y$ becomes $y + b$; that is, $y$ always increases by $b$ whatever $x$ is. This is essentially interpreted as the same effect of $x$ on $y$ over the entire range of values of $x$ under observations, and not that $y$ values increase at a faster rate for some values of $x$ and at a slower rate for other values, or increase for some values of $x$ and decline for other values of $x$. If the risk of myocardial infarction (MI) increases by 1% as diastolic pressure increases by 5 mmHg, beyond 90 mmHg, it would be called linear when this rise in risk remains the same 1% whether the diastolic pressure increases from 100 to 105 mmHg or from 130 to 135 mmHg. If the risk of MI rises by 2% when the diastolic pressure increases from 130 to 135 mmHg but by only by 1% when it increases from 100 to 105 mmHg, the effect is not linear.

To understand linearity more clearly, consider Figure L.4a. This contains a graphical depiction of three linear equations: $y = -x$, $y = 5 + 2x$, and $y = 5 + 3x$. In the first equation, $a = 0$ and $b = -1$; in the second, $a = 5$ and $b = 2$; and in the third, $a = 5$ and $b = 3$. All these three are lines. The gradient or slope of the line is determined by the coefficient $b$ of $x$ and the intercept on the $y$ axis is the constant $a$ in the equation.

The steepest slope among these three lines is in the equation $y = 5 + 3x$. For this equation, if $x = 0$, $y = 5$; if $x = 1$, $y = 8$; if $x = 2$, $y = 11$; and if $x = 3$, $y = 14$. As $x$ increases by 1, $y$ always increases by the same quantity—in this case by 3. This is what defines line. The line $y = -x$ has a negative slope: as $x$ increases by 1, $y$ decreases by 1.

For more clarity, let us give it a twist and introduce a square term to the equation. This is called a *quadratic equation* and plots into a curve. Thus,

$$\text{Quadratic curve: } y = a + b*x + c*x^2.$$

For example, the graphical depiction of the equation $y = 5 + 3x - \frac{1}{3}x^2$ is shown as a dark curve in Figure L.4b. This is no longer a line because of the square term. Technically, this is called a *parabola*. The values of $y$ increase for $x$ values between –1 and 5 and decrease thereafter. This contradicts linearity. The variable $x$ has a nonlinear effect on $y$ in this case.

The linear effect of a characteristic on the outcome can also be understood as that part of the effect that can be explained by a line. In other words, this is the *component* that remains the same for each unit increase in the value of the characteristic. The linear effect from the just mentioned quadratic curve can be extracted in two different ways: (i) Best line that can represent the trend in the curve as shown in Figure L.4b. The linear component of the curve is best

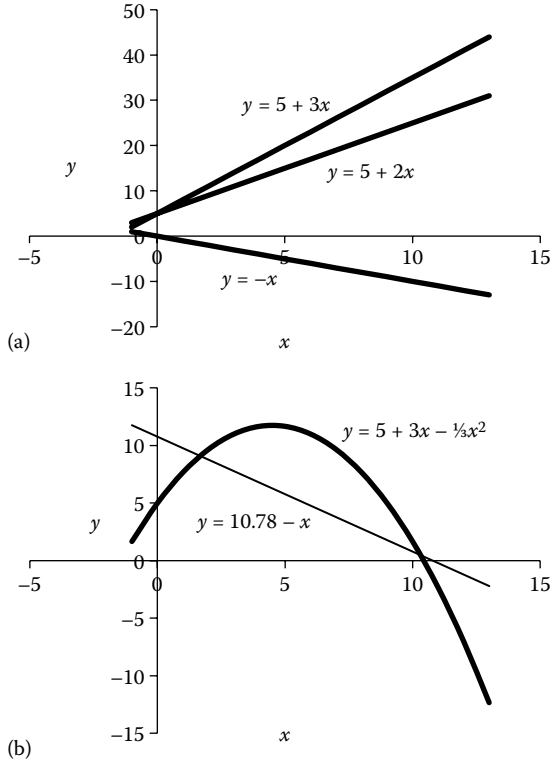



**FIGURE L.4** (a) Three lines and their equation. (b) Quadratic curve and best line.

represented by the line $y = 10.78 - x$, which shows that the linear component of the curve on the whole over the range ($x = -1$ to $+13$) is declining $y$ as $x$ increases. (ii) Although rarely used, the second is to ignore the quadratic term in the equation and say that the linear component of this curve is $y = 5 + 3x$ (deleting the term $-\frac{1}{3}x^2$). This would be difficult to interpret and not valid.

We have used easy examples of two variables $x$ and $y$ to explain linear effect, but the concept can be easily extended to many variables. The multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_3 x_K + \varepsilon$$

is a popular example. This model has $K$ regressor variables and each is considered to have a linear effect on $y$. This graphically depicts a hyperplane in multidimensions, which, in three dimensions (two variables on the left-hand side), is reduced to a plane sheet. Such a linear regression model is not without limitations as most medical parameters have a nonlinear effect, but it does capture the linear effect of the $x$'s on $y$. In the context of regression models, the term *linear* does not have this narrow sense as made out in this section but has a wider meaning since linearity in regression refers to linearity in parameters. This linearity includes curvilinear models such as quadratic curve discussed in this section. For a brief on this, see the topic **linearity (test for)**.

## linearity (test for), *see also* linear effect



For the meaning of linear, see the topic **linear effect**. There is no doubt that hardly any relationship in medicine is truly linear. Yet, a linear relationship is the most commonly studied form of relationship in health and medicine. This simplification seems to work fairly well in many situations but can destroy an otherwise very clear relationship in others. Square or such other terms of the regressors are rarely included in medical investigations and this can cause fallacies in some situations. Such terms should be included where needed.

In a narrow sense of linearity, it defines a line (or a hyperplane in multiple dimensions), but broadly, and strangely, statistical linearity also includes a curve such as a parabola. This is so when the square term is considered just another variable; that is, $x^2$ is denoted by $x_2$. For example, the quadratic equation $y = 5 + 3x - \frac{1}{3}x^2$ that is graphically a parabola (see Figure L.4b) can also be written as $y = 5 + 3x_1 - \frac{1}{3}x_2$, where $x_2$ is $x^2$, and this is linear in $x_1$ and $x_2$. This is categorized as curvilinear instead of nonlinear. Statistically, a relationship is nonlinear when any coefficient of the regression equation is nonlinear. Thus, the equation $y = a + bx_1 + b^2x_2$ is statistically nonlinear since the coefficient $b^2$ is nonlinear. This equation puts the restriction that the coefficient of $x_2$ is the square of the coefficient of $x_1$.

Researchers tend to look for one equation—either linear, curvilinear, or whatever—for the entire range of a measurement. Left ventricular ejection fraction (LVEF) ranges from 10% to 90%—less than 50% is considered a coronary risk and more than 60% does not provide any benefit. If coronary risk is plotted against LVEF, it will be flat after 60%. One model over the entire range 10% to 90% would not capture this phenomenon. For measurements that show one pattern for $x \le a$ and another for $x > a$, a **spline** function is needed. Without this, the results could be fallacious. Realize however that this could be used only if you know, or at least suspect, that the patterns could be different over different ranges of $x$. If this is not known, the fallacy can continue for a long time.

How do you find out whether the relationship between two or more medical variables can be considered linear or not? In the case of two variables, you can plot the scatter diagram of $y$ versus $x$, overlay a trend if needed, and examine if this trend is close to a straight line. However, this is inexact and can be subjective. Other methods can be used depending on the context. In a simple case when there are only two variables and the model has a square or any other term, just test the statistical significance of such a term. This can be done by the **Wald test** just as for any **regression coefficient**. If this is statistically significant, conclude that the relationship is not linear (in the narrow sense as earlier mentioned) but is curvilinear or nonlinear. In this case, the usual (Pearsonian) **correlation coefficient** would be low since this measures the strength of a *linear* relationship. If the correlation coefficient between $y$ and $x$ is high, you can be fairly confident that the relationship is mostly linear. Some statisticians may advise testing the statistical significance of the correlation coefficient but that is not a right procedure in this case since a low correlation in a clearly nonlinear relationship could also be statistically significant when the sample size is large. Another method is to divide the range of $x$ into three or four plausible categories and see if the regression coefficient for each range is nearly the same.

In the broad statistical sense, and when many variables are involved, an easy method to check linearity is to look at the value of the square of the **multiple correlation** coefficient ($R^2$). Note that this is calculated by using a multiple linear regression in the broad sense that all regression coefficients are linear. If the value of $R^2$ is high, say, exceeding 0.7, you can be quite confident that the relationship is approximately linear. Examples are available where this is not true, but those are exceptions. The other is to construct a model with whatever nonlinearity you suspect and compare the **coefficient of determination** ($\eta^2$) from this model with the $R^2$ obtained from the corresponding linear model and consider that the relationship is approximately linear if the difference is minor. For a formal statistical test, use the **likelihood ratio test** by using likelihoods of the two models.

A continuous regressor is generally entered as such in a **logistic regression** that implies that the effect of this continuous $x$ is linear on the **logit** of the outcome. Few researchers care to test if this is really so. Minor deviations do not matter, but U-shaped or J-shaped relationships can drastically affect the regression equation. For example, the regression between logit of incidence of coronary heart disease with post-glucose insulin in the general population is J-shaped in the sense that the incidence is high (even when converted to logit) both at low values of insulin and at very high and high values of insulin. To test linearity, check the plot of residuals against the regressor $x$. If the relationship is genuinely nonlinear, divide the values of $x$ into few rational categories and use appropriate scores, or use splines as already advised and fit separate regression for low values, middle values, and high values.

If you have proportions of "positive" response in different categories, such as percentage anemic in various parity women, linearity of trend in proportions with increasing parity can be tested by the **Cochran test for linear trend**.

## linear models, *see* general linear models

## linear regression, *see also* simple linear regression, multiple linear regression

Prediction of a dependent variable $y$ by a set of regressors (or independent variables) may take any of several different forms. One that is relatively simple to comprehend and most commonly studied is the linear form. For $K$ regressors, this is expressed as

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_K x_K,$$

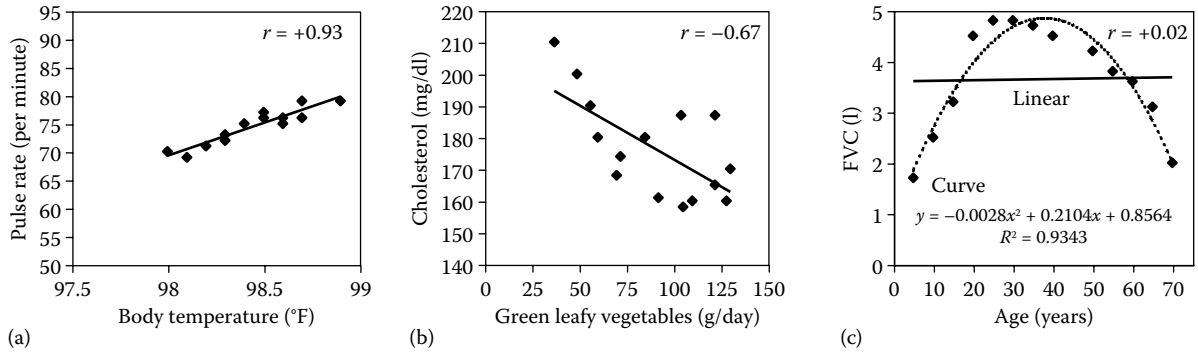



**FIGURE L.5** (a) Good linear regression of pulse rate on body temperature; (b) linear regression of cholesterol level on intake of green leafy vegetables due to two specific points; and (c) inappropriate linear regression of forced vital capacity (FVC) on age.

where $\hat{y}$ is the predicted value. This is linear in the coefficients $b_1$, $b_2$, …, $b_K$. Geometrically, this equation defines a hyperplane in $(K + 1)$-dimensional space and reduces to a line when $K = 1$. When there is only one independent variable $x_1$, the equation is $\hat{y} = b_0 + b_1 x_1$, which depicts a line. It is more convenient to write it as $\hat{y} = a + bx$, which is the usual equation for a **simple linear regression**. With $K$ independents, this defines **multiple linear regression**. The constants $b_1$, $b_2$, …, $b_K$ in this equation are called regression coefficients, and $b_0$ is called the intercept. These in fact are estimates of actual regression coefficients in the target population. The corresponding regression parameters in the population are denoted by $\beta_0$, $\beta_1$, $\beta_2$, …, $\beta_K$, respectively. It is important to realize that a linear model is linear in parameters and not necessarily linear in $x$'s. For details, see the topic **linear effect**. In the conventional regression setting, both the dependent set and the set of independents are quantitative, but that is not necessary for independents. Consider the following two examples that illustrate how simple and multiple linear regressions are used in practice.

Body surface area (BSA) determination is useful in several applications related to the body's metabolism such as ventilation, fluid requirements, extracorporeal circulation, and drug dosages. Current [1] found that this could be adequately estimated by using weight (Wt) in well-proportioned infants and children of weight between 3 and 30 kg. The regression equation obtained is

$$\text{BSA} = 1321 + 0.3433(\text{Wt}),$$

where BSA is in square centimeters and Wt is in grams. This is a simple linear regression and shows that as weight increases by 1 g, BSA increases on average by 0.3433 cm. The left-hand side is in fact an estimate, although the hat (^) sign is removed for simplicity. Here, $b_0 = 1321$ and $b_1 = 0.3433$. As a side note: according to the authors, a simplified version of this regression is BSA = (Wt + 4)/30, where BSA now is in square meters and Wt is in kilograms.

Marquis et al. [2] studied length (in centimeters) of 15-month-old (L15m) Peruvian toddlers living in a shantytown of Lima. This length was studied in relation to length at 12 months (L12m), time interval (TI) in months between 12- and 15-month measurements (as per the authors), breast-feedings (BFs) per day between 12 and 15 months of age, and its interaction with a low diet/high diarrhea (LDHD) combination if present. The following equation was obtained:

$$\text{L15m} = 74.674 + 0.976(\text{L12m}) + 0.860(\text{TI}) + 0.043(\text{BF}) - 0.157(\text{BF}*\text{LDHD}),$$

where the last term represents an **interaction**. This also is a linear regression. The authors also studied other regressors, but this equation includes only those that are relevant for our illustration. Of these, BF was not significant ($P > 0.3$). The associated signs show that L12m, TI, and BF had a positive contribution to L15m, but the interaction BF*LDHD had a negative contribution; that is, when low diet/high diarrhea is present with a high number of breast-feedings, the length of the child is less. This was part of an investigation to show that increased breast-feeding did not lead to poor growth but poor health led to increased breast-feeding—an example of what authors called **reverse causality**.

Linear regression is an overused model in health and medical setups. Whereas the relationship between pulse rate and body temperature is indeed linear (Figure L.5a), this may not be so for other measurements. In Figure L.5b, the trend looks like a decline in cholesterol with increased intake of green leafy vegetables just because of the first two points in the scatter. If these two points are ignored, the other points are randomly scattered with no visible trend. This highlights the need to critically examine statistically obtained relationships and not take on its face value. Even more daunting is the relationship between forced vital capacity (FVC) and age from 5 to 70 years (Figure L.5c). FVC increases up to the age of 35–40 years in this figure and declines thereafter. If you try to fit a linear regression of FVC on age 5 to 70 years, it will have a nearly zero slope, indicating that as age increases, FVC remains constant. This absurd result is due to a completely inappropriate application of linear regression. The scatterplot clearly indicates that a curve should be fitted to these data instead of a line. This comes under the rubrics of **curvilinear regression**. This illustrates why there is a need to distinguish between linear models in a narrow sense (linear in $x$'s) and linear models in a broad sense, which includes curvilinear relationships (linear in parameters).

1. Current JD. A linear equation for estimating the body surface area in infants and children. *Internet J Anesthesiol* 1997;2(2). https://ispub.com/IJA/2/2/10302
2. Marquis GS, Habicht J, Lanata CF, Black RE, Rasmussen KM. Association of breastfeeding and stunting in Peruvian toddlers: An example of reverse causality. *Int J Epidemiol* 1997;26:349–56. http://ije.oxfordjournals.org/content/26/2/349.full.pdf+html

## line diagrams

A line diagram is used to show the trend of one variable over another by means of a line or a set of lines. For example, Figure L.6a shows the trend of infant mortality rate against the socioeconomic status of

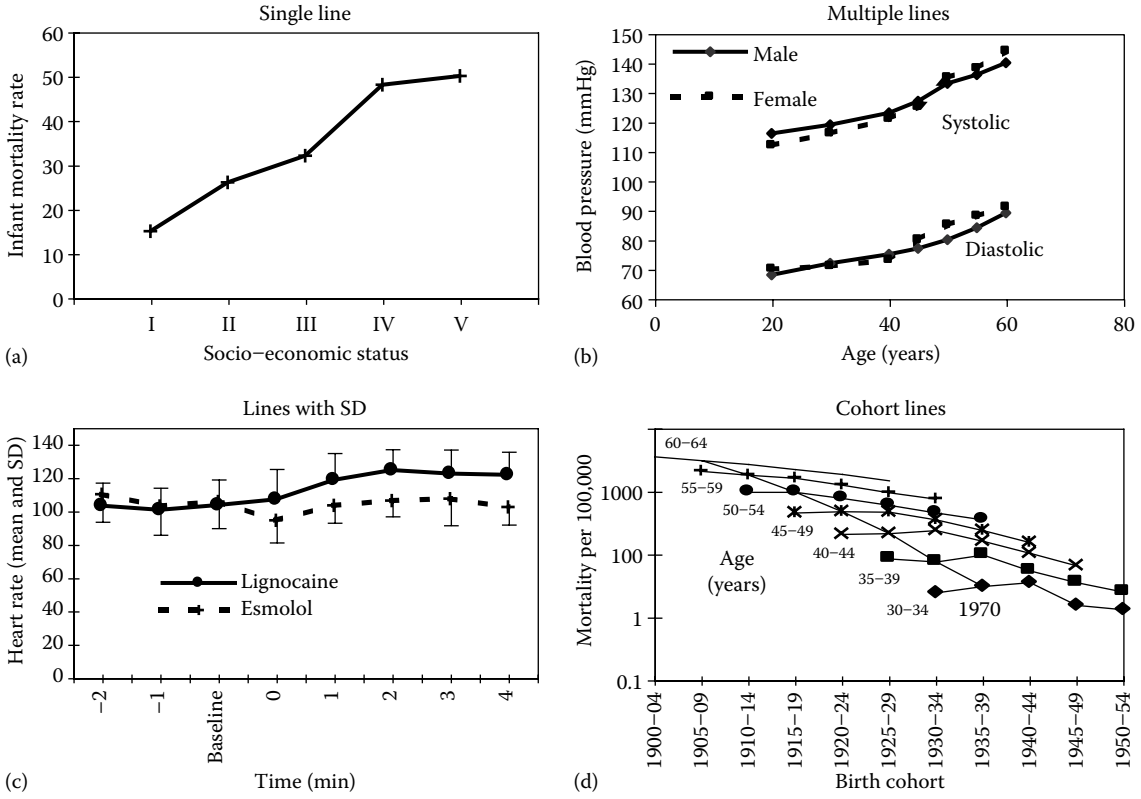



**FIGURE L.6** Different types of line diagram: (a) Infant mortality rate per 1000 live births by socioeconomic status in an area. (b) Trend of average systolic and diastolic BP over age in males and females in a general population of adults. (c) Heart rate in women with pregnancy-induced hypertension undergoing cesarean section—esmolol and lignocaine groups. (d) Trend in age-specific coronary heart disease mortality per 100,000 in Australia from 1960 to 1992 within 5-year birth cohorts for females aged 30–64 years (*y* axis is on logarithmic scale). (From Wilson A, Siskind V. *Int J Epidemiol* 1995;24:678–84. http://ije.oxfordjournals.org/content/24/4/678.full.pdf. Permission to be obtained from Oxford University Press.)

families in a developing country. The plotted points are joined by a line. It can be called a time plot when the *x*-axis is time. It continues to be called a line diagram even when the representation is by a curve instead of a straight line. The variable on the *y*-axis could be an average as in Figure L.6b between age and mean blood pressure. Although not clearly brought out in this diagram, a line diagram can plot the points for unequally spaced values on the *x*-axis. This facility is not available in a **bar diagram**. Such adjustment can be done when the *x*-axis is on a metric scale but not when it is on an ordinal scale. The question of trend does not arise in the case of the nominal scale because then the categories do not have any particular order.

Two or more lines can be drawn in one diagram (Figure L.6b), but more than five lines make it cluttered and difficult to understand. Another complexity arises when the standard deviation (SD) of the variable for each point on the *x*-axis is shown on the *y*-axis by vertical lines on either side as in Figure L.6c. This gives an indication of the variation present in the data. Realize for this that representation of 1*SD on either side can be fallacious because the actual variation is much more.

An interesting representation by line diagram could be of mortality in several cohorts. Figure L.6d shows the age-specific coronary heart disease mortality rate in different 5-year birth cohorts for females in Australia from 1910–1914 to 1950–1954 [1]. Note that the cross-sectional lines can be used to show the age trend in mortality in a particular group of years. One such line is drawn for the years 1970 ± 5. This is because various cohorts attain different ages in the same calendar year.

1. Wilson A, Siskind V. Coronary heart disease mortality in Australia: Is mortality starting to increase among young men? *Int J Epidemiol* 1995;24:678–84. http://ije.oxfordjournals.org/content/24/4/678.full.pdf

## linkage, *see* record linkage

## link functions

This topic is directly related to **generalized linear models**. You may be aware that ordinary quantitative regression covered by **general linear regression** models (which includes regression, analysis of variance, and analysis of covariance) requires that the residuals have a Gaussian (Normal) distribution. This requirement is dispensed with in generalized linear models where most other distributions can be studied through the link functions that tend to convert the residuals to a Gaussian form, at least approximately. These link functions also tend to make linear relationship more plausible and also help in stabilizing variances.

Consider, for example, the number of deaths in a critical care unit of a large hospital in a day, which is a discrete variable with a small number of possible values and is likely to have a highly skewed distribution in the sense that the number of deaths would be 0, 1, or 2 on most days but can even go up to 5 or 6 on isolated days. If this is the dependent variable in a regression, the residuals are not likely to follow a Gaussian distribution. This kind of count quite often follows a

**TABLE L.3**
**Link Functions for Various Distributions**

| Distribution | Type and Range of $y$ | Name of Link Function | Link Function |
|---|---|---|---|
| Normal | Real $-\infty < y < \infty$ | Identity | $\mu$ (No transformation) |
| Poisson | Positive integer 0, 1, 2, … Or, rate ($y/t$) | Log | $\ln(\mu)$, $\sqrt{n}$ |
| Binomial | Binary 0,1 | Logit | $\ln\left(\frac{\pi}{1-\pi}\right)$ |
| | Binary $<a$, $\geq a$ (underlying Gaussian) | Probit | $\Phi^{-1}(\pi)$ |
| | Binary 0,1 (highly skewed) | Complementary log–log | $\ln(-\ln(1-\pi))$ |
| Multinomial | Nominal | Logit | $\ln\left(\frac{\pi}{1-\pi}\right)$ |
| Gamma | Real but positive $0 < y < \infty$ | Inverse | $\mu^{-1}$ |

*Note:* Link functions are expressed for the population mean of the dependent variable instead of the variable itself.
$\pi$ is the probability of "success."
$\Phi$ is the notation for the Gaussian probability.

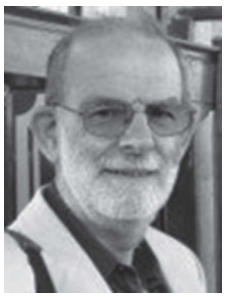
Poisson distribution with the variance nearly the same as the mean. When this is so, use transformation ln($y$), where $y$ is the number of deaths in a day. You can obtain $y$ for several hospitals or several days in a hospital, but the condition of independence must be fulfilled for regression model to be valid. Suppose the regressor of interest is the severity of the condition at the time of admission. The objective is to find how much the number of deaths is affected by the different combination of severities of the patients. On any day, there might be $x_1$% serious cases, $x_2$% critical cases, and $x_3$% gravely critical cases. This relationship can be examined by ln (natural logarithm) link as just stated. Similar links are available for many other distributions (Table L.3). While all other links may be readily clear, a word about probit link will help. This link is used when the underlying variable is continuous and has a Gaussian distribution but is observed in two exclusive categories. This may be like the bilirubin level in healthy subjects and the cutoff such as 20 mg/dL is used to dichotomize the subjects. Complementary log–log is used for highly skewed binary variables such as smokers and nonsmokers among adolescents where nonsmokers hugely outscore smokers.

## listwise deletion, *see* casewise, pairwise, and listwise deletion

## literacy rate, *see* education indicators

## LMS method

Lambda–Mu–Sigma (LMS) is a popular method of constructing smooth **centile** curves describing growth in children and such other parameters that change with time. The LMS method was initially developed by Tim Cole in 1990 [1]. We explain this for growth curves. Application to other setups such as bone mineral density at different ages is similar.


Tim Cole

The construction of **growth charts** involves two distinct steps: (i) correctly finding various percentiles at different ages and (ii) converting the plot of percentiles versus age to a smooth curve. The first becomes difficult because most growth parameters (such as weight and height) rarely follow a Gaussian pattern; thus, finding the correct percentile is complicated. The second is challenging because smoothing can lose the real points of troughs and peaks that define children's growth pattern. The LMS method is designed to overcome most of these problems. "Most," first because this method fails to address the problem of deviations in **kurtosis** that also occurs in the distribution of many growth parameters, and second because now better methods are available for the smoothing than used in the LMS method. For these, an improvement called the **BCPE method** is used as separately described. Indrayan [2] has written a seminal paper on the LMS and BCPE methods and how they are used, with the aim of demystifying the statistics around these methods.

A large number of other methods are also available for constructing centile curves. Borghi et al. [3] have reviewed 30 such methods, but the LMS methods became kind of the standard method until the BCPE method appeared. The BCPE method is more complicated and many researchers continue to use the LMS method because of its relative simplicity.

The fundamental quantity in finding the centiles is the ***Z*-score**. This is the deviation of the value from its mean in standard deviation units. In notations, this is $Z = (y - \mu)/\sigma$. If the distribution is **Gaussian** (Normal), $Z = 1.96$ corresponds to the 97.5th percentile and $Z = 1.28$ corresponds to the 90th percentile. Software packages give these percentiles easily for any value of $Z$. The difficulty arises when the distribution is non-Gaussian. Several types of deviations from Gaussianity can occur, but most growth parameters are nice to follow a unimodal pattern for each age. That is, if you measure the weight of 400 healthy children of age 3 years, the distribution will have one peak. However, the distribution is not likely to be symmetric and it may not have normal kurtosis. Generally speaking, **skewness** |Sk| > 0.5 and kurtosis |Kurt| > 1.0 are considered high. The statistical significance of these values can also be tested. In these situations, $Z$-scores do not have a valid interpretation. The LMS method takes care of the skewness but not of the kurtosis.

The conventional method that deals with skewness is the use of transformations such as square root or square, generally after dividing each value $y$ of the measurement under consideration by its central value $\mu$, that is, after calculating $y/\mu$. This central value could be mean or median or any other value. Now, the square-root transformation is $(y/\mu)^{0.5}$ and the square transformation is $(y/\mu)^2$. In general, the power transformation is $(y/\mu)^\lambda$, $\lambda > 1$ is for correcting left skewness, and $\lambda < 1$ is for correcting right skewness; the exact value of $\lambda$ depends on the extent of skewness. If the distribution is already Gaussian, no correction is required, and then $\lambda = 1$.

The LMS method is an improvement over these transformations and uses the following:

$$\text{(A)} \quad Z_{\text{LMS}} = \frac{1}{\sigma_L \lambda}\left[\left(\frac{y}{\mu}\right)^{\lambda} - 1\right], \ (y, \lambda, \mu, \sigma L \neq 0),$$

where $\sigma_L$ is a measure of dispersion ($\sigma_L$ is generally the **coefficient of variation**: $\sigma/\mu$). The original measurements such as weight in our child growth example may have any skewed distribution with single mode; the distribution of $Z_{\text{LMS}}$ with this transformation will be standard normal and this will give the correct $Z$-score for calculating the percentile provided the kurtosis is already zero. Note the involvement of **l**ambda ($\lambda$), **m**u ($\mu$), and **s**igma ($\sigma_L$), making it an LMS method. The rationale of $(y/\mu)^{\lambda}$ is already explained and $\sigma_L$ is in the denominator just as is $\sigma$ in $z = (y - \mu)/\sigma$. Note that when $\sigma_L = \sigma/\mu$, and $\lambda = 1$, the equation above reduces to the usual $Z$-score $(y - \mu)/\sigma$. The reason that coefficient of variation (CV) is used in place of the usual $\sigma$ is that $\sigma$ tends to rise with age for many measurements, whereas CV remains nearly constant for different age-groups.

The first difficult part is to estimate the values of $\lambda$, $\mu$, and $\sigma_L$. Explicit forms for estimating these parameters do not exist, and special software (LMSchartmaker of the Medical Research Council, UK, is the software package of choice for this purpose) is used to find the values of these parameters that maximize the likelihood of the transformed values of the sample that have come from a standard Gaussian distribution with mean = 0, SD = 1, and skewness = 0. Software will find those values of $\lambda$, $\mu$, and $\sigma_L$ that make the distribution of $Z_{\text{LMS}}$ closest to standard normal. However, the estimates of $\lambda$, $\mu$, and $\sigma_L$ are iteratively revised by the method of penalized likelihood proposed by Cole and Green [4] so that the curves for these estimates against age are smooth. The **Akaike information criterion** is used to test goodness of fit so that a higher number of parameters are penalized and the model can remain relatively simple. Sometimes, age is also transformed into $\sqrt{\text{age}}$, which stretches younger (<1 year) and shrinks older (>1 year) age and takes care of the steep rise in growth parameters during the infantile period. The software LMSchartmaker also has provision for this and to provide curves for different percentiles.

There is a mathematical relationship between usual $Z$ and LMS percentile:

$$\text{(B)} \quad p\text{th percentile of } y = \mu(1 + \lambda\sigma_L Z_p)^{1/\lambda},$$

where $Z_p$ is the usual value from Gaussian distribution corresponding to the $p$th percentile. For example, for the 75th percentile, $Z_p = 0.675$. If the mean weight of children of age 2 years is 13.6 kg, and if $\sigma_L = 0.147$ and $\lambda = 0.30$ from the LMSchartmaker, then the 75th LMS percentile = $13.6 \times (1 + 0.30 \times 0.147 \times 0.675)^{1/0.30} = 13.6 \times 1.103 = 15.0$ kg. Once estimates of $\lambda$, $\mu$, and $\sigma_L$ are obtained, Equation B shows that calculating percentiles at any particular age for a skewed distribution is very easy. Since the curves of the estimates of $\lambda$, $\mu$, and $\sigma_L$ are already smoothed, the percentiles so obtained are likely to follow a smooth curve. To overcome the possibility of having not-so-smooth growth curves via the LMS method, Centers for Disease Control and Prevention of the United States used a type of inverse method. They first obtained the agewise percentiles, smoothed these curves, and then estimated the LMS parameters based on the smoothed percentile curves [5].


1. Cole TJ. The LMS method for constructing normalized growth standards. *Eur J Clin Nutr* 1990 Jan;44(1):45–60. http://www.ncbi.nlm.nih.gov/pubmed/2354692
2. Indrayan A. Demystifying LMS and BCPE methods of centile estimation for growth and other health parameters. *Indian Pediatr* 2014 Jan;51(1):37–43. http://www.indianpediatrics.net/jan2014/jan-37-43.htm
3. Borghi E, de Onis M, Garza C, Van den Broeck J, Frongillo EA, Grummer-Strawn L, Van Buuren S et al. Construction of the World Health Organization child growth standards: Selection of methods for attained growth curves. *Stat Med* 2006;25:247–65. http://www.stefvanbuuren.nl/publications/Construction%20WHO%20-%20Stat%20Med%202006.pdf, last accessed January 29, 2015.
4. Cole TJ, Green PJ. Smoothing reference centile curves: The LMS method and penalized likelihood. *Stat Med* 1992;11(10):1305–19. http://www.ncbi.nlm.nih.gov/pubmed/1518992
5. Flegal KM, Cole TJ. Constriction of LMS parameters for the Centers for Disease Control and Prevention 2000 growth charts. *National Health Stat Rep* 2013 Feb 11;13:1–3. http://www.cdc.gov/nchs/data/nhsr/nhsr063.pdf


## LMSP method, *see* Box-Cox power exponential (BCPE) method

## loadings, factor

The concept of loadings arises while performing **factor analysis**. The statistical purpose of factor analysis is to obtain each observed variable as a combination of a few unobservable factors, that is,

observed value of a variable = linear combination of factors + error.

If the observed variables are $x_1, x_2, \ldots, x_K$, the factor analysis seeks the following:

$$\left.\begin{aligned} x_1 &= a_{11}F_1 + a_{12}F_2 + \cdots + a_{1M}F_M + U_1, \\ x_2 &= a_{21}F_1 + a_{22}F_2 + \cdots + a_{2M}F_M + U_2, \\ &\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ &\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ x_K &= a_{K1}F_1 + a_{K2}F_2 + \cdots + a_{KM}F_M + U_K, \end{aligned}\right\},$$

where $F_1, F_2, \ldots, F_M$ are the $M$ unobservable factors common to $x_k$'s, and $U_k$'s ($k = 1, 2, \ldots, K$) are called unique factors. Unobservable common factors are also called *constructs*.

The coefficients $a_{km}$ ($k = 1, 2, \ldots, K$; $m = 1, 2, \ldots, M$; $M << K$) are estimated by the factor analysis procedure. These coefficients are called loadings and measure the importance of the factor $F_m$ in the variable $x_k$ on a scale zero to one when sign is ignored. The loadings close to −1 or +1 indicate that the factor strongly affects the variable, and the loadings close to zero indicate that the factor has a weak effect on the variable. When a loading is very small, say less than 0.20, the corresponding factor can be dropped from the above equation. Thus, different variables may contain different sets of factors. Some factors would overlap and would be present in two or more variables.

The factor's equations differ from the usual multiple regression equations because the $F_m$'s are not single independent variables. Instead, they are labels for a combination of variables that characterize these constructs. They are obtained in such a manner that they are uncorrelated with one another. The statistical method of **principal components** is generally used for this purpose. You may like to see Kline [1] for details. A relevant statistical software package would easily provide an output, but the software package may ask you to specify different aspects of the methodology. This exercise should be undertaken by only those who understand the intricacies. Do not hesitate to obtain the help of a biostatistician when needed.

Kim et al. [2] studied reliability and validity of the Korean version of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire to assess chemotherapy-induced peripheral neuropathy (CIPN). This questionnaire had 20 items, but the factor analysis confirmed three dimensions of CIPN, namely, the sensory, motor, and autonomic. These are the factors in this example. The factor loadings in these 20 items ranged from 0.38 to 0.85. Thus, none of the factors was unimportant for any item in this questionnaire.

1. Kline P. *An Easy Guide to Factor Analysis*. Routledge, 1994.
2. Kim HY, Kang JH, Youn HJ, So HS, Song CE, Chae SY, Jung SH, Kim SR, Kim JY. Reliability and validity of the Korean version of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire to assess chemotherapy-induced peripheral neuropathy. *J Korean Acad Nurs* 2014 Dec 31;44(6):735–42. http://synapse.koreamed.org/DOIx.php?id=10.4040/jkan.2014.44.6.735, last accessed January 30, 2015.

## location (measures of)

In statistics, location refers to the physical location of the **distribution** of the variable under consideration on the *x* axis. In a statistical distribution, the *x*-axis is the one that plots the values of the characteristic. For example, direct bilirubin is lower than indirect bilirubin; thus, the distribution of direct bilirubin would be located to the left of the distribution of indirect bilirubin. Figure L.7 shows two distributions that differ in location but are otherwise identical.

Measurement of location requires that we concentrate on some specific values or some representative value. If we restrict ourselves to central values, the statistical measures are **mean, median, and mode**. **Geometric mean** and **harmonic mean** are also measures of central location, which are separately explained. However, the interest sometimes is not in any central value but in a threshold below which, say, 90% of values lie. This is the 90th **percentile** and is also a measure of location, though not of central location. Similarly, tertiles, quartiles, and deciles also are measures of location of a distribution. All these are explained under the topic **quantiles**. Figure L.7 shows how 90th percentiles of the two distributions differ when they differ in location and not in any other feature.

## logarithmic scale/transformation

In case you are not aware, logarithm (abbreviated as log) is a mathematical function that converts a very large number to a very small number. It depends on the "base" used for this conversion. If nothing is stated, base is taken as 10. If $y = 10^x$ then $\log y = x$. This is the definition of logarithm. Thus, log of 100 is 2, log of 1000 is 3, and log of 1,000,000 is 6. See how logarithm sharply attenuates the number as the number becomes large. Also note that log of 1 is zero and the log of values less than 1 is negative, reaching $-\infty$ for log of zero. Thus, values between 0.001 and, say, 1000 are converted to –3 to +3. Also, $\log(a \times b) = \log a + \log b$. Thus, multiplicative values become additive—a great mathematical convenience. The other log required in biostatistics is to the base *e*. This is called the Naperian base and the log taken with this base is called the *natural log*, denoted by ln. This base helps in mathematical manipulations such as integration and differentiation that base 10 does not do. This also has similar properties, and neither of them can be used for negative values. If you have negative values and you think that log transformation would help, add slightly more than the maximum negative value (minimum value in mathematical language) to all the values so that none is zero or negative. This is readjusted to the original values at the time of making inference.

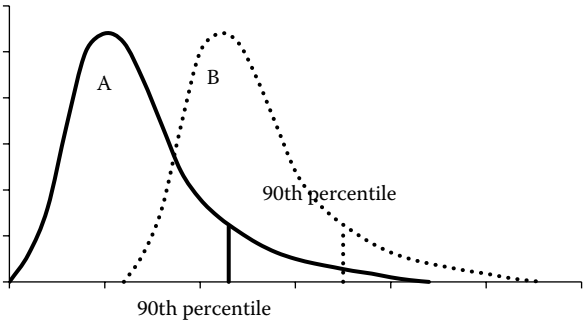



**FIGURE L.7** Different locations of two distributions.

Where do we use log-scale in health and medicine? It is used for any characteristic that is multiplicative in nature. Radiation dose is a common example. This is just approximately 5 μSv for dental x-ray, 100 μSv for chest x-ray, 4000 μSv for mammogram, and 10,000 μSv for average computed tomography scan. In conventional radiotherapy, this can go up to 2,000,000 μSv. Acidity measured in pH (e.g., in blood) also has this feature. Each whole pH value below 7 is 10 times more acidic than the next higher value; for example, a pH of 6 is 10 times more acidic than a pH of 7. The same holds true for pH values above 7, each of which is 10 times more alkaline than the next lower whole value; for example, a pH of 9 is 100 times more alkaline than a pH of 7. Titers also have a similar feature. All these are expressed in log units.

In biostatistics, log-scale is used for values that have an extremely wide range. Generally, these are those values that steeply go up or steeply go down in different subjects or in different situations. Logarithmic transformation in these setups reduces very large numbers to smaller numbers that can be easily represented or talked about. See the figure in the topic **line diagram** where mortality rate on the vertical axis is shown in log-scale. This rate in that figure has steeply come down from what it was in 1900. In this kind of figure, if we have to show a mortality rate of 1000 (per 100,000) and also of 1 by the usual (linear) scale, we need an extremely long vertical axis that may not even fit a page. Log-scale is a great convenience in such cases. This is illustrated in Figure L.8 that shows Ebola virus cases in West Africa in the first 9 months of epidemic on the usual (linear) scale in the left panel and on logarithmic scale in the right panel. The exponential rise is clear. If the same trend continued, the number of cases after 1 year of epidemic was projected to be nearly 80,000 [1]. To show this, the vertical axis on the linear scale in the left panel will have to be extended to a steep height but is easily shown on log-scale in the right panel. Although the log-scale conceals the high speed with which the cases were growing and expected to grow, the shape becomes amenable to modeling.

Suh et al. [2] used logarithmic transformation of the thickness of the retinal nerve fiber layer to study relationship with visual field indices. In this case, log-scale helped to linearize the relationship. Tomitaka and Furukawa [3] also used log-scale for major depressive disorder durations that ranged from a month to several years—again to get a linear relationship. There are many such examples in the literature.

There are several other statistical uses of log-scale. The variables that have highly **skewed** distribution on the positive side are transformed on log-scale to achieve relative symmetry, possibly Gaussian distribution. This helps in using usual statistical methods on the transformed values. The second common use is in studying ratios and proportions. Log of odds ratio and log of relative risk linearize them and make them additive—thus raising the specter of applicability of the **central limit theorem**. This theorem is for additive indices

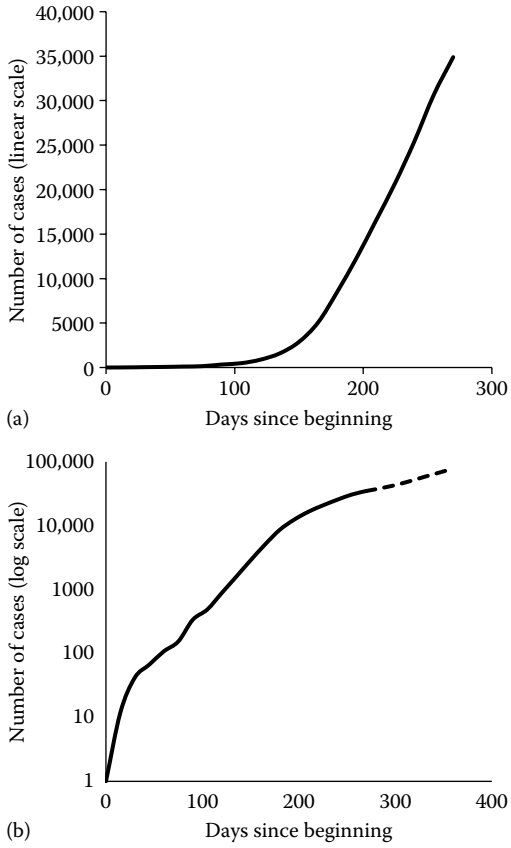



**FIGURE L.8** Ebola virus cases in the first 9 months in West Africa: (a) linear scale (smoothened curve); (b) log-scale with projection at 1 year since inception.

and helps us to legitimately fall back on Gaussian distribution. Third, proportional hazards such as in a **Cox regression** can be studied as difference (in place of ratio) after log transformation. Fourth, the Cox regression also assumes that the covariates affect the hazard in a multiplicative manner. This means that when two factors are simultaneously present, the hazard multiplies instead of adding. This property (multiplication) amounts to additivity in logarithm terms.

The following precautions are required when interpreting the log-transformed data:

- If $\ln(y)$ has a Gaussian distribution, its mean is $\ln(\mu + \frac{1}{2}\sigma^2)$ and not $\ln(\mu)$. Thus, antilog of the mean of $\ln(y)$ will not give you the mean of $y$. For log-transformed values, use geometric mean (GM) in place of arithmetic mean; that is, the arithmetic mean of log values is the log of GM. If the interest in arithmetic mean persists, correct the mean by adding $\frac{1}{2}\sigma^2$.
- When the dependent variable in a regression is log-transformed, the regression coefficient $\beta$ measures approximate *proportionate* change (better understood as percentage change when multiplied by 100) in $y$ per unit change in $x$. This can be seen as follows: When $y$ is log-transformed, the simple linear regression equation is $\ln(y) = a + bx$. When 1 is added to $x$, that is, $x_1 = x_0 + 1$, then $\ln(y_0) = a + bx_0$ becomes $\ln(y_1) = a + bx_0 + b = \ln(y_0) + b$, and $\exp(b) = y_1/y_0$.
- If interested in percentage change in $y$ when $x$ increases by 1, this is $100*(y_1 - y_0)/y_0 = 100*y_1/y_0 - 1 = 100*[\exp(b) - 1]$. Thus, it is not directly obtained by the regression coefficient $b$. However, in most situations, the coefficient $b$ would be extremely small because of log transformation, and in that case, $\exp(b) - 1 = b$ approximately, and you would be safe in making the usual conclusion.
- If both $x$ and $y$ are log-transformed, the regression will give an approximate percentage change in $y$ for a percentage change in $x$.
- All these arguments can be extended to multiple linear regression while considering individual variables.

1. WHO Ebola Response Team. Ebola virus disease in West Africa—The first 9 months of the epidemic and forward projections. *N Eng J Med* 2014;371:1481–95. http://www.nejm.org/doi/full/10.1056/NEJMoa1411100
2. Suh W, Lee JM, Kee C. Depth and area of retinal nerve fiber layer damage and visual field correlation analysis. *Korean J Ophthalmol* 2014 Aug;28(4):323–9. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120353/
3. Tomitaka S, Furukawa TA. Mathematical model for the distribution of major depressive episode durations. *BMC Res Notes* 2014 Sep 12;7:636. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4246456/

## logistic coefficients (CI and test of $H_0$)

This section presumes familiarity with the basics of **logistic regression**. If not, review that topic in this volume. For $K$ regressors, this regression takes the form

$$\hat{\lambda} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_K x_K,$$

where $\hat{\lambda} = \ln[p/(1-p)]$, and $p$ is the observed proportion of occurrence of the event of interest in the sample. Statistical software packages easily provide the values of $b_0$, $b_1$, $b_2$, …, $b_K$, which are the estimates of the corresponding parameters $\beta_0$, $\beta_1$, $\beta_2$, …, $\beta_K$, respectively. These are called logistic coefficients. You may also like to review the topic **logistic coefficients (interpretation of)** to understand the meaning of these coefficients. Since the $b$'s are estimates, we can find the confidence interval (CI) for the corresponding $\beta$'s and can also test hypothesis on them. The CIs and $P$-values for significance of the logistic coefficients are standard features in reporting the results of a logistic regression.

When all of the specified regressor variables are forced into the model and not selected by any of the **stepwise** algorithms, there is a need to test the statistical significance of each logistic coefficient. For this, one method is to refer the difference between $-2\ln L$ without and with the variable in question to **chi-square** with one df (see **likelihood ratio test**). The other method is to refer $z = b/\mathrm{SE}(b)$ to the Gaussian distribution, where $\mathrm{SE}(b)$ is the standard error of $b$. $\mathrm{SE}(b)$ has a complex expression, but its value is routinely provided by the statistical software. Both chi-square and Gaussian $z$ require large $n$. Most statistical software perform the Gaussian test; its square, called the **Wald** statistic, is referenced to chi-square with one df for testing significance. Consider the following elaborate example that illustrates some of these facets of logistic regression.

Among a group of Peruvian children studied by Marquis et al. [1], 127 were still breast-feeding at the end of their 13th month of age. These children were thus at *risk* of being weaned. The objective was to find support for the contention that poor health of the child determined maternal breast-feeding practice and maternal breast-feeding is not a precursor to poor health. The following variables were investigated for their role in determining the weaning: (i) $Z$-score

of 12-month weight-for-age (WA), (ii) complementary food intake (FOOD) between 9 and 12 months of age in terms of number of foods (out of 27 groups), and (iii) change (CH) in diarrheal incidence from 9–12 months to 12–15 months of age. As these are to be considered precursors for weaning during the 14th month, consider CH from 9–11 to 11–13 months for our example. The response variable is the probability of weaning during the 14th month of age. All variables were centered on the mean, and the results so obtained by logistic regression are summarized in Table L.4. The table contains the estimated logistic coefficients, their SE values, and the two-tailed *P*-values for assessing their statistical significance. All these easily come from a statistical software package.

The aim of a logistic model is to predict or explain $\lambda = \ln[\pi/(1-\pi)]$, and thereby $\pi$, by a linear combination of the regressor variables. In this example, $\lambda$ is the logarithm of odds of weaning. Using the regression coefficient in Table L.5, the logistic equation obtained is

$$\ln\frac{p}{1-p} = -3.303 - 0.024(\text{WA}) + 0.093(\text{FOOD}) + 0.542(\text{CH}) + 0.046(\text{WA}\times\text{FOOD}) + 0.336(\text{WA}\times\text{CH}) - 0.143(\text{CH}\times\text{FOOD}) - 0.202(\text{WA}\times\text{CH}\times\text{FOOD}).$$

A coefficient *b* close to 0 implies that the corresponding explanatory variable is not a good predictor. Whether it is significantly different from 0 can be tested by referring [*b*/SE(*b*)] to the standard Gaussian or its square to the Wald statistic as mentioned earlier. The *P*-values thus obtained are given in the last column of the table. Diarrheal change (CH) is statistically significant at $\alpha = 0.10$ but not at the conventional $\alpha = 0.05$, where $\alpha$ is the **level of significance**. The only regressor variable that shows significance ($P < 0.05$) in this case is in the last row. This is the **interaction** between WA, FOOD, and CH. The coefficient *b* is negative for this interaction, which indicates that logarithm of odds of weaning *declines* as WA, FOOD, and CH change *together* such that WA × FOOD × CH increases. When log-odds decreases, the corresponding probability also decreases. The authors of this investigation explained that the probability of weaning decreases especially when low FOOD and low WA are accompanied by an increase in diarrheal incidence. These three together indicate poor health of the child. Thus, the inference is that a child who is breast-fed until 13 months of age is less likely to be weaned during the 14th month if its health condition is poor, hence the conclusion that health condition determines maternal breast-feeding practice.

**TABLE L.4**
**Results of Logistic Regression of Weaning on Health of the Children**

| Variable | *b* | SE(*b*) | *P*-Value |
|---|---|---|---|
| Constant | –3.303 | 0.644 | 0.000 |
| 12-month *Z*-score—WA | –0.024 | 0.446 | 0.958 |
| 9–12 months—FOOD | 0.093 | 0.267 | 0.733 |
| Diarrheal change—CH | 0.542 | 0.322 | 0.095 |
| Interaction 1—WA × FOOD | 0.046 | 0.211 | 0.833 |
| Interaction 2—WA × CH | 0.336 | 0.179 | 0.064 |
| Interaction 3—FOOD × CH | –0.143 | 0.144 | 0.323 |
| Three-factor interaction—WA × FOOD × CH | –0.202 | 0.096 | 0.038 |

**TABLE L.5**
**Classification Table for the Subjects**

| Observed Group | Predicted by Logistic Model | | | |
|---|---|---|---|---|
| | Control | Case | Total | Percent Correct |
| Control | 23 | 7 | **30** | 100 × 23/30 = 76.7 |
| Case | 8 | 22 | **30** | 100 × 22/30 = 73.3 |
| | | | **Overall** | **100 × 45/60 = 75.0** |

The example illustrates the method to test the significance of individual coefficients. If there are many regressors as in this example, an alpha level of 0.05 for each individual predictor may be too high because all tests are based on the same data. Exercise caution and see if you would like to use a **Bonferroni** type of procedure to adjust the alpha level instead of the one used in Table L.5.

Some researchers prefer providing results of logistic regression for only those regressors that turn out to be statistically significant. However, the results of logistic regression must also contain a list of other regressors that were considered but were not found statistically significant so that a comprehensive view is available to the reader.

Ninety-five percent **confidence interval (CI)** for each logistic coefficient is obtained as $b \pm 1.96\text{SE}(b)$ using Gaussian approximation. This is converted to OR by computing its exponent. This procedure is valid only for large *n*. When such a CI is obtained, there is no need to perform the test of hypothesis $H_0$: $\beta = 0$ on individual coefficients since CI for any regressor containing zero indicates that the regressor is not significantly affecting the odds of the dependent variable when the values of other regressors are fixed. Thus, that factor is not a useful predictor or as explanatory. When there are a large number of regressors, CIs on individual coefficients may not be very useful for any *joint* conclusion because joint probability may be very different.

1. Marquis GS, Habicht J, Lanata CF, Black RE, Rasmussen KM. Association of breastfeeding and stunting in Peruvian toddlers: An example of reverse causality. *Int J Epidemiol* 1997;26:349–56. http://ije.oxfordjournals.org/content/26/2/349.full.pdf+html

## logistic coefficients (interpretation of), *see also* logistic coefficients (CI and test of $H_0$)

To understand this section, we presume that you are aware of **logistic regression**. Briefly, logistic regression is the regression between logit of *p* defined as $\lambda = \ln[p/(1-p)]$, where *p* is the observed proportion of occurrence of the event of interest in the sample and the regressors. This takes the form

$$\hat{\lambda} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_Kx_K$$

for *K* regressors. This is also called the logistic regression equation. Your software easily provides the values of $b_0, b_1, b_2, \ldots, b_K$, which are the estimates of the corresponding parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_K$, respectively. These are now called logistic coefficients. Note that logit in the above equation is actually log of **odds**.

The coefficients $b_1, b_2, \ldots, b_K$ in the logistic regression formula given in the above equation have an extremely useful interpretation in terms of odds ratios (ORs). Guard against the temptation to interpret an OR as relative risk. The two are not the same, but OR is nearly equal to relative risk when the prevalence of disease, *p*, is small, say less than 5%. Other precautions as mentioned under the topic **odds ratio** should also be observed.

The meaning of logistic coefficients in the case of a dichotomous regressor is slightly different from the meaning in the case of a polytomous or continuous regressor.

### *Dichotomous Regressors*

Consider a simple logistic equation with only one regressor, $x_1$. Then, $\hat{\lambda} = b_0 + b_1 x_1$. Since $p/(1-p) = e^{\hat{\lambda}}$, it implies that $p/(1-p) = e^{b_0 + b_1 x_1}$. Suppose $x_1$ is a *binary* regressor with value 1 when present and value 0 when absent. Refer to these as exposure present and exposure absent. Since $p/(1-p)$ is odds for positive response, odds for exposed subjects $(x_1 = 1) = e^{b_0 + b_1}$ and odds for unexposed subjects $(x_1 = 0) = e^{b_0}$. Thus,

$$\text{odds ratio (OR)} = \frac{e^{b_0 + b_1}}{e^{b_0}} = e^{b_1} \text{ and } \ln(\text{OR}) = b_1.$$

It is now clear that $b_1$ is the log of the OR corresponding to the variable $x_1$ when $x_1$ is binary. This is the OR with respect to $x_1 = 0$, which is called the reference category.

In general, when several regressors are present in the logistic model, $e^{b_1}$ is an *independent* contribution of one unit of $x_1$ to the OR when other $x$'s remain the same. Then, this is also called the **adjusted odds ratio**. While reporting results of a logistic regression, many researchers give both adjusted and unadjusted ORs, where unadjusted ORs are based on logistic regression with only one regressor at a time and adjusted when all the regressors are considered together.

Consider a case–control study of adult males to assess the relative importance of various risk factors of benign prostate hyperplasia (BPH). For convenience of illustration, restrict risk factors to only three: $x_1$ for age, categorized as 50–59, 60–69, or 70+; $x_2$ for self-reported sexual activity, categorized as mild, moderate, or heavy; and $x_3$ for diet, categorized as vegetarian or nonvegetarian. These are coded as (0, 1, 2), (0, 1, 2), and (0, 1), respectively. The controls in this study are non-BPH males from the same social milieu, and let there be 30 subjects in each group.

The data are entered into a computer and the logistic regression is obtained. Suppose the estimated value of the intercept (sometimes called constant) is –2.65 and the regression coefficients are +0.50, +0.78, and +0.22 for $x_1$, $x_2$, and $x_3$, respectively. Thus, the following logistic regression is obtained:

$$\hat{\lambda} = -2.65 + 0.50x_1 + 0.78x_2 + 0.22x_3.$$

For a person with the highest risk factors, $x_1 = 2$ (age 70+ years), $x_2 = 2$ (heavy sexual activity), and $x_3 = 1$ (nonvegetarian diet). Thus, for this person,

$$\hat{\lambda} = -2.65 + 0.50 \times 2 + 0.78 \times 2 + 0.22 \times 1 = 0.13.$$

Since $\hat{\lambda} = \ln[p/(1-p)]$, the estimate of the odds for this person with highest rank is $p/(1-p) = e^{0.13} = 1.1388$. For a person with the lowest risk, $x_1 = 0$, $x_2 = 0$, and $x_3 = 0$, $\hat{\lambda}$ is –2.65. This gives odds $e^{-2.65} = 0.0707$. Therefore, the **odds ratio (OR)** is 1.1388/0.0707 = 16. Thus, the odds of BPH for the first person with all the three risk factors are nearly 16 times compared with a person with least risk (of age 50–59 years, with mild sexual activity, and on vegetarian diet). ORs are not probabilities, for example, for the first odds, $p/(1-p) = 1.1388$ gives $p = 0.5325$, and for the second odds, $p/(1-p) = 0.0707$ gives $p = 0.0660$. The first probability is about 8 times the second, whereas the OR was nearly 16 times. Thus, do not interpret odds as probabilities.

How do you interpret the other logistic coefficients such as 0.22 for $x_3$ in this equation? In our example, for a person with $x_1 = 0$, $x_2 = 0$, and $x_3 = 0$, odds $= e^{-2.65} = 0.0707$. If the person is nonvegetarian ($x_3 = 1$) with other $x$'s same, then $x_1 = 0$, $x_2 = 0$, and $x_3 = 1$, and odds $= e^{-2.65+0.22} = 0.0880$. These values give an OR for BPH for a nonvegetarian against a vegetarian of 0.0880/0.0707 = 1.25, which means that a nonvegetarian diet increases odds by 25%. This is the adjusted OR and measures the independent contribution of $x_3$ when $x_1$ and $x_2$ are in the model. Again, since ln(1.25) = 0.22, the logistic coefficient ($b_3 = 0.22$ in this case) is the logarithm of the OR for $x_3$. When logistic coefficients are available, the corresponding OR relative to a reference can be immediately computed. It is this property of the logistic regression that has made this method so popular. The category coded as $x = 0$ serves as reference and $e^b$ is the ratio of odds for category $x = 1$ relative to the reference category. As another example, if $b = 1.3$ for a positive family history in case of breast cancer, OR $= e^{1.3} = 3.67$. This means that the odds of breast cancer in those with a positive family history are 3.67 times the odds in those without a positive family history. (This statement is not for chance: chance has to be worked out separately as illustrated earlier.) Thus, $e^b$ is the predicted change in odds of the dependent for a unit increase in the corresponding predictor. An OR of less than one for any $x$ corresponds to a decrease in odds and an OR of more than one corresponds to an increase. An OR close to one indicates that the change in the regressor does not affect the response, and thus the regressor has no predictive utility.

The preceding explanation is in terms of OR because this is how a logistic regression is widely understood. However, in the case of prospective studies, the interpretation is in terms of **relative risk** and not OR. Also, realize that $\pi$ ranges between (0, 1), OR ranges between (0, ∞), and logit ranges between (–∞, +∞).

### *Polytomous and Continuous Regressors*

There is no restriction in logistic regression for any predictor to be binary and can be polytomous or continuous. Interpretation of logistic coefficients for regressors with three or more categories depends on the coding system adopted. If a particular regressor is polytomous ordinal such as disease severity categorized as none, mild, moderate, serious, or critical, many types of coding can be done. The simplest can be 0, 1, 2, 3, 4 when the number of categories is five. We have used this kind of coding for age-group and sexual activity in our BPH example, which quantifies the regressor values and works like a score. In this case, the corresponding $e^b$ is the factor by which OR multiplies when the category moves one up in terms of this score. If $e^b = 1.15$, this means that the odds for category $x = 1$ (mild disease) are 1.15 times the odds for category with $x = 0$ (no disease), and the odds for category $x = 2$ (moderate disease) are 1.15 times the odds for category $x = 1$ and $1.15 \times 1.15 = 1.32$ times the odds for $x = 0$ (no disease). This will change if the scoring is not 0, 1, 2, 3, 4, but something else.

This interpretation of the logistic coefficient is valid for ordinal $x$, which can be given scores. For real nominal categories, suitable contrasts of interest should be defined. They are generally defined in terms of the difference of one group from one or more of the others. If the objective is to examine 5-year survival with site of malignancy, you can have one particular site, say lung, as the reference category and compare this with oral, prostate, esophagus, and so on, by forming contrasts such as (lung–oral), (lung–prostate), and so on. Each of these contrasts would appear in the predictor set of the logistic model. If "prostate cancer" is the reference category, all other sites will be compared with this cancer with suitable coding. The advantage with these kinds of contrasts is that each contrast

will have its own logistic coefficient, and the coefficient for one contrast can differ from that of another contrast. Thus, differential effects of the categories on the response, if present, will emerge. It is possible to find out if one particular category, say category 2, is a significantly higher contributor to the response than category 3 or category 1. The serial coding 0, 1, 2, 3, 4 mentioned in the preceding paragraph does not have this feature because this works as a score.

If $x_1$ is *continuous*, odds $(x_1 = a+1) = e^{b_0+b_1a+b_1}$, and odds $(x_1 = a) = e^{b_0+b_1a}$. Thus,

$$\text{odds ratio} = \frac{e^{b_0+b_1a+b_1}}{e^{b_0+b_1a}} = e^{b_1}.$$

Again, $b_1$ is the log of the odds ratio. Whether $x_1$ is binary or continuous, $b_1$ is the log of OR when $x_1$ is increased by one.

If a predictor is continuous such as age in years, a logistic coefficient of 0.15 would imply OR = $e^{0.15}$ = 1.16, indicating that each year increase in age increases OR by a factor of 1.16. A 10-year increase in age would increase OR by a factor of $(1.16)^{10}$ = 4.41, and you can say that a 10-year increase in age increases OR by 341%. In this case, the OR that may look only slightly more than 1.0 can translate into an enormous effect.

For some predictors, a unit change can mean enormous change—even impossible in practice. Waist–hip ratio (WHR) is a variable that generally varies from 0.7 to 1.5. It looks preposterous to think of a unit change in WHR such as from 1.2 to 2.2. For this kind of variable, it is prudent to consider 0.1 as the unit while running the regression.

Logistic regression is linear, and interpretation of logistic coefficients for a metric (continuous or not) predictor assumes that the effect of increase from, say, 5 to 10 in the value of $x$ is the same on the outcome as an increase from 70 to 75. This should be tested for its validity as per the procedure laid down under the topic **linearity (test for)** in this volume. This may not be valid for many clinical variables. For example, a blood pressure rise from 120 to 130 mmHg may not have the same effect on outcome as an increase from 170 to 180 mmHg. In such a situation, you may like to include square or any other term that can realistically depict the relationship.

## logistic discriminant functions

If needed, see the topic **discriminant function** to understand what this is and where this can be used. Briefly, this is used to find functions (generally linear) of the predictor variables that can adequately classify the subjects into known groups. The procedure mentioned in that topic is for quantitative predictors and is best suited for the variables that have a Gaussian distribution. Logistic discriminant function is the counterpart used when classification decision is to be based on qualitative or categorical variables, although this can include quantitative predictors as well.

For outcome in two categories, the preferable method is **logistic regression**, where predictors could be quantitative or qualitative. We have a **polytomous** extension of logistic, which answers the same research question for multiple categories, but the logistic discriminant function is simpler, is more practical, and considers all categories together. Also, it does not require as big a sample as polytomous logistic regression does. As in the case of the usual discriminant functions, the logistic discriminant functions help in predicting the category of the new set of values of the predictors, for example, to predict the type of liver disease (cirrhosis, malignancy, or hepatitis) on the basis of signs–symptoms and laboratory investigations.

The only requirement for logistic discriminant functions is that the logarithms of **likelihood ratios** of all pairs of categories are **linear** in the observed values. This helps in using a known algorithm for finding the discriminant functions that maximize the chance of correct classification. The number of discriminant functions is one less than the number of categories under prediction just as in the case of the usual discriminant function. For predicting that a patient has liver cirrhosis, malignancy, or hepatitis on the basis of various clinical features and investigations, only two discriminant functions are needed since there are three categories. Also, as in the case of the usual discriminant functions, you can use a variable selection method such as **stepwise** so that the discriminant functions do not use the variables that do not contribute significantly to the discrimination, which can help in obtaining simpler discriminant functions. For further details, see Albert [1] and Anderson [2].

Rossi et al. [3] used logistic discriminant functions to conclusively diagnose primary aldosteronism or idiopathic hyperaldosteronism in hypertensive patients in Italy. The predictors they used are measurement of Na+ and K+ in serum and 24-h urine, sitting plasma renin activity, and aldosterone at baseline and after 50 mg of captopril. Mwangi et al. [4] used logistic discriminant analysis to assess the diagnostic performance of zinc protoporphyrin, alone and in combination with hemoglobin concentration in those without inflammation, *Plasmodium* infection, or HIV infection in rural Kenyan women.


1. Albert A. *Multivariate Interpretation of Clinical Laboratory Data*. CRC Press, 1987.
2. Anderson JA. Diagnosis by logistic discriminant function: Further practical problems and results. *J Royal Statist Soc Series C* (*Appl Statist*) 1974;23(3):397–404. http://www.jstor.org/discover/10.2307/2347131?sid=21105264169451&uid=4&uid=2
3. Rossi GP, Bernini G, Caliumi C, Desideri G, Fabris B, Ferri C, Ganzaroli C et al. A prospective study of the prevalence of primary aldosteronism in 1,125 hypertensive patients. *J Am Coll Cardiol* 2006 Dec 5;48(11):2293–300. http://www.sciencedirect.com/science/article/pii/S0735109706023321, last accessed February 4, 2015.
4. Mwangi MN, Maskey S, Andang OP, Shinali NK, Roth JM, Trijsburg L, Mwangi AM et al. Diagnostic utility of zinc protoporphyrin to detect iron deficiency in Kenyan pregnant women. *BMC Med* 2014;12(1):229. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4276103/


## logistic models/regression (adequacy of)

This section presumes that you are familiar with the basics of logistic models. If not, review the topic **logistic models (basics of)**.

Before using any model for inferential purpose, it is necessary to be convinced of the statistical adequacy of the model in addition to the biological plausibility, which is assessed separately as a prime consideration. Statistical adequacy means whether the model is able to explain the data adequately. Although several criteria such as generalized $R^2$, likelihood ratio, and Wald statistic are available to check the statistical adequacy of a logistic model, **log-likelihood** seems to have better appeal and wider applicability. The other commonly used criteria are the classification accuracy and the area under the receiver operating characteristic curve. Each of these three is explained below. The **Hosmer–Lemeshow test** can also be used as described separately under that topic. For other criteria, see the book by Hosmer et al. [1]. All these methods assess adequacy for the data on which the model was developed, but for real applicability, the model should be assessed on new set of data.

### *Log-Likelihood Method for Assessing Overall Adequacy of a Logistic Model*

The **likelihood**, $L$, is the probability of obtaining the values observed in the sample when the model is correct. Since a probability is necessarily a small number, that is, less than 1, and your high school math tells you that the logarithm of a number less than one is negative, it is helpful to use $-L$, in fact $-2\ln L$ instead, because the distribution of $-2\ln L$ for large $n$ has been found to follow chi-square under $H_0$ in fairly general conditions. The $H_0$ in the case of logistic is that there is no relationship between the dependent and regressor variables. If $H_0$ is true, then all the regression coefficients $\beta_1, \beta_2, \ldots, \beta_K$ are zero, and

$$\hat{\lambda} = b_0'.$$

This is the reduced model. Compare this with the fitted model to determine if this is adequate or not.

Denote likelihood for the reduced model by $L_0$ and for the fitted model by $L_1$. The value of $-2\ln L_0$ for the reduced model would invariably be more than $-2\ln L_1$ for the fitted model. The difference between these two also follows a chi-square distribution with $K$ degrees of freedom, where $K$ is the number of parameters in the fitted model. This deviance is called the model chi-square. Most standard statistical software packages give the value of $-2\ln L$ for the model under consideration. For example, if $-2\ln L_0 = 83.18$ for the reduced model and $-2\ln L_1 = 61.01$ for the fitted model (with, say, $K = 3$ regressors), then model chi-square = 83.18 − 61.01 = 22.17. At $K = 3$ df, this is highly significant ($P < 0.001$). This shows that $x_1$, $x_2$, and $x_3$ (together) in this example are useful in predicting the outcome.

The following comments explain some of the implications:

- If the model is a perfect fit, then the likelihood is 1 and $-2\ln L = 0$. Note that since probability $L < 1$, $\ln L$ is always negative and the deviance $-2\ln L$ is always positive.
- One measure of the adequacy of a fitted logistic model is the extent of decrease in the value of $-2\ln L$ relative to the value for the reduced model. This can be calculated as follows:

$$\text{Contribution of the model: } C = \frac{(-2\ln L_0) - (-2\ln L_1)}{-2\ln L_0},$$

  where $L_0$ corresponds to the reduced model and $L_1$ corresponds to the fitted model as before. The role of $-2\ln L$ is similar to that of $\boldsymbol{R^2}$ in the ordinary linear regression but has a negative meaning. In the case of $R^2$, a larger value is better, but in the case of $-2\ln L$, a smaller value is better. Also $-2\ln L$ does not fall between 0 and 1 as $R^2$ does. For the purpose of $C$ in the formula just given, $\ln L$ can be used if your software provides log-likelihood instead of $-2\ln L$.
- All models can be improved by adding more regressor variables. This will invariably decrease the value of $-2\ln L$, but this decrease may or may not be statistically significant. To test this, the decrease in $-2\ln L$ is again referred to chi-square to obtain a $P$-value. A nonsignificant decrease indicates that adding those regressors is not helpful. Similarly one or more regressors can be dropped in search of a more parsimonious model. A nonsignificant increase in $-2\ln L$ justifies dropping because the model still works almost just as well without those predictors. Also, as always in a regression setup, an increase in number of parameters decreases $-2\ln L$ and improves fit in absolute terms, but sometimes the df's increase relatively faster and statistical significance declines.
- Beware of outliers since these can vitiate any model, and logistic is no exception. Parameter estimates generally incorporate all values but are affected disproportionately by such extreme values. In the process, the whole model can shift its position. As a result, the likelihood $L$ and the deviance may be substantially affected. At the time of scrutiny of data, try to locate outliers and take a conscientious decision to include or exclude some or all of them from the analysis. This should not be done after you have seen the final results.
- All statistical tests, including the one described here, are heavily dependent on sample size. A large sample will almost invariably yield statistical significance whether the model is good or not. Thus, use this method for checking overall adequacy only as a preliminary test. If the overall adequacy is not significant on the basis of this test, there is no need to proceed further, but if it is significant, assess its actual utility by one or more of the methods subsequently described in this section.
- The logistic method for small $n$ in case of categorical dependent variables is too complex—interested readers may see Mehta and Patel [2] for an exact logistic regression valid for any $n$, including small $n$.

When the number of possible regressor variables is large, the computer program can be asked to identify and include only the significant variables in the logistic model. This can be done by one of the following three **variable selection** algorithms: (i) forward selection, (ii) backward elimination, and (iii) stepwise algorithms. These are the same as for ordinary quantitative regression, the only difference being that instead of $R^2$, $-2\ln L$ is now used as the criterion for selection. However, these algorithms are not as well defined for logistic regression as they are for ordinary quantitative regression. None of these algorithms may yield the best model, and this may lead to three different models. It is a good idea to examine several possible models and choose on the basis of interpretability, parsimony, and convenience in obtaining the data.

### *Classification Accuracy Method for Assessing Overall Adequacy of a Logistic Model*

Another way of assessing the adequacy of a logistic model is by finding what percentage of subjects is correctly classified. For this, the model is used on the same set of data on which the model was built, and an estimated probability $p$ is obtained for each subject. A subject is predicted to belong to the "control" if $p < 0.5$ and to the "case" if $p > 0.5$. Any other cutoff point can be used if that is more justified; otherwise, 0.5 seems like a reasonable cutoff. The terms *case* and *control* used here are in a very general sense for the dichotomous group—a case could be either a subject with disease or the one with exposure. The probability $p$ is calculated to several decimal places and $p = 0.5$ is unlikely to occur. If it does, the subject is not classified into either group. The actual status of all the subjects is already known; thus, the predicted grouping can be compared with the observed grouping, and a table similar to Table L.5 can be constructed.

Consider 30 cases with benign prostatic hyperplasia (BPH) and 30 non-BPH controls as shown in Table L.5. When the probability of BPH is estimated for these subjects based on the fitted logit model, suppose 23 out of 30 controls have $p < 0.5$ and 22 out of 30 cases have $p > 0.5$. This information can be used to calculate both positive

and negative **predictivity** type of measures; the two are generally considered together. In Table L.5, a total of 45 out of 60 subjects (75%) were correctly classified by the model, which is not high, indicating a room for improvement. This can be done either by including more regressors in the model or by considering an alternative set of variables as regressors that can really predict BPH status. The higher the correct classification, the better the model. Generally, a correct classification between 80% and 89% is considered good and 90% or more is considered excellent.

One difficulty with this procedure is that the middling probability 0.51 is treated the same way as the high probability 0.97. Also, a subject with a probability of 0.49 goes to another group compared to a subject with a probability of 0.51 despite the minimal difference. The classification is strictly with respect to the cutoff point 0.5 and ignores the magnitude of the difference from 0.5. Another difficulty arises when one group is extremely large and the other group is too small. Suppose 600 pregnant women are followed for eclampsia, and only 30 of them develop this complication. The model may be able to correctly classify 560 of the 570 who did not develop eclampsia and only 2 out of 30 who develop eclampsia. The overall correct classification is 562/600 = 93.7%, which may give an impression that the model is good, ignoring that it actually is good for ruling out the disease but not for confirming. It has correctly classified only 2 out of 30 (6.7%) with eclampsia, and these are crucial for the success of the model. Thus, classification accuracy should be correctly interpreted.

### *ROC Method for Assessing the Overall Adequacy of a Logistic Model*

A popular method for checking the adequacy of a logistic model is by drawing a **receiver operating characteristic (ROC)** curve. This is typically used when the logistic model is used to develop a scoring system. This curve is drawn between **sensitivity** and (1 − **specificity**) for various cutoffs by considering positives and negatives in the sample as "gold" and by working out the sensitivity–specificity of the model by comparing the predicted categories with the actually observed ones in the sample. Sensitivity and specificity are obtained for different values of the regressors so that a range of values is available for drawing the ROC curve. The area under this curve provides a measure of the adequacy of the model; if it exceeds, say, 0.80 out of possible 1.0, the model is usually considered adequate.

Wang et al. [3] used multiple logistic regression to develop a scoring system for identifying individuals with impaired fasting glucose tolerance in the southern Chinese population, with age, waist circumference, body mass index, family history of diabetes, and so on as regressors. They evaluated the adequacy of the scoring system by the area under the ROC curve, which was 0.70 in their subjects. This is not particularly high to inspire confidence in the logistic model–based scoring system developed by them.


1. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*, Third Edition. Wiley, 2013.
2. Mehta CR, Patel NR. Exact logistic regression: Theory and examples. *Stat Med* 1995;14:2143–60. http://www.cytel.com/pdfs/Logistic-Regression—MEHTA-PATEL-Exact-Logistic-Regression-Theory-and-Examples-STATISTICS-IN-MEDICINE-1995.pdf, last accessed February 5, 2015.
3. Wang H, Liu T, Qiu Q, Ding P, He YH, Chen WQ. A simple risk score for identifying individuals with impaired fasting glucose in the southern Chinese population. *Int J Environ Res Public Health* 2015 Jan 23;12(2):1237–52. http://www.mdpi.com/1660-4601/12/2/1237/htm, last accessed February 5, 2015.


## logistic models/regression (basics of)

A logistic model is typically suitable in situations where the response or the **dependent variable** is dichotomously observed; that is, the response is **binary** and the objective is to determine how the response is affected by a set of given **regressors** or to assess the net effect of one particular regressor when all others are fixed. The ordinary regression is for quantitative outcomes and cannot be used for binary outcomes. As explained for **models**, the logistic model can be causal, predictive, or explanatory depending on the objective. It can also be used for developing scores such as **APACHE** that classify the subjects into severity groups as per the details given for **scoring system**.

Consider occurrence of prostate cancer in males of age 60 years or more. This is dichotomous (yes or no) and may be investigated to depend on, say, vasectomy status (yes or no) and dietary habits as the primary variables of interest. Possible **confounders** such as smoking and covariates such as age at vasectomy can be the other regressors in the model. In another setup, hypothyroidism versus euthyroidism can be investigated for dependence upon the presence or absence of signs or symptoms such as lethargy, constipation, cold intolerance, or hoarseness of voice, and upon serum levels of thyroid-stimulating hormone, triiodothyronine, and thyroxine. A binary variable is also obtained when a continuous variable is dichotomized, such as diastolic BP <90 and ≥90 mmHg with names such as normotension and hypertension. However, there are extensions of logistic regression for polytomous and ordinal data. For these, refer to **logistic models/regression (multinomial, ordinal, and conditional)**. The present section is for dichotomous outcomes.



In all dichotomous situations, the dependent variable $y$ is given a value 0 for a negative response or 1 for a positive response, and nothing in between. Statistically, the dependent variable is the proportion $p$ of subjects providing a positive response. This proportion is converted to $\text{logit}(p) = \ln[p/(1-p)]$ for fitting a logistic model. This is denoted by $\hat{\lambda}$ as this is the estimate of $\lambda = \text{logit}(\pi) = \ln[\pi/(1-\pi)]$, where $\pi$ is the probability of the positive response in the target population. Justification and other details of this transformation are in the topic **logit**. It has been observed that the results not only lend themselves to easy and useful interpretation in terms of **odds ratio** when the binary dependent variable is transformed to logit but also help retain the sigmoid shape of the probabilities that are bound by 0 and 1, thus doing away with the requirement of the Gaussian pattern of the responses and homoscedasticity that do not hold in this setup. The logistic model takes the form

$$\hat{\lambda} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_K x_K$$

for $K$ regressors. The regressors can be quantitative or qualitative—there is no restriction. This is also called the logistic regression equation. Statistical software easily provides the values of $b_0$, $b_1$, $b_2$, …, $b_K$, which are the estimates of the corresponding parameters $\beta_0$, $\beta_1$, $\beta_2$, …, $\beta_K$, respectively, in the population. These are now called logistic coefficients. See the topic **logistic coefficients (interpretation of)** to understand their practical utility and interpretation in terms of odds ratio. The software obtains the estimates by using a re-weighted **least squares method** that requires iterations.

There are several advantages of logistic regression compared with the ordinary quantitative regression:

1. There is no need to worry about Gaussian pattern, even for testing of hypothesis. It is a nonparametric procedure.
2. **Homoscedasticity** is not a requirement.

3. $\pi/(1-\pi)$ represents the odds for positive response in the subjects, which is a useful statistical quantity. Logit mentioned earlier is actually log of odds.
4. Logistic regression does not explicitly require a *linear* relationship between the dependent $\pi$ and the regressors. However, logistic regression does require a linear relationship between the predictors and the log of odds of the dependent variable, as should be clear from the logistic equation given earlier. This is especially important for quantitative regressors.

However, independence of various values of the dependent variable is still needed. This is fulfilled when the response is assessed for different subjects, and where response of one does not affect the response of another subject. It may be violated in repeated measures such as in placental abruption or placenta previa in pregnancies in the same set of women, which are known to recur with greater frequency in subsequent pregnancies. Exercise due caution in such cases.

Although the real objective is to estimate the probability $\pi$ on the basis of a set of predictors, $\lambda$ is predicted instead. The logistic relationship recognizes that the probability $\pi$ is always between 0 and 1, and this probability is given by

$$\pi = \frac{1}{1+e^{-\lambda}}.$$

In practice, $\pi$ is replaced by its estimate $p$ in logit function, and correspondingly in the formula just mentioned where $\lambda$ is replaced by $\hat{\lambda}$.

A useful application of logistic regression is in **case–control studies**. Such studies have a retrospective design, but for the purpose of the logistic model, case–control studies are considered a binary response of the subjects. The case group may comprise those who have the disease and the control group may comprise those who do not have the disease. When the cases and controls are one-to-one matched, a modification called *conditional logistic* is used.

As in the case of the ordinary quantitative regression, the regressor could be one variable $x$, providing a simple logistic setup, or a set of many predictors providing a multiple logistic setup. This is also called *multivariable logistic regression*, although the term *multiple* looks better. Many such similarities are present between the logistic regression and the usual quantitative regression. For example, interaction can be considered by including the product of the values of the concerned predictors, **multicollinearity** affects the reliability of the estimates in the case of logistic as well, and outliers can throw this out of gear. The logistic model is connected with **generalized linear models** through the logit **link function**. In the literature, multivariable logistic is sometimes described as multivariate logistic, but that is not an adequate description.

One basic difficulty with logistic regression is that it demands a large sample. The tests of significance of the regression coefficients and confidence intervals (see the topic **logistic coefficients (CI and test of $H_0$)**) are based on Gaussian distribution: a condition fulfilled with large sample only. Simulation results show that at least 10 positive and 10 negative responses (called the limiting sample size) are required per regressor. If there are 6 regressors, and if only 5% are positive responses, a sample of 1200 subjects is needed to give 60 positive responses. Our experience suggests that this requirement is even more for few regressors but less when the number of regressors is large.

## logistic models/regression (multinomial, ordinal, and conditional), *see also* logistic models/regression (basics of)

The conventional regression models are for binary outcome as described under the topic **logistic models/regression (basics of)**. However, there are extensions for the setups where the outcome is **polytomous** or **ordinal**.

### *Multinomial Logistic Models*

**Nominal** categories cannot be ordered, and each category must be studied as stand-alone. Logistic regression is straightforward if a category is to be compared with the rest since, in this case, the probability $\pi$ is for the category of interest and $(1-\pi)$ is for the rest. In the case of liver diseases, compare hepatitis with (cirrhosis + malignancy), cirrhosis with (hepatitis + malignancy), and malignancy with (hepatitis + cirrhosis) by separate logistics when these are the only possibilities under consideration. A joint model can be made for studying all three comparisons together that will reduce the number of parameters, but the interpretation of the logistic coefficients becomes complex. Calculations also become more complex although appropriate software can help.

Methods that compare one category with the other without the restriction of two probabilities adding to one are available. Thus, category 1 can be compared with category 3, and category 2 can be compared with category 3 by running two separate logistic regressions. In this case, it is helpful to consider one particular category as reference just as is done for regressors. For the comparisons just cited, since both comparisons are with category 3, this serves as the reference. In our example on liver disease, any one category can be considered the reference. If cirrhosis were the reference, both hepatitis and malignancy would be compared with cirrhosis, and there will be two logistics—one for hepatitis and one for malignancy. For a practical example, see Mohaghegh et al. [1] who used this for breast cancer staging in Iran. Breast cancer stages are the multiple categories of the response in this example.

### *Ordinal Logistic Models*

Several possibilities exist in the case of ordinal dependent categories.

- Compare each category with each of the others. If the dependent variable is severity of illness categorized as none, mild, moderate, and serious, the possible comparisons are mild, moderate, and serious each with none; moderate and serious each with mild; and serious with moderate. This would require six different logistic runs. This is the same as considering them nominal.
- Compare each category with the preceding or succeeding category—mild with none, moderate with mild, and serious with moderate. These can be called *adjacent categories logits*. This is one of the **logit models** used in ordinal categories.
- Compare each with the combination of the others. This means comparing none with (mild + moderate + serious), mild with (none + moderate + serious), and so on. This is the easiest but may not be sensible in some cases.
- Compare each category with the combination of the preceding or succeeding categories. This means comparing mild with none, moderate with (none + mild), and serious with (none + mild + moderate). Similarly, comparison with the succeeding categories can also be done if it is considered more meaningful. This uses what is called

*cumulative logit* and is one of the other logit models used for ordinal categories.
- Compare each with a reference. If "no disease" is considered the reference, compare mild with no disease, moderate with no disease, and serious with no disease category.

Perhaps some other comparisons can also be thought of. In all these cases, it is possible again to prepare a joint model, although this may make the model too complex for interpretation. A joint model is relatively easy for cumulative logits when odds for successive categories are proportional. That is, the coefficient that describes the relationship between one category versus all higher categories of the dependent variable is the same as the one that describes the relationship between the next category versus all its higher categories. All such possibilities should be considered only for a small number of categories of the response variable. If the number of ordinal categories is large, say, more than four, consider running the usual quantitative regression (instead of logistic) after giving suitable scores to the categories.

For ordinal logistic in action, see Sharma et al. [2].

### *Conditional Logistic*

**Before–after studies** always provide matched data but one-to-one matching can also happen in case–control studies where the subjects are matched for background characteristics such as age-group, sex, and body mass index (BMI). You may be aware that matching is done only for those characteristics that are not under investigation for their influence of the outcome: that is, only the unmatched characteristics can be the regressors. Matching obviously produces a high degree of correlation and special methods are needed to analyze data from such studies. The **McNemar test** for matched data cannot be used here because the aim of logistic is to study the form of relationship of regressors with the outcome, and not just for presence or absence of relationship. One-to-one matching gives rise to $n$ strata corresponding to $n$ matched pairs, each comprising only two individuals. This requires generation of $(n - 1)$ **dummy variables**, and each will have a corresponding logistic coefficient. Thus, the number of parameters for estimation becomes too many for $n$ pairs of subjects. Increasing the sample size does not help in this case because the number of dummy variables also correspondingly increases as the number of pairs increases. Each stratum will have a parameter dedicated to the stratum that statisticians call **nuisance parameter** because it does not help but instead hurts the interpretation.

A conditional logistic method eliminates this nuisance parameter, which is the intercept $b_0$ in this case, and thus simplifies the estimation method and protects against potential bias that arises in the regular unconditional logistic model when used for matched data. Only the parameters corresponding to the predictors are estimated, and "no intercept" is specified so that a logistic model is fitted with no $b_0$. In conditional logistic, matching variables are not included in the model, and the entire pair is discarded when information on one part of the pair is missing.

Many statistical software packages provide option for running the conditional logistic, and some allow even more than one matched control per case. You must ensure that a suitable package and the correct option in the package are used for analyzing matched data.

1. Mohaghegh P, Yavari P, Akbari ME, Abadi A, Ahmadi F. The correlation between the family levels of socioeconomic status and stage at diagnosis of breast cancer. *Iran J Cancer Prev* 2014 Fall;7(4):232–8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4307106/
2. Sharma A, Bynum SA, Schneider MF, Cox C, Tien PC, Hershow RC, Gustafson D, Plankey MW. Changes in body mass index following HAART initiation among HIV-infected women in the Women's Interagency HIV Study. *J AIDS Clin Res* 2014;5. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4285631/

## logistic regression (requirements for)

This section presumes familiarity with the basics of **logistic models**. These models are obtained by running the regression of a function of probability of occurrence of an event on a set of regressors. Thus, the **dependent** variable is **qualitative**, mostly **binary**. The regressors can be binary, polytomous or quantitative—there is no restriction.

Logistic regression is among the most commonly used statistical methods in health and medicine. Because of the easy and useful interpretation of **logistic coefficients** and the extensive availability of the software packages, the method has been indiscriminately used forgetting that the results of logistic regression are valid only under some regularity conditions. Logistic regression does not require Gaussianity or **homogeneity of variances** for the variables, but it does require the following.

The first is regarding the selection of the regressors both in terms of number and in terms of relevance. You should have a clear idea of what might be influencing the chance of occurrence of the event of interest (the dependent variable). For this, use your experience, review the literature, and consult the experts. If the potential number of regressors is large, examine how to reduce them, since the large number of regressors compromises the parsimony of the model and the computation time and computer memory requirement also become enormous. Remember that the estimation of the logistic coefficients is computation-intensive. If good biological reasons are not available to reduce the number of regressors, use a statistical method such as **stepwise** to select the significant variables. Also ensure that the regressors included in a logistic regression do not have much **collinearity** among them because that can produce very unreliable estimates.

The second requirement is for sample size and is intimately related to the number of regressors. The requirement is in terms of the number of rarest responses, called the *limiting sample size*. The norm for logistic regression is that the limiting sample size should be at least 10 times the number of regressors. If you are investigating the relationship of the chance of eclampsia in pregnant women, which is supposedly seen in 10% of women, on 6 regressors, you need a minimum sample size of 600 women so that at least 60 women with eclampsia are available in your data set. Our experience suggests that you need higher than 10 times when the number of regressors is small and lower than 10 times when the number of regressors is large, but a substantial sample size is needed in any case.

The third requirement is with regard to the coding of the qualitative regressors. Whereas dichotomous regressors are invariably coded as 0 for absent and 1 for present, care is required for **polytomous** regressors. These could be nominal or ordinal. For ordinal categories such as disease severity into none, mild, moderate, serious, and critical, you can think of using codes 0, 1, 2, 3, and 4 as quantitative variables but with precaution that such coding actually is scoring and implies that a moderate disease with code 2 is twice of a mild disease with code 1, and so on. If these do not reflect the actual situation, any other appropriate scores can be used. For nominal categories such as blood group O, A, B, and AB, such codes cannot be used and the representation should be in terms of indicator variables: (0, 0, 0), say, for blood group O, (1, 0, 0) for blood group A, (0, 1, 0) for blood group B, and (0, 0, 1) for blood group AB. This

coding is valid when blood group O is the reference category since this is coded as (0, 0, 0). The regression coefficients for all others will give log of odds with this category as the reference.

The fourth requirement is linearity of relationship of log of odds with the quantitative regressor ($x$) if that is included in your model. Note that this linearity is not between the regressor and the chance of outcome but with the logarithm of odds. To check this, plot the log of odds versus the values of $x$ and see if it is close to a line, although this can be done only when many observations for each of several values of $x$ are available. If the plot is not nearly linear, this can give an idea of whether to include a square term, a square-root term, or any other term. The second method is to divide the range of quantitative $x$ into three or four appropriate intervals, fit a logistic to each interval separately, and check whether or not nearly the same logistic coefficient is obtained for each interval of $x$. This is possible when sufficient data are available in each interval of $x$ values. If the relationship is not linear, it is advisable to divide range of $x$ into as many plausible intervals as required (generally three or four) and use these as categories in the logistic regression.

Last is the most important requirement of independence of errors. In effect, this means that the subjects of research should be such that one does not influence the values in any other. That is, for example, they should not belong to the same family where the values tend to be similar.

For further details of requirements of logistic regression, see Kumar et al. [1].

1. Kumar R, Indrayan A, Chhabra P. Reporting quality of multivariable logistic regression in selected Indian medical journals. *J Postgrad Med* 2012;58(2):123–6. http://www.jpgmonline.com/article.asp?issn=0022-3859;year=2012;volume=58;issue=2;spage=123;epage=126;aulast=Kumar

## logit, *see also* logistic models/regression (basics of)

For a probability $\pi$ of occurrence of an event, the logit in statistics is defined as

$$\text{Logit: } \lambda = \ln \frac{\pi}{1-\pi}.$$

Note that $\pi/(1-\pi)$ is the **odds** of the occurrence of the event, and logit is the logarithm of odds. This is an acronym for *logistic integral transformation*.

Why do we need this kind of weird transformation? The need arises from the fact that any probability necessarily lies between 0 and 1—and this is highly restrictive for statistical modeling. Probabilities cannot go in a straight line forever. If the probability increases with value of a variable, such as probability of death of a critical patient within a week increasing with increasing severity of the condition, the relationship generally takes the form of an S-shaped (called **sigmoidal**) curve (Figure L.9a). The probability of death is almost zero in patients with no disease, increases slightly when the disease is mild, but shows a sharp increase as the severity increases. After the probability reaches something like 90% for critical disease, increase in chance of death is slow even when the disease is more severe than critical. This is because the probability cannot exceed 1 and has to flatten. This shape is natural because of restriction of (0,1) on $\pi$. It should be clear from this figure that very small values of $x$ make little contribution to the probability and so do the very large values, whereas an increase in value of $x$ toward the middle steeply increases the probability in a pretty straightforward fashion. This is typical for most medical events.

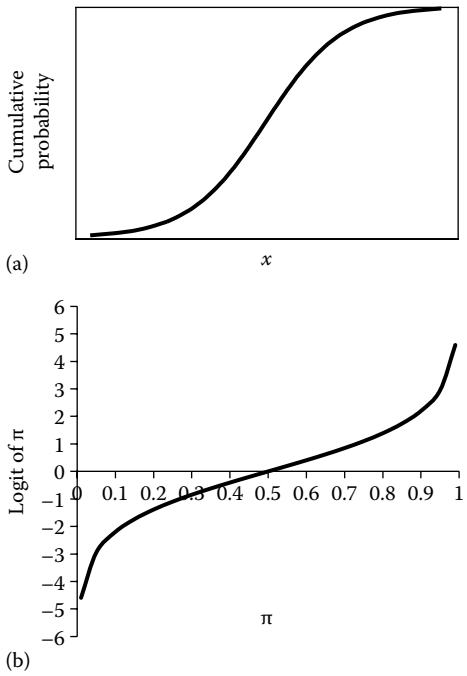



**FIGURE L.9** (a) S-shape of the relationship of cumulative probability with the value of the variable positively affecting the probability. (b) Plot of logit($\pi$) for $\pi = 0.01$ to 0.99.

The logit transformation removes the (0, 1) restriction since $\lambda$ can now be between $-\infty$ for $\pi = 0$ to $+\infty$ for $\pi = 1$ (Figure L.9b). It is nice to note that $\lambda = 0$ for $\pi = ½$, logit is negative for $\pi < ½$, and this is positive for $\pi > ½$. Thus, $\lambda$ is statistically more suitable for studying the relationship with the regressors. The most important use of this transformation is in **logistic regression**, where logit of probability of occurrence of an event such as death is expressed as a linear combination of the regressors. This helps in delineating the role of each regressor in determining or explaining the chance of occurrence of the event. Luckily, in this case, logit helps in a very useful interpretation of the **logistic coefficients** in terms of logarithm of **odds ratio**. The odds ratio is estimated as $\text{OR} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$, and $\ln(\text{OR}) = \text{logit}(p_1) - \text{logit}(p_2)$. Thus, a ratio of ratios is converted to a linear function that can be easily handled.

The inverse of logit is $\pi = 1/(1 + e^{-\lambda})$, and this is called *logistic* of $\pi$. Thus, if the log of odds for any probability is known, the corresponding probability can be obtained.

In some isolated cases where the binary dependent variable has underlying continuum that follows Gaussian (Normal) distribution, **probit transformation** may perform better than logit transformation, especially when the variable values are divided as $<a$ and $\geq a$ categories. For example, diastolic level can be divided as <90 mmHg and ≥90 mmHg, and if the diastolic level in the subjects has a Gaussian distribution, probit may be more appropriate.

## log-likelihood, *see* likelihood, log-likelihood, and maximum likelihood

## log-linear models

Log-linear models use the linearizing property of logarithm on the expected frequencies (see **cell frequency**), which otherwise

are multiplicative, in a contingency table, and help in finding how cell counts are affected by the levels of factors. A unique feature of these models is that there is no dependent or independent variable. All characteristics are treated statistically the same way and no distinction is made between an antecedent and outcome. This is the basic distinction between log-linear models and logistic models. These models are best suited for data obtained from **cross-sectional studies** divided by three or more qualitative characteristics and are applicable to **categorical data** only. When male cancer cases are divided by site of cancer and age-group, the interest may be to know which site is more common in which age-group, without antecedent–outcome implications. Biologically, one characteristic can be dependent on the other but log-linear models accord them the same status. In the usual log-linear setup, the variables are treated on a nominal scale—thus, order in age-groups in our example is ignored. The subjects must be independent and randomly selected as in most statistical setups. Just as for almost any categorical data, the test of hypothesis in the case of log-linear models is also based on **chi-square**, and this requires a large sample. The historical account given by Fienberg and Rinaldo [1] attributes the development of log-linear models to a series of efforts of Goodman and others.

As explained next, the logarithm of expected frequencies in a contingency table can be expressed in terms of additive factors relating to the variables under study, hence the term *log-linear*. The objective is to find whether the categories of different variables, individually or jointly, are especially contributing to determining the cell frequency.

## Log-Linear Model for Two-Way Tables

Log-linear models are easily understood with the help of two-way tables, although their application is more useful for higher-order tables. Suppose the number of rows in this table is $R$ and the number of columns is $C$. The expected frequency under $H_0$ of independence in case of two-way tables is

$$E_{rc} = (O_{r\bullet}O_{\bullet c})/n; \quad r = 1, 2, \ldots, R; \quad c = 1, 2, \ldots, C,$$

where $O_{r\bullet}$ And $O_{\bullet c}$ are the totals in the $r$th row and $c$th column, respectively, in the contingency table and $n$ is the grand total. This gives

$$\ln(E_{rc}) = -\ln(n) + \ln(O_{r\bullet}) + \ln(O_{\bullet c})$$

$$= \mu + \alpha_r + \beta_c,$$

where $\mu = -\ln(n)$, $\alpha_r = \ln(O_{r\bullet})$, and $\beta_c = \ln(O_{\bullet c})$. These are interpreted as the general mean and the main effects of rows and columns, respectively, just as we do for ANOVA. They can be redefined to satisfy conditions such as $\Sigma_r\alpha_r = 0$ and $\Sigma_c\beta_c = 0$ that can help make the quantities $\alpha_r$ and $\beta_c$ more interpretable as deviation from the average. The model in the last equation describes the logarithm of expected cell frequency as a linear combination of an overall effect, the effect of the $r$th category of the first variable, and the effect of the $c$th category of the second variable. It is this feature of log-linear models that has made these so popular.

The actually observed frequency $O_{rc}$ will not be the same as expected from this equation but will be something else. Let the difference be denoted by $\theta_{rc}$. Thus, in general,

$$\text{Log-linear model (two-way): } \ln(O_{rc}) = \mu + \alpha_r + \beta_c + \theta_{rc},$$

provided the observed cell frequency is not zero. Log-linear is essentially a large sample procedure where no $O_{rc}$ is likely to be zero.

The component $\theta_{rc}$ measures **interaction** and $\theta_{rc} = 0$ for all $(r, c)$ implies that the observed frequencies conform to the hypothesis of independence. The higher the value of $\theta_{rc}$, the stronger the interaction or dependence. The test of the hypothesis of no interaction ($\theta_{rc} = 0$ for all $r$, $c$) is the usual chi-square, but it is customary in the case of log-linear models to use another criterion called $G^2$ that maximizes the likelihood. For a two-way table, this is defined as

$$G^2 = 2\Sigma_{rc}O_{rc}\ln(O_{rc}/E_{rc}).$$

This also follows a chi-square distribution with $(R - 1)(C - 1)$ df when expected frequencies in at least four-fifths cells of the contingency table are 5 or more and $n$ is large. In other words, $\chi^2$ and $G^2$ are equivalent for large $n$. $G^2$ is called the likelihood ratio chi-square test.

## Log-Linear Model for Three-Way Tables

The preceding discussion for two-way tables was just to explain the concept of log-linear models, but these models are more effective for three-way or higher-way tables. An explanation similar to that for a two-way table can also be given for a three-way table. In this case,

$$\text{Log-linear model (three-way): } \ln(O_{rcl}) = \mu + \alpha_r + \beta_c + \gamma_l + \theta_{rc} + \omega_{rl} + \delta_{cl} + \phi_{rcl},$$
$$r = 1, 2, \ldots, R; \; c = 1, 2, \ldots, C; \; l = 1, 2, \ldots, L,$$

where $\mu$ is the general mean, $\alpha_r$, $\beta_c$, and $\gamma_l$ are the main effects, $\theta_{rc}$, $\omega_{rl}$, and $\delta_{cl}$ are the two-factor interactions, and $\phi_{rcl}$ is the three-factor interaction. Thus, it contains all possible interactions for a three-way table, which will always be an exact fit, called a *saturated model*. The adequacy of fit under the null, or its lack, is tested by $G^2$ in the equation above. The primary interest in log-linear models is to test significance of interaction terms because those indicate dependence. In the site of cancer and age-group example, the third variable could be type of diet (veg/nonveg), and the interest among others could be to find if diet and site of cancer are related. In terms of notation, this transforms to $H_0$: $\omega_{rl} = 0$. If this null is not rejected, this interaction term can be deleted from the model without significantly affecting the utility of the model. The less the number of interaction terms, the more is the parsimony of the model.

The procedure is to calculate the expected cell frequencies under $H_0$ and obtain $G^2$ by a three-way analog of the formula given earlier, and use the chi-square distribution to check whether or not $P < 0.05$. If it is, reject $H_0$ and conclude that $\omega_{rl} \neq 0$, or that age-groups and site of cancer are related in our example. The category actually dominating is identified by further analysis. Interested readers may consult Knoke and Burke [2] for details of this kind of analysis. The model can be easily extended to four or more variables although the interpretation becomes increasingly complex as the number of variables increases. However, log-linear may still turn out to be the method of choice for multidimensional tables because the row effects and column effects can be expressed in terms of marginal totals.

## Issues with Log-Linear Models

Different statistical software packages give different types of output, but everything starts to make sense after some experience. A log-linear model is generally specified by leaving out the interaction or the main effect under test from the model given in the equation just mentioned and the $P$-value corresponding to the value of $G^2$ obtained from the data. A series of models may have to be fitted to come to a focused conclusion. The software may also help in identifying particular cells that are most divergent and cause the significance.

Among many uses of log-linear models, one is to evaluate the net association between two variables after removing the effect of the others. Tiensuwan et al. [3] used log-linear models to conclude that the site of cancer is related to marital status, diagnostic evidence, and treatment in males as well as in females. This example illustrates the use of log-linear models but does not demonstrate the utility. The following remarks may be helpful in understanding the implications of log-linear models:

- Simultaneous consideration of three or more factors raises the question of mutual independence, partial independence, and conditional independence. For details, see Le [4].
- Log-linear models can be adjusted for structural zeroes in cells of the contingency tables if present. These are not zero frequencies by chance in the sample but present by design. In such a case, fit a model to the subset of cells that remains.
- Just as in the case of classical $\chi^2 = \Sigma[(O-E)^2/E]$, log-linear models disregard the order, if any, in the categories of the variables. Each variable is considered to be on a nominal scale. If an ordinal scale is present and is to be given due consideration, different models are required. Ordinal and metric categories allow investigation of the presence or absence of a trend or a gradient in proportion as well as its nature similar to the one discussed under chi-square for trend. See Agresti [5] for details of these models.
- One way to assess the goodness of fit of log-linear models is to compute the standardized deviate $z = (O-E)/\sqrt{E}$ for each cell, where $O$ is the observed frequency and $E$ is the expected frequency under the model. When the cell frequency is large, this $z$ follows an approximately Gaussian pattern. If almost all values of $z$ are between –2 and +2, the fit can be considered good. When the value of $z$ is large for one or more cells, those cells can be considered to contribute significantly to the association. If there are many such cells, the value of each $|z|$ should be more than 3 as is generally required on **Bonferroni** considerations.


1. Fienberg SE, Rinaldo A. Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. Research Showcase@CMU, 2006. http://repository.cmu.edu/cgi/viewcontent.cgi?article=1122&context=statistics
2. Knoke D, Burke PJ. *Log–Linear Models (Quantitative Application in Social Sciences)*. Sage, 1980.
3. Tiensuwan M, Yimprayoon P, Lenbury Y. Application of log-linear models to cancer patients: A case study of data from the National Cancer Institute. *Southeast Asian J Trop Med Public Health* 2005; 36:1283–91. http://www.ncbi.nlm.nih.gov/pubmed/16438159
4. Le CT. *Applied Categorical Data Analysis and Translational Research*, Second Edition. Wiley, 2009.
5. Agresti A. *Analysis of Ordinal Categorical Data*, Second Edition. Wiley, 2010.


## log-normal distribution

As the name implies, a log-normal distribution is an attempt to Gaussianize a highly **skewed** distribution on the positive side by taking logarithm of the variable under study. As explained for **logarithmic scale/transformation**, with base 10, logarithm of 1 is 0, that 10 is 1, that of 100 is 2, that of 1000 is 3, and so on. Also, logarithm of 0.1 is –1, that of 0.01 is –2, that of 0.001 is –3, and so on. Thus, log transformation drastically attenuates the large values and inflates values less than 1 with a negative sign. Thus, a highly skewed distribution tends to become symmetric. We have explained this with base 10 for simplicity but log-normal distribution uses Naperian base $e$, which is also called *natural log* (written as ln), because of its nice mathematical properties. Also note that log-normal distribution is applicable to only those variables that cannot be negative. Luckily, hardly any medical measurement has negative value. However, while dealing with differences, such as change in cholesterol level before and after a treatment, be careful since some of these changes can be negative. In case negative values are present in the data and log-normal still looks like a plausible distribution, add slightly more than the highest negative value (in absolute value) to all the values in the data set so that none is zero or negative, and readjust the results to account for this transformation.

Figure L.10a shows the distribution of fasting blood glucose (FBG) level in a general population of adults. In this population, most have FBG between 80 and 140 mg/dL but some have a high level and a few have a very high level of FBG. These are the people who are not on any treatment or have a high level despite treatment. This distribution is highly skewed on the positive side and has a sharp peak. When we plot the same values after taking the natural logarithm of FBG values, the distribution tends to become symmetrical and the peak becomes close to what we expect in a Gaussian distribution (Figure L.10b). The right tail remains but is now much shorter and the skewness is mild. Most statistical methods are robust to this kind of mild skewness and they can be safely used for this transformed distribution. Since ln(FBG) has close to Gaussian distribution, the distribution of FBG itself is log-normal.

The most common medical example of the variable with log-normal distribution is the duration—the duration of survival after detection of a terminal disease, the duration of hospital stay after a particular surgery, latent period of infectious diseases, and so on. Almost all durations concentrate at around a specific value, but the duration for some subjects tends to become long, and for some others, it tends to become very long because of complications or otherwise. Because of this feature, the duration usually will have a log-normal distribution. However, there are other applications as well. Karulin et al. [1] found it useful in counting ELISPOTs produced by CD8 and CD4 cells as the size of the spots follows a log-normal distribution ranging from micrometers up to a millimeter

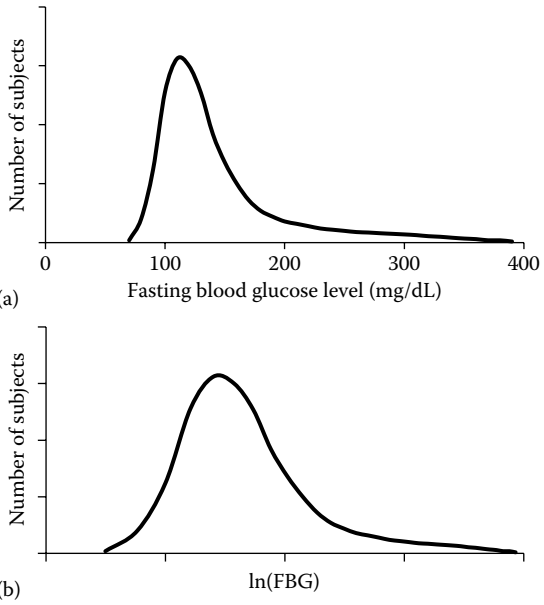



**FIGURE L.10** (a) Distribution of fasting blood glucose (FBG) level in a general population of adults. (b) The same distribution after log transformation of FBG level.

in diameter. Tomitaka et al. [2] observed that depression score in a Japanese population has a log-normal distribution. They used this feature to compare depression scores in various age-groups.

When $y = \ln(x)$ has a log-normal distribution, the mean of $y$ is $\ln(\mu + \frac{1}{2}\sigma^2)$, where $\mu$ is the mean of $x$ and $\sigma^2$ is the variance of $x$. Thus, the antilog of the mean of $y$ does not give you the mean of $x$ but a correction by $\frac{1}{2}\sigma^2$ is needed.


1. Karulin AY, Karacsony K, Zhang W, Targoni OS, Moldovan I, Dittrich M, Sundararaman S, Lehmann PV. ELISPOTs produced by CD8 and CD4 cells follow log normal size distribution permitting objective counting. *Cells* 2015 Jan 20;4(1):56–70. http://www.mdpi.com/2073-4409/4/1/56/htm, last accessed February 11, 2015.
2. Tomitaka S, Kawasaki Y, Furukawa T. Right tail of the distribution of depressive symptoms is stable and follows an exponential curve during middle adulthood. *PLoS One* 2015 Jan 14;10(1):e0114624. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4294635/


## log-rank test (Mantel–Cox test)

First proposed by Nathan Mental in 1966 [1], the log-rank test is designed to compare the overall pattern of two **survival curves** without reference to any particular point of time. This is a large sample test and uses chi-square criterion to arrive at significance or nonsignificance of the difference. Being a nonparametric test, this is not affected by the shape of the survival curves or distribution of survival times. It requires that any **censoring** in both the groups is independent of the survival process. If patients getting two treatments are under comparison, the patients getting the worse or the better treatment should not drop out quicker than others. Also, the log-rank method is valid when all time points are equally important. If the time points with a higher number of subjects at risk are given more weight, another method such as the **Breslow test** should be used. As a middle path between log-rank and Breslow, some prefer the **Tarone–Ware test** that gives weight proportional to $\sqrt{n_t}$, where $n_t$ is the number at risk at time point $t$.


Nathan Mantel

The null hypothesis for the log-rank test is that the survival curves are identical in the two populations under comparison. This implies that the probability of survival (or of death) at *each* point of time is the same in one group as in the other. Under this null hypothesis, the expected number of deaths is calculated for each time using the combined experience in the two groups.

$$E_{1t} = \frac{n_{1t}}{n_{1t} + n_{2t}}(d_{1t} + d_{2t}) \quad \text{and} \quad E_{2t} = \frac{n_{2t}}{n_{1t} + n_{2t}}(d_{1t} + d_{2t}); t = 1, 2, \ldots, T,$$

where $n_{1t}$ is the number of subjects at risk at the $t$th time point in the first group, $n_{2t}$ is the number of subjects at risk at the $t$th time point in the second group, and $d_{1t}$ and $d_{2t}$ are the observed number of deaths at the $t$th time point in the first and the second group, respectively.

When any survival time is censored, that individual is considered to be at risk of dying in the time of censoring but is ignored for subsequent time points as done for obtaining the **Kaplan–Meier** curve. The sums $\Sigma E_{1t}$ and $\Sigma E_{2t}$ give the expected frequencies for use in chi-square as illustrated in the following example.

In a study of early posttransplant phase of renal allografts, peak antibody level was recorded to assess if it affects survival in the first 2 years [2]. Out of a total of 216 patients, the peak antibody level did not reach 15% in 137 patients (group I). In the other 79 patients (group II), the antibody level exceeded 15%.

Table L.6 contains calculations for the two groups. Note how subjects with censored observations are excluded from the calculations. For the purpose of completeness, time = 16 months is shown, but it does not contribute to the calculation as all the three subjects are censored at this time. A relevant software package will automatically do this for you. Under the null, the expected frequency (number of deaths) for group I is 12.39 and that for group II is 6.61. The observed frequencies for groups I and II are 3 and 16, respectively. Subtract the expected frequencies from the observed frequencies for the two groups and get

$$(O_1 - E_1) = -9.39 \text{ and } (O_2 - E_2) = +9.39$$

$$\chi^2 = \frac{(-9.39)^2}{12.39} + \frac{(+9.39)^2}{6.61}$$

$$= 20.46.$$

This has one df and a software gives $P < 0.001$ for this value of $\chi^2$, leading to the conclusion that the difference in the survival pattern in the two groups is highly significant. The two survival curves are shown in Figure L.11 and the difference is easily seen.

The test can be easily extended to more than two groups. The validity conditions of the log-rank test are the same as for the Kaplan–Meier curve as follows:

- Censored values have the same pattern of survival as the uncensored values; that is, they have the same prognosis.
- In case the subjects are enrolled sequentially over a period, those enrolled later have the same survival pattern as those enrolled early. This means that no innovation has occurred during the period of enrollment that may have a case-mix.
- The time of occurrence of events is exactly recorded and not in intervals.

The log-rank test is most likely to succeed in detecting a difference between survival patterns in two groups when the chance of survival in one group is consistently higher or consistently lower at each time point than the other. If the survival curves cross one another, as can happen while comparing medical treatment with surgical operation, then the difference in the overall pattern may be masked. Also, this test only reveals whether the curves are different, without any implication regarding the magnitude of difference. One method for assessing the magnitude of difference is to compute the ratio of the **hazards** in the two groups at the time point of interest.

For further details of the log-rank test, see Hosmer et al. [3].


1. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemo Rep* 1966;50(3):163–70. http://garfield.library.upenn.edu/classics1983/A1983QB30100002.pdf, last accessed February 11, 2015.
2. Jeyaseelan L, Walter SD, Shankar V, John GT. Survival analysis: An introduction. *Natl Med J India* 1999; 12:172–7. http://nmji.in/archives/Volume-12/issue-4/clinical-research-methods.pdf
3. Hosmer DW, Lemeshow S, May S. *Applied Survival Analysis: Regression Modeling of Time to Event Data*, Second Edition. Wiley Interscience, 2008.

**TABLE L.6**
**Calculations for the Log-Rank Test**

| | | Number of Deaths | | | Number Censored | | Number at Risk | | | Expected Frequency | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Time (Months)** | **Time Point[a]** | **Group I** | **Group II** | **Total** | **Group I** | **Group II** | **Group I** | **Group II** | **Total** | **Group I** | **Group II** |
| ***t*** | ***i*** | $d_{1i}$ | $d_{2i}$ | $d_{1i} + d_{2i}$ | $c_{1i}$ | $c_{2i}$ | $n_{1i}$ | $n_{2i}$ | $n_{1i} + n_{2i}$ | $E_{1i}$ | $E_{2i}$ |
| 1 | 1 | 0 | 1 | **1** | 0 | 0 | 137 | 79 | **216** | (137/216) × 1 = 0.634 | (79/216) × 1 = 0.366 |
| 2 | 2 | 0 | 1 | **1** | 0 | 0 | 137 | 78 | **215** | (137/215) × 1 = 0.637 | (78/215) × 1 = 0.363 |
| 3 | 3 | 1 | 0 | **1** | 0 | 0 | 137 | 77 | **214** | (137/214) × 1 = 0.640 | (77/214) × 1 = 0.360 |
| 4 | 4 | 0 | 1 | **1** | 0 | 0 | 136 | 77 | **213** | (136/213) × 1 = 0.638 | (77/213) × 1 = 0.362 |
| 5 | 5 | 0 | 3 | **3** | 0 | 0 | 136 | 76 | **212** | (136/212) × 3 = 1.925 | (76/212) × 3 = 1.075 |
| 8 | 6 | 0 | 1 | **1** | 0 | 0 | 136 | 73 | **209** | (136/209) × 1 = 0.651 | (73/209) × 1 = 0.349 |
| 9 | 7 | 0 | 1 | **1** | 0 | 0 | 136 | 72 | **208** | (136/208) × 1 = 0.654 | (72/208) × 1 = 0.346 |
| 10 | 8 | 0 | 1 | **1** | 0 | 0 | 136 | 71 | **207** | (136/207) × 1 = 0.651 | (71/207) × 1 = 0.343 |
| 12 | 9 | 2 | 0 | **2** | 2 | 0 | 136 | 70 | **206** | (136/206) × 2 = 1.320 | (70/206) × 2 = 0.680 |
| 13 | 10 | 0 | 1 | **1** | 4 | 2 | 132 | 70 | **202** | (132/202) × 1 = 0.653 | (70/202) × 1 = 0.347 |
| 14 | 11 | 0 | 1 | **1** | 0 | 2 | 128 | 67 | **195** | (128/195) × 1 = 0.656 | (67/195) × 1 = 0.344 |
| 16 | 12 | 0 | 0 | **0** | 3 | 0 | 128 | 64 | **192** | – | – |
| 17 | 13 | 0 | 2 | **2** | 2 | 0 | 125 | 64 | **189** | (125/189) × 2 = 1.323 | (64/189) × 2 = 0.677 |
| 18 | 14 | 0 | 1 | **1** | 0 | 0 | 123 | 62 | **185** | (123/185) × 1 = 0.665 | (62/185) × 1 = 0.335 |
| 19 | 15 | 0 | 0 | **0** | 1 | 0 | 123 | 61 | **184** | – | – |
| 20 | 16 | 0 | 1 | **1** | 2 | 2 | 122 | 61 | **183** | (122/183) × 1 = 0.667 | (61/183) × 1 = 0.333 |
| 21 | 17 | 0 | 0 | **0** | 2 | 0 | 120 | 58 | **178** | – | – |
| 24 | 18 | 0 | 1 | **1** | 0 | 0 | 118 | 58 | **176** | (118/176) × 1 = 0.670 | (58/176) × 1 = 0.330 |
| **Total** | | **3** | **16** | **19** | **16** | **6** | | | | **12.390** | **6.610** |

*Source:* Adapted from Jeyaseelan L, Walter SD, Shankar V, John GT. *Natl Med J India* 1999; 12:172–7. http://nmji.in/archives/Volume-12/issue-4/clinical-research-methods.pdf.

[a] At death or censoring in any of the two groups.

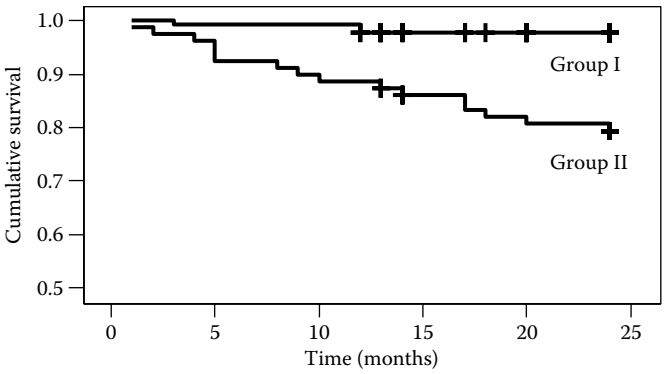



**FIGURE L.11** Survival curves for patients of renal allograft with peak antibody level <15% (group I) and ≥15% (group II) (+ sign for censored values).

## longitudinal data (analysis of)

Longitudinal data arise from **longitudinal studies** where a characteristic is measured on a number of occasions for each subject. The measurements can be qualitative or quantitative, but for simplicity, most of what we state in this section applies to quantitative data for one variable. Among many issues that make analysis of longitudinal data complicated, the most important is correlation among the serial values. This violates the basic requirement of independence that we need for using ordinary regression, ANOVA, and most other statistical methods. Ignoring this correlation provides invalid results of estimation and hypothesis testing.

When the data are obtained for the same fixed points of time for each subject, the analysis can be done in many situations by using **repeated measures ANOVA** as explained for that topic. This requires that if you are measuring the creatinine level of kidney transplant patients, this should be measured at the same time points, say, 1 week, 1 month, 3 months, and 6 months for each patient. The time intervals may or may not be equal; that does not matter, but the time points should remain the same. However, this method is only able to provide statistical significance or nonsignificance of differences in means across groups and not about trend. Also, there must not be many missing values—a common malady of longitudinal studies—for repeated measures ANOVA to be valid.

The longitudinal data are generally obtained from studies that measure different subjects at different points in time. For example, child 1 may have measurements at 0, 7, 10, 18, and 23 months and child 2 may have measurements at 0, 3, 8, 14, and 19 months depending on when they visit the well-baby clinic. The objective in this case is generally to find the time trend such as growth curves for children and bioavailability of drugs in pharmacokinetic studies. The observations are not necessarily forward in time and can also be from past records.

If the objective of longitudinal measurements is in finding the time point that reaches a predefined end point, such as the time taken by body functions to reach the normal physiological levels, **survival analysis** may be the right procedure to use. Some subjects will reach the end point at 15 days, some at 2 months, some at 4½ months, and so on, and these data can be modeled as survival function. Survival pattern in two or more groups can be compared by the **log-rank test** or its extension. If the interest is in finding the effect of various correlates on the duration (of reaching to the predefined end point), you can use the **Cox model**. This can incorporate both time-invariant (e.g., sex and blood group) and time-dependent (e.g., blood pressure, liver functions, etc., that can vary at different points in time under observation) covariates.

If the objective is to look at the mean values at different time points for the entire follow-up, consider whether **area under the concentration curve** is an appropriate measure for the comparison of pattern in different groups. This will work well when the values in one group are generally lower than the other. If they crisscross, this area can be misleading. Concentration curves do not require that the same time points are observed for each subject. If the objective is to compare peak (or lowest) value obtained, the method used in **pharmacokinetic studies** can be used. Both these will be based on average at different time points, and these averages can be problematic in some situations as anywhere else.

If the series is long, something like observations at 50 time points on each subject, **time series** methods can be used for analysis of data. Typically, though, in the present context, longitudinal data pertain to four to six repeated measurements where time series methods will not apply.

**Mixed effects models** that include random effects are particularly suitable for analyzing correlated continuous outcomes in longitudinal studies. These model group effect as fixed and individual effects as random. Several observations for the same individual at different time points or otherwise will constitute within-subject values. Mixed models can be used even when the observation time points vary from subject to subject. This might be the method of choice if many repeated measures are missing, which makes uneven observation points for different subjects, provided the missing values are random and do not create a biased set of data. Mixed model is a fairly general method but confidence intervals and test of hypotheses generally require that the error term follows a Gaussian distribution. Other distributions can also be studied through **link functions** available in **generalized linear models**. This includes categorical responses and count data. However, the computations are complex in this case and setting up the analysis properly in statistical software requires considerable expertise. This should be done only by competent statisticians.

There are several other approaches such as those based on serial differences or differences from base value, but none is simple. For details, see Diggle and Heagerty [1].

1. Diggle P, Heagerty P. *Analysis of Longitudinal Data*, Second Edition. Oxford, 2013.

## longitudinal studies, *see* prospective studies

## lot quality assurance scheme

Hospital officials look for assurance that the incoming material such as drugs, chemicals, and blood products is of good quality. When purchased from a reputable company, these may have already been subjected to a rigorous quality check, yet it is sometimes desirable to also check their quality at the time of receiving them. This can be done by using a lot quality assurance scheme (LQAS) when bulk purchases are made. Blood testing strips and skin patches are the other examples where this can be used. This scheme is also applied to the assessment of health services and to disease surveillance. Details of LQAS have been provided by Montgomery [1]. A brief is as follows.

### *LQAS in a Laboratory Setup*

The total material to be inspected for quality is first divided into lots of nearly equal size, and a small sample of units from each lot is checked for quality. If the number of defective units exceeds a predetermined tolerance threshold, the whole lot is rejected; otherwise, it is accepted.

Suppose a hospital received 80 boxes, each box containing 10 laboratory kits. These can be divided into five lots of 16 boxes each. Each lot will now have 160 kits. Take a random sample of, say, $n$ = 15 kits from each lot, and check these for a predetermined quality standard. All kits cannot be inspected as that can render them unusable, or it may be too expensive to do 100% checking. If the number of kits not meeting the standard exceeds the tolerance level, reject the whole lot. The tolerance level for large $n$ (based on Gaussian approximation) is determined by an expression similar to the following equation depending on the acceptable rate of error.

$$\text{upper limit of tolerance} = n\pi + 1.645*\text{SE},$$

where $\text{SE} = \sqrt{n\pi(1-\pi)}$ and $\pi$ is the proportion defective. If the proportion defective is expected to be at most $\pi = 0.2$, that is, 2%, the upper limit of tolerance for $n$ = 15 is $15 \times 0.02 + 1.645 \times \sqrt{(15 \times 0.02 \times 0.98)} = 1.19$. Thus, if the number of defectives is more than 1 in a sample of 15, you can be confident that the percentage of defectives is more than 2% and the lot can be rejected. We have illustrated this using Gaussian approximation, but for small $n$ and small $p$ such as this example, an exact threshold [1] based on the **binomial distribution** should be used.

### *LQAS in Health Assessment*

The lot quality method has been used in disease surveillance, nutrition programs, assessing women's health services, and, most of all, immunization coverage assessment [2].

Suppose coverage of at least 90% by polio vaccine is required to build herd immunity so that the transmission is throttled and the disease is not able to propagate itself. Thus, not more than 10% of children should remain uncovered ($\pi = 0.10$). Let the target population be children below the age of 3 years in a district. The area of the district is divided into, say, 20 zones that will serve as "lots" in LQAS. Within each zone, suppose a random sample of $n$ = 80 children is examined for noncoverage by polio vaccine. The tolerance threshold in this case is

$$80 \times 0.10 + 1.645\sqrt{80 \times 0.10 \times 0.90} = 12.4.$$

If the number of nonimmunized children is 13 or more in any sample of size 80, reject that lot. That lot—zone in this case—can be considered to have almost surely not reached the level of 90% coverage required for herd immunity. This is because the probability of 13 or more nonimmunized out of 80 is less than 0.05 when the coverage is 90%. When the problem is stated this way, the conclusion is that 90% (or more) coverage is unlikely. For more assurance, the problem can be stated in a reverse manner with threshold $y = n\pi - 1.645*\sqrt{n\pi(1-\pi)}$ for the maximum nonimmunized. In this example, this is 8.0 − 4.4 = 3.6; that is, if there are 3 or less unimmunized in a lot of 80, you are confident that the level of immunization is at least 90%. Note the reverse nature of this conclusion.

The following comments regarding LQAS may be helpful:

- The division of the total material into lots should preferably be such that each lot is homogeneous while the lots themselves are different from one another. In the example above, the zones may be such that one consists mostly of a slum or underprivileged population and another consists of a population of high socioeconomic status. LQAS can then better identify the zones with low coverage.
- Sometimes, lots with a high proportion of defectives can also be wrongly accepted under this scheme (for that matter, in any statistical decision). This is the same as **Type II error** under the testing of hypothesis procedure. Methods are available to devise an LQAS such that the probability Type II error is under control. For details, see Robertson et al. [2].

1. Montgomery DC. *Introduction to Statistical Quality Control*, Seventh Edition. Wiley, 2012.
2. Robertson SE, Anker M, Roisin AJ, Macklai N, Engsirom K, LaForce FM. The lot quality technique: A global review of applications in the assessment of health services and disease surveillance. *World Health Stat Q* 1997; 50:199–209. https://www.globalhivmeinfo.org/Gamet/Gamet%20Library/1336_LQAS%20-%20global%20review%20of%20applications.pdf, last accessed February 12, 2015.

## LOWESS plot

This is an acronym for "LOcally WEighted Smooth Scatter" plot. This method finds smoothed predicted value for each point on the basis of **linear regression** fit to the points in the neighborhood, giving more weight to the closer points and less weight to the farther points. Thus, the predicted value of dependent $y$ for any $x_i$ is based on linear regression that gives higher weight to $x_{i-1}$ and $x_{i+1}$ and less weight to the distant values. Higher weight to the immediate neighbor and lower weight to the farther values make it a locally weighted method. This differential weighting requires the weighted least square method in place of the regular **least square method**. The procedure is repeated to obtain the smoothed predicted values for each value of the regressor $x$, which means that a separate weighted regression is performed for every point in the data. When quadratic or other regression is used in place of linear regression for each $x$, the method is called *LOESS* (Local regrESSion) instead of LOWESS.

Figure L.12 shows our version of a LOWESS plot between increasing 25-hydroxyvitamin D levels and the cumulative frequency of community-acquired pneumonia [1]. The straight lines (linear regression) at different points can be distinctly seen. This

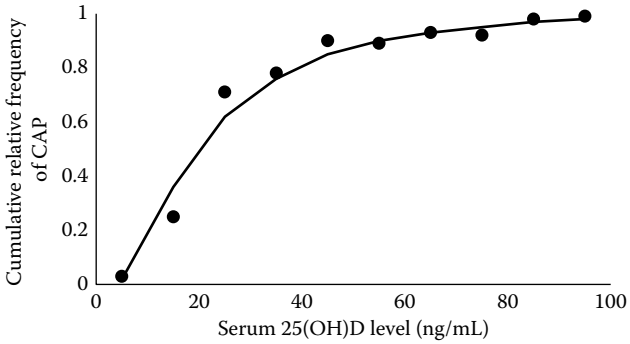



**FIGURE L.12** Locally weighted scatterplot smoothing (LOWESS) plot between increasing 25-hydroxyvitamin D levels and the cumulative frequency of community-acquired pneumonia. (Adapted from Quraishi SA, Bittner EA, Christopher KB, Camargo CA Jr. *PLoS One* 2013 Nov 15;8(11):e81120. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3829945/.)

lacks smoothness that we generally expect in a LOWESS plot. Smoothness is achieved when the number of $x$ values is large.

Consider the regression of bone mineral density (BMD) on age in apparently healthy females of age 40 to 80 years. The objective may be to establish normal values for your population. Under the LOWESS method, the predicted value of BMD for age, say, 52 years will be based on linear regression of BMD on age between, say, 47 and 57 years, giving more weight to the BMD values at age 51 and 53 years (which are close to 52 years) but less weight to the BMD values at age 47 and 57 years. Predicted BMD at 53 years will be based on linear regression of BMD values at age 48 and 58 years by giving more weight to the values at age 52 and 54 years and less weight to the ages far from 53 years. Note how the weight shifts from one value of $x$ (age in this example) and also note that many linear regressions are obtained. This is what makes this method computation intensive. Such a method can be conceived now with computer to our assistance but was unthinkable a few decades ago.

Important features of this method are as follows. (i) It requires that the span of values to be considered for each fit is specified. In our BMD example, this is 10 years—from 47 to 57 years for predicting the BMD at age 52 years, from 48 to 58 years for predicting BMD at age 53 years, and so on. Each smoothed value is determined by neighboring data points defined within this span. Another term for this span is bandwidth. (ii) Because of local weights, the values farther away do not influence the regression as much, and the possibility of some unusual values at the extremes distorting the regression is minimized. (iii) The method would give valid results only when a big series of data points is available with diverse values of the regressor $x$. In our BMD example, if the age is restricted to the narrow range of 40 to 49 years in place of 40 to 80 years, the LOWESS method will not succeed. (iv) The method is easily affected by outliers. If BMD values at the hip for age 65 years are mostly 0.7 to 0.9 g/cm$^2$ and one value is 1.2 g/cm$^2$, the regression could be severely affected. Smoothing is likely to take care of any aberration that might occur somewhere in the middle, but for end point outliers, a robust weight function is available [2]. (v) The method does not produce a regression function that can be represented by a mathematical equation; instead, the plot itself is utilized as the output in most cases.

LOWESS has overriding advantages in some situations, not just in correctly fitting a complex trend that can go up and down, but because it does not require the specification of the equation to fit a trend. That, however, is also a disadvantage since an equation is very helpful in understanding the features of the trend.

1. Quraishi SA, Bittner EA, Christopher KB, Camargo CA Jr. Vitamin D status and community-acquired pneumonia: Results from the Third National Health and Nutrition Examination Survey. *PLoS One* 2013 Nov 15;8(11):e81120. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3829945/
2. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc* 1979;74(368):829–36. http://www.jstor.org/discover/10.2307/2286407?sid=21105329903581&uid=2&uid=4

# M

## Mahalanobis distance

Mahalanobis distance is the usual **Euclidean distance** between two **multivariate** observations but adjusted for variances and covariances. This does not answer the question whether the two observations are sufficiently different or not; it answers how much is the difference. This distance was developed by Prasanta Mahalanobis in 1930 [1] and further explained in 1936 [2]. A simple explanation has been provided by McLachlan [3].

Prasanta Mahalanobis

In a simple situation, the Euclidean distance between two multivariate observations of $K$ component is measured by $\sqrt{\Sigma_k (x_k - y_k)^2}$, where $k = 1, 2, \ldots, K$; and

$$\text{Mahalanobis distance: } d = \sqrt{\sum_k \left( \frac{x_k - y_k}{s_k} \right)^2},$$

where $s_k$ is the standard deviation (SD) of the $k$th component, provided that the components are independent. In a multivariate setup, this will seldom be the case, and we will come back to this setup.

Consider values of hemoglobin level (Hb) (g/dL), body mass index (BMI) (kg/m$^2$), total cholesterol level (TCL) (mg/dL), and plasma creatinine level (PCL) (mg/dL). Let these be (12.4, 23.7, 189, 0.6) for one person and (13.6, 21.8, 211, 0.8) for another person. How much different are these values? The Euclidean distance between these two is $\sqrt{(12.4-13.6)^2 + (23.7-21.8)^2 + (189-211)^2 + (0.6-0.8)^2} = \sqrt{489.09} = 22.12$. If SD of Hb is 1.7 mg/dL, of BMI 2.4 kg/m$^2$, of TCL 15.2 mg/dL, and of PCL 0.15 mg/dL, the Mahalanobis distance between these two persons (assuming independence) is

$$\sqrt{\frac{(12.4-13.6)^2}{(1.7)^2} + \frac{(23.7-21.8)^2}{(2.4)^2} + \frac{(189-211)^2}{(15.2)^2} + \frac{(0.6-0.8)^2}{(0.15)^2}} = 2.23.$$

Large values in absolute sense such as TCL in our example are divided by the corresponding SD that will also be large, and small values such as PCL in our example are divided by the corresponding small SD. Dividing by the respective SDs tends to equalize the weight of the variables irrespective of the unit of measurement and the extent of variability.

The calculation of Mahalanobis distance just mentioned is valid when the Hb level, BMI, TCL, and PCL are independent of one another. Since this would seldom be the case in a multivariate setup, we need to consider the **covariances** among these variables. Whereas exact formula for Mahalanobis distance will be in terms of matrix notations that we wish to avoid, suffice it to say that it will involve inverse of the **covariance matrix** of the variables. Inverse of the covariance matrix corresponds to the division by the respective SDs in the independent variables setup.

Mahalanobis distance can be used to answer the questions of the type whether average metabolic characteristics of male diabetics are more different from normal than of female diabetics. It is mainly used in classification problems where there are several groups and the concern is with affinity among groups [3]. This distance is also used in **discriminant analysis**.


1. Mahalanobis PC. On tests and measures of group divergence. *J Asiatic Soc Bengal* 1930;26:541–88.
2. Mahalanobis PC. On the generalised distance in statistics. *Proc Natl Inst Sciences India* 1936;2(1):49–55. http://www.new.dli.ernet.in/rawdataupload/upload/insa/INSA_1/20006193_49.pdf
3. McLachlan GJ. Mahalanobis distance. *Resonance*, June 1999:20–6. http://www.ias.ac.in/resonance/Volumes/04/06/0020-0026.pdf, last accessed February 15, 2015.




## main effects and interaction effects in ANOVA

We explain these two in separate subsections.

### *Main Effects*

Main effect is the average effect of a level of **factor** in **analysis of variance (ANOVA)**. This is generally measured as the difference from the overall mean but can also be measured in other manners such as the difference from the control or from the group with minimum level. Suppose the mean increase in hemoglobin (Hb) level in anemic after regimen A is 2.3 g/dL, after regimen B is 4.1 g/dL, and after regimen C is 1.7 g/dL. The average of these three is 2.7 g/dL. Thus, the main effect of regimen A can be considered to be (2.3 – 2.7 =) –0.4 g/dL, of regimen B (4.1 – 2.7 =) +1.4 g/dL, and of regimen C (1.7 – 2.7 =) –1.0 g/dL. Under this formulation of the main effects, their sum is zero and that is a convenience in interpretation and calculations. However, if there is a control group and the mean increase in this group is –0.1 g/dL, the main effect can be measured from this value as 2.4 g/dL for regimen A, 4.2 g/dl for regimen B, and 1.8 g/dL for regimen C.

Similar explanations can be given for the main effects of the levels of two or more factors in an ANOVA setup. Consider the following example of two factors that will also help us in explaining interactions later in this section. A study was done by Wang et al. [1] on the effect of maternal smoking on birth weight in the United States. For our purpose, vary this study and assume that some women, who are otherwise habitual smokers, give up smoking off and on during pregnancy but could not give up altogether. Nonsmokers are excluded in this example. Suppose that such women are asked at the time of their last antenatal visit just before birth about the duration of smoking (factor 1) and the amount of smoking (factor 2). At the end of the

pregnancy, the former is categorized as <18, 18–31, and ≥32 weeks. These are the three levels of factor 1. The amount of smoking is categorized as mild (1–9 cigarettes per day), moderate (10–19 cigarettes per day), and heavy (20+ cigarettes per day). The days in this calculation are only those when at least some smoking was done, and the amount of smoking is average per smoking day. These are the three levels of factor 2. The outcome of interest (response variable) is birth weight of the children born to these women. Let the mean birth weight in different groups be as given in Table M.1.

There are, for instance, 15 women who smoked an average of 1–9 cigarettes per day for a total of less than 18 weeks during the entire pregnancy. Average birth weight of their babies was 3.45 kg. There are 8 women who smoked on average of 1–9 cigarettes per day for a total of 18–31 weeks, and the average birth weight of their babies was 3.38 kg, and so on. Because of unequal $n$, the design is unbalanced. Although in our example, factors 1 and 2 are both ordinal, this feature is overlooked in the present analysis since ANOVA considers all categories nominal.

Main effects and interactions are easy to explain with the help of notations. Let the levels of factor 1 be identified by subscript $j$ ($j$ = 1, 2, …, $J$), of factor 2 by subscript $k$ ($k$ = 1, 2, …, $K$), and subjects within each group by subscript $i$ ($i$ = 1, 2, …, $n$). In our example, $J = 3$ and $K = 3$, but $n$ in different groups is not equal. For the purpose of explaining the concepts, it is useful to keep the notation relatively simple and assume that all groups have the same number $n$ of subjects. The response of the $i$th subject belonging to the $j$th level of factor 1 and the $k$th level of factor 2 is denoted by $y_{ijk}$. If birth weight of the child born to the fourth woman in the group comprising those smoking mildly (first group of amount of smoking) for 18–31 weeks (second group of duration of smoking) is 3.49 kg, then $y_{421} = 3.49$ kg. With these notations, the mean for the $j$th level of factor 1 is $\bar{y}_{.j.}$, and that for the $k$th level of factor 2 is $\bar{y}_{..k}$; the overall mean of all $nJK$ observations is $\bar{y}_{...}$. When measured from the overall mean,

estimated main effect of the $j$th level of factor 1:

$$\hat{\alpha}_j = (\bar{y}_{.j.} - \bar{y}_{...}); \quad j = 1, 2, \ldots, J;$$

and estimated main effect of the $k$th level of factor 2:

$$\hat{\beta}_k = (\bar{y}_{..k} - \bar{y}_{...}); \quad k = 1, 2, \ldots, K.$$

The main effect of smoking for less than 18 weeks in our example is 3.44 – 3.38 = +0.06 kg, and that of moderate smoking is 3.40 – 3.38 = +0.02 kg. The positive effect is not surprising since these are the effects of those categories on the birth weight *compared with the overall mean* of all categories that include the heavy and long-duration smokers. It is easy to show by some algebra that under this definition of main effects, $\Sigma_j \alpha_j = 0$ and $\Sigma_j \beta_j = 0$. Consequently, only $(J - 1)$ α's and $(K - 1)$ β's are independently determined. One each is automatically determined by these conditions.

### Interaction Effect

Variation in the effect of one level of a factor with the levels of the other factor is called **interaction**. The amount of this variation is measured by interaction effect. This is obtained separately for each combination of the levels of the two factors. This is the excess mean after adjustment for the main effects of the concerned level of factors 1 and 2. Thus, the estimated interaction effect between the $j$th level of factor l and the $k$th level of factor 2 is

$$\hat{\theta}_{jk} = (\bar{y}_{.jk} - \bar{y}_{...}) - (\bar{y}_{.j.} - \bar{y}_{...}) - (\bar{y}_{..k} - \bar{y}_{...}) = (\bar{y}_{.jk} - \bar{y}_{.j.} - \bar{y}_{..k} + \bar{y}_{...}),$$

where $\bar{y}_{.jk}$ is the mean of $n$ subjects in the $(j, k)$th group. For example, the estimate of the interaction effect between moderate smoking and smoking for 32+ weeks in Table M.1 is 3.30 – 3.25 – 3.40 + 3.38 = +0.03 kg. Thus, this combination of duration and amount of smoking increases birth weight by 30 g on average in this sample *relative to the means in respective categories*. Note, again, that the overall mean is based on all women including those who are heavy smokers and long-duration smokers. Mathematically, relative to the respective means implies $\Sigma_j \theta_{jk} = 0$ and $\Sigma_k \theta_{jk} = 0$. As always, lack of statistical significance of interaction does not mean interaction is absent. It only means that this could not be detected from the available data. This might be more so in case of ANOVA since the sample size is generally planned to detect main effects, whereas a higher sample size is required for detection of interaction.

The calculations are done on the original values of birth weight in 75 children in our example. Statistical significance of the main and interaction effects is tested by ***F*-test**. Suppose a software package reveals $P > 0.05$ for $F$ when calculated for amount of smoking (factor 2) and $P < 0.01$ for $F$ when calculated for duration of smoking (factor 1). The first is not statistically significant, but the second is. The conclusion then is that this sample of women does not provide sufficient evidence to conclude that the amount of smoking makes a difference in birth weight, but the duration of smoking in the pregnancy does make a difference. Let $P < 0.05$ for $F$ for interaction. This indicates that an interaction between amount of smoking and duration of smoking is present and implies that the effect of duration of smoking on birth weight is not uniform in the three categories of amount of smoking. This can also be easily seen in Figure M.1, which is drawn from

**TABLE M.1**
**Average Birth Weight (kg) of Children Born to Women with Different Amounts and Duration of Smoking**

| Duration of Smoking in Pregnancy | Amount of Smoking | | | |
|---|---|---|---|---|
| | Mild | Moderate | Heavy | All[a] |
| –18 weeks | 3.45 ($n$ = 15) | 3.42 ($n$ = 12) | 3.43 ($n$ = 7) | 3.44 ($n$ = 34) |
| 18–31 weeks | 3.38 ($n$ = 8) | 3.40 ($n$ = 10) | 3.39 ($n$ = 6) | 3.39 ($n$ = 24) |
| 32+ weeks | 3.35 ($n$ = 25) | 3.30 ($n$ = 23) | 3.18 ($n$ = 29) | 3.27 ($n$ = 77) |
| All | 3.39 ($n$ = 48) | 3.35 ($n$ = 45) | 3.25 ($n$ = 42) | 3.33 ($n$ = 135) |

*Note:* Entries are average birth weight in kilograms.

[a] Weighted average because of varying $n$.

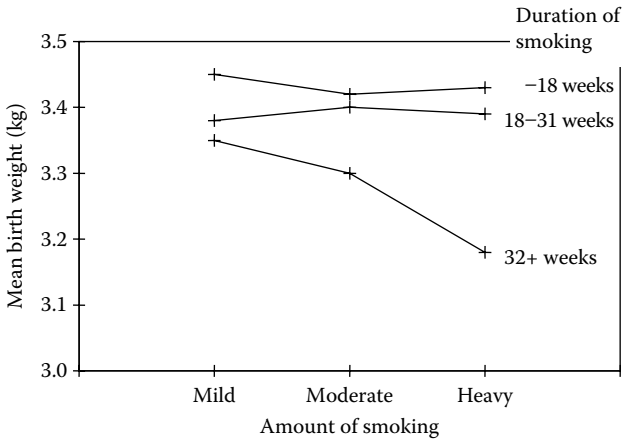



**FIGURE M.1** Interaction between duration and amount of smoking.

the data in Table M.1. If the lines for the three duration groups are parallel, an interaction is said to be absent. In this example, the line is flatter for –18 and 18–31 weeks but shows an unusual dip for smoking heavily for 32+ weeks. This indicates that an interaction is present. Smoking heavily for 32+ weeks steeply reduced the mean birth weight compared with, for example, heavy smoking for 18–31 weeks or moderate smoking for 32+ weeks. This example also illustrates that if the interaction is significant, you should condition your conclusion about one factor's effect on the level of the other factor. When the interaction is not significant, the focus is on main effects. In that case, factor 1 levels should not be compared within factor 2 levels because factor 1 effects are not significantly different. Only the average is adequate.

The equations we have mentioned define the main effects and the interaction effects relative to the overall mean. This is the most commonly used definition, but main effects and interactions can also be defined relative to any other base value. In our example, the base can be changed to the category with least smoking where the mean birth weight is 3.45 kg. In that case, for example, the estimated main effect of heavy smoking would be 3.32 – 3.45 = –0.13 kg. That is, heavy smoking reduces birth weight by 0.13 kg on average compared with the mild smokers for less than 18 weeks. A similar interpretation can be given for the other main effects and interactions when measured from a base value.

1. Wang X, Tager IB, van Vunakis H, Speizer FE, Hanrahan JP. Maternal smoking during pregnancy, urine cotinine concentrations, and birth outcomes: A prospective cohort study. *Int J Epidemiol* 1997;26:978–88. http://ije.oxfordjournals.org/content/26/5/978.full.pdf

## MANOVA, *see* multivariate analysis of variance (MANOVA)

## Mann–Whitney–Wilcoxon test, *see* Wilcoxon rank-sum test

## Mantel–Cox test, *see* log-rank (Mantel–Cox) test

## Mantel–Haenszel procedure

This is a procedure for combined inference from the stratified **categorical data** in observational studies that have done stratification of the subjects to control confounding or for some other reason. Strata could be different subgroups but can also be different times, different places, etc. Mantel–Haenszel (M-H) procedure is most commonly used for testing the hypothesis of association in stratified data and for obtaining pooled estimate of **odds ratio** (OR) and **relative risk** (RR) in this setup. A prominent application of this procedure is in **meta-analysis** where results from different studies are combined to get firmer evidence. This is also referred to as *Cochran–Mantel–Haenszel (CMH) procedure*.

Consider recurrence of eclampsia in second or subsequent pregnancies when it once occurred at primigravida stage. It is widely suspected that occurrence at the time of first pregnancy is related to a positive family history of eclampsia, and so is the recurrence. But the risk of recurrence could differentially increase depending on the sex of the family members with hypertension, if any. This could be a **confounding factor**. To examine if it really is, the association between recurrence of eclampsia and family history should be examined separately in the groups with (i) no family history of hypertension, (ii) only female members with history of hypertension, (iii) only male members with history of hypertension, and (iv) both male and female members with history of hypertension. Thus, the data on recurrence of eclampsia are stratified into four groups based on family history of hypertension. We can answer the following three questions:

(i) Is there any association at all in any stratum?
(ii) If there is, is their degree the same in all the strata (if the degree of association in each of these groups is the same, it can be concluded that family history of hypertension has no role)?
(iii) If the degree is the same, what is the pooled estimate of the odds ratio or relative risk?

Stratified analyses are considered fundamental to causal thinking. When confounders exist, overall results are not valid and stratum-specific results should be reported, whereas unadjusted associations are acceptable in the absence of confounding. For adjustment, methods such as the M-H chi-square could be used to get a combined chi-square that would find if there is any association. To get the pooled estimate, the association in different strata should be homogenous, which can be tested, for example, by the **Breslow–Day test** for odds ratio. We explain the M-H procedure for 2 × 2 tables stratified into $K$ strata. Thus, this is limited to subjects cross-classified by two binary characteristics observed for $K$ groups. This incidentally is also the most common and most simple application of the M-H procedure. For further details of the M-H procedure, see Agresti [1].

### *Mantel–Haenszel (M–H) Chi-Square*

Add a subscript $k$ to the observed frequencies in a 2 × 2 table for the $k$th strata ($k = 1, 2, \ldots, K$) so that these are now denoted by $a_k$, $b_k$, $c_k$, and $d_k$, respectively, in place of $a$, $b$, $c$, and $d$. The null hypothesis is that there is no association between the two binary characteristics in any stratum. This can also be stated as OR = 1 for all $k$. The alternative is that OR ≠ 1 for at least one $k$. For cross-sectional studies, the corresponding pooled chi-square is given by

M-H chi-square:

$$\chi^2_{MH} = \frac{\left[\Sigma\left\{(a_k d_k - b_k c_k)/n_k\right\}\right]^2}{\Sigma\left[(a_k + b_k)(c_k + d_k)(a_k + c_k)(b_k + d_k)/\left\{(n_k - 1)n_k^2\right\}\right]},$$

where $a_k$, $b_k$, $c_k$, and $d_k$ are the cell frequencies in the $k$th 2 × 2 table, and $n_k$ is the total number of subjects in the $k$th stratum, that is,

$n_k = (a_k + b_k + c_k + d_k)$. This chi-square has 1 df. It can be shown with some algebra that the numerator of this expression is the square of the (observed – expected) frequencies, and the denominator is the estimated variance of these squared differences.

As for almost any chi-square, M-H also requires a large sample size for all strata together. For small samples, an exact procedure is also available in some statistical software packages. This also requires that all observations are independent, which effectively means that the different subjects do not belong to any specific cluster (such as belonging to one family). Another important prerequisite is that the two binary characteristics under study should have the same chance of being positive in each subject. The association between the two binary characteristics under study should be either positive in all strata or negative in all strata. If it is positive in some and negative in others, the M-H chi-square may not be able to detect it. Statistical significance of M-H chi-square is evidence that the association is consistently in one direction in almost all strata. The one difficulty with the M-H procedure is that only one confounder can be conveniently studied at one time, and the procedure fails if there are many confounders. Also the confounder must be categorical. If it is continuous, categorize this first into meaningful groups and then use the M-H test.

### *Pooled Relative Risk*

In case the M-H test finds significant association, the next step is to use the Breslow–Day test for checking their homogeneity across strata. If yes, a pooled estimate can be obtained. For **prospective studies**, the adjusted M-H pooled RR is given by

$$\text{pooled RR: } \mathrm{RR}_{\mathrm{MH}} = \frac{\sum_k \left[ a_k (b_k + d_k)/n_k \right]}{\sum_k \left[ b_k (a_k + c_k)/n_k \right]}.$$

In case any cell frequency is zero, 0.5 is added to all cell frequencies for computing pooled RR. The sum is over the values of $k$.

The calculations for $\mathrm{RR}_{\mathrm{MH}}$ are illustrated next with the help of an example that also shows how it can help in practical applications.

Consider two strata (males and females) and frequencies of antecedent and outcome characteristics as in Table M.2. Using the usual formulas for **relative risk**, the results for this table are as follows:

$$\mathrm{RR}_1 \text{ for males} = \frac{3/100}{13/450} = 1.04$$

from the first two columns of Table M.2

$$\mathrm{RR}_2 \text{ for females} = \frac{97/400}{12/50} = 1.01$$

from the middle two columns of Table M.2.

**TABLE M.2**
**Cell Frequencies in Two Strata and Combined**

| | Males | | Females | | Combined | |
|---|---|---|---|---|---|---|
| | Antecedent (Exposure) | | Antecedent (Exposure) | | Antecedent (Exposure) | |
| **Outcome** | **Present** | **Absent** | **Present** | **Absent** | **Present** | **Absent** |
| Present | 3 | 13 | 97 | 12 | 100 | 25 |
| Absent | 97 | 437 | 303 | 38 | 400 | 475 |
| **Total** | **100** | **450** | **400** | **50** | **500** | **500** |

These two are not much different.

$$\mathrm{RR}_c \text{ for combined} = \frac{100/500}{25/500} = 4.00$$

from the last two columns of Table M.2.

The combined RR is so different from each of the strata in this example mainly because in males, only 100 are exposed out of 550, but in females, 400 are exposed out of 450. This imbalance is a situation where the M-H procedure is very effective. In this example, by the formula given earlier,

$$\mathrm{RR}_{\mathrm{MH}} = \frac{\left(\frac{3\times 450}{550}\right) + \left(\frac{97\times 50}{450}\right)}{\left(\frac{13\times 100}{550}\right) + \left(\frac{12\times 400}{450}\right)} = \frac{13.2323}{13.0303} = 1.02.$$

The M-H procedure restores the right value, which was masked in the combined data. This is adjusted for the confounding effect of sex.

### *Pooled Odds Ratio*

For **retrospective studies**, the M-H adjusted

$$\text{Pooled OR: } \mathrm{OR}_{\mathrm{MH}} = \frac{\sum_k (a_k d_k / n_k)}{\sum_k (b_k c_k / n_k)}$$

For a simple explanation and an example of use of pooled OR, see Le [2]. As in the case of RR, this pooling also is admissible when the homogeneity of ORs in different strata is not refuted by the Breslow–Day test.


1. Agresti A. *Categorical Data Analysis*, Third Edition. Wiley, 2012.
2. Le CT. *Applied Categorical Data Analysis and Translational Research,* Second Edition. Wiley, 2009.


## maps (statistical), *see* choroplethic map, cartogram, spot map, thematic map

## marginal distribution, *see* bivariate distribution, multivariate distribution

## Markov process, *see* stochastic process

## masking, *see* blinding, masking and concealment of allocation

## matched pairs, *see also* matching

Matched pair is a set of two subjects whose characteristics that can influence the outcome but are not of our interest are deliberately matched. One subject of the pair goes to one group under study and the second subject to the other group, such as one gets test regimen and the other gets placebo. This helps in ruling out the effect of such extraneous factors on the results. Identical twins are a good example of matched pair. Statistically, matched observations can also arise naturally in some setups such as in before–after experiments and

crossover designs where the subjects are not matched but the values are pairs as they belong to the same subjects. Matched pairs also arise when the same specimen or its parts are sent to two different laboratories for evaluation, or the same image is examined by two radiologists.

Deliberate matching can be done in any setup but is common in the **clinical trials** that have a small number of subjects in the treatment and the control group. The method generally advocated for equivalence of the two groups in clinical trials is to **randomize** the subjects. This indeed is the gold standard but works well for large samples and is feasible when an adequate number of eligible subjects willing to be randomized are available. Randomization in case of small samples can easily fail to yield two groups with equivalent baseline. Thus, matching is preferred for small samples.

If the number of cases is not as large, examine if matched pairs are available and if the two persons forming a pair can be randomized to receive the test and the control regimen. Thus, randomization and matching can go on simultaneously—they are not mutually exclusive. Matched pair design may be suitable for acute rather than chronic conditions. If the number of eligible subjects is even less, controls may have to come from elsewhere. In this situation, matching becomes even more important. Experiments using matching instead of randomization are called **quasi-experiments**. Evidence from such experiments is not considered as strong as from randomized trials.

It is possible in some situations to simultaneously give a different treatment to known pairs such as two eyes or two limbs of the same persons. Randomization can be done within each pair to determine which one will receive the test regimen and the control regimen. If the trial is on comparison of methods such as pulse oximeter and sphygmomanometer, blood-pressure readings can be taken at the same time in the two arms, and many pairs would be easily available. If the trial is for treatment regimen, it could be extremely difficult to find suitably matched paired organs in some situations, such as of the same severity of glaucoma in the two eyes, or both limbs with same degree of paralysis.

In a **case–control study**, the control group should ideally include exactly similar subjects except for the disease and the antecedent factors under study so that it is parallel in a true sense. Ensuring exact similarity may not be feasible, but it is easy to understand that the cases and controls should be matched with respect to all those factors that do not fall into the set as hypothesized risk factors but can influence the outcome.

Consider a trial on vitamin D supplementation for severe pneumonia in under-five children. It is suspected that the effect of vitamin D could be influenced by the preexisting nutritional status of the children. One strategy to rule out this confounding effect is to choose two children with nearly identical nutritional status, and assign one to the treatment group and the other to the control group. Now move on to select another pair with the same nutritional status and assign to the two groups under study, and so on. The nutritional status of the second pair need not be the same as that of the first pair. Both the groups receive regular treatment, but one group gets vitamin D supplementation and the other gets placebo. This is an example of matched pairs, but the results would still be affected if, for example, well-nourished children are overrepresented in both the groups compared with the target population.

Matching tends to take away the independence of the values that is so commonly needed in the analysis of data, and separate methods are used in case of matched pairs. In case of quantitative data, these methods generally use the difference in the values of the pairs rather than the values themselves. If the recovery time of one subject of a pair is 5 days and that of the other is 7 days, the difference is –2 days. Such differences are obtained for each pair, and the analysis is done of these differences. Realize that these differences belong to different pairs of subjects—and thus are independent. Usual statistical methods can be used on these differences. This is the method followed, for example, in the **Student *t*-test** for paired data.

In the case of qualitative response, the magnitude of difference cannot be obtained; instead concordant and disconcordant pairs are counted. Both subjects of the pair with the same response are called concordant, and the pairs with one subject's response different from the other are called disconcordant. For **dichotomous** outcomes, we have methods such as **McNemar chi-square** for testing the statistical significance of the association, and for polytomous outcomes, methods such as **Cohen kappa** are used to assess the degree of association. See those topics for examples.

## matching, *see also* matched pairs

Choosing subjects of a study that are similar with respect to the extraneous characteristics that can affect the outcome and allocating them to different groups for some intervention, such as test regimen and placebo, is called matching. This is a popular statistical strategy for ruling out the effect of some extraneous factors that can influence the results. Matching helps in minimizing the bias in study results, but the matching criteria should be identified before conducting the study and should be stated in the protocol. Matching after seeing the data can produce biased results. Matched data can also arise naturally in some setups such as in **before–after** studies, **repeated measure** studies, and **agreement** studies.

In addition to selection, matching should also be in ascertainment. Consider whether the cases and controls in a study are likely to respond to the questionnaire in a similar manner, and no bias is likely to creep in due to the differential pattern of responses not related to the factors under study. The controls must be assessed with the same keenness and with the same methodology as for the cases. Similar proforma and procedures should be used as far as possible. If cases are being interviewed in a clinic, the controls should also be interviewed in a clinic. Controls may be healthy subjects, but interviewing them at home or by telephone can alter the response. Cases with disease may be more motivated, but try to extract the same cooperation from the controls as well. Controls should be able to provide an equally correct estimate of the rate of occurrence of antecedents in subjects without the disease.

Ideally, all relevant characteristics that might influence the outcome, except those under study, should be matched. This does not stop at age and sex as is sometimes done. Nonetheless, comprehensive matching for all prognostic factors may not be feasible in all situations, and some constraints on conclusions may become necessary. For example, in a trial of a new oral antidiabetic drug, the subjects in the test and the control group could be matched for age, sex, and perhaps obesity, but it may be difficult to match for genetic factors and stress conditions. These two factors can also influence the outcome. The other important prognostic factors in this setup are severity of the disease and any coexisting disease. All such factors may have to be adjusted at the time of analysis. Note that matching characteristics cannot be studied later on for finding their effect on the outcome—thus, the matching criteria should be carefully chosen.

The other limitation is that matching can be tried only for the known factors. There might be other factors in the **epistemic** domain about which nobody knows yet—an uncertainty that still remains. Note that randomization has the advantage of giving chance to

known as well as unknown factors to be equally distributed; matching does not have this feature.

You can see that matching may mean incurring extra cost due to baseline investigation on a large number of subjects, many of which may be discarded as unmatchable. Generalizability suffers as the control group is somewhat distorted and interactions cannot be properly assessed. Also, matched data require special methods of statistical analysis such as **repeated measures ANOVA** and **Cohen kappa**.

**Matched pairs** are discussed as a separate topic in this volume. The following are the details of other types of matching.

### *Baseline Matching*

Baseline matching applies to **prospective studies** including before–after studies. For valid comparison in these studies, the exposed and unexposed groups must be similar at baseline, particularly with regard to the factors that can influence the outcome and are not of interest. If the objective is to study the effect of recently acquired central obesity on the electrocardiogram changes over time, factors such as age, gender, personality traits, stress conditions, and smoking need to be matched between the study group (with central obesity) and the control group (without central obesity). If complete matching is not possible, as would generally happen in practice, statistical methods are used to do the required adjustment at the time of analysis. Such an adjustment can become incomprehensible if done for a large number of factors. If exposure is by choice such as smoking and taking aspirin, the unexposed group may have to be matched also for factors affecting such exposure.



### *One-to-One and One-to-Many Matching*

The results of **case–control studies** are much more valid if there is one-to-one matching between cases and controls, i.e., each case should have a corresponding one matched control. These are also called **matched pairs** and the process is called matching. The attempt should be to simulate an identical twins situation. The purpose is to be able to conclude that any difference between the cases and controls is attributable to the antecedent under study and to no other factor. Thus, the role of extraneous factors is minimized. In the case of near-perfect one-to-one matching, the statistical analysis should consider two groups as paired and not independent samples.

It may not be possible to find a control of age 62 years for matching with a case of 62 years. In most situations, matching within ±2 years for adults is considered adequate. Such relaxation can be possibly allowed for other factors as well. An alternative is to divide age into age groups and match for age group.

In most cases, one matched control is included in the study for each case, but when controls are easy to find and are less expensive, you can include two or more controls for each case. This increases the sample size and helps to increase the reliability of the conclusion. These controls can also be matched with respect to all the extraneous factors of no interest to the investigation. This gives rise to one-to-many matching. Depending upon how many controls are available, two—sometimes even three or four—controls are included for each case.

### *Frequency Matching or Group Matching*

Matching more than two or three confounders is an uphill task because of limited availability of eligible subjects. Generally, matching stops at age and sex, which are influential factors in almost every medical setup. If matching of many confounders is required, which indeed is desirable, an acceptable but less valid procedure is *group matching* (also called *frequency matching*). Under this scheme, controls are matched with the cases on average or with regard to pattern of presence of the extraneous factors. If obesity is such a factor, and if 35% of cases are obese, then nearly the same percentage of controls should also be obese for group matching. Since finding controls with many matching characteristics can be difficult, some factors may have to be adjusted at the time of analysis even when group matching is done.

### *Overmatching*

We have made out a case for matching for all the factors that can affect the outcome except the ones under study. The last qualifier is the actual operational clause. This is sometimes ignored. The real question is what to do with factors that affect the outcome as well as the risk factors under study—called **confounders**. Matching for these can sometimes result in overmatching, and biased results toward the null are obtained. For example, in a study on the effect of postmenopausal estrogens on uterine cancer, matching on uterine bleeding can present invalid findings since uterine bleeding is associated with uterine cancer also. Any matching on such outcome-related variable can distort the results. Look also at the following example.

Marsh et al. [1] describe a study of relation between radiation exposure and mortality from leukemia in workers at a nuclear reprocessing plant. The matching factors in this study were site, sex, work status, age (within 2 years), and date of entry (within 2 years). Note that risk of leukemia varies with these factors, and they looked like valid factors for matching. Date of entry was considered necessary as the risk of leukemia changes with calendar time. Examination of data since 1950 shows that the radiation dose steeply declined after 1980. Matching for date of entry unwittingly also matched for radiation exposure, which was the antecedent under study. Such overmatching obscured the relationship between radiation dose and risk of leukemia. Likewise, exposure-related variables that can make risk factor distribution similar in cases and controls should not be matched as overmatching can reduce the statistical significance of the results.

1. Marsh JL, Hutton JL, Binks K. Removal of radiation dose response effects: An example of over-matching. *BMJ* 2002;325:327–30. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1123834/

## maternal mortality ratio, *see also* mortality rates

The term *maternal mortality* applies when the cause of death in women is related to or aggravated by pregnancy or its management. For practical reasons, the World Health Organization (WHO) definition includes only maternal deaths that occur either during pregnancy or within 42 days after the termination of pregnancy. The cause may be direct such as hemorrhage, sepsis, eclampsia, and abortion, or indirect such as heart diseases in pregnancy and hepatitis. However, accidental or incidental deaths are excluded. Maternal mortality is measured as follows:

$$\text{Maternal mortality ratio: MMR} = \frac{\text{maternal deaths}}{\text{live births}} * 100{,}000.$$

The denominator in this formula is live births, but multiple births are counted as one. Realize that maternal mortality can also occur at the time of stillbirth or abortion. The data on stillbirths may be available; it is extremely difficult to get the count of abortions even

in the best of conditions. The group at risk comprises women of reproductive age, but this is not used as the denominator for the events in the numerator—thus, MMR is called a ratio and not a rate. If the denominator is replaced by the number of women of reproductive age group, this becomes a rate.

Maternal deaths in developing countries are now receiving more attention. Many of these countries do not have a sufficiently strong system to monitor every maternal death; some deaths escape attention. Thus, the reported rate is low compared with the actual rate. WHO, United Nations Population Fund, the World Bank, and United Nations Children's Emergency Fund made joint efforts to arrive at more realistic estimates of MMR in all countries. They used surrogates such as proportion of births with different types of birth attendants and general fertility rate to arrive at these estimates [1]. Data on these surrogates are more easily available and more accurate. According to these estimates, the global MMR declined by 45% between 1990 and 2013—from 380 deaths to 210 deaths per 100,000 live births. The minimum MMR in 2013 was nearly 1 in Belarus, and the maximum was 1100 in Sierra Leone per 100,000 live births [2].


1. WHO. *Trends in Maternal Mortality: 1990 to 2013: Estimates by WHO, UNICEF, UNFPA, The World Bank and the United Nations Population Division*. World Health Organization, 2014. http://data.unicef.org/files/MMR2013.pdf
2. UNICEF. Maternal mortality has declined steadily since 1990, but not quickly enough to meet the MDG target. http://data.unicef.org/maternal-health/maternal-mortality


## mathematical models, *see* models (statistical)

## Mauchly test for sphericity

The Mauchly test assesses the validity of the sphericity assumption that underlies univariate **repeated measures analysis of variance** (ANOVA). It was developed by John Mauchly in 1940 [1]. We first explain sphericity and then describe the Mauchly test.

### *Sphericity in Repeated Measures*

In repeated measures, sphericity is the equality of variances of the differences in values observed at different points in time and covariances of these differences being zero. As usual, this refers to the population parameters rather than to the sample values. Repeated measures ANOVA uses these differences, and such homogeneity is required for valid results. Note that the concept of sphericity applies only when you have three or more repeated values. If you have just two repetitions, there is only one difference, and the question of sphericity does not arise.

Suppose the brain size of pregnant women [2] is measured by magnetic resonance imaging at 3, 6, and 9 months of pregnancy to find if pregnancy affects the brain size. Let the values obtained be as in Table M.3. The differences and their variances are also shown in the table. Sphericity is about equality of variances at the bottom of the last three columns. In this example, the variance of the difference in brain size between 3 and 9 months, and between 6 and 9 months are nearly the same, but the variance of the difference between 3 and 6 months is large. Thus, the sphericity condition is not likely to be valid for repeated measures ANOVA in these data.

The sphericity can also be studied by the correlation structure of the values obtained at different time points. This is sometimes confused with *compound symmetry* where not just the covariances between values at different time points are the same, but the variances of values at different time points are also the same. This is more than required for sphericity. Sphericity requires same variances of the differences but not the same variances of the values at different time points. In other words, compound symmetry is sufficient for sphericity but not necessary.

**TABLE M.3**
**Sample Variances of Differences in Brain Size of Pregnant Women at Different Periods of Gestation**

| | Brain Size ($cm^3$) | | | Difference | | |
|---|---|---|---|---|---|---|
| Subject No: | t1 = 3 Months | t2 = 6 Months | t3 = 9 Months | t1-t2 | t1-t3 | t2-t3 |
| 1 | 1260 | 1273 | 1232 | –13 | 28 | 41 |
| 2 | 1159 | 1143 | 1123 | 16 | 36 | 20 |
| 3 | 1232 | 1223 | 1221 | 9 | 11 | 2 |
| 4 | 1342 | 1302 | 1295 | 40 | 47 | 7 |
| 5 | 1167 | 1198 | 1163 | –31 | 4 | 35 |
| 6 | 1252 | 1241 | 1232 | 11 | 20 | 9 |
| **Mean** | **1235.33** | **1230.00** | **1211.00** | **5.33** | **24.33** | **19.00** |
| **Variance** | **4554.27** | **3159.20** | **3617.20** | **603.47** | **254.67** | **254.80** |

Let us also mention as a side note that sphericity is the requirement for repeated measures ANOVA when univariate methods are used and not for multivariate methods. A univariate method is advised when the sample size is relatively small (say, $n < K + 10$, where $K$ is the number of repeated measures), unequal from group to group, and sphericity is not violated. There is another twist to the story. When sphericity is violated, the univariate solution is to reduce the degrees of freedom (df) by what is called **Huynh–Feldt correction** or similar other corrections. This correction is a multiplier less than or equal to 1, called ε, and depends on the extent of lack of sphericity in the data—the more the departure, the more the correction. The mean squares and other calculations including the actual value of the $\boldsymbol{F}$-statistic remain the same; only the df changes. This reduction in df increases the critical value and thus adjusts the Type I error to the correct level for lack of sphericity. Since this correction to the df is available, you can directly use this correction without going through the process of checking the validity of sphericity in your data. This makes the Mauchly test redundant but is still commonly used for medical data.

Repeated measures can also be analyzed by **MANOVA** where sphericity is not a strict requirement. But MANOVA requires a large sample and still depends on the covariance structure. MANOVA is recommended when sphericity is seriously violated (say, $\varepsilon < 0.7$), but group sizes should be equal.

### *Mauchly Test*

The computational formula for the Mauchly test is a complex one that we wish to avoid. However, this is the default test of sphericity in several common statistical software programs. You should not have any difficulty in using this test when needed. This follows a chi-square distribution with $\left(\frac{K(K-1)}{2}-1\right)$ df, where $K$ is the number of repeated measures. Note that $K(K-1)/2$ is the number of pairwise differences when the measurement is repeated $K$ times.

The validity condition of the Mauchly test includes multivariate Gaussian (normal) distribution of the data, and the test is not robust to non-Gaussianity. If the data do not have a Gaussian pattern, think

of some transformation before using the Mauchly test. In that case, the repeated measures ANOVA will also be on the transformed values. The other limitation of the Mauchly test is that it fails when the sample size is too small and overdetects sphericity in very large samples. A significant Mauchly test result indicates that the assumption of sphericity is untenable, and the df in repeated measures ANOVA $F$-test needs correction as already suggested.


1. Mauchly JW. Significance test for sphericity of a normal $n$-variate distribution. *Ann Math Stats* 1940;11:204–9. http://projecteuclid.org/download/pdf_1/euclid.aoms/1177731915, last accessed February 20, 2015.
2. Redelmeier D, May S. Caution: Baby on board. *Significance* 2014;11(5):20–5. http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2014.00779.x/full

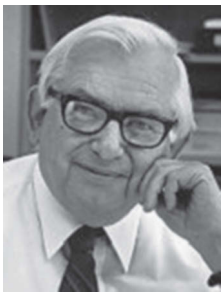

## maximum likelihood method, *see* likelihood, log-likelihood, and maximum likelihood estimates

## McNemar–Bowker test

This test is for internal symmetry in a $C \times C$ (square) **contingency table**, and checks whether the frequency in the $i$th column and the $j$th row is different or not from the frequency in the $j$th column and the $i$th row ($i = 1, 2, \ldots, C$; $j = 1, 2, \ldots, C$) in that population. This kind of problem arises in matched pairs when the assessment is in **polytomous** categories in place of the usual **dichotomous** categories. Thus, this is an extension of the **McNemar test** from 2 to many categories. If the corresponding probabilities in the population are denoted by $\pi_{ij}$, the null hypothesis for the McNemar–Bowker test is $H_0$: $\pi_{ij} = \pi_{ji}$ for all $(i, j)$, and the alternative hypothesis is $H_1$: $\pi_{ij} \neq \pi_{ji}$ for at least one pair of $(\pi_{ij}, \pi_{ji})$. Albert Bowker [1] provided this extension, and the test is also referred to simply as the **Bowker test** in the literature. This is computed as

$$\text{McNemar} - \text{Bowker test: } Q_{\text{MB}} = \sum_{i<j} \frac{(O_{ij} - O_{ji})^2}{O_{ij} + O_{ji}},$$

where $O_{ij}$ is the observed frequency in the $(i,j)$th cell and $O_{ji}$ in the $(j,i)$th cell. This has **chi-square** distribution with $C(C-1)/2$ degrees of freedom (df) for large samples. Thus, the ***P*-value** can be easily obtained.



Albert Bowker

A good example of an application of the McNemar–Bowker test is comparing a health parameter of the right side with the left side of the same person, such as cataract (nuclear, cortical, and subscapular) in the left eye and the right eye, and size (shrunk, normal, and enlarged) of the left kidney and the right kidney. There are many paired organs in our body where this kind of test would be needed when the measurement is qualitative in more than two categories.

**TABLE M.4**
**Follow-Up Clinical and MRI Results for Spinal Infections Patients**

| | Improved Clinically | No Change Clinically | Worse Clinically | Total |
|---|---|---|---|---|
| MRI improved | 33 (33.7%) | 12 (12.2%) | 0 (0.0%) | **45 (45.9%)** |
| MRI unchanged | 16 (16.3%) | 5 (5.1%) | 3 (3.1%) | **24 (24.5%)** |
| MRI worse | 20 (20.4%) | 4 (4.1%) | 5 (5.1%) | **29 (29.6%)** |
| **Total** | **69 (70.4%)** | **21 (21.4%)** | **8 (8.2%)** | **98 (100%)** |

*Source:* Baxi S et al., *Infect Dis Clin Pract (Baltim Md)* 2012 Sep 1;20(5):326–9. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3989101/

It is not necessary that the row and column categories are same. However, the characteristic in the row should have exactly the same number of categories as the characteristic in the column, and both characteristics should be assessed for the same subjects. Baxi et al. [2] compared the follow-up clinical status and magnetic resonance imaging (MRI) of patients with spinal infections in the United States and obtained the data on 98 patients as given in Table M.4. The objective is to evaluate the concordance between clinical improvement and MRI improvement. In this table, $C = 3$ as there are three rows and three columns. Thus, the df for the McNemar–Bowker test is $3 \times 2/2 = 3$. The authors report the McNemar–Bowker test highly significant ($P < 0.001$), indicating that MRI improvement did not correspond with clinical improvement. They concluded that the use of MRI without new clinical indications in routine follow-up testing should be interpreted with caution.


1. Bowker AH. A test for symmetry in contingency tables. *J Amer Stat Assoc* 1948;43:572–4. http://www.jstor.org/discover/10.2307/2280710?sid=21105915621563&uid=4&uid=2
2. Baxi S, Malani PN, Gomez-Hassan D, Cinti SK. Association between follow-up magnetic resonance imaging and clinical status among patients with spinal infections. *Infect Dis Clin Pract (Baltim Md)* 2012 Sep 1;20(5):326–9. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3989101/

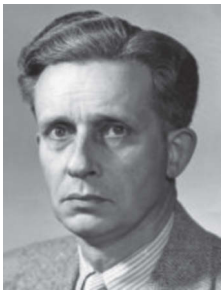

## McNemar test

Introduced by Quinn McNemar in 1947 [1], this test is used to assess statistical significance of association between two attributes when a pair of observations is taken on the same group of subjects. The usual procedure described under the topic **chi-square test for 2 × 2 tables** is valid only when the two groups of subjects are independent. Independence is lost when there is one-to-one matching or pairing, a frequently adopted mechanism in medical studies. Pairing also occurs when the same group of subjects is observed before and after therapy. A matched pair is considered one unit, and the contingency table contains the number of units or pairs with both elements positive, the first element positive and the second negative, the first negative and the second positive, and both negative. These can be denoted respectively by $a$, $b$, $c$, and $d$ as in Table M.5. This table is arranged for a prospective study with exposure–outcome format. There are $n$ matched pairs in all in this table, and the cell frequencies are the numbers of the pairs in different categories. For example, $b$ is the number of pairs in which the exposed partner develops

**TABLE M.5**
**Matched Pairs with Dichotomous Antecedent and Dichotomous Outcome: Prospective Study**

| | Partner 1 Antecedent Not Present (Not Exposed or Control) | | |
|---|---|---|---|
| **Partner 2 Antecedent Present (Exposed or Experiment)** | **Positive Outcome (Disease+)** | **Negative Outcome (Disease–)** | **Total** |
| Positive outcome (disease+) | *a* | *b* | ***a + b*** |
| Negative outcome (disease–) | *c* | *d* | ***c + d*** |
| **Total** | ***a + c*** | ***b + d*** | ***n = a + b + c + d*** |

the disease and the nonexposed partner does not develop the disease. This is the number of disconcordant pairs of one type—the other being *c*. *a* and *d* are the numbers of concordant pairs.


Quinn McNemar

A very popular criterion for testing association in case of matched pairs is as follows:

$$\text{McNemar test: } \chi^2_M = \frac{\left(|b-c|-1\right)^2}{b+c},$$

where *b* and *c* are as given in Table M.5. For large *n*, this is referred to as a chi-square distribution with 1 df for obtaining the *P*-value. The restriction of no expected cell frequency less than five applies to the cells containing *b* and *c*. The subtraction of one in the numerator of this formula represents a **continuity correction** similar to Yates correction for chi-square. Note that the concordant pairs *a* and *d* do not contribute to the decision, and it is based solely on the number of discordant pairs of the two types. Significance would mean that the disconcordance is not symmetric, and the pairs do not match with respect to the outcome. The numbers *b* and *c* can be large, but if they are equal, the test will not give significance. This is illustrated in the example below.

To evaluate the role of a therapy in relieving common cold within a week, suppose 50 cases underwent the therapy and another group of 50 cases served as controls. The experimental and control cases were matched one-to-one for age, gender, and body mass index (BMI) so that these do not act as confounders. Matching on BMI was done largely to rule out nutritional status as a confounder. The results obtained are summarized in Table M.6. There are 22 pairs in which both types of subjects—with therapy and without therapy—felt relieved in 1 week's time. In 15 pairs, the subject with therapy felt relieved, but the subject without therapy did not feel so. The

**TABLE M.6**
**Trial for Therapy for Common Cold: Matched Pairs**

| | Without Therapy (Control Group) | | |
|---|---|---|---|
| **With Therapy (Experimental Group)** | **Relieved within 1 Week** | **Not Relieved within 1 Week** | **Total** |
| Relieved within 1 week | 22 | 15 | **37** |
| Not relieved within 1 week | 5 | 8 | **13** |
| **Total** | **27** | **23** | **50** |

frequencies in the second row can be similarly explained. In this table, $b = 15$ and $c = 5$. Therefore,

$$\text{McNemar } \chi^2_M = \frac{\left(|15-5|-1\right)^2}{15+5} = 4.05.$$

A software package gives $P = 0.044$, which is less than 5% for this value of chi-square. The null hypothesis in this case is that the therapy has no effect. But the likelihood of this sample coming from this null is extremely small—less than 5%. Thus, reject $H_0$ and conclude that the therapy is helpful in relieving common cold within 1 week. Note that the number of those relieved by therapy (15 subjects) is much more than those relieved without therapy (5 subjects) among the discordant pairs.

The McNemar test is valid for large *n*. For small *n*, there is an exact test for association in matched pairs as described under the topic **binomial test**.

1. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947;12(2):153–7. http://link.springer.com/article/10.1007%2FBF02295996

## mean (arithmetic, geometric and harmonic), *see also* mean, median, and mode (calculation of)

### *Arithmetic Mean*

The usual mean, called the average in day-to-day language, is statistically called the arithmetic mean. This is calculated as $\Sigma x/n$ in the case of ungrouped data and as $(\Sigma f_k x_k)/n$ in the case of grouped data, where $f_k$ is the frequency (number of subjects) in the $k$th class interval and $x_k$ is the midpoint of the class interval. For further details, see the topic **mean, median, and mode (calculation of)**. The arithmetic mean is used where the numbers are additive. This does not apply to multiplicative values such as dilution and pH values. This is also not used for values measured in logarithmic units. For these, use the geometric mean (GM).

### *Geometric Mean*

In place of sum of the values used in arithmetic mean, the GM uses the multiplication of the values, and in place of dividing by the number of values *n*, it takes the *n*th root of the values. Thus,

$$\text{GM} = (x_1 * x_2 * \cdots * x_n)^{1/n}.$$

This shows that $\log\text{GM} = (1/n)*\Sigma \log x_i$. Thus, the GM is the antilogarithm of the arithmetic mean of logarithms. In the case of grouped data, this becomes

$$\text{GM (grouped data): } \log\text{GM} = (1/n)*\Sigma f_k \log x_k,$$

where $x_k$ is the midpoint of the $k$th ($k$ = 1, 2, …, $K$) interval that has $f_k$ frequency, and $n = \Sigma f_k$.

Since the $n$th root is involved, the GM can be calculated only when all $x$'s are positive. The GM can be very different from arithmetic mean; for example, the GM of 4, 16, and 64 is 16, whereas the arithmetic mean of these values is 84/3 = 28.

The GM is used extensively in microbiological and serological research where the observations are often expressed as titers: the dilutions of certain suspensions or reagents at which a specified phenomenon (for example, agglutination of red blood cells) first takes place. With repeated observations, the possible values of a titer will usually be multiples of the same dilution factor, for example, 2, 4, 8, 16, etc., for twofold dilutions. Denote the titer by $x$ and the log titer by $u$ (= log $x$). The arithmetic mean of $u$ is the logarithm of the GM—take the antilog to get the GM. Another common use of this measure is in tissue attenuation correction for gastric emptying time. The GM of gastric counts is considered the **gold standard** for this correction. Gastric counts generally have a multiplicative feature so that the GM is more appropriate than the arithmetic mean. The GM is also used for gamma radiation count since these also are multiplicative rather than additive.

Situations in which logarithmic transformation of values is helpful in achieving certain desirable statistical properties are discussed under the topic **logarithmic scale/transformation**. The GM can be gainfully used in those cases. A statistical property of the GM is that it can never be greater than the arithmetic mean, and the two means are equal only if all the $x$'s are the same. This property can be used to find if the values are nearly the same or not. If the GM and the arithmetic mean are given and the actual values are not known, you can roughly assess the variation in the values by the closeness of the GM to the arithmetic mean.



### *Harmonic Mean*

This type of mean transforms the data to reciprocals, calculates the arithmetic mean on the transformed scale, and then converts back to the original scale by taking the reciprocal again. The resulting quantity is known as the harmonic mean (HM). In simple terms, the HM is the reciprocal of the mean of the reciprocals. It is rare to find this being used but is useful for averaging the rates.

Sometimes rates are stated in a reciprocal manner. An example is the average population served per doctor. If this is 1000 for the rural area in a district and 500 for the urban area, what is the average for the district as a whole? Even if the populations in rural and urban areas are equal, the average is not 750. The data in Table M.7 explain this anomaly.

When rural and urban areas are combined, the average population served per doctor is 100,000/150 = 667 in area 1 and not 750, the simple average of 1000 and 500. This correct average can be computed as follows:

$$\text{Average population served per doctor} = \frac{50{,}000 + 50{,}000}{\frac{50{,}000}{1000} + \frac{50{,}000}{500}} = 667.$$

This is the HM. If the populations in rural and urban areas are unequal as in area 2 in Table M.7, the average population served per doctor is 100,000/125 = 800. This also can be obtained as the HM:

$$\text{Average population served per doctor} = \frac{75{,}000 + 25{,}000}{\frac{75{,}000}{1000} + \frac{25{,}000}{500}} = 800.$$

The general procedure for calculating the HM for grouped data is as follows:

$$\text{Harmonic mean} = \frac{\Sigma f_k}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \cdots + \frac{f_K}{x_K}},$$

where $x_k$ is the value in $f_k$ subjects. In case of categorized metric data, $x_k$ is the midpoint of the $k$th class interval ($k$ = 1, 2, …, $K$). For ungrouped data, each $f_k$ = 1 for all $k$ and $K = n$. Note the following:

- The GM gives relatively small weight to the large values, and the HM gives relatively large weight to the small values. This may become necessary in some cases as illustrated above. However, the use of these two averages is relatively rare and rightly so since they are difficult to calculate and understand, and are not desirable unless really needed.
- Just as the GM, the HM too cannot be computed if one or more values are zero or negative.

## mean deviation, *see* variation (measures of)

## mean, median, and mode (calculation of), *see also* central values (understanding and which one to use)

Mean, median, and mode are three statistical measures of central values in the data. Statistically, mean is of three different types as separately described under the topic **mean (arithmetic, geometric, and harmonic)**. When not so qualified, mean is the arithmetic mean

**TABLE M.7**
**Population Served per Doctor in Rural and Urban Areas: Equal and Unequal Rural–Urban Population**

| | | Area 1 (Equal Rural–Urban Distribution) | | Area 2 (Unequal Rural–Urban Distribution) | |
|---|---|---|---|---|---|
| Area | Population Served per Doctor | Population | Number of Doctors | Population | Number of Doctors |
| Rural | 1000 | 50,000 | 50 | 75,000 | 75 |
| Urban | 500 | 50,000 | 100 | 25,000 | 50 |
| | **Total** | **100,000** | **150** | **100,000** | **125** |

and is the usual average. Mean still is the most useful statistical tool ever invented and perhaps the most widely used.

The average is a very popular measure of central value. If $x_1$, $x_2$,..., $x_n$ are $n$ observations in our sample, the sample mean is $\bar{x} = \Sigma x_i/n$. The median is the middle value, which is obtained as the $\left(\frac{n+1}{2}\right)$th value if $n$ is odd, after arranging in ascending order, and the average of $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th if $n$ is even. The mode is the most common value. Their calculation in a simple data is illustrated next.

The following are the days of immobility in 38 women with acute polymyositis of the back:

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | 5 | 9 | 7 | 36 | 4 | 6 | 7 | 5 | 8 | 3 | 6 | 5 |
| 7 | 8 | 10 | 7 | 14 | 10 | 9 | 4 | 6 | 11 | 9 | 6 | 5 |
| 8 | 8 | 6 | 7 | 5 | 5 | 12 | 3 | 5 | 9 | 10 | 7 | |

Mean = 7.9 days
Median = Average of the 19th and 20th values after rearranging in ascending order
= (7 + 7)/2 = 7 days
Mode = 5 days and 7 days, occurring in 7 patients each

Mean is denoted by $\bar{x}$ for sample and by μ for population. No such universally accepted notations are available for median and mode.

Mean should not be calculated for codes, but when the measurement is **dichotomous** such as code 0 for absence of a characteristic and 1 for presence of the characteristic, the average has a valid physical meaning—it is the proportion of cases with the presence of the characteristic. Mean, median, and mode can be calculated for **ordinal** data such as 0 for no disease, 1 for mild, 2 for moderate, 3 for serious, and 4 for critical disease; but these are then used as scores and not codes in the sense that the sum of one mild with score 1 and one moderate case with score 2 is equal to score 3, which is the score for serious disease. Be careful since this kind of implication is not valid for many ordinal characteristics.

### *Calculations in the Case of Grouped Data*

When the exact values are available, the calculations must invariably be based on such values. However, for reasons given for **categories of data values**, the data are sometimes available in grouped form. In that case, the calculations become approximate. In many applications, the midpoint of the class intervals containing the required central value is considered adequate. However, the best approximations are obtained as follows.

Let the class intervals in ascending order be $(a_0, a_1)$, $(a_1, a_2)$, …, $(a_{K-1}, a_K)$ and their midpoints $x_k = (a_{k-1} + a_k)/2$, $k = 1, 2, \ldots, K$. Note that these are continuous intervals, and there is no gap between $a_{k-1}$ and $a_k$. Be careful of this requirement. If age intervals are like 10–14, 15–19, etc., they have to be considered as 10–15, 15–20, etc. This actually is so since age 14 in completed years is the same as age <15 years. However, the frequency can be zero in any interval—that does not make them discontinuous. Let the number of subjects or the frequency in the first interval be $f_1$, the second interval $f_2$, and so on. Then,

$$\text{grouped data: mean} = \frac{\Sigma f_k x_k}{n}, \text{ where } n = \Sigma f_k,$$

$$\text{grouped data: median} = a_{m-1} + \frac{n/2 - C}{f_m} * h_m,$$

where
$f_m$ is the frequency in the interval containing the $(n/2)$th observation (called the median interval).
$a_{m-1}$ is the lower limit of the median interval.
$C$ is the cumulative frequency preceding the median interval.
$h_m$ is the width of the median interval.

$$\text{Grouped data: mode} = a_{M-1} + \frac{f_M - f_{M-1}}{2f_M - f_{M-1} - f_{M+1}} * h_M,$$

where
$f_M$ is the frequency in the interval with the highest frequency (modal class).
$f_{M-1}$ and $f_{M+1}$ are the frequencies in the preceding and succeeding intervals.
$a_{M-1}$ is the lower limit of the modal class.
$h_M$ is the width of the modal class.

The last considers adjacent frequencies to identify the peak that determines the mode. The following example illustrates the calculations. These are the same data as in the previous example in this section, but they are now grouped.

The continuous intervals formed in Table M.8 are a natural consequence of the observed durations. The duration of immobility would rarely be exactly 6 days or exactly 8 days. When it is noted as 6 days, it is likely to be anywhere between 5.5 and 6.5 days. If the duration is noted in terms of completed days, then 6 days is really between 6 and 7 days. In that case, the continuous intervals would be (3–6), (6–9), etc., and the mean, median, and mode would change accordingly. For the data in Table M.8,

$$\text{mean} = (11 \times 4 + 16 \times 7 + 8 \times 10 + 2 \times 13 + 1 \times 36)/38 = 7.8 \text{ days.}$$

For the median, the interval containing the 19th and 20th observations is (5.5–8.5) days. Thus, $f_m = 16$, $a_{m-1} = 5.5$, and $C = 11$. Also, $h_m = 3$. Thus,

$$\text{median} = 5.5 + \frac{19-11}{16} \times 3 = 7.0 \text{ days.}$$

**TABLE M.8**
**Grouping of Data in Duration of Immobility in Cases of Acute Polymyositis**

| Group ($a_{k-1}$, $a_k$) | 2.5–5.5 | 5.5–8.5 | 8.5–11.5 | 11.5–14.5 | 36 | Total |
|---|---|---|---|---|---|---|
| Midpoint ($x_k$) | 4 | 7 | 10 | 13 | 36 | |
| Frequency ($f_k$) | 11 | 16 | 8 | 2 | 1 | **38** |
| Cumulative frequency | 11 | 27 | 35 | 37 | 38 | |

Note that 19 is away from cumulative 11, and it is reasonable to divide this difference proportionately. For the mode, the interval containing the highest frequency is again (5.5–8.5) days. Thus, $f_M$ = 16, $a_{M-1}$ = 5.5, $f_{M-1}$ = 11, and $f_{M+1}$ = 8.

$$\text{Mode} = 5.5 + \frac{16-11}{32-11-8} \times 3 = 6.7 \text{ days.}$$

In case the graphical representation indicates that there are two or more modes, similar calculations will be required for each mode, after identifying its modal class. As explained for **bimodal distribution**, one mode may be for smaller peak and the other for the higher peak.

### *Features of Mean, Median, and Mode*

The values of mean, median, and mode sometimes change because of grouping. The values obtained on the basis of the ungrouped data in our example are exact. The approximation in grouped data occurs because all observations in an interval are assumed to be at its midpoint. The magnitude of error depends mostly on the width of the intervals. You may wish to try another grouping and see how the values of the mean, median, and mode are affected.

Mean, median, and mode are proportionately affected by change of origin and scale. For example, if you add 6 to each value of $x$, the mean will also be 6 more. If you multiply each value of $x$ by 3, the value of the mean will also be thrice as much as the original. That is,

$$\text{mean}(a + bx) = a + b^*\text{mean}(x).$$

If body temperature is measured in Fahrenheit and subsequently required to be converted to Celsius, there is no need to calculate the mean again. Since C = 5/9*F – 160/9, (mean in °C) = 5/9 × (mean in °F) – 160/9. This property is also true for median and mode.

Compared to the mean, the median has a number of disadvantages.

(i) It wastes information in the sense that it takes no account of the precise magnitude of most of the observations. This, in fact, is a strength of the median as much as a weakness. Strength arises due to its insensitivity to outliers or extreme values—thus, it could be a suitable measure when such values are present. The weakness is that if one value changes from 3 to 5 days or from 12 to 14 days in our example, the median will not change.

(ii) If two groups of observations are pooled, the median and mode of the combined group cannot be expressed in terms of the medians or modes of the two component groups. This is not so true with the mean. If the mean of one group of $n_1$ subjects is $\bar{x}_1$ and that of the second group with $n_2$ subjects is $\bar{x}_2$, the combined mean is $\frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$.

(iii) Median is not much amenable to mathematical manipulations as mean is.

Whereas mean and median are unique in the sense that you cannot have two means or two medians for the same set of data, you can have two modes. For details, see the topic **bimodal distributions**. In addition, sometimes a dataset may have two modes just by chance due to sampling, which may not really exist in the population from which this sample was drawn. Thus, extra care is needed when dealing with modes.

## mean squares due to error (MSE), *see* mean squares in ANOVA

## mean squares in ANOVA

Like **variance**, mean square is the average of sum of squares of deviations from mean. The difference is that variance is obtained for the data as a whole, whereas mean square is obtained after breaking the total sum of squares into its various components such as between groups and within groups. Consequently, the mean from which deviations are taken and the divisor also change. Mean squares are used in the **analysis of variance (ANOVA)** to find the contribution of various factors to the total variance and to assess their statistical significance. This is applicable when the outcome is quantitative and the groups in statistical terms are the categorical **independent variable**.

We explain the mean squares with the help of a simple setup of **one-way ANOVA** with three or more groups. The objective is to find whether or not the means in these groups are same in the population. We also assume that the number of subjects in each group is the same $n$ indexed by $i$ ($i$ = 1, 2, …, $n$). This can be easily extended to unequal $n$ and two-way and higher-way ANOVA, but the notations become complex.

Let the $i$th value in the $j$th ($j$ = 1, 2,…, $J$) group be denoted by $y_{ij}$, the mean in the $j$th group by $\bar{y}_{\cdot j}$, and the combined mean for all groups together by $\bar{y}_{\cdot\cdot}$. Now the total sum of squares for all groups combined is the numerator of the usual variance given by $\Sigma_i\Sigma_j(y_{ij} - \bar{y}_{\cdot\cdot})^2$. With little algebra, it can be shown that

$$\Sigma_i\Sigma_j(y_{ij} - \bar{y}_{\cdot\cdot})^2 = \Sigma_i\Sigma_j(y_{ij} - \bar{y}_{\cdot j})^2 + \Sigma_i\Sigma_j(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2.$$

Or total sum of squares = within-groups sum of squares + between-groups sum of squares.

Within-groups sum of squares is also called the **error sum of squares** or *sum of squares due to error*. The term *error* here is just for deviation from the respective mean and not for a mistake. Similarly, the total degrees of freedom (df) can also be broken as $(nJ - 1) = J(n - 1) + (J - 1)$, i.e., total df = error df + between-groups df. When divided by the respective df, we get the following.

$$\textbf{mean squares due to error (MSE)} = \frac{\Sigma_i\Sigma_j(y_{ij} - \bar{y}_{\cdot j})^2}{J(n-1)},$$

$$\text{between-groups mean squares} = \frac{\Sigma_i\Sigma_j(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2}{J-1}.$$

The MSE is also called **error variance**. This should make it clear what mean squares are. As stated earlier, these are the average sum of squares of deviations from the respective means. Similar mean squares can be obtained for several other setups such as two-way ANOVA, random effects models, and regression. These mean squares are used to test the significance of the difference between group means as well as for other such purposes depending on the setup. If between-groups mean squares is substantially greater than the within-groups mean squares (MSE), the means in the groups are considered statistically significantly different. This is checked by ***F*-test** using the ratio of these two in place of the difference as the ratio follows an $F$-distribution. For this reason, this is also called the ***F*-ratio**.

## measurement errors, *see also* errors

When the measurement of an object is not what it actually is, error is said to have occurred. Error is the difference between the obtained value and the actual value. For example, consider an ideal situation where each birth is immediately registered so that the baby's age at any time would be exactly known. If you ask the person 41 years later and he/she reports an age of 40 years, there is an error of 1 year. If a person's exact age is 41 years, 8 months, and 21 days, and the reported age at the time of an interview is 42 years, the error is 3 months and 9 days. It is a different matter whether this error is of any consequence or not. In many cases, the error can be so small that it can be safely ignored, or in some cases, our instrument may not be able to detect it.

Measurement error could be systematic or random. In case you are measuring blood pressure (BP) by mercury sphygmomanometer and there is a bubble in the mercury column, this will always give BP more than the actual value. This is a systematic error. The number of births and deaths in an area with incomplete registration system will always be underreported—it cannot be overreported. This is a systematic error although it can vary from area to area. The average underreporting could be considered as an estimate of the systematic error, although this requires that the correct value in at least some sample subjects is obtained by intensive effort. Thus, this estimate can be obtained only when a **gold standard** is available against which the error is measured. This is rarely available in health and medical setup, and the best available method tends to be recognized as gold for this purpose.

When the same object measured twice in the same state by the same method yields two different values, the difference of these values is also called the measurement error. This error occurs in nearly all situations, although there are some situations in which it is so minor that it can be ignored. This error is random in the sense that it will be negative in some subjects and positive in some, and the average tends to become zero in the long run. Random error looks unpredictable by its very nature, and it is really so in individual subjects; but the beauty is that these errors tend to follow **Gaussian distribution** when considered for a large number of subjects—thus, their long-term pattern is predictable. We know that if the errors are truly random, their mean would be zero, most errors would be clustered around zero with equal proportion of positives and negatives, and a few can be large.

How do we know whether the measurement error is systematic or random? If the mean is zero, it is random, and if the mean is significantly different from zero, at least part of it is systematic. This can be easily checked by the **Student *t*-test**, particularly for large samples. Systematic errors generally have small standard deviation (SD) because this is mostly constant from subject to subject. Random errors may have a relatively large SD. Both would generally have a Gaussian distribution. In situations where the error can become large for large values, the distribution would be positively **skewed**. An example of this is duration of disease. If a disease is present for several years, it could be reported as 10 years instead of 11 years—a difference of 365 days; but if it is present just for 11 days and reported as 10 days, the error is just 1 day. Both look the same in terms of percentage, but they are very different in absolute value. If the variable is duration of disease comprising all diseases, the distribution of errors would be skewed to the right. If the measurement errors have both systematic and random components, just subtract the mean error from each measurement and what is left is the random component.

No matter what the errors are, they always breed uncertainty. If these are more than the ignorable threshold, relationships and dependences could be misinterpreted. The confidence intervals would be inflated, and tests of statistical significance can call a significant effect as nonsignificant. This happens because the presence of errors would inflate the **standard error (SE)** of the estimate. Just becoming aware of these implications may make you more alert to adopt methods that have less measurement errors.

## measures of central tendency, *see* central values (understanding and which one to use)

## measures of dissimilarity and similarity

These measures quantify the extent of similarity or dissimilarity between two or more observations. In the case of univariate values such as a brain size of 1254 $cm^3$ in one person and 1267 $cm^3$ in another person, we immediately know that the difference is 13 $cm^3$ and this difference is just about 1% of the usual size. If a third person's brain size is 1256 $cm^3$, we know that it is "similar" to that of the first person and dissimilar to the brain size of the second person. For quantitative values, the dissimilarity could be in terms of absolute difference or relative difference. In case of qualitative data, if two persons have cancer at the same site, we know that these persons are similar with respect to the site of cancer. In case of ordinal data also, we know that mild disease is more similar to moderate disease compared with serious disease. Thus, measurement of similarity in univariate observations is pretty much straightforward.

The question of dissimilarity and similarity assumes importance in case of multivariate observations. If one person is hypertensive, is male, is 58 years old, and has body mass index [BMI] of 32 $kg/m^2$ and total cholesterol of 220 mg/dL, and the second person is diabetic, is male, is 67 years old, and has BMI of 28 $kg/m^2$ and total cholesterol of 184 mg/dL, how similar or dissimilar are these persons with respect to these five measurements? The answer is not easy for multivariate observations that comprise both qualitative and quantitative values but can be answered in a variety of ways when all measurements are either quantitative or all qualitative.

### *All Measurements Quantitative*

Consider the diastolic blood pressure level of $n$ pairs of persons where each person of a pair has been on a similar diet (such as a husband and his wife eating the same food for 10 years). This gives rise to bivariate measurements. Dissimilarity between the values in each pair can be assessed in terms of the difference in diastolic levels, and that for the group by, say, the average difference. But it might be more useful to find the correlation coefficient between these values. This can also be obtained when you measure systolic and diastolic levels of $n$ persons or even when you measure age and BMI. While this measure has considerable merits, there are a lot of demerits also as mentioned under the topic **correlation coefficient**. This is obtained as **Spearman rank correlation** for rank data. Similarity as measured by correlation coefficient is just for pattern (when one is increasing, how consistently the other is also increasing) and not for closeness of the values, and it is restricted to linear relationship. Correlation coefficient will be perfect if one value is always, say, six less than the other. For one-to-one matching, explore if you can use **agreement** methods. If several quantitative measurements are available on each person, **intraclass correlation** might be appropriate.

For absolute values (and not trend), a popular measure of dissimilarity in multivariate quantitative measurements is the **Euclidean distance** as presented under that topic. A variation of this is the **Mahalanobis distance** that adjusts the Euclidean distance by the covariance structure among them.

For measuring dissimilarity in a **clustering** setup, where the same quantitative characteristic is measured for two groups of subjects, we have indices such as nearest neighbor as used in the **single linkage method**, farthest neighbor used in **complete linkage**, distance between **centroids**, distance between **medians**, etc. All these methods of clustering are discussed under the respective topics.

### *All Measurements Qualitative*

Similarity between qualitative characteristics is generally measured in terms of degree of association. For this, a host of indices are available. For **association between dichotomous characteristics**, we have measures such as relative risk, odds ratio, Jaccard coefficient, Yule Q, etc. For measuring the degree of **association between polytomous characteristics**, choose among phi coefficient, contingency coefficient, and Cramer *V*. In case of matched data such as agreement between two raters, **Cohen kappa** measures the degree of concordance. For **association between ordinal characteristics**, we have Kendall tau, Goodman–Kruskal gamma, and Somer *d*. See these topics for details.

## median, *see* central values (understanding and which one to use); mean, median, mode (calculation of); confidence interval (CI) for median



## median effective dose, *see* $ED_{50}$

## median method of clustering

See the topic **clustering** to understand what it is. In brief, this is the process to group the observations into clusters such that they are similar to one another within the clusters but different from other groups—a property called internal homogeneity and external isolation. This requires a metric that can measure similarity (or its opposite, distance) between groups of values that can be used to assign values to the most similar cluster. This is not easy when the observations are multivariate—a situation where clustering is most commonly adopted. For **hierarchical clustering**, several metrics are available such as **single linkage**, **complete linkage**, **average linkage**, **Ward**, and **centroid**. All these are discussed in this volume under the respective topics.

The median method of clustering is intimately related to the centroid method. The centroid method considers the distance between the means of two clusters as the distance between the clusters. When a big cluster is merged with a small cluster in hierarchical clustering, the centroid method gives proportionately more weight to the mean of the larger cluster. Gower [1] developed an alternative strategy—give equal weight to the clusters being merged irrespective of their size. This is called the median method of clustering. This allows small and big groups to have an equal effect on the characterization of larger clusters into which they are merged. However, this method can result in a **dendrogram** that is hard to interpret. For this reason, this is rarely used in medical sciences.

1. Gower JC. A comparison of some methods of cluster analysis. *Biometrics* 1967;23:623–8. http://www.jstor.org/stable/2528417

## median test, *see also* sign test, Wilcoxon signed-ranks test

The median test is used for testing equality of medians in two or more populations. This is used for quantitative data only where median has a meaning. You may be aware that medians are used as a measure of central value for highly skewed values since mean is not a good representative in this situation. But the requirement of independence of values in different populations remains. The requirement of same distribution pattern in the populations under comparison is not so strict for this test. This is also called the *Mood median test*.

The median test uses **chi-square** for the data divided into two sets in each group with common median as the cutoff. Suppose you have sample values from $K$ populations, referred to as $K$ groups in the literature. The procedure is to find the combined median of all these $K$ samples. The null hypothesis is that the median in each group is the same. If this null is true, the number of observations greater than the common median would be the same in each group as the number less than the common median. (In a continuous distribution, values exactly equal to the median are not expected—if they occur in a sample, they can be ignored and $n$ correspondingly reduced.) These would be the expected frequencies ($E_{ik}$, $i = 1, 2$; $k = 1, 2, \ldots, K$) for calculating the chi-square. Find the actual frequencies greater than or equal to and less than the common median in each group. These would be the observed frequencies ($O_{ik}$, $i = 1, 2$; $k = 1, 2, \ldots, K$). Display the data with a $2 \times K$ two-way contingency table, and calculate chi-square as usual by $\chi^2 = \Sigma[(O_{ik} - E_{ik})^2/E_{ik}]$ with $(K - 1)$ df. Ignore the group if the sample size is 2 or less. A large value of $\chi^2$ would indicate that the null hypothesis is false.

Consider the ambulation time of bariatric surgery patients divided into three body mass index (BMI) (kg/m$^2$) groups: 40–49, 50–59, and 60+. The number of patients in these BMI groups is 40, 33, and 27, respectively, for a total of 100 patients. The distribution of ambulation time is highly skewed on the positive side with some patients taking a very long time to ambulate. The objective is to find whether or not the median ambulation time is the same in each of these BMI groups.

Suppose the median ambulation time of these 100 patients is 120 h. This is the combined median with 50 patients less than 120 h and 50 patients greater than or equal to 120 h. When the patients in each BMI group are divided by this combined median, the numbers observed in the samples are as shown in the left panel of Table M.9. If all groups have the same median, this would be the same as the combined median. In that case, the numbers in each group would be equally divided above and below the combined median. These expected frequencies are shown in the right panel of the table. These values give $\chi^2 = 6.87$. At 2 df, this gives $P < 0.05$. Thus, the median ambulation time in the three BMI groups cannot be considered equal.

For two groups, the Wilcoxon ranks sum test is widely believed to test the equality of medians. It ranks all of the observations from both groups and then sums the ranks from one of the groups. It is possible, although not common, for groups to have different rank sums and yet have equal or nearly equal medians. Thus, strictly speaking, the Wilcoxon ranks sum test requires that the distribution pattern of the populations under comparison is the same; otherwise, it can give statistical significance even when the medians are nearly the same. This is not so for the median test. The median test requires a large sample to be able to use chi-square, but the Wilcoxon ranks sum test is basically for small samples. The median test can be used with three or more groups, whereas the Wilcoxon test is for two groups only. The corresponding nonparametric test for

**TABLE M.9**
**Observed Frequencies Above and Below the Common Median and Expected Under the Null Hypothesis of Equal Medians**

| Ambulation Time | BMI Group (kg/m²) | | | | BMI Group (kg/m²) | | | |
|---|---|---|---|---|---|---|---|---|
| | 40–49 | 50–59 | 60+ | Total | 40–49 | 50–59 | 60+ | Total |
| ≥120 h | 14 | 18 | 18 | **50** | 20.0 | 16.5 | 13.5 | **50** |
| <120 h | 26 | 15 | 9 | **50** | 20.0 | 16.5 | 13.5 | **50** |
| **Total** | **40** | **33** | **27** | **100** | **40.0** | **33** | **27** | **100** |
| | Observed | | | | Expected under the null hypothesis | | | |

three or more groups is **Kruskal–Wallis**, but that also is basically for small samples and requires the same distribution pattern in different groups under comparison.

## mediators and moderators of outcome

In a cause–effect setup, mediators and moderators are those extraneous factors that can alter the nature and extent of relationship. Clarity regarding distinction between these two types of factors can help in substantially improving the understanding of how a particular effect has emerged. Kraemer et al. [1] have discussed such factors in detail.

Moderators are those factors that are present before the cause starts to operate such as age, sex, and nutritional status (when these are not the causes of interest such as in smoking–lung cancer relationship), and mediators are those that emerge after the cause starts to operate such as switching to filter cigarettes, treatment taken for cough, and dietary supplements for weakness in smoking–lung cancer relationship. In this example, mediators involve steps that are taken after smoking starts to affect the lungs. In an experimental setup, such as in clinical trials, preexisting factors such as age, sex, and nutritional status are moderators again since these can affect the outcome, and mediators are nursing care, compliance with the regimen, diet modifications, etc.

Moderators precede the intervention and are not correlated with the intervention. They either amplify or attenuate the effect or the outcome on their own. Randomization is supposed to equalize the moderators between different intervention groups—thus, their effect can be minimized by random allocation. Otherwise, the subjects can be stratified by the level of moderators, e.g., divide the subjects by age and sex if the outcome is anticipated to be affected by these factors, and then allocate. Comparison of like with like tends to eliminate the effect of the moderators. Such stratification will also help in identifying the subgroup that has more favorable outcome. Proper stratification of moderators before the study helps in improving the study design for controlling their effect more effectively.

Mediators emerge during the course of the intervention and affect the progress from the intervention to the outcome. Thus, these naturally depend on the intervention. Mediators also either amplify or attenuate the outcome, but they are in the causal pathway and are not the cause of the effect. Knowledge of mediators does not affect the design. Their control is only in terms of close supervision so that all the subjects follow the routine as planned, and by handling them at the time of analysis of data if they continue to have potential to alter the outcome. If a mediator is found to enhance the positive outcome, this can be a good indicator of how to modify the intervention for better outcome.

For more details of mediators and moderators, see Shephard [2].

1. Kraemer HC, Stice E, Kazdin A, Offord D, Kupfer D. How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *Am J Psychiatry* 2001 Jun;158(6):848–56. http://ajp.psychiatryonline.org/doi/full/10.1176/appi.ajp.158.6.848
2. Shephard D. *How Investigating Mediators and Moderators Helps Explain Intervention Effects*. Academia.edu, 2010. https://www.academia.edu/1482786/How_investigating_mediators_and_moderators_helps_explain_intervention_effects, last accessed May 25, 2015.

## medical care errors (statistical control of)

Errors in the context of medical care are those that mostly occur due to carelessness and thus are avoidable. These do not include errors beyond human control as per the existing norms such as imperfect tools (lack of sensitivity and specificity of medical tests), nonavailability of the best of otherwise available tools, universal lack of knowledge (epistemic gaps), and nonavailability of human and other resources. Medical care errors include medication errors (deviation from the prescriptions), improper application of tools such as incorrect use of medical devices and disproportionate radiation or anesthetic dose, inadequate nursing care, nosocomial infections, etc. These lead to preventable adverse events. According to an estimate provided by MacDonald [1], hospital medical errors are the third leading cause of death in the United States. Theoretically, all these can be eliminated; practically, they can be minimized.

While a large number of care-providing steps can be suggested to prevent medical errors, our concern is with their statistical control. This in effect means that all the errors of each type are tracked and counted to see if they exceed a predefined acceptable level. Thus, the emphasis is on their control rather than on elimination or even minimization.

The first step in controlling errors remains acknowledging that an error has occurred wherever it does. This is easier said than done since underreporting is common. Granting that they are reported nearly as much as they occur, a control chart as presented under the topic **quality control** can be prepared. This requires setting up a tolerance limit considering the practical limitations. Statistically, the tolerance limit is +2SE beyond the average seen in an excellent hospital, where SE is the standard error. The errors occurring each day or each week are plotted on the control chart. As and when the tolerance limit is crossed, an alert is generated that triggers concerned authorities to investigate why this has occurred and to take action accordingly. This chart should be prepared separately for each kind of error—(i) for medication errors that result into adverse drug reactions; (ii) for anesthesia errors that can result in death; (iii) for sloppy surgeries that can result in resurgery; and (iv) for inadequate nursing care that can increase ambulation time, etc. **Cusum charts** can also help in detecting an unfavorable trend in errors.

1. MacDonald I. Hospital medical errors now the third leading cause of death in the U.S. *Fierce Health Care*, 2013. http://www.fiercehealthcare.com/story/hospital-medical-errors-third-leading-cause-death-dispute-to-err-is-human-report/2013-09-20, last accessed February 26, 2015.

## medical decisions (statistical aspects of)

It may not be apparent to many clinicians how biostatistics plays a substantive role in their day-to-day decisions regarding managing patients. Decisions regarding diagnosis, treatment, and prognosis are rarely absolute and depend on the confidence a clinician places on the accurate assessment of the status of the patient. A measure of this confidence is the probability as discussed in the topic **probabilities in clinical assessment**. This includes how statistical probabilities of misdiagnosis and missed diagnosis help in devising strategies for better outcomes.

Medical tests contribute significantly to the medical decision process. Their validity indices such as sensitivity and specificity are important parameters for factoring into this process. These indices are biostatistical in nature. The **receiver operating characteristic (ROC)** curve helps in identifying the best cutoff with highest sensitivity and specificity of quantitative tests, and in comparing overall performance of one test with the other.

Clinicians like dichotomous categories much more than measurements in continuum. For example, they like to know what levels are normal and what are not normal that require intervention. Gray areas and consequent indecision are sometimes considered sign of incompetency. For decision regarding treating or not, such cutoffs are indeed helpful, and these cutoffs almost invariably are based on statistical rather than clinical considerations (see **normal range**).

Clinicians also like to divide, for example, body mass index (BMI) into normal, overweight, obese, morbidly obese, and super obese categories (rather than the BMI itself), say for deciding the dietary and exercise regimen. Most researchers realize that convenient categorization such as considering BMI 40–49 as morbid obesity and 50+ as super obesity is arbitrary, but many clinicians prefer categories. Some objectivity in such categorization can be achieved through the statistical method of clustering, although this method is rarely used in this context.

We have described only some of the statistical aspects of medical decisions—they, in fact, pervade deeply. For some other aspects, see **decision analysis tree**, **scoring systems**, and **clinimetrics**.

## medical experiments, *see* experimental studies

## medically important difference (tests for detecting), *see also* medically important effect (the concept of)

The null hypothesis commonly used for statistical tests is of no difference. When this is rejected, the only conclusion reached is that a difference is present, without any implication on the magnitude of the difference. The difference could be very small with no clinical implication or could be large enough to be medically important. This uncertainty is tackled by setting up an $H_0$ that specifies the magnitude of the medically important difference.

In such problems, the medical profession needs to decide the minimum acceptable or tolerable difference that justifies the intervention under consideration. The specification could be either in terms of proportion or in terms of mean. The statistical methods used to detect medically important differences are discussed in this section, first for the proportions and then for the mean. Similar tests can be devised for other effects such as odds ratio and relative risk.

### *Detecting a Medically Important Difference in Proportions*

You may be aware that a small difference between groups can become statistically **significant** when the sample size is large. If the cure rate after 1 month of a particular new therapy is 40% in a sample of subjects versus 30% with the existing therapy, the difference would be statistically significant if the number of subjects in each group is 123 or more. But this small difference may not be worth the trouble of switching over to the new therapy if it is relatively more difficult to implement. A difference of even 1% can be statistically significant for a sufficiently large sample, but very few clinicians, if at all, would change their practice for 1% gain. Thus, the medical significance of the difference is always a potent consideration.

In view of the importance of **medical significance** of the result vis-à-vis statistical significance, the concern could be to fairly ensure that a medically important difference is not missed when present. It is for the medical profession to specify the minimum difference that would be considered medically important. The biostatistician has little role.

If a difference of more than 20% is considered to be of some consequence, then $H_0$: $(\pi_1 - \pi_2) = 0.20$. The conventional null is $(\pi_1 - \pi_2) = 0$ (nil), but $(\pi_1 - \pi_2) = 0.20$ is an equally valid null despite not being nil. If this is rejected, the alternative hypothesis $H_1$: $(\pi_1 - \pi_2) > 0.20$ is accepted provided that the sample difference is more than 0.20. If the difference in sample proportions is 0.20 or less, there is no scope for $H_1$ to be true. There is no way, then, that $H_0$ can be rejected in favor of the alternative that says that the difference is more, and there is no need to carry out the test of statistical significance in that case. The question of testing arises only when the sample difference is more than 0.20, and the intention is to find whether it can still be 0.20 (or less) in the target population. This type of argument applies to most **one-tailed tests**. In some situations, the argument can be reverse. The sample difference is less, and you want to know if it still can be a higher value in the population.

For two independent samples of large size, the criterion to test $H_0$: $(\pi_1 - \pi_2) = \pi_0$ is only slightly different from the usual criterion. This is

$$z = \frac{(p_1 - p_2) - \pi_0}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}},$$

where the denominator is the estimated standard error (SE) of $(p_1 - p_2)$ when $\pi_1$ and $\pi_2$ are different. The notations $p_1$ and $p_2$ are for the sample proportion in the two groups as usual. The usual criterion is based on the null hypothesis that $\pi_1 = \pi_2$, which is no longer true in this new setup since the null is not of equality. If $p_1 = 0.40$, $p_2 = 0.30$, each based on a sample of size 50, and $\pi_0 = 0.20$, $H_0$ cannot be rejected in favor of $H_1$: $(\pi_1 - \pi_2) > 0.20$ since the sample difference is less than 0.20. Now consider $H_1$: $(\pi_1 - \pi_2) < 0.20$. Since $n_1 = n_2 = 50$, the value of the criterion is

$$z = \frac{(0.40 - 0.30) - 0.20}{\sqrt{\frac{0.40 \times 0.60}{50} + \frac{0.30 \times 0.70}{50}}} = -1.05.$$

For $H_1$: $(\pi_1 - \pi_2) < 0.20$, smaller values of $z$ would favor $H_1$, and we need to find $P(z \le -1.05)$. Note that $H_1$ and the $P$-value both have

the same direction: in this case, the less-than type. From Gaussian distribution, this is 0.1469, which is not small, i.e., $H_0$ cannot be rejected. The chance that the samples have come from the populations with difference in proportions = 0.20 is not small, and the likelihood of this cannot be denied. However, this does not mean that the difference is 0.20.

In this example, the therapy group is labeled as the first group and the control group as the second group, and $(\pi_1 - \pi_2)$ was expected to be +0.20 under $H_0$. If the labels were reversed, $H_0$ would specify this difference to be –0.20. When everything else is accordingly changed, the conclusion would remain the same.

### *Detecting Medically Important Difference in Means*

Let the minimum acceptable (or tolerable) difference be $\mu_0$. Then $H_0$: $\mu_1 - \mu_2 = \mu_0$ and $H_1$: $\mu_1 - \mu_2 > \mu_0$. The test criterion for independent samples (pooled variance setup) is

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2 - \mu_0}{s_p\sqrt{1/n_1 + 1/n_2}},$$

where $s_p$ is the pooled standard error (SE) of difference in means. This is same as the usual **Student *t*-test** for unpaired samples except for extra $\mu_0$ in the numerator. Similarly, the criterion for a paired setup would also be the same as in a paired *t*-test except for the difference $\mu_0$ in the numerator. For a right-sided $H_1$, the *P*-value would be the probability that *t* is *more* than the value obtained for the sample. Reject the null hypothesis when the *P*-value so obtained is less than the predetermined level of significance. The conditions for application of the criterion just given are the same as for the two-sample *t*-test such as Gaussian conditions, equality of variances, and independence. The following example illustrates the procedure.

It is now well known that an increased level of homocysteine is a risk factor for myocardial infarction and stroke. It has been observed in some populations that vitamin B converts homocysteine into glutathione, which is a beneficial substance, and reduces homocysteine level. To check whether this also happens in nutritionally deficient subjects, a trial was conducted in 40 adults with Hb < 10 g/dL. They were given vitamin B tablets for 3 weeks, and their homocysteine level before and after was measured. Vitamin B has almost no known side effect, but the cost of administering vitamin B to all nutritionally deficient adults in a poor country with a great deal of undernourishment can be enormous. The program managers would consider this a good program if the homocysteine level were reduced by at least 3 μmol on average. In this sample of 40 persons, the average reduction is 3.7 μmol with SD = 1.1 μmol. Can the mean reduction in the target population still be 3 μmol or less?

This is a paired setup, and the medically important difference is a minimum of 3 μmol. Thus, $H_0$: $\mu_1 - \mu_2 = 3$ μmol and $H_1$: $\mu_1 - \mu_2 >$ 3 μmol. Thus,

$$t_{39} = \frac{3.7 - 3}{1.1/\sqrt{40}}$$

$$= 4.02.$$

The table of the Student *t* gives $P(t > 4.02) < 0.01$, which is extremely small, and the null hypothesis is rejected in favor of the alternative. The conclusion reached is that the mean reduction in homocysteine level is most likely more than 3 μmol, and the evidence from this sample is sufficient to conclude that running this program will provide the minimum targeted benefit.

## medically important effect (the concept of)

This is the **effect** that has potential to change the management of a health condition. View this against the annoying feature of statistical methods to discover *significance* in the case of large samples even when the effect of an intervention or of a factor is minimal. For a debate on this aspect, see **medical significance versus statistical significance**.

The malady lies in the **null hypothesis** that is set in testing hypothesis situations and commonly set as no effect. When this null is rejected, the conclusion is that *some* effect is present. The problem starts when this *some effect* is interpreted as *significant* for managing the health condition. The fact is that the statistical significance in this setup only means that an effect is most likely present and that it is not zero. If a new convenient and inexpensive regimen changes HbA1c (glycated hemoglobin) level in diabetics from an average of 9.8% to an average of 9.6% after use for 1 year, which happened to be statistically significant because this was tried on a large number of subjects, the medical question is that whether or not you will like to prescribe this new regimen to your patients. In view of extremely minor effect, perhaps this regimen is not worth adopting despite its convenience and low cost as the current one is probably fully established and tried. That is, the effect of 0.2% in terms of average reduction in HbA1c level is not a medically important effect even if it is statistically significant.

Consider the following examples:

1. When can a new antihypertensive drug be considered clinically effective in postsurgical cases: (i) average decrease in diastolic blood pressure (BP) by at least 2, 5, 8, or 10 mmHg; or (ii) achieving a threshold diastolic BP such as 90 mmHg in a large percentage of subjects—60%, 70%, or 90% of subjects?
2. An iron supplementation program in female adolescents is organized in a developing country. This raises the mean hemoglobin (Hb) level from 13.6 to 13.8 g/dL after intake for 30 days. Is this gain of 0.2 g/dL in the average sufficient to justify the program to all the adolescents in that area? What gain can be considered enough to justify the expenditure and efforts in running the program—0.5 g/dL, 1 g/dL, or more?
3. The prognostic severity of bronchiolitis in children can be assessed either by respiration rate (RR) alone or by using a bronchiolitis score (BS) comprising RR, general appearance, grunting, wheezing, etc. Suppose the former can correctly predict severity in 65% of cases and the latter when considered together in 69% of cases. BS is obviously more complex to implement. Is it worth the trouble to use BS in place of RR for a gain of nearly 4%? What percentage gain in predictivity warrants adopting a more complex BS instead of simple RR—5%, 10%, 15%, or higher?
4. Suppose the normal intraocular pressure (IOP) is 15.8 (SD = 2.5) mmHg in healthy subjects when measured by applanation tonometry. In glaucoma, it is elevated. In a group of 60 patients with primary open-angle glaucoma, suppose the average was 22.7 (SD = 4.5) mmHg. After treatment with a new beta-adrenergic blocker, it came down to 19.5 (SD = 3.7) mmHg. This reduction is statistically significant, but the reduced level in the treatment group is still higher than 15.8 mmHg seen in the healthy subjects. Is this reduction still clinically important? What kind of difference from the normal level of 15.8 mmHg is clinically tolerable—1/2 mmHg, 2 mmHg, or higher?

Depending on the setup, the medically important effect can be the difference in means between two (or more) groups as in our example, difference in percent efficacy, odds ratio, relative risk, etc. For details, see the topic **effect size**. This does not depend on your present data but depends on previous data, experience, and clinical assessment. Thus, it can vary from physician to physician. One may consider that improvement of 1% in HbA1c is quite a gain by such a regimen, but some other may opine that at least a decline of 2% is meaningful when the baseline level is 9.8%. Consensus among medical community is difficult but can be possibly arrived at for such parameters. Specification of medically important effect is a clinical decision and not statistical. An investigation can be planned that is capable of detecting any specified effect size if present by doing the **power analysis** for determining the sample size.

When the medically important effect is known, you can test the null hypothesis that this much effect is present. The methods for doing so are given under the topic **medically important difference (test for detecting)**.

## medical records

Records in any setup are a source of authentic information. Written records of events are probably worth much more than the ones passed orally. They also tend to be permanent and a source of remembrance and reference whenever needed. With online systems in place in most health facilities across the world, medical records tend to be automatically created. These records can provide invaluable longitudinal history of patients when proper **linkages** are done. When complete for all individuals in a community, medical records provide all-important age–sex wise incidence and prevalence rate for all diseases, and trends over time. This information is extremely useful not just to track the health of the people but also to identify the conditions that need priority attention, and to devise strategies for control with a focus on the groups with special needs. Medical records can also help in monitoring the efficacy of various treatment regimens, particularly in a hospital setup, and in identifying more effective regimens for different segments of the patients. The latter comes through when the outcomes with different alternative regimens are compared with one another. Records also help in assessing the quality of medical care provided in different departments and in assessing how much and in what cases clinical findings correlate with laboratory and radiological investigations. Medical records in hospitals are also the source for calculating various administrative indices such as bed occupancy and length of stay in different departments that help to optimize the allocation of beds and services.

A problem with most routine medical records is that they are not complete, they have almost no linkages, and they are maintained without sufficient care that takes away their authenticity. Some of these deficiencies could be because the patients do not provide complete information or sometimes provide even wrong information, but much of it is because the data entry is not done with the care it deserves. Nonetheless, the records tend to be unbiased as they are for past events—thus a good source for research. When the data are freshly collected for a particular research, there is a distinct possibility of bias in collecting the data or of chances of giving more attention to the data that tend to support a particular hypothesis. This bias is not present in the records as they are generated before the study is planned. Thus, records-based studies can provide more valid results provided the records are correct. However, the bias resulting from selected patient profile from records as opposed to random patients remains a severe bottleneck. The findings may be valid for the kind of patients seen in the hospital and may not extend to other types of patients.

With modern methods of **data mining** and **data analytics**, hospital records have become a useful resource for inventing new relationships and new patterns, and for confirming or refuting the existing ones. This is now possible as large hospitals have accumulated enormous data over a decade or so when the online system was introduced in many hospitals across the world.

Public health records, such as of medical certification of cause of death, disease registries, and surveillance systems, also have similar merits and demerits. They may be unlinked anonymously yet still provide useful information on health trends in different segments of population when properly collected and maintained with quality checks.

## medical research (types of)

Before the types of research, understand the nuances of medical research as provided under the topic **research**. Functionally, medical studies can be divided into basic and applied types. Basic research, also termed as pure research, involves advancing the knowledge base without any specific focus on its application. The results of such research are utilized somewhere in the future when that new knowledge is required. Applied research, on the other hand, is oriented to an existing problem. In medicine, basic research is generally done at the cellular level for studying various biological processes. Applied medical research could be on the diagnostic and therapeutic modalities, agent–host–environment interactions, or health assessments.

We would like to classify applied medical studies into two major categories (Figure M.2). The first category can be called primary

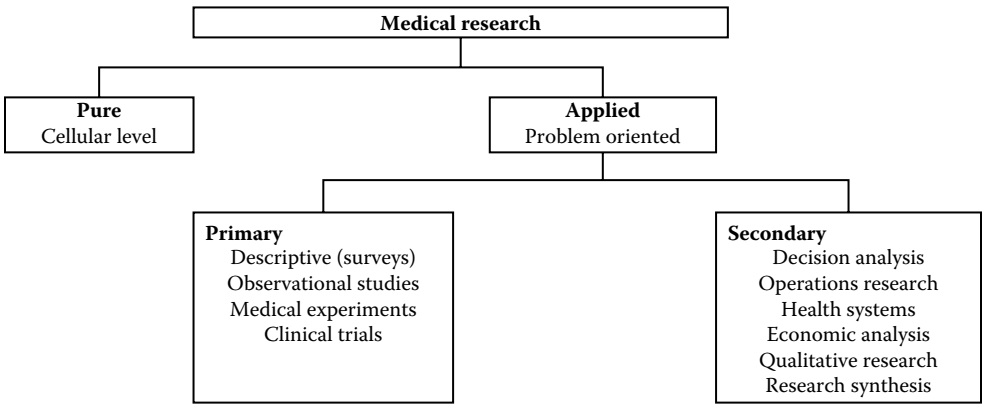



**FIGURE M.2** Types of medical studies.

research that includes **descriptive studies** (**sample surveys**), **observational studies** (**prospective**, **retrospective**, and **cross-sectional**), and **experiments** (in the laboratory on animals and biological specimens, and **clinical trials** on humans). This forms the bulk of medical research today. The second category is secondary research, quite common these days, that includes **decision analysis** (risk analysis and decision theory), **operations research** (prioritization, optimization, simulation, etc.), evaluation of health systems (assessment of achievements and shortcomings), economic analysis (cost-benefit, cost-effectiveness, etc.), **qualitative research** (focus group discussion), and research synthesis (**systematic reviews**, **meta-analysis**). Most of these are discussed in this volume under the respective topic.

## medical significance *versus* statistical significance

The term *significant* in common parlance is understood to mean noteworthy, important, or weighty as opposed to trivial or paltry. **Statistical significance** has the same connotation, but it can sometimes be at variance with medical significance. A statistically significant result may be of no consequence in the practice of medicine, and a medically significant finding may sometimes fail a test of statistical significance. In view of its importance, we explain this distinction in detail in this section.

Some medical professionals consider statistical methods notorious for discovering significance where there is none and not discovering where one really exists. The difficulty is that statistical methods give high importance to the number of subjects in the sample. If a difference exists in 12 cases of chronic cirrhosis of liver and 12 cases of hepatitis with respect to average aspartate aminotransferase (AST) levels, it can be considered a fluke because of the small size of the groups. This is like weak evidence before the court of law. But if the same difference is exhibited in a study on 170 cases of each type, it is very likely to be real. The conventional statistical methods of testing hypothesis only tell whether a difference is unlikely or not without saying how much. The difference in average AST levels between cirrhosis and hepatitis cases could be only 3 units/mL, which has no clinical relevance, but it would turn out to be statistically significant if this occurs in large groups of subjects. Medical significance of such a small difference should be separately evaluated using clinical criteria, and it should not depend exclusively on statistical significance. However, admittedly, clinical criteria could be very subjective in many situations.

Suppose it is known that 70% of cases with sore throat are automatically relieved within a week without treatment because of a self-regulating mechanism in the body. A drug was tried on 800 patients and 584 (73%) were cured in a week's time. Thus, for $H_0$: $\pi = 0.70$,

$$z = \frac{0.73 - 0.70}{\sqrt{0.70 \times 0.30 / 800}} = 1.85.$$

In this case, suppose that the drug is pretested and there is no stipulation that the drug can make the relief rate even less than 70%. Then, the alternative hypothesis is one-sided, $H_1$: $\pi > 0.70$. For this $H_1$, a one-tailed probability is required, and

$P$-value = $P(z \geq 1.85) = 0.0322$ from Gaussian distribution.

This probability is very low and certainly small in comparison with the conventional level of significance of 0.05. The sample is extremely unlikely to be in consonant with this null, and the null is rejected. Statistical significance is achieved, and the conclusion is reached that the 73% cure rate observed in the sample is really more than 70% seen otherwise. But is this difference of 3% worth pursuing the drug? Is it medically important to increase the chance of relief from 70% to 73% in case of sore throat by introducing a drug? Perhaps not. Thus, a statistically significant result can be medically not significant. A remedy is that a medically significant difference such as 10% is specified, and then a statistical test is used to check whether the difference is beyond this threshold. This aspect is discussed under the topic **medically important difference (tests for detecting)**.

The inverse possibility is not so convincing, but that can also occur. A medically important difference should be statistically significant for it to be acceptable. Consider the same setup as in our example where sore throat is assumed to be relieved in a week in 70% of cases without any intervention. Suppose now that the drug is tried in a sample of $n = 50$ patients and 40 ($p = 0.80$) respond in a week's time. Note that $np$ and $n(1 - p)$ are both at least 8, so the **Gaussian conditions** are met even with $n = 50$. Now, under $H_0$: $\pi = 0.70$,

$$z = \frac{0.80 - 0.70}{\sqrt{0.70 \times 0.30/50}} = 1.54.$$

This gives $P$-value = $P(z \geq 1.54) = 0.0618$.

If the probability of **Type I error** to be tolerated is less than 0.05, then the $H_0$ cannot be rejected. Despite 10% improvement shown by the drug in the sample, the sample still cannot be considered to provide sufficient evidence in favor of the alternative $H_1$: $\pi > 0.70$. Thus, this sample of size 50 does not favor a recommendation of use of the drug for patients with sore throat.

The gain of 10% in efficacy, from 70% to 80%, may be considered medically important, and sufficient to pursue the drug, but this rise in this sample of 50 subjects could well have arisen due to chance—due to **sampling fluctuations**. There is quite some likelihood that this rise would fail to be reproduced in another sample of 50 subjects or that it would fail to persist in the long run. Despite this negative finding, the results could be considered to justify a bigger trial because the $P$-value is only slightly more than 0.05.

### *Considerations in Proper Interpretation of Statistical Significance vis-à-vis Medical Significance*

Besides the usual considerations in statistical significance, the following considerations also help in determining medical significance of a result.

- *Whether or not a statistically significant result has any medical significance*. Perhaps it needs to be emphasized that the effect must be statistically significant for it to be medically relevant. If it is not statistically significant, nobody can be confident that the effect is actually present. Thus, the first step is to assess statistical significance. If not significant and the statistical power is adequate, there is no need to worry about its medical relevance because this effect is likely to be there by chance. If significant, further statistical testing is used to judge if it reaches a medically relevant threshold. This threshold comes from medical acumen (see **medically important effect**), and a value judgment is still required. In a randomized controlled trial, if 3 deaths occur in 100 subjects in the placebo group and none out of 100 in the treatment group, the difference is not statistically significant; but would you not try this treatment for your family if everything else has failed? Additional factors such as environment, family condition, and availability of health infrastructure are also considered while taking a final decision regarding the

management of a patient, and statistical significance alone is not enough.

- *Whether or not a plausible medical reason is available for the observed difference*. In many situations, it can be safely concluded that the increase in relief rate is due to the effect of the drug, but in some other situations, the difference is difficult to explain. Consider a random sample of 24 male and 15 female patients with leukemia, of which 4 males and 7 females survive for 5 years. The difference in their survival rate is statistically significant because $P$ would be less than 0.05 for one-sided $H_1$. No worthwhile reason may be available for this difference in their survival rate. When the level of significance is 5%, there is a 1 in 20 chance that false significance is obtained. On the other hand, there might be hitherto unknown factors that could account for such a difference between survival in the two sexes, and the difference could be real. For example, a suspected inborn resistance in females, which is surmised to partially contribute to their greater longevity, may be an explanation. Statistical significance without proper medical explanation is rarely useful with the caveat that such an explanation may not be immediately available in some situations and may emerge in the future.

It should be clear with this that a regimen can be abandoned if sufficiently **powered** study tells you that the effect is not statistically significant. At the same time, do not forget that *P*-value alone is rarely enough to draw a valid conclusion. Previous knowledge, biological plausibility, and your intuition that would also incorporate epistemic gaps must remain the guiding factors. A valid conclusion is reached when all these are considered together wherein statistical evidence plays a role though not as dominant as made out by some statisticians.



## medical uncertainties (types of), *see also* aleatory uncertainties, epistemic uncertainties

Uncertainties pervade all walks of life and medical uncertainties seem omnipresent—their unfailing presence in practically all medical situations can be easily seen. All scientific results are susceptible to error, but uncertainty is an integral part of medical framework. The realization of the enormity of uncertainty in medicine may be recent, but the fact is age-old. Also, our knowledge about biological processes still is extremely limited. These two aspects—variation and limitation of knowledge—throw an apparently indomitable challenge, sometimes becoming so profound that medicine transgresses from a science to an art.

Uncertainty is not something that doctors are used to handling; many find it hard to communicate and understand risks that pervade so much in the practice of health and medicine. Many clinicians deal with these uncertainties in their own subjective ways. Some are very successful, but most are not as skillful. To restore a semblance of science, methods are needed to measure these uncertainties, to evaluate their impact, and of course to keep their impact under control. All these aspects are primarily attributed to the domain of biostatistics. Biostatisticians make a living out of medical uncertainties.

Management of some medical uncertainties is easier when these are divided into appropriate compartments. Major categorization from the biostatistics point of view are **aleatory** and **epistemic uncertainties**. These were first highlighted by Indrayan [1] in medical context in the first edition of his book in 2001. These uncertainties are described under those topics. In addition, the following categorization may also help.

### *Diagnostic, Therapeutic, and Prognostic Uncertainties*

Diagnostic uncertainties arise because the tests or assessments used for diagnosis do not have 100% **predictivity**. An electrocardiogram can give false-positive or false-negative results. None of the procedures, e.g., fine needle aspiration cytology, ultrasonography, and mammogram, are perfect for identifying or excluding breast cancer. Cystic fibrosis is difficult to evaluate.

No therapy has ever been fully effective in all cases. Therapeutic uncertainties are particularly visible in surgical treatment of asymptomatic gland confined prostate cancer and in medical treatment of benign prostatic hyperplasia. Many such examples can be cited. In addition are substances such as combined oral pill where long-term use may increase the risk of breast, cervical, or liver cancer but reduce the risk of ovarian, endometrial, and colorectal cancer. Such anomalies also cause treatment uncertainties.

Prognostic uncertainties due to lack of knowledge exist in sudden severe illness. Such illness can occur due to a variety of conditions, and its cause and outcome are difficult to identify. Nobody can predict occurrence or nonoccurrence of irreversible brain damage after ischemic stroke. Prognosis of terminally ill patients is also uncertain. Method of care for women undergoing hysterectomy is not standardized. Degree might vary from situation to situation, but uncertainties are present everywhere in all clinical decisions including prognosis.

### *Predictive and Other Clinical Uncertainties*

Besides prediction of diagnosis, prediction of outcome of treatment, and prediction of prognosis, there are many other types of medical predictions for which knowledge barriers do not allow certainty. Look at the following examples:

- Gender of a child immediately after conception
- Survival duration after onset of a serious disease
- Number of hepatitis B cases that would come up in the next year in a country with endemic affliction
- Number of people to die of various causes in the future

These examples may be convincing to realize that uncertainties are prominent in all aspects of health and disease, and medical decisions should consider these as much as possible.

### *Uncertainties in Medical Research*

The discussion so far is restricted mostly to the uncertainties present in day-to-day clinical problems. But they are more conspicuous in a research setup than in everyday practice. The backbone of such research is **empiricism**. An essential ingredient of almost all primary medical research is observation of what goes on naturally, or how status changes after an intervention. Because of aleatory and epistemic uncertainties, such observations seldom provide infallible evidence. For example, laboratory experiments are many times replicated to gain confidence as the results of one single experiment are doubted. Results of clinical trials are almost always presented with precautions so that nothing is inferred as absolute. In epidemiological research, a strict distinction is made between an association and cause–effect. Risk factors remain probabilistic in all setups despite convincing evidence, since the results can fail in some cases.

A large number of steps are taken in a research setup to control the effect of uncertainties. Uncertainties remain, but their effect on the results is minimized by a host of biostatistics tools such as appropriate **designs**, adequate **sample size**, validated **data collection** instruments, proper data entry, scrutiny of data, expert statistical

**analysis**, and correct interpretation. Even after such extensive care, the results still are presented in terms of probability.

### *Uncertainties in Health Planning and Evaluation*

Medical care is just one component of the health care spectrum. Prevention of diseases and promotion of health are possibly more important. These can be done at the individual level but are most cost-effective when done at the population level. The first step toward this is **health situation analysis**. This contains steps such as identification of the specifics of the problem, size of the target population, magnitude of the problem, available health infrastructure, and feasibility of remedial steps. Because of interindividual, environmental, and such other variations, uncertainty remains at each step despite using the best methods. Similarly, **evaluation of health programs** also suffers from uncertainty due to better performance by some segments of population despite equal exposure and equitable allocation of resources. Also, the tools to implement the programs are never perfect. Varying motivation of the officials also contributes to the level of uncertainty.

For further details, see Indrayan [1] who has meticulously discussed medical uncertainties at length and has put up convincing arguments how biostatistics can help in managing certain aspects of medical uncertainties. Keep in mind the famous saying that if you begin with certainties, you will end in doubts, but if you begin with doubts, you can end up with almost certainties.

1. Indrayan A. *Medical Biostatistics*, Third Edition. Chapman & Hall/CRC Press, 2012.

## medical uncertainties (sources of), *see* aleatory uncertainties, epistemic uncertainties

## mental health (indicators of)

Mental health is considered to be related to satisfaction and happiness—thus is mostly abstract. Although there is a degree of stigma attached to mental disorders that inhibit valid measurement, attempts have been made to develop indicators that can measure mental health both at individual and population levels. These indicators are used for assessing the level and trends, and to develop strategies for improvement of mental health.

Individual level positive indicators of mental health for adults comprise self-assessment of being useful to the family and the society; feeling optimistic, good, cheerful, confident, and loved; interest in new things and surroundings; creativity; successful coping with day-to-day challenges; and the like. You can ask people to rate their last week for each of such indicators on a 0 to 5 scale, where 0 is for never during the last week and 5 for always in the last week. Thus, mental health of individuals can be measured on a defined scale. However, the instrument so developed may have to be checked for its validity in local setting. The items should be chosen in a manner that the instrument reflects current concepts, yet are sustainable. When the same set of questions are asked year after year in a randomly selected group of people from the same population, the average scores can provide a good understanding of whether mental health of people is improving or declining. Note that this set of indicators does not include negative feelings such as neglect, failures, tension, stress, and anxiety. But these too can be included. Tannenbaum et al. [1] used self-reported subthreshold mental health symptoms, self-reported full diagnostic disorders such as depression and anxiety disorders, physicians' billings for outpatient mental health visits, and psychotropic medications for measuring mental health of Canadians. Their emphasis was on using the existing databases. They found that a maximum of 20% of women and 14% of men in Canada had mental health problems.

Beside percentage of population with mental disorders, epidemiological negative indicators of mental health can be aberrations such as crime rate, suicide rate, homicide rate, accidents, and divorces. You may like to include smoking and alcohol and drug addictions also. Perhaps a combination of these can be devised that can provide a single index. These data are easily available in most countries at national and subnational levels. On the positive side are availability and utilization of sports and entertainment facilities; education levels; availability and utilization of telephone, banks, and post offices; per capita income; etc.

1. Tannenbaum C, Lexchin J, Tamblyn R, Romans S. Indicators for measuring mental health: Towards better surveillance. *Health Policy* 2009 Nov;5(2):e177–86. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2805147/

## meta-analysis, *see also* funnel plot, $I^2$ (index of homogeneity)

Popularly known as doing statistics on statistics, meta-analysis is a synthesis technique of pooling varying results of various studies on the same parameter after a **systematic review**. For this, studies meeting prespecified quality criteria are selected after a comprehensive search of literature. A particular relevant parameter that measures effect size such as odds ratio (OR), relative risk, or mean difference is chosen, and its value with confidence interval (CI) is extracted from each selected study. Sometimes you may have to calculate CI yourself, and sometimes median/IQR type of CI may have to be converted to mean ± 2SE when that information is not directly available. The values and their CIs are examined to come up with a pooled estimate as per the procedure described later in this section. According to O'Rourke [1], the term *meta-analysis* was introduced in 1976 by Glass [2].

Now that so much research is going on each topic around the world, it is difficult to decipher a common message. Meta-analysis helps in combining the diverse results. Pooling of samples in different studies also increases the reliability of the result. Once the pooled result is available, there is rarely any need to study so many different articles on the same topic, thus saving a lot of time. Most prominent among meta-analyses are **Cochrane reviews**, whose findings are many times considered gold standard for actual implementation in health care.

Care required in selecting the articles for meta-analysis can hardly be overemphasized. Generally, databases such as PubMed, EMBase, and Google Scholar are searched for relevant terms. Some articles that do not use those terms may still escape—thus, selection of terms is crucial to spread the net without losing the focus. Secondly, it is customary to use the PECOS system, which stands for **Population, Exposure, Control, Outcome, and Study design**, to filter the relevant articles. There is a practice to pick up only the credible and relevant articles and exclude those that follow different methodology or different definitions not fitting into the preset criteria. Thus, the findings are generally unbiased for published results with a caution as described next.

There is a big drawback of meta-analysis results, however. The studies included tend to self-select because of **publication bias**. There is a **file-drawer effect**, which says that much of research with statistically nonsignificant results remains in a drawer, and studies with

negative results and neutral outcomes often go unpublished. Thus, the published articles provide a biased picture. When the results from the published articles are pooled, the bias inadvertently increases instead of being reduced. We tend to ignore such bias because perhaps nothing can be done to alleviate this problem. One way out is to give extra weight, say double, to negative results assuming that only one of the two negative findings finds its way into the literature [3].

There are statistical precautions also. Results based on a small sample size or with high standard error (SE) in different studies will obviously spread across a broad range of values compared with the results based on large samples. If you plot ORs in three studies each with a small sample size, they are likely to be far apart from one another compared with ORs in another set of three studies each with a large sample size. If you are reviewing a large number of studies—some of small size and some of large size—and plot the OR on the horizontal axis and the sample size on the vertical axis, you will get what is called a **funnel plot** because of its resemblance with an inverted funnel. If your values follow this pattern, you can feel more confident that the results included in meta-analysis conform to this requirement. An asymmetric shape of the funnel plot raises suspicion over the results since the selected studies may suffer from publication bias, favoring either higher or lower effect size. It also suggests the possibility of a systematic bias in smaller studies. If most of smaller studies tend to give larger (or smaller) effect size compared to larger studies, the bias is evident and the results of meta-analysis would not be valid. When biased studies are not included, heterogeneity among results of various studies does not cause much of a problem, and the final CI would reflect this. If needed, heterogeneity can be assessed by an index as explained under the topic ***I*² (index of homogeneity) in meta-analysis**. This measures the percentage variance attributed to between-study variation and replaces analysis of variance because the original data are not available in this case. $I^2 = 0.25$ is low, 0.50 is middling, and 0.75 is high. If this is high, identify studies causing this high value and exclude them from the meta-analysis. Both funnel plot and index of homogeneity are obtained before meta-analysis is done to get convinced that the values being used are consistent across different studies.

A meta-analysis conventionally comprises a plot of the effect size and their CI that provides graphical summary view of the varying results obtained in different studies, called a **forest plot**. For an example, see Figure M.3 for ln(ORs) for patients free of diarrhea in probiotic groups compared with the control groups. This plot provides a great deal of information about the studies included and the results obtained along with the sample size of each. Horizontal lines in Figure M.3 are the 95% CI, and the numerical values are stated in a column on the right side. For pooling the results, the average is calculated, but all studies are not given equal weight in this average. Generally, studies with a larger sample size or with a smaller SE get more weight. The area of the black square represents this weight, and this is also mentioned in another column on the right side in this figure. You may use any other weight that is uniformly available or can be obtained for all the studies. The pooled result is shown by a diamond touching the *x*-axis. The width of this diamond is the width of the pooled CI. If this touches or crosses the line of no effect (OR = 1 in this example), the pooled conclusion is that the effect is not statistically significant. In Figure M.3, the diamond is on the left side of OR = 1; thus, the protective effect of probiotics for diarrhea is statistically significant.

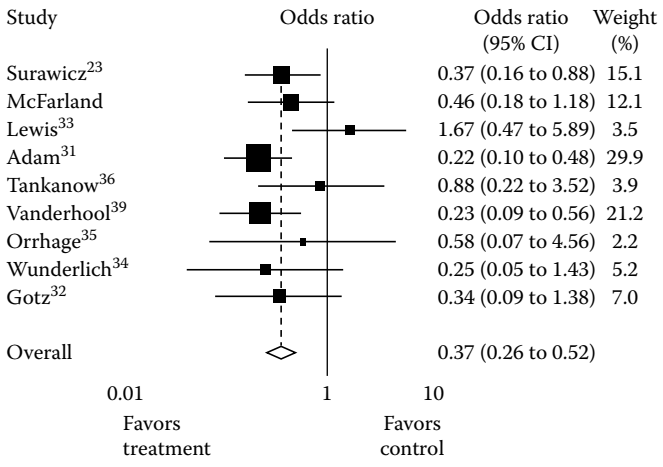



**FIGURE M.3** Plot of the log of odds ratios for the proportion of patients free of diarrhea in probiotic groups compared with control groups. (From D'Souza AL et al., *BMJ* 2002(8 June);324:1361. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC115209/.)

In place of aggregate results of studies, the emphasis now is on individual participant data. Since almost all studies around the world have data on individual subjects in electronic form, they can be easily pooled in a collaborative effort that would provide a direct estimate based on a large number of subjects. Thompson et al. [5] have provided this kind of analysis based on 154,211 participants in 31 studies on hazard ratio for coronary heart disease per 1 g/L raised baseline fibrinogen. Care is needed while pooling in this case also because the participants within each study form a cluster with shared similarities, and the effect of clustering is factored into pooling.

The type of meta-analysis we have described typically considers one particular outcome measure. Studies that do not report this outcome cannot be included, and this can introduce bias. An extension is available, called *network meta-analysis*, which can include two or more end points. Woods et al. [6] have described a meta-analysis that combines count and hazard ratio of survival in **multiarm trials** into single analysis.

Although meta-analysis seems to obviate the need to do large-scale multicentric studies, there are legitimate questions regarding meta-analysis being a proper replacement. Probably this is just a makeshift strategy to arrive at a more reliable conclusion to meet the pressure of quick yet informed decisions.


1. O'Rourke K. An historical perspective on meta-analysis: Dealing quantitatively with varying study results. *J R Soc Med* 2007 Dec;100(12):579–82. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2121629/
2. Glass GV. Primary, secondary, and meta-analysis of research. *Educ Res* 1976:5(10):3–8. http://www.jstor.org/stable/1174772
3. Champkin J. "We need the public to become better BS detectors" Sir Iain Chalmers. *Significance* July 2014; 25–30. http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2014.00751.x/pdf
4. D'Souza AL, Rajkumar C, Cooke J, Bulpitt CJ. Care of the Elderly Section, Faculty of Medicine, Imperial College School of Medicine, Hammersmith Hospital, London W12 0NN. Probiotics in prevention of antibiotic associated diarrhoea: Meta-analysis. *BMJ* 2002(8 June);324:1361. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC115209/
5. Thompson S, Kaptoge S, White I, Wood A, Perry P, Danesh J. Emerging Risk Factors Collaboration. Statistical methods for the time-to-event analysis of individual participant data from multiple epidemiological studies. *Int J Epidemiol* 2010 Oct;39(5):1345–59. http://ije.oxfordjournals.org/content/39/5/1345.full
6. Woods BS, Hawkins N, Scott DA. Network meta-analysis on the log-hazard scale, combining count and hazard ratio statistics accounting for multi-arm trials: A tutorial. *BMC Med Res Method* 2010;10:54. http://www.biomedcentral.com/1471-2288/10/54/

## meta-analysis of observational studies in epidemiology (MOOSE) guidelines

**Meta-analysis** is increasingly used to arrive at an evidence-based reliable result after pooling the information available in publications on different segments of population. Many of these publications are based on **observational studies** since deliberate intervention is not feasible for harmful risk factors such as for smoking. Such studies are especially susceptible to inherent bias and varying designs. MOOSE is a set of guidelines for reporting of meta-analysis of results from observational studies so that the reporting is done with care and nothing is left out. This set was initially developed in a workshop held in Atlanta, Georgia, in 1997 under the aegis of Center for Disease Control and Prevention of the United States [1].

These guidelines are in the form of a checklist as given in Table M.10. If you want to follow these guidelines, see that your manuscript has all these components. If any is missing, either include that or justify why this cannot be included in the manuscript. Readers, authors, reviewers, and editors can use this checklist to confirm the quality and completeness of the meta-analysis. For further details, see Stroup et al. [1].

1. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, Moher D, Becker BJ, Sipe TA, Thacker SB. Meta-analysis of observational studies in epidemiology: A proposal for reporting. *JAMA* 2000 Apr 19;283(15):2008–12. http://jama.jamanetwork.com/article.aspx?articleid=192614

## meta-regression, *see also* meta-analysis

Similar to **meta-analysis**, meta-regression is a method used to combine results from several studies, but here **regression analysis** is used to come to a unified result. The **regressors** are study characteristics such as year of study, type of participants (e.g., mean age, percentage of males, and ethnicity), method of data collection (cross-sectional, longitudinal, length of follow-up, etc.), and assessment of the quality of the study. These regressors can be quantitative or qualitative, fixed or random. Studies can be weighted to reflect the sample size or the precision in terms of the inverse of variance. The dependent could be the effect size as in meta-analysis, and the unit of analysis is a study instead of an individual. Meta-regression is one of the methods of **systematic reviews** that provides an insight of how study characteristics are affecting the effect size. When studies are done at different points in time, say in the years 1980, 1995, 2003, and 2010, their combination after accounting for other factors can provide a legitimate trend of change in the effect size over time.

As in the case of meta-analysis, the studies for meta-regression too are identified after intensive search of the literature for a set of predefined inclusion criteria. These are appraised to check that they indeed deserve to be included or not. Because of **publication bias** and **file-drawer effect**, the results of meta-regression can also suffer from bias just as the results of meta-analysis do.

The analysis is done by following the **generalized linear regression** approach depending on the type of regressors (quantitative, qualitative, or mixed; **fixed or random effects, or mixed effects**) and the **scale** of the dependent (quantitative, ordinal, or nominal) variable. All these terms are explained in this book in easy language for medical professionals who are not familiar. For quantitative dependent and fixed effect regressors, it can be analysis of covariance; for nominal or ordinal dependent, it will be logistic regression; and for counts, it will be Poisson regression. One big problem with meta-regressions is that the number of eligible studies may be small, whereas the number of regressors is large that can overfit the model. Make sure that an enough

**TABLE M.10**
**Reporting Checklist for Meta-Analyses of Observational Studies (MOOSE)**

| MOOSE Checklist |
|---|
| Reporting of background should include |
| Problem definition |
| Hypothesis statement |
| Description of study outcome(s) |
| Type of exposure or intervention used |
| Type of study designs used |
| Study population |
| Reporting of search strategy should include |
| Qualifications of searchers (e.g., librarians and investigators) |
| Search strategy, including time period included in the synthesis and keywords |
| Effort to include all available studies, including contact with authors |
| Databases and registries searched |
| Search software used, name and version, including special features used (e.g., explosion) |
| Use of hand searching (e.g., reference lists of obtained articles) |
| List of citations located and those excluded, including justification |
| Method of addressing articles published in languages other than English |
| Method of handling abstracts and unpublished studies |
| Description of any contact with authors |
| Reporting of methods should include |
| Description of relevance or appropriateness of studies assembled for assessing the hypothesis to be tested |
| Rationale for the selection and coding of data (e.g., sound clinical principles or convenience) |
| Documentation of how data were classified and coded (e.g., multiple raters, blinding, and interrater reliability) |
| Assessment of confounding (e.g., comparability of cases and controls in studies where appropriate) |
| Assessment of study quality, including blinding of quality assessors; stratification or regression on possible predictors of study results |
| Assessment of heterogeneity |
| Description of statistical methods (e.g., complete description of fixed or random effects models, justification of whether the chosen models account for predictors of study results, dose–response models, or cumulative meta-analysis) in sufficient detail to be replicated |
| Provision of appropriate tables and graphics |
| Reporting of results should include |
| Graphic summarizing individual study estimates and overall estimate |
| Table giving descriptive information for each study included |
| Results of sensitivity testing (e.g., subgroup analysis) |
| Indication of statistical uncertainty of findings |
| Reporting of discussion should include |
| Quantitative assessment of bias (e.g., publication bias) |
| Justification of exclusion (e.g., exclusion of non-English language citations) |
| Assessment of quality of included studies |
| Reporting of conclusions should include |
| Consideration of alternative explanations for observed results |
| Generalization of the conclusions (i.e., appropriate for the data presented and within the domain of the literature review) |
| Guidelines for future research |
| Disclosure of funding source |

number of eligible studies are available so that the results are reliable. For further details of this method, see Thompson and Higgins [1].

A typical example of use of meta-regression is in obtaining the estimates of prevalence rates of various diseases in different countries by combining similar studies in nearby areas. These estimates are used, for example, in burden of disease studies. For an example, see Kassebaum et al. [2] in which this method, along with others, has been used to estimate the prevalence, incidence, and trend over the years 1990–2010 of untreated caries in all countries of the world on the basis of 192 studies comprising 1,502,260 children aged 1–14 years in 74 countries, and 186 studies on a total of 3,265,546 individuals aged 5 year or older in 67 countries for untreated caries in deciduous and permanent teeth, respectively. Tsivgoulis et al. [3] used meta-regression for studying the effect of disease-modifying therapies on brain atrophy in patients with relapsing–remitting multiple sclerosis based on just four studies.


1. Thompson SG, Higgins JPT. How should meta-regression be undertaken and interpreted? *Stat Med* 2002 June 15;21(11);1559–73. http://www.ncbi.nlm.nih.gov/pubmed/12111920
2. Kassebaum NJ, Bernabé E, Dahiya M, Bhandari B, Murray CJ, Marcenes W. Global burden of untreated caries: A systematic review and meta-regression. *J Dent Res* 2015 May;94(5):650–8. http://www.ncbi.nlm.nih.gov/pubmed/25740856
3. Tsivgoulis G, Katsanos AH, Grigoriadis N, Hadjigeorgiou GM, Heliopoulos I, Kilidireas C, Voumvourakis K. The effect of disease modifying therapies on brain atrophy in patients with relapsing–remitting multiple sclerosis: A systematic review and meta-analysis. *PLoS One* 2015 Mar 10;10(3):e0116511. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0116511




## metric categories, *see* categories

## metric scale, *see* scales of measurement (statistical)

## misclassification

Misclassification occurs when an object is assigned to a category to which it does not belong. This includes misdiagnosis and missed diagnosis in a clinical setup, but the term is mostly used when a predictive **statistical model** wrongly tells you that the subject is in particular category A while it actually is non-A. You will know about misclassification only if you know the correct class; otherwise, it can go unnoticed.

Misclassification in our context primarily happens with two statistical methods, namely, the **logistic regression** and **discriminant analysis** as per the details provided for those topics. Briefly, in logistic regression, the category of a dependent binary variable, such as presence or absence of memory complaints in geriatric subjects, is sought to be predicted or identified on the basis of a set of regressor **variables**. These variables may be able to correctly identify presence of memory complaints in, say, 80% of the subjects with such complaints and correctly identify absence of complaints in 95% of subjects without these complaints. If so, the total misclassifications are not 25% but would depend on how many subjects were in each group. If the number of subjects in the two categories are the same, the total misclassification would be (20 + 5)/2 = 12.5%. See Table M.11 for unequal numbers where the category misclassifications are again 20% and 5% but the total misclassification is (40 + 3)/260 = 16.5%.

As evident from Table M.11, there are two types of misclassifications: one when absence is misclassified as presence and the other when presence is misclassified as absence. A model may be extremely good in correctly excluding the disease but poor in confirming its presence, and another model may be extremely good in confirming the presence but poor in correctly excluding it. Thus, these two types of misclassifications should be separately interpreted. In our example in the preceding paragraph, the model is good in correctly identifying absence of complaints but not so good in correctly identifying presence of complaints.

**TABLE M.11**
**Classification Table in Logistic Regression—Misclassifieds Are in Bold Italics**

| | Predicted | | |
|---|---|---|---|
| **Observed↓** | **Presence** | **Absence** | **Total** |
| Presence | 160 | ***40*** | **200** |
| Absence | ***3*** | 57 | **60** |

The setting is the same in discriminant analysis, but the exercise is for predicting the category of future observations on the basis of the known categories of the available observations. The classification table remains the same as in Table M.11 for two categories. Discriminant analysis can handle many categories simultaneously and would mostly require that the regressors are quantitative, while logistic does this by handling binary outcome at a time, though without restriction on the scale of the regressors.

## missing data, *see also* imputation for missing values and nonresponse

Missing data is a frequent problem in medical and health setup. In a clinical setup, decisions regarding diagnosis, treatment, and prognosis are taken despite incomplete history of the patients and incomplete reporting of complaints by the patients. Sometimes the required investigation cannot be done because of the absence of facility for that investigation, or because there is no time to wait. Sometimes the patient is in a coma or in aberrant condition, and cannot adequately describe his or her sufferings, while secondary reporting by the attendants may be incomplete, even misleading. Missing data in this setup compromise the quality of decisions. Except exercising care, probing, and reprobing, possibly nothing better can be done in this setup.

Our concern in statistics is primarily with missing values that occur due to nonavailability of information after a subject is included in a study. Suppose in a sample of 150 subjects from the population of eligible subjects with consent, some are not available at the appointed time and place, or fail to report at the clinic due to sudden illness or any other exigency in the family. Some may fail to recall the event of your interest. This problem is particularly acute in clinical trials where follow-up is required. The difficulty is that the nonrespondents are rarely random—thus, their exclusion can introduce bias. Some patients may withdraw after consent as they no longer wish to continue. This can happen when they find that it is too much of a hassle for them to be under trial or that the treatment to which they are allocated is not to their liking. Such missing data should be counted among the failures of the treatment and should not be neglected. Ignoring missing data may introduce bias in the results and may lead to invalid conclusions. Statistical methods such as **imputations** have limitations—they cannot compensate for missing data—only that they intend to provide best approximations. Careful planning and meticulous execution of the study are the only answer

to ensure that the missing values do not occur or occur with ignorable frequency. For more details, including aspects of trial design that limit the chance of missing data, see Singhal and Rana [1].

1. Singhal R, Rana R. Intricacy of missing data in clinical trials: Deterrence and management. *Int J Appl Basic Res* 2014;4(3):2–5. http://www.ijabmr.org/article.asp?issn=2229-516X;year=2014;volume=4;issue=3;spage=2;epage=5;aulast=Singhal, last accessed March 11, 2015.

## misuse of statistical tools, *see also* fallacies (statistical)

If a child cuts his/her finger with a sharp knife, should you blame the child or the knife? If a person takes an excessive dose of sleeping pills and dies, the pills cannot be blamed. Abuse or misuse of statistics has precisely the same consequence. An *abuse* occurs when the data or the results are presented in a distorted form with the intention to mislead. Sometimes part of the information is deliberately suppressed to support a particular hypothesis. A *misuse* occurs when the data or the results of analysis are unintentionally misinterpreted because of lack of comprehension. The fault in either case cannot be ascribed to statistics; it lies with the user. Correct statistics tell nothing but the truth. The difficulty, however, is that the adverse effects of wrong statistical methods are slow to surface, and this makes these methods even more vulnerable to wrong use. The other difficulty is that the learning curve of medical professionals for statistical methods, particularly for intricate methods, is shallow and daunting—thus, errors become inevitable. Few researchers, if any, are able to engage a qualified statistician from beginning to end. They tend to depend on their colleagues, their own experience, and software *Help* files. This support is incomplete, to say the least. At the same time, many biostatisticians also are not above the board. Besides biostatistical competence, their appreciation of medical implications is a severe constraint. The following are some instances of misuse of statistical tools. These include abuses also, although we call them misuses in the hope that these are not intentional.

### *Misuse of Percentages and Means*

Percentages can mislead if calculations are (i) based on small $n$ or (ii) based on an inappropriate total. If two patients out of five respond to a therapy, is it correct to say that the response rate is 40%? In another group of five patients, if three respond, the rate jumps to 60%. A difference of 20% looks like a substantial gain, but in fact, the difference is just one patient. This can always occur due to sampling fluctuation, and the percentage based on small $n$ can easily mislead. It is preferable to have $n \geq 100$ for valid percentages, but the following is our subjective guideline in this regard.

1. State only the number of subjects without percentages if $n < 30$.
2. For $n \geq 30$, percentages can be given, but $n$ should always be stated.

The second misuse of percentages occurs when they are calculated on the basis of an inappropriate group. The following example illustrates this kind of misuse.

Consider a group of 134 cases of heart bypass surgery who are followed for postsurgical complications. The data obtained are presented in Table M.12.

**TABLE M.12**
**Complications in Cases of Heart Bypass**

| Complication | Number of Cases | Wrong Percentage (Out of 44) | Correct Percentage (Out of 127) | |
|---|---|---|---|---|
| Excessive bleeding | 9 | 20.5 | 7.09 | |
| Chest wound infection | 15 | 34.1 | 11.81 | |
| Other infections | 8 | 18.2 | 6.30 | |
| Breathing problems | 10 | 22.7 | 7.87 | |
| Blood clot in the legs | 13 | 29.5 | 10.24 | |
| Others | 11 | 25.0 | 8.66 | |
| Any complication | 44 | 100.00 | 34.65 | |
| No significant complication | 83 | | 65.35 | |
| Total (data available) | 127 | | 100.00 | (94.78) |
| Data not available | 7 | | | (5.22) |
| **Grand total** | **134** | | | **(100.00)** |

Information was not available for seven patients. Since 83 did not experience any significant complication, it would be wrong to calculate the complication rate on the basis of the 44 patients who experienced complications. It unnecessarily magnifies the problem. The correct base for the complication rate is 127. The fact is not that 20.5% had excessive bleeding but that 7.09% of the patients had this problem. It would also be wrong to include seven patients in this calculation for whom the data are not available. For a nonresponse rate, however, the correct base is 134. This can be separately stated as shown in parentheses in the table. This example also illustrates the calculation of percentages in the case of multiple responses where a patient can have two or more complications. Thus, the percentages are not additive in this case.

For misuse of means, consider the popular saying by detractors of statistics: Head in an oven, feet in a freezer, and the person is comfortable, on average! There is no doubt that an inference on mean alone can sometimes be very misleading. It must always be accompanied by the standard deviation (SD) so that an indication is available about the dispersion of the values on which the mean is based. Sometimes the standard error (SE) is stated in place of the SD. This also may mislead unless its implications in the context are fully explained. If smoking prevalence in percent with the SE in male students of fifth grade is stated as $0.93 \pm 0.47$ [1], what kind of message do you get? Also, $n$ must always be stated when reporting a mean. These two, $n$ and the SD, should be considered together when drawing any conclusion based on mean. Statistical procedures such as confidence intervals and test of significance have a built-in provision to take care of both of them. A mean based on large $n$ naturally commands more confidence than the one based on small $n$. Similarly, a smaller SD makes the mean more meaningful.

You should also evaluate whether the mean is an appropriate indicator for a particular data set. Averages are not always what they seem. If in a group of 10 persons, 9 do not fall sick and 1 is sick for 40 days, how correct is it to say that the average duration of sickness in this group is 4 days per person? If extreme values or outliers are present, mean is not a proper measure—either use the median or recalculate the mean after excluding the outliers. If exclusion is done, this must be clearly stated. Else, consider if proportions are more adequate than mean as in our example of duration of sickness.

### *Misuse of Graphs*

As mentioned for **graphs and diagrams**, some misuse of graphs can occur by choosing an inappropriate scale. A steep slope can be

represented as mild and vice versa. Similarly, a wide scatter may be shown as compact. In addition, variation is sometimes shown as ±1 SD in graphs, whereas actually it is much more. Also, means in different groups or means over time can be shown without corresponding SDs. They can be shown to indicate a trend that really does not exist or is not statistically significant.

One of the main sources of misuse of graphs is their insensitivity to the size of $n$. A mean or a percentage based on $n = 2$ is represented the same way as the one based on $n = 100$. The perception, and possibly cognition, received from a graph is not much affected even when $n$ is explicitly stated. One such example is box-and-whiskers plots drawn by Koblin et al. [2] for the time elapsed between cancer and AIDS diagnoses among homosexual men with cancer diagnosed before or concurrently with AIDS in San Francisco during 1978–1990. Five lines (minimum, $Q_1$, median, $Q_3$, and maximum) are shown on the basis of only four cases of anal cancer. It was concluded on the basis of the figure that "Hodgkin's disease cases occurred relatively close to AIDS diagnoses." This may have merit but is stated without checking statistical significance and in disregard of the small number of available cases.

### *Misuse of* P-*Values*

Statistical ***P*-values** seem to be gaining acceptance as a gold standard for data-based conclusions. However, biological plausibility should not be abandoned in favor of *P*-values. Inferences based on *P*-values can also produce a biased or incorrect result.

A threshold of 0.05 (called the **level of significance**) for a Type I error is customary in health and medicine. Except for convention, there is no specific sanctity of this threshold. There is certainly no cause for obsession with this cutoff point. A result with $P = 0.051$ is statistically almost as significant as one with $P = 0.049$, yet the conclusion reached would be very different if $P = 0.05$ is strictly used as the threshold. Borderline values always need additional precaution.

A value close to the threshold such as $P = 0.06$ can be interpreted both ways. If the investigator is interested in showing the presence of difference, (s)he might argue that this $P$ approaches significance. If the investigator is not interested, this can be easily brushed aside as not indicating significance at $\alpha = 0.05$. It is for the reader to be on guard to check that the interpretation of such borderline *P*-values is based on objective consideration and not driven by bias. We generally interpret $P = 0.06$ or 0.07 as encouraging though not quite good enough to conclude a difference. They can be called marginally significant. If feasible, wait for some more data to come and retest the null in this situation.

The second problem with a threshold of 0.05 is that it is sometimes used without flexibility in the context of its usage. In some instances, as in the case of a potentially hazardous regimen, a more stringent control of Type I error may be needed. Then $\alpha = 0.02$, 0.01, or 0.001 may be more appropriate. It is not necessary to use $\alpha = 0.01$ if a value less than 0.05 is required, and the value $\alpha = 0.02$ can also be used. In some other instances, as in concluding the presence of differences in social characteristics of the subjects, a relaxed threshold $\alpha = 0.10$ may be appropriate. For most physiological and pathological conditions, however, the conventional $\alpha = 0.05$ works fine—that is why it has stayed as a standard for so long.

There are situations where a one-tailed test is appropriate, but a two-tailed test is unnecessarily used that makes the test conservative. In case of iron supplementation for increasing hemoglobin (Hb) level, biological knowledge and experience confirm that iron supplementation cannot reduce the Hb level. Use of the two-tailed test in this case makes it unnecessarily restrictive and makes rejection of $H_0$ more difficult. Most statistical packages provide two-tailed *P*-values as a default, and many workers would not worry too much about this aspect. Scientifically, a conservative test does not do much harm, although some real differences may not be detected when a two-tailed test is used instead of a one-tailed test. Our advice is to use a one-tailed test only where a clear indication is available for one-sided gain or loss, but not otherwise. In most medical situations, assertion of one-sided alternative is difficult and a two-sided test is needed.

See the topic **multiple comparisons** for adjustment of *P*-values while comparing several groups. In practice, you may have tried many statistical tests before reaching to the final ones you decide to report. Hardly anybody makes adjustment for such "behind-the-scene" statistical tests, although they also affect the final *P*-value. See the topic ***P*-values** for examples on how this is dramatized by some authors and how this is calculated for nonrandom samples where this is not applicable.

### *Misuse of Statistical Packages*

Computers have revolutionized the use of statistical methods for empirical inferences. Methods requiring complex calculations are now done in seconds. This is a definite boon when appropriately used but is a bane in the hands of nonexperts. Understanding of the statistical techniques has not kept pace with the spread of their use. This is particularly true for medical and health professionals. The danger arises from misuse of sophisticated statistical packages for intricate analysis without fully appreciating the underlying principles. The following are some of the common misuses of statistical packages.

Popularly termed as torturing until it confesses, data are sometimes overanalyzed, particularly in the form of post-hoc analysis. A study may be designed to investigate the relationship between two specific measurements, but correlations between pairs of a large number of other variables, which happen to be available, are calculated and examined. This is easy these days because of the availability of computers. If each correlation is tested for statistical significance at $\alpha = 0.05$, the total error rate increases enormously. Also, $\alpha = 0.05$ implies that 1 in 20 correlations can be concluded to be significant when actually it is not. If measurements on 16 variables are available, the total number of pairwise correlations is $16 \times 15/2 = 120$. At the error rate of 5%, 6 of these 120 can turn out to be falsely significant. Hofacker [3] has illustrated this problem with the help of randomly generated data. This underscores the need to consider any result from post-hoc analysis as indicative and not conclusive. Further study should be planned to confirm such results. This also applies to post-hoc analysis of various subgroups that were not part of the original plan, such as trying to find the age–sex or severity groups that benefited more from the treatment than the others. Numerous such analyses are sometimes done using a drop-down menu in the hope of finding some statistical significance somewhere. Again, such detailed analysis is fine for searching a ground to plan a further study but not for drawing a definitive conclusion. However, this is not to deny existence of Americas because it was not in Columbus' plan: evidence as hard as this must be given credence. In most empirical cases though, it may be sufficient to specifically acknowledge that the results are based on post-hoc analysis and possibly need to be confirmed.

1. Park SW, Kim JY, Lee SW, Park J, Yun YO, Lee WK. Estimation of smoking prevalence among adolescents in a community by design-based analysis. *J Prev Med Pub Health* 2006;39:317–24. http://jpmph.org/journal/view.php?year=2006&vol=39&page=317

2. Koblin BA, Hessol NA, Zauber AG, Taylor PE, Buchbinder SP, Katz MH, Stevens CE. Increased incidence of cancer among homosexual men, New York City and San Francisco, 1978–1990. *Am J Epidemiol* 1996;144:916–23. http://aje.oxfordjournals.org/content/144/10/916.full.pdf
3. Hofacker CF. Abuse of statistical packages: The case of the general linear model. *Am J Physiol* 1983;245:R299–302. http://ajpregu.physiology.org/content/245/3/R299

## mixed diagram, *see also* graphs and diagrams

A diagram is called mixed when two or more diagrams are shown together as one diagram. This is used when it is advisable to simultaneously show the variation in two or more characteristics, each measured on the same *x*-axis. This definitely saves space and helps to provide a more comprehensive view. Figure M.4 shows the age-wise prevalence of cataract blindness in a community and the percentage operated (surgical coverage). The former is shown by bars and the latter by a line. Thus, this is an example of a mixed diagram.

Although both measurements on *y*-axis in Figure M.4 are percentages, they are on a different scale. This itself makes reading difficult. Moreover, if the units are different, then different measurements on the *y*-axis can create more confusion. Prepare a mixed diagram with abundant precautions and interpret it with greater care.

## mixed effects models

This is an extension of the **analysis of variance (ANOVA)** method that includes random effects in addition to fixed effects. In case you are not familiar with **fixed and random effects**, see that topic. The conventional ANOVA considers that all the factors have fixed effect; that is, if you have studied three hospitals in ANOVA, the results would be valid for these three hospitals only and cannot be extended to any other hospital. When the studied hospitals are random samples from all the hospitals, they will be studied as factors with random effect, and then the results can be extended to other hospitals not in your sample. For mixed effects, there must be at least one factor with fixed effect and at least one factor with random effect. The random effects factor can be for **longitudinal data**, where some specified points in time are included in the analysis, but the results are sought for other time points as well. ANOVA would generally include only the factors with **nominal** categories, but mixed models can include continuous **covariates** (just as we do in **analysis of covariance**) also. The mixed models can accommodate both time-dependent covariates (e.g., dose levels that can vary from week to week) and static covariates (e.g., sex and age that would not change if the follow-up is a few weeks). The dependent in all of this is a quantitative variable. In short, the term *mixed models* is a synonym for **general linear models** where there is no restriction on the regressors.

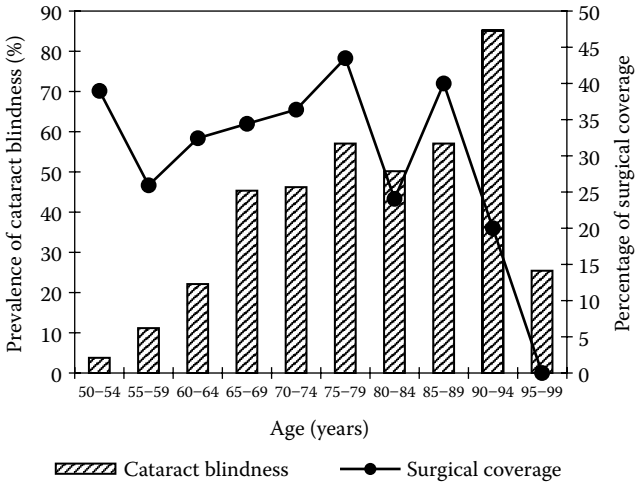



**FIGURE M.4** Mixed diagram: age-wise prevalence of cataract blindness (one or both <6/60) and the percentage of surgical coverage.

Consider a simple study on the effect of lifelong fully vegetarian and partially vegetarian diet on body mass index (BMI) of eighth grade students in a city. Suppose that, out of 42 such schools in the city, 4 are randomly selected for the study. Thus, there are two factors in this study—the type of diet (factor A) is fixed with two levels and school (factor B) is random with four levels. BMI is the dependent variable. The sum of squares and **mean squares** would be obtained as in the usual ANOVA, but the ***F*-test** changes for diet because of random effect of schools. If diet is indexed by $j$ ($j = 1, 2$), school by $k$ ($k = 1, 2, 3, 4$), and subject by $i$ ($i = 1, 2, \ldots, n$), the usual two-factor ANOVA model is

$$\text{fixed effects model: } y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk},$$

where $\mu$ is the overall mean, $\alpha_j$ is the main effect of the $j$th level of factor A, $\beta_k$ is the main effect of the $k$th level of factor B, $(\alpha\beta)_{jk}$ is the interaction term, and $\varepsilon_{ijk}$ is the error, which is the only random component in this model. The last is generally considered to follow a **Gaussian distribution** with mean zero and variance $\sigma^2$, written as $\varepsilon_{ijk} \sim N(0, \sigma^2)$. In a mixed effects model, if factor B is random, as are schools in our example, $\beta_k$ will be replaced by $b_k$ and this will also be random following, say, $N(0, \sigma_B^2)$ distribution. Thus,

$$\text{mixed effects model (factor A fixed, factor B random): } y_{ijk} = \mu + \alpha_j + b_k + (\alpha b)_{jk} + \varepsilon_{ijk}.$$

Because of this new formulation, the test of significance for fixed effect factor A would be $F = \text{MSA/MSAB}$, where MSA is the **mean square** due to factor A and MSAB is the mean square due to the interaction. This is different from what we do in the usual fixed effects ANOVA where each mean square is divided by mean square due to error (MSE). Tests for the random factor B and for interaction AB remain the same, $F = \text{MSB/MSE}$ and $F = \text{MSAB/MSE}$, respectively, as in the fixed effects ANOVA.

We have explained just about the simplest situation with a mixed effects model. But that should be adequate to clarify how mixed effects models are different from the usual fixed effects models and how they affect the statistical tests of significance in ANOVA. As mentioned earlier, this model is applicable to a variety of situations. For an application to longitudinal data analysis, see Kim et al. [1]. While working with statistical software, caution is required in specifying the model so that you get correct results. Our advice is that this should be done by only those who have sufficient expertise. Presence of random effects in a mixed effects model also involves issues such as **intraclass correlation** and **components of variance**. These are as explained separately. For further details of mixed effects models, see West et al. [2].

1. Kim JH, Park EC, Lee S. The impact of age differences in couples on depressive symptoms: Evidence from the Korean longitudinal study of aging (2006–2012). *BMC Psychiatr* 2015 Feb 5;15(1):10. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4330941/
2. West BT, Welch KB, Galecki AT. *Linear Mixed Models: A Practical Guide Using Statistical Software*, Second Edition. Chapman & Hall/CRC, 2014.

**mode,** *see* **mean, median, mode (calculation of),** *see also* **bimodal distribution**

## models (statistical)

A model, by definition, is a simple version of a complex process. By its very nature, it is imperfect, but the attempt is to be able to capture the essentials of the process so that it can be understood, explained, or predicted with reasonable accuracy. Statistical models aim to do the same by means of equations with a dependent or the response **variable** on one side and the independents or the explanatory variables on the other side. The objective is again to explain or predict the response or outcome on the basis of some ancillary information. The explanation or prediction is never perfect but can have high degree of **validity** and **reliability**. Popular examples of statistical models are ordinary **regression models**, **logistic models**, and **Cox models**, but there are other models such as **infectious disease models**, **pharmacokinetic** models, **time-series** models, and **stochastic** models. All these models are also referred to as mathematical models in the literature, though hardly any model in medicine would be strictly mathematical with no probability component.

You may have guessed from the preceding paragraph that statistical models are etiologically of two types (we will come to the third type a little later): the predictive and the explanatory. The distinction is thin but it is there, at least in concept. There is a considerable overlap in the usage of these terms in the medical literature. The nature of both the models is the same, but the objective is different. Statistical methods for arriving at and evaluating the utility of both are the same.



In a **predictive model**, the focus is on correctly reaching at the outcome irrespective of the predictor set. Two different sets of predictors can be equally good in predicting the outcome. You may have a model to predict hospital survival of a critical patient on the basis of APACHE score at the time of admission, and another model to predict survival on the basis of age, hemoglobin level, and the care he/she gets. The choice of the predictor set is not important, and the model is considered good as long as it is able to predict the correct outcome in a large percentage of cases. There is no issue of how the predictor set and the outcome are biologically related, or not related.

In the **explanatory model**, on the other hand, the choice of regressors is important as the objective is to understand the mechanism of how these regressors are leading to the outcome. All those variables that can biologically affect the outcome are considered, and the significant ones are retained. van Oostveen et al. [1] investigated cost of care in hospitalized surgical patients in a Dutch university hospital and found through linear regression analysis that medication during hospitalization, complications, comorbidity, medical specialty, age, undergoing surgery, and length of stay are the significant factors that explain the cost of care. If the outcome is the incidence of swine flu, the explanatory variables would be different for bed planning than for stopping the epidemic, despite the fact that the outcome in both the models would be incidence of swine flu.

Both predictive and explanatory models are associational in nature and cannot be interpreted to describe the **cause–effect relationship**. As explained in that topic, cause–effect relationship is much more serious and requires special investigations. In the context of models, as for clinical trials, a **control group** is of tremendous help to infer cause–effect relationship. Models based on case–control studies where controls are matched with the cases except for the purported cause can provide a convincing evidence of a factor being the cause and not merely a correlate. This is the third etiological type of the statistical models referred to earlier.

There is always a question about evaluating the performance of a model. Most researchers would check it on the same dataset that is used to develop the model. Remember that the performance on the same dataset (called internal validity) could be extremely good since the model is based on that dataset. At best, this provides evidence that the model is not wrong. If it is not good on this dataset, the model should be immediately discarded after learning the lessons for such exercise. But if it is good on that dataset, the question still remains if the model is good for other subjects. Thus, external validation is required. Statistically, it is advised to test the model on another sample. Steps such as splitting the available sample into two parts—use one for developing the model and the other to test it—are suggested. Sometimes **simulations** are advised, but they rarely serve as a good approximation of reality. In any case, the work is not fully done unless the model has been put to real-life test.

1. van Oostveen CJ, Vermeulen H, Gouma DJ, Bakker PJ, Ubbink DT. Explaining the amount of care needed by hospitalised surgical patients: A prospective time and motion study. *BMC Health Serv Res* 2013 Feb 4;13:42. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3599528/

## monotonic relationship

A monotonic relationship is said to exist between two variables when the value of one variable increases (decreases) or remains the same as the value of the other increases (decreases). Thus, values of both the variables change in the same direction, but the relationship continues to be called monotonic even if one remains stationary. The probability of death after the age of 5 years is monotonic despite being nearly constant between the ages 5 and 40 years. However, if the age less than 5 years is also included, the relationship in many populations is not monotonic between the age 0 and 100 years since the probability of death is generally high in infant period, declines till the age 5 years, remains constant for some time, and rises thereafter. The shape of this relationship is as shown in the figure in the topic **bathtub curve**. However, you can say that this relationship is monotonically decreasing from the age 0 to 5 years and monotonically increasing from the age 5 to 100 years. Figure M.5 illustrates another kind of monotonic relationship where the value of one variable increases with the other, remains the same for some time, and then rises again.

The knowledge of whether a relationship is monotonic or not can be helpful for modeling. For example, a **linear** relationship with nonzero slope is always monotonic. In fact, it is what is called strictly monotonic since there are no constant or stationary values in linear relationship when the slope is not zero. The relationship of the type

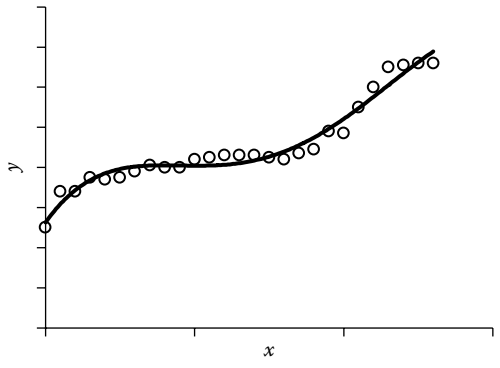



**FIGURE M.5** Monotonic relationship.

shown in Figure M.5 may require a fourth degree polynomial—it is **curvilinear**. Secondly, monotonic relationship also helps in investigating the biological reasons for stationarity if that happens. If the relationship is of the type shown in Figure M.5, it would be helpful to know why for some middling values of $x$, the variable $y$ does not show an increase, while it shows so for small and large values of $x$.

For examples of the use of monotonic relationships in medicine, see Ruckart et al. [1] who observed a monotonic relationship between benzene exposure during the entire pregnancy and term low birth weight in North Carolina. Kirli et al. [2] found a monotonic relationship between N-methyl-D-aspartate (NMDA) conductance onto the pyramidal cells with network gamma in schizophrenia, whereas this relationship was inverted U (not monotonic) with conductance onto fast-spiking interneurons.


1. Ruckart PZ, Bove FJ, Maslia M. Evaluation of contaminated drinking water and preterm birth, small for gestational age, and birth weight at Marine Corps Base Camp Lejeune, North Carolina: A cross-sectional study. *Environ Health* 2014 Nov 20;13:99. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4247681/
2. Kirli KK, Ermentrout GB, Cho RY. Computational study of NMDA conductance and cortical oscillations in schizophrenia. *Front Comput Neurosci* 2014 Oct 17;8:133. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4201161/


## Monte Carlo methods

Monte Carlo methods are algorithms that use deliberately generated random observations to derive conclusions. Originally, the method was developed and used for those data that are difficult to study through mathematical formulations, but now it is used in a variety of setups such as for assessing the effect of violation of the requirements of a statistical model. Eckhardt [1] explains that the method was secretly developed by von Neumann and Ulam in 1947 and codenamed Monte Carlo.

Suppose you suspect that a coin is biased and has a tendency to show tails more often than heads. If you toss it 1000 times and note how many times it shows tails for assessing bias, this would be a form of a Monte Carlo experiment. If you are able to analyze the same blood sample 1000 times for, say, blood sugar level using a machine, you will have a good idea of the reliability of the machine. You can find how many times the value found by the machine is within 1% of the mean. Without this, it is difficult to assess the reliability of a machine. Such opportunity to generate a large number of observations is rare in practice, particularly in statistical modeling, and computer **simulations** are done to circumvent the problem. This requires generating a large number of observations by a repetitive process through computer algorithm. Simulation is now the most common tool used in Monte Carlo methods. This can be used, for example, to generate 10,000 random observations from any model, examine the pattern, and draw conclusions.

For an illustration of Monte Carlo simulations in biostatistical applications, consider a group of stroke patients who either survived with full health, survived with disabilities, or died. Suppose the interest is in finding whether the type of outcome has anything to do or not with blood groups O, A, B, and AB. The data will give **cell frequencies** in a $3 \times 4$ table for which an established test for association is **chi-square**. But if the total sample is small, say, just 40 subjects, many cell frequencies would be less than 5, and chi-square cannot be used. One way out is to use the Monte Carlo method and generate thousands of $3 \times 4$ tables with same row and column totals, and find the percentage of tables that give cell frequencies equal to the observed or more tilted toward the **association**. This will give the ***P*-value** required to decide statistical significance of the association between the type of outcome and the blood group pattern in stroke patients.

Pérez-Pitarch et al. [2] have described a study based on a Monte Carlo–generated sample of 2000 treatment-naive patients of ulcerative colitis based on a validated population pharmacokinetic model. Six dosing strategies for maintenance therapy were simulated on this population. Strategy of fixed dose of 5 mg/kg and individualized interdose intervals proved most effective. Zhang et al. [3] evaluated the efficiency of skull optical clearing solution-induced method with Monte Carlo simulations based on the measurements of divergence of beam spot and collimated transmittance of skull. Thus, the method has wide applicability in medical setups.


1. Eckhardt R. Stan Ulam, John von Neumann, and the Monte Carlo method. *Los Alamos Science, Special Issue* 1987;15:131–7. http://library.lanl.gov/cgi-bin/getfile?15-13.pdf
2. Pérez-Pitarch A, Ferriols-Lisart R, Alós-Almiñana M, Mínguez-Pérez M. A pharmacokinetic approach to model-guided design of infliximab schedules in ulcerative colitis patients. *Rev Esp Enferm Dig* 2015 Mar;107(3):137–42. http://www.grupoaran.com/mrmUpdate/lecturaPDFfromXML.asp?IdArt=4621125&TO=RVN&Eng=1, last accessed June 12, 2015.
3. Zhang Y, Zhang C, Zhong X, Zhu D. Quantitative evaluation of SOCS-induced optical clearing efficiency of skull. *Quant Imaging Med Surg* 2015 Feb;5(1):136–42. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4312305/


## MOOSE guidelines, *see* meta-analysis of observational studies in epidemiology (MOOSE) guidelines



## morbidity indicators, *see also* incidence, prevalence, and prevalence rates

These are those indicators that assess the extent and magnitude of morbidity. Morbidity indicators are separate for individuals compared to those for community. Individual level indicators are duration of disease and severity (see **disability weight**), whereas community level indicators are **incidence**, **prevalence**, average duration of disease, and severity distribution. Severity distribution tells us what percentage of affected subjects have mild, moderate, serious, or critical condition.

Incidence and prevalence are generally computed separately for each disease and health condition, and for each age group and sex, using tools such as the International Classification of Diseases. Duration of disease can be combined for all conditions together and can be calculated as disability days per year of life at different ages separately for each sex, or as total disability days during the life so far. When worked out for the entire life for all persons in a population, and adjusted by disability weight, this can give the **disability-adjusted life-years** lost per 1000 population during the year under study.

Morbidity rates across populations are seldom comparable because of several variations. First, the definition of sickness can vary from population to population and from age to age. For working population and school students, absenteeism due to sickness can be a useful indicator. But many people work or go to school even when mild sickness is present. This raises the question of whether or not conditions such as cold and cough should be counted. Similarly, mild muscle spasm or cramp and mild injury can be classified either way. Second, no globally acceptable scale is available to categorize disease as mild, moderate, serious, or critical. This varies from one condition to another and from population to population. It may also

depend on what part is affected and also the mood of the person at the time this assessment is being made. In addition, what is mild for one may be serious for someone else.

Individual level morbidity can come from clinics presuming that nobody is taking home remedy. That is a big if, though. Surveys are done to assess morbidity, but they too can mislead since the reporting is based on perception of the affected person rather than the actual physical examination. Then there are issues with failure to recall since mild conditions tend to be ignored. Old-age persons may not report vision and breathing problems considering that these are normal at their age. On the other hand, some old-age persons can feel harassed by a small problem and magnify its reporting. Such aberrations should be kept in mind when interpreting morbidity survey data.

Morbidity data are rarely complete for the entire population of any country and have to be estimated. The estimation method can become complex because of poorly understood geographic and socioeconomic differentials that can be substantial in addition to the known age–sex differentials. For the purpose of planning, projections can be done by synthesizing data for different individual years. For one such exercise for coronary diseases and diabetes in India, see Indrayan [1]. For global and regional estimates of morbidities with uniform methodology, see WHO [2].


1. Indrayan A. Forecasting vascular disease cases and associated mortality in India. Background Papers: *Burden of Disease in India, National Commission on Macroeconomics and Health*, Government of India, 2005:197–215. http://www.searo.who.int/india/topics/cardiovascular_diseases/Commision_on_Macroeconomic_and_Health_Bg_P2_Forecasting_vascular_disease_cases_and_associated_mortality_in_India.pdf
2. WHO. *Health Statistics and Information Systems, Estimates for 2000–2012: Burden of Disease*. http://www.who.int/healthinfo/global_burden_disease/estimates/en/index2.html




## mortality rates, *see also* death rates

For overall deaths, see the topic **death rates**. The present section is devoted to the mortality indicators for specific age groups. Among these, fetal deaths are discussed separately as a part of **abortion rate/ratio** and **still birth rate/ratio**. We begin this section with perinatal deaths that include late fetal and early neonatal period (Figure M.6) and move up to the adults. **Maternal mortality ratio** in adult women is also discussed separately.

### *Perinatal and Neonatal Mortality*

Late fetal deaths are those that occur after 28 weeks of gestation. Early neonatal deaths are those occurring during the first week of life (<7 days). With these, the following can be defined:

$$\text{perinatal mortality rate (PMR)} = \frac{\text{late fetal deaths} + \text{early neonatal deaths}}{\text{late fetal deaths} + \text{live births}} * 1000$$

$$\text{perinatal mortality ratio (PM ratio)} = \frac{\text{late fetal deaths} + \text{early neonatal deaths}}{\text{live births}} * 1000.$$

A more precise definition considers births weighing at least 1000 g because those weighing less have poor prognosis anyway. This gives

$$\text{PM ratio} = \frac{\text{late fetal deaths} + \text{early neonatal deaths weighing} \geq 1000\ \text{g}}{\text{live births weighing} \geq 1000\ \text{g}} * 1000.$$

Even though the frequency of occurrence is the essential feature of a rate, note that the preceding formulas use the term **rate** when the numerator is part of the denominator; otherwise, it is considered a **ratio**. This distinction is not universal though. These are generally calculated on annual basis for each calendar year.

The measures of mortality after a live birth has taken place are the following:

$$\text{neonatal mortality rate (NMR)} = \frac{\text{neonatal deaths}}{\text{live births}} * 1000$$

$$\text{postneonatal mortality rate} = \frac{\text{postneonatal deaths}}{\text{live births}} * 1000.$$

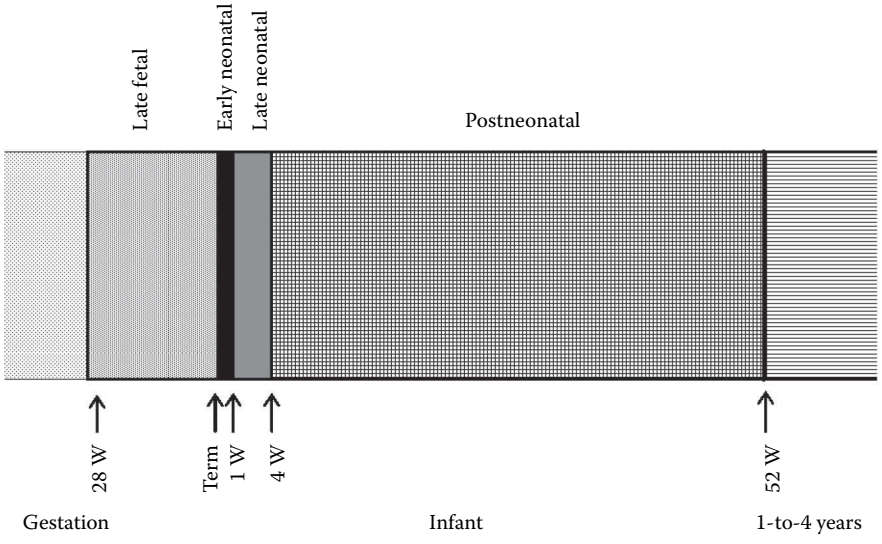



**FIGURE M.6** Childhood age groups important for study of mortality in a community.

The neonatal period is the first 27 days of life, and the postneonatal period is 28–364 days. In the case of rate in the formula just given, the postneonatal deaths can be calculated out of those who survive the neonatal period. This could be substantially different in areas where neonatal deaths are high.

Neonatal mortality is generally determined by the health of the mother and the adequacy of services available at the time of birth. The main causes of neonatal mortality are prematurity and birth asphyxia. Postneonatal mortality is mostly due to infections and undernourishment since birth. International efforts have succeeded in controlling much of the postneonatal mortality, but the neonatal mortality is slow to respond to these efforts. Emphasis is now given to maternal nutrition and skilled attendance at birth in developing countries to control neonatal mortality, and that is paying dividends.

### *Child Mortality*

Neonatal and postneonatal periods together form the infantile period. Note that an infant is a child *less* than 1 year of age, and a neonate is *less* than 4 weeks of age, which is the same in completed days as just stated. The sum of the neonatal and postneonatal mortality is the infant mortality when their denominator is the same live births:

$$\text{infant mortality rate (IMR)} = \frac{\text{deaths of infants}}{\text{live births}} * 1000.$$

A popular mortality rate for children is the under-5 mortality rate (U5MR). It is calculated as

$$\text{under-5 mortality rate (U5MR)} = \frac{\text{deaths of children} < 5 \text{ years}}{\text{live births}} * 1000.$$

The U5MR is also called the child mortality rate, although this term is sometimes used for mortality in the 1–4 years age group. These rates are different from the respective **age-specific death rates** because the denominator in these mortality rates is the number of live births and not the population of that age group. It is customary to use the term *mortality rate* when the denominator is live births and the term *death rate* when the denominator is population.

The mortality count used in the numerator in these formulas is not necessarily out of the live births in that year. A child born in the month of October may die in the month of January in the next calendar year. The birth and the death of the same child, thus, would be counted in different years. Despite this anomaly, these are called rates and the terms are retained as such because of simplicity in counting. The births are not required to be followed up to count the deaths for these rates. The effect of this anomaly on the rate is minimal because of the averaging-out phenomenon.

Some professionals modify the denominators for calculating some of these rates. For example, for calculating postneonatal mortality rate, they use (live births – neonatal deaths) as the denominator. Although this can be justified, such a modification introduces complexity and interferes with the feature of additivity of some of these rates. For this reason, we advocate simple denominators as stated.

All these childhood mortality rates are considered very sensitive indicators of population health in the global context. Attention is particularly paid to IMR because this is relatively easily understood and is now almost universally available. A large number of background and proximate factors such as education, affluence, nutrition, availability and utilization of health care facilities, and cultural practices affect IMR. This rate is affected rather quickly by health programs in developing countries and so is also used to measure the effectiveness of such programs.

### *Adult Mortality*

Child mortality is a huge concern in some populations, but those populations who have low child mortality shift their focus to adult mortality. Since geriatric mortality does not contribute much to the study of adverse health, adult mortality is concerned with deaths between the age of 15 and 59 years and ignores deaths at age 60 years and beyond. Thus,

$$\text{adult mortality rate} = \text{probability of death between 15 and 59 years per 1000 population of this age,}$$

and can be obtained through a **life table**. Adult mortality rate arouses passion as it concerns the most productive segment of the population and is supposed to be the healthiest period of life. Adult mortality is generally higher in males than females. This differential is attributed to greater exposure of males to hazards at work, stress, and strain, and also to their biological vulnerability.

See the topic **maternal mortality ratio** also for another aspect of adult mortality.

## moving averages

Moving averages is a smoothing mechanism for fluctuating serial data—mostly **time series** data. This is done to extract trend in the series. If there is a series of, say, 50 years from 1965 to 2015 of incidence rate of cataract for a population of age 60 years and above, which hardly fluctuates somewhat from year to year, moving averages, say for each consecutive 5 years, can give a somewhat smooth trend. These would be the average of incidence rates in the years 1965–1969, 1966–1970, 1967–1971, and so on, up to 2011–2015. These moving averages would take away much of the fluctuations but would still preserve the trend, including any persistent slowdown or reversals. Such averages are then assumed to be for the midpoint of the periods successively averaged. In our example, the first moving average will be considered to be for the year 1967, the second for the year 1968, and the last for the year 2013. Thus, initial and last few time points will not have any moving average. This loss is not much if you have a long series.

The method of moving averages is evidently applicable to only large series, and it requires a fair assessment of the period for averaging every time. The two, in fact, are related since the period cannot be large for relatively small series. For example, for only 50 data points as in our cataract example, the period cannot be 20 years, but for 200 data points, a 20-point moving average can be tried. This periodicity also depends on whether the objective is to find a short-term trend or to assess a long-term trend. Depending on how many time points are available, a 5-time point moving average will give a short-term trend, whereas a 50-time point moving average will give a long-term trend. A 10- or 20-time point moving average may provide a medium-term trend.

In a medical setup, suppose you are monitoring every 5 min the systolic blood pressure (BP) level of a patient in critical care. A short-term moving average may be good to assess the effect of each dose of drug, whereas a long-term trend would assess the improvement of the patient. Figure M.7 illustrates 30-min (6 time points at 5-min interval) and 120-min (24 time points at 5-min interval) moving averages. A short-term trend based on 6-point moving averages

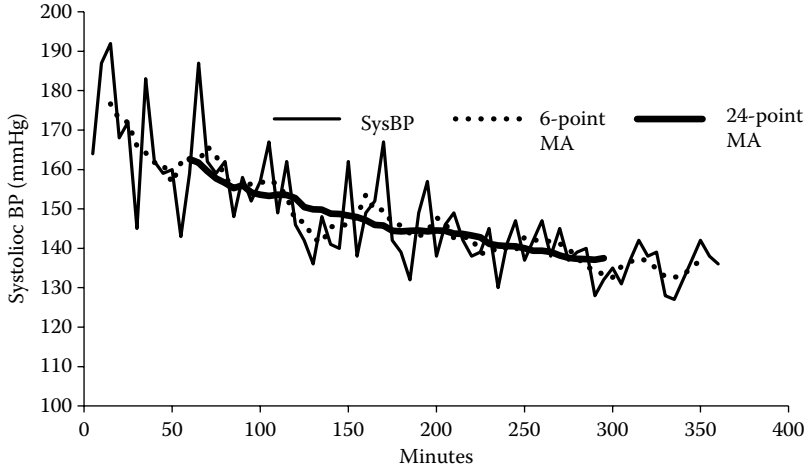



**FIGURE M.7** Systolic BP in a critical patient at 5-min intervals and moving averages (MAs).

is still wavering, but a 24-point trend is solid evidence of declining systolic BP. The last of course is not available for the first and last 12 time points (first and last 1 h). This may not be important too for the patient management.

If the periodicity is known, such as seasonal disease each year, you will be better off using the 12-month moving averages on monthly data. If a high afternoon temperature is expected each day for some patient, a 24-h moving average of hourly (or 6-point four-hourly) temperature could provide the trend of clinical importance. Bedside monitors can be programmed to provide trends based on such moving averages. If two or more series are available such as systolic and diastolic BP that move in the same direction, but are observed to move in reverse directions, such as one moving downward and the other moving upward after reaching a particular threshold, one can provide useful clues of what possibly is going on, say, by way of body reactions, and set alarm for doing something to arrest this conflicting trend.

Having provided clinical examples, let us say that such applications of moving averages in hospitals are not widespread yet. The method is mostly used for time series data on incidence and prevalence rates. In this case, it might be worthwhile also to plot the **confidence intervals** for each moving average. This plot will give you a confidence band.

## MSE, *see* mean squares in ANOVA

## multiarm trials

This is a trial that seeks to compare several regimens in one go. Conventional trials have one group that receives the test regimen and one more group that receives the control. Multiarm trials will have two or more regimens besides the control. These regimens could be different doses of the drug, different modes of administration of the same drug, or different regimens altogether such as comparing laparoscopic and open tension-free inguinal hernia repair with Shouldice operation [1].

Because of the high pace of development in medicine, sometimes many promising competing regimens are simultaneously ready for testing. Multiarm trials provide the opportunity to test all of them together. Otherwise also, multiarm trials are always cheaper than the conventional case–control trials because now at least the control group is shared. Administrative logistics are also shared, and the requirements of the subjects are also fewer. The latter happens because the same subjects on, say, regimen 1 are used to compare with regimen 2 and the control. You can see how the same trial is used for multiple comparisons, obviating the need to conduct several independent trials. Subjects may be relatively easy to recruit as a wide spectrum of treatment choices is available in this format. Thus, multiarm trials are efficient and cheaper, save time and resources, and are recommended for setups where several regimens are to be tested.

There is a flip side too. Multiarm trials are difficult to plan and execute as many regimens are involved, and there is a risk of overlapping of concepts. These are relatively bigger than the conventional two-arm trials, and it may not be easy to find so many subjects at the same time. They also involve issues of **multiple comparisons** and statistical control of **Type I error**. Real problem arises when the main outcome of interest varies from one regimen to another. For example, it could be recovery for one regimen and arrest of the disease for the other. Even if it is the same as recovery for all the regimens, the ancillaries such as speed of relief, cost of treatment, and adverse side effects may differ, and the comparison becomes complicated. The researcher must anticipate these variations and decide beforehand on how these would be reconciled to come to a unified conclusion.

There are not many examples of multiarm trials, although their design has been discussed. Royston et al. [2] described designs for comparing four new chemotherapy regimens for advanced ovarian cancer. Magirr et al. [3] presented flexible **sequential designs** for multiarm trials, and Wason et al. [4] discussed Type I error correction for multiple testing in such trials.

1. Zieren J, Zieren HU, Jacobi CA, Wenger FA, Müller JM. Prospective randomized study comparing laparoscopic and open tension-free inguinal hernia repair with Shouldice's operation. *Am J Surg* 1998 Apr;175(4):330–3. http://www.americanjournalofsurgery.com/article/S0002-9610(98)00004-X/abstract
2. Royston P, Parmar MKB, Qian W. Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Stat Med* 2003;22:2239–56. http://onlinelibrary.wiley.com/doi/10.1002/sim.1430/abstract;jsessionid=ABFB7B00A87089A5508D61DCC0BD5778.f03t01
3. Magirr D, Stallard N, Jaki T. Flexible sequential designs for multi-arm clinical trials. *Stat Med* 2014 Aug 30;33(19):3269–79. http://onlinelibrary.wiley.com/doi/10.1002/sim.6183/full
4. Wason JM, Stecher L, Mander AP. Correcting for multiple-testing in multi-arm trials: Is it necessary and is it done? *Trials* 2014 Sep 17;15:364. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4177585/

## multicentric trials

As the name implies, these are **clinical trials** that are simultaneously conducted at several centers following a common **protocol**. The goals of these trials are (i) to get more cases as one center may not have sufficient cases, (ii) to have a better cross-section of cases as different centers can have different patient profiles, (iii) to study the impact on efficacy and side effects of the varying environments prevailing in different centers, and (iv) to have a more realistic assessment of the reliability of results not just because of more accrual of cases but also by checking that different centers provide consistent results despite each being based on a relatively small sample. All of these can help to be more confident of the results and can also help in the better generalizability of these results.

Multicentric trials pose tough challenges. The most prominent among them is to devise a common protocol—different centers may have different opinions and different problems. Protocol development for a multicentric trial is a long and arduous task, and requires extensive consultation with investigators at different centers. The collaborators must have a flexible attitude and should come to a consensus to realize a common goal. The second challenge is in executing the protocol uniformly. Some centers may experience unforeseen problems that require deviation from the agreed protocol, and the investigators at some centers may want deviation after initially agreeing to participate. Interest of the patients is commonly cited to justify such deviations. Some of these deviations can be serious so as to contaminate the results. Another challenge is in arranging huge funds that a multicentric trial needs. The fourth challenge is in the analysis of data. The data from different centers are sent to one place where they are scrutinized again for inconsistent entries even if they have already been checked at each center. Pooling can be done only after ensuring that the data from different centers are not heterogeneous. This means that center-wise data are analyzed separately and checked for consistency in results for all the centers. The method of stratified analysis such as the Mantel–Haenszel chi-square may be used for analyzing data from multicentric trials. If the results differ despite sticking to the uniform protocol, perhaps a new hypothesis can be forwarded to explain the differences.

Although varying conditions are an essential ingredient of multicentric trials, they tend to exacerbate due to varying interpretations of the protocol requirements. The difficulty arises when such variations are not acknowledged and go on inadvertently, finally damaging the comparability. Thus, these trials need periodic review—much more than single-center trials do. Reviews such as those by the **Data Safety and Monitoring Board** help in quality assurance and uniformity across sites. Complex regimens that require intensive training may not be appropriate for multicentric trials.

Despite all the problems just stated, multicentric trials are commonly conducted in view of their positive features. Chen et al. [1] have reported a multicentric trial conducted in the United States on inactivated monovalent influenza virus vaccine. Du Toit et al. [2] have described a multicentric trial in the United Kingdom on different strategies of peanut consumption and avoidance to determine which strategy is most effective in preventing the development of peanut allergy in infants at high risk for the allergy.

1. Chen WH, Jackson LA, Edwards KM, Keitel WA, Hill H, Noah DL, Creech CB, Patel SM, Mangal B, Kotloff KL. Safety, reactogenicity, and immunogenicity of inactivated monovalent influenza A(H5N1) virus vaccine administered with or without AS03 adjuvant. *Open Forum Infect Dis* 2014 Oct 8;1(3):ofu091. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4324222/

2. Du Toit G, Roberts G, Sayre PH, Bahnson HT, Radulovic S, Santos AF, Brough HA et al. Randomized trial of peanut consumption in infants at risk for peanut allergy. *N Engl J Med* 2015 Feb 26;372(9):803–13. http://www.nejm.org/doi/full/10.1056/NEJMoa1414850

## multicollinearity

Multicollinearity among a set of variables is said to exist when some of them are highly correlated with each other. Some statistical methods such as ordinary and logistic **regressions** can give weird results when such collinearity is present among the **regressors**. The estimates of the **regression coefficients** are still valid, but they are not reliable. The **standard errors** of the estimates will be unduly large in case of multicollinearity, indicating that the estimates are unstable. Consequently, the **confidence intervals** and the **tests of significance** will not be valid. But multicollinearity can be ignored in some situations. We will come to the solution, but first, one needs to know how to assess whether multicollinearity really exists. The common methods used for detecting multicollinearity are as follows:

- Examine the biological relationship among the regressors. For example, it is well known that systolic and diastolic blood pressures have strong **correlation**. Similarly, hemoglobin level and mean corpuscular hemoglobin concentration are highly correlated, so are skinfold thickness and waist–hip ratio.
- Obtain the correlation coefficient among all pairs of the regressors and check if one or more correlations are high, say >0.7. Whereas high correlation is a good indication of presence of multicollinearity, small correlations are no guarantee that multicollinearity does not exist.
- Examine the **variance inflation factor** (VIF) of each suspected variable. This may be calculated for any regressor by doing a linear regression of that variable on all the other regressors, and then obtaining the **multiple correlation** coefficient $R^2$ from that regression. The VIF = $1/(1 - R^2)$. High VIF (say, more than 2.5) indicates presence of multicollinearity since a variable then can be reasonably predicted by others.
- Statistical significance of the overall ***F*-test** but no significance of any of the individual **regression coefficients** also indicates multicollinearity.
- Nonsignificance of a regression coefficient of a variable that you know on theoretical grounds should contribute significantly is also an indication that multicollinearity exists.
- Drop the variable that is suspected to be collinear with one or more of the others from the regression model and see if the regression coefficients change drastically. If this happens, there is evidence in favor of the suspicion.
- A regression coefficient is found negative when theoretically the dependent should increase with increasing values of that regressor or vice versa. That is, if a variable is known to have positive contribution to the response yet the regression gives negative contribution, suspect multicollinearity.
- If there are many similar datasets or splits of the same dataset, and if the estimates of the regression coefficients substantially vary in models obtained by different datasets, suspect multicollinearity.

Multicollinearity among covariates that are there in the regression just for control and not of direct interest to the research objective does

not matter much. Similarly, high correlation due to one variable being the power or the product of the others (such as $x_3$ as $x_1^2$, and $x_4$ and $x_{1*}x_2$) should also not cause worry. Multicollinearity among indicator variables for categories that add up to 100% also is natural since the increase in the percentage of one category will automatically decrease the percentage in the other categories. However, if it is there among variables of primary interest, a solution must be found and implemented.

The easiest solution is to drop one variable in a pair of variables that are highly correlated; you should use your judgment to decide which one should be dropped. Do not drop the variable that you know has biological justification to be considered for inclusion in the regression. The second solution is to increase the sample size. As mentioned earlier, multicollinearity affects reliability and not validity, and reliability can be taken care of by increasing the sample size. The third solution is to try to combine correlated variables into an **index** when this index can be given a biological meaning. The fourth solution is what is called centering by subtracting the mean. Centering of the offending variables may reduce multicollinearity, but the interpretation of the corresponding regression coefficients changes. When centering is done for all the variables, the intercept becomes zero. There are other complex methods based on interactions and variances, but we leave them out from our purview. If you are interested, consult the book by Mendenhall and Sincich [1].

1. Mendenhall W, Sincich TT. *A Second Course in Statistics: Regression Analysis*, Seventh Edition. Pearson, 2011.

## multilevel models/regression, *see* hierarchical models/regression

M

## multinomial distribution/test

Multinomial distribution is an extended version of the binomial distribution that is used for multiple categories instead of two categories. The topic **binomial distribution** gives details of the situation where the variable is observed in just two categories such as death/survival, pain/no pain, male/female, agree/disagree, and test positive/negative. Now suppose it is not just death/survival but also survival with disability, making three categories, namely, death, survival with disability, and survival without disability. A disease severity can be divided into categories none, mild, moderate, serious, and critical, and a person can belong to any one of these categories and nothing else. These categories are **polytomous** and **mutually exclusive and exhaustive**. Multinomial distribution is built up precisely for this setup. This distribution gives the probability of $O_k$ subjects falling into the $k$th ($k = 1, 2, \ldots, K$) category whose probability is $\pi_k$. The probability in this distribution can be calculated as

$$\text{Multinomial distribution: } P = \frac{n!}{O_1!O_2!\cdots O_K!} * \pi_1^{O_1}\pi_2^{O_2}\cdots\pi_K^{O_K};$$

$$\Sigma O_k = n \quad \text{and} \quad \Sigma\pi_k = 1.$$

If the probability of survival without disability after stroke is $\pi_1 = 0.50$, that of survival with disability is $\pi_2 = 0.30$, and that of death is $\pi_3 = 0.20$, the probability that out of 8 such patients, 6 will survive, 1 will die, and 1 will survive with disability with this formula is $P = 8!/(6!\times1!\times1!)\times(0.5)^6\times(0.3)^1\times(0.2)^1 = 0.0525$. Thus, there is nearly a 5% chance that the outcome of 8 stroke patients will be in this pattern. Such probabilities can be used to build a statistical test of significance for polytomous data.

### *Multinomial Test*

The major application of multinomial distribution is in testing of the hypothesis for multiple categories. This is approximated by chi-square for large $n$, but for small $n$, exact probability based on multinomial distribution as just mentioned should be obtained. For this, a null hypothesis $H_0$ is set for the category probabilities, and the plausible alternative hypothesis $H_1$ comprises all the probabilities other than the ones specified under the null but favoring the alternative. The $P$-value is obtained after summation of $P$ just mentioned over the configurations of the frequencies that are as observed in the sample or more extreme favoring $H_1$. Manual computation can become too complex even for moderate $n$. It is advisable to use a software package for calculating this probability. The following example illustrates the calculations that may help in understanding how these probabilities are obtained.

Suppose a regimen for control of angina pectoris is considered effective if at least 60% of patients on this regimen do not have any attack in 1 year of follow-up and not more than 10% have two or more attacks. Thus, the desired ratio is as follows:

| Number of angina attacks | 0 | 1 | 2+ |
|---|---|---|---|
| Desired percentage of patients | 60 | 30 | 10 |

These define the null hypothesis—$H_0$: $\pi_1 = 0.60$, $\pi_2 = 0.30$, $\pi_3 = 0.10$.

A higher percentage of patients with a lower number of attacks is even better. A lower percentage in these categories is $H_1$. The null hypothesis in this case is for the effectiveness of the regimen, whereas generally it is stated for ineffectiveness. The regimen under trial is lifestyle changes such as yoga, dietary changes, and physical exercise. After excluding other causes, only six eligible volunteers could be followed up for 1 year. The data obtained are as follows:

| Number of angina attacks | 0 | 1 | 2+ |
|---|---|---|---|
| Observed number of patients | 2 | 3 | 1 |

The observed ratio is loaded more toward a higher number of attacks than postulated under $H_0$. Is the regimen ineffective according to the criterion?

In this case, $O_1 = 2$, $O_2 = 3$, and $O_3 = 1$. The configurations adverse to $H_0$ and favoring $H_1$, beginning from the observed, are shown in Table M.13. This includes all configurations adverse to $H_0$ that have two or fewer patients with no attack. (It can be debated whether configuration (2, 2, 2) is adverse to (2, 3, 1). In the first case, two patients have one attack and another two have two or more attacks. In the second case, three patients have one attack and one patient has two or more attacks.) There are 18 such configurations. From the equation given earlier, the probability of observing the first configuration under $H_0$ is

$$P_1(O_1 = 2,\ O_2 = 3,\ O_3 = 1) = \frac{6!}{2!3!1!}(0.60)^2\,(0.30)^3\,(0.10)^1 = 0.058.$$

Similar probabilities can be calculated for the other configurations. But there is no need to do so here because $P_1$ itself is more than 0.05. The sum of the probabilities for these 18 configurations is going to be higher in any case. Since this $P$-value is not sufficiently small ($P_1$ itself is more than 0.05), the null hypothesis cannot be rejected. The evidence is not sufficient to call the regimen

**TABLE M.13**
**Configurations Adverse to $H_0$ and Favoring $H_1$ in Our Example**

| No. of Angina Attacks | Notation | Configurations Favoring $H_1$ (as Extreme as or More Extreme than the Observed) | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** |
| 0 | $O_1$ | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | $O_2$ | 3 | 0 | 1 | 2 | 4 | 0 | 1 | 2 | 3 | 4 | 5 | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 2+ | $O_3$ | 1 | 4 | 3 | 2 | 0 | 5 | 4 | 3 | 2 | 1 | 0 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |

ineffective in controlling angina attacks. Note again the reverse nature of the null we have considered in this example.

You may have noticed that with only $n = 6$ subjects and just $K = 3$ categories in our example, the number of configurations is already large. If, for example, $n = 12$ and $K = 4$, the number of configurations, even those adverse to $H_0$, may become enormous. This is the reason for the advice to use a software package to calculate this probability. The calculations in this example are given only to enhance your understanding of the underlying procedure.

The categories in this example are metric. But the method of computing probabilities considers them **nominal**. The only use made of the metric scale of categories is in identifying the configurations adverse to $H_0$.

## multiple comparisons, *see also* Bonferroni procedure (test), Dunnett test, least significant difference, Tukey test for multiple comparisons

These are comparisons of statistical summaries (such as means and proportions) in different groups, particularly where a group is used more than once. For example, comparisons of the mean in group 1 with the mean in group 2 and the mean in group 1 with the mean in group 3 are multiple comparisons since the mean in group 1 has been used twice. This affects the statistical results even when the groups are independent. Repeated use of the same data for statistical significance inflates the probability of **Type I error**, more so if it involves the same group—thus, special methods are needed that can keep this error within control.

Multiple comparisons most frequently arise in **analysis of variance (ANOVA)**. Overall significance of difference among group means is indicated by the ***F*-test** in ANOVA, significance of which tells that there is some difference somewhere. Once this significance is obtained, the next step is to identify the groups that are different from one or more of the others. For this reason, this is also called **post-hoc comparisons**. This requires several comparisons. In case of pairwise comparisons, for example, if there are 4 groups, the comparisons are group 1 with group 2, 1 with 3, 1 with 4, 2 with 3, 2 with 4, and 3 with 4. There are a total of 6 comparisons, and these are called multiple comparisons. Means of two groups are generally compared by the Student *t*-test, but repeated application of this test at, say, 5% level of significance on the same data blows up the total probability of Type I error to an unacceptable level. If there are 15 tests on the same data, each done at the 5% level, then the overall (*experiment-wise*) Type I error could be as high as $1 - (1 - 0.05)^{15} = 0.54$. Compare this with the desired threshold of 0.05.

The Type I error allowed for each individual comparison is called the *comparison-wise error rate*, whereas the total error for all the comparisons together is called the *experiment-wise error rate*. To keep the probability of a Type I error within a specified limit such as 0.05 for an experiment, many procedures for multiple comparisons are available. Each of these is generally known by the name of the scientist who first proposed it. Among them are Bonferroni, Tukey, Scheffe, Fisher's LSD, Newman–Keul, Duncan, and Dunnett. These methods differ in how they take care of inflated Type I error and are suitable in different setups. The **Bonferroni** can be used in a variety of setups, and the **Tukey** procedure is for pairwise comparisons. These two are commonly used in medical and health literature and are also the most suitable ones. The **Dunett test** is used specifically when each group is to be compared with the control only. These are described under these topics in this volume. For other procedures, see Klockars and Sax [1]. Note that all these methods are for quantitative data, particularly the mean. Bonferroni can be used for qualitative data as well where chi-square or any other test has been used.

The issue of multiple comparisons also arises in interpreting the results reported in publications. Many *P*-values may have been considered to arrive at a conclusion. Researchers may have tried many statistical tests before reaching to the final ones that they decide to report. Hardly anybody will make adjustment for such *behind-the-scene* statistical tests, although they also affect the final *P*-value. Also, several tests of statistical significance, each at $\alpha = 0.05$, in the same article for various comparisons would make the total error much higher than acceptable. A wiser approach is to limit the calculation of *P*-values to the original question and not give too many *P*-values that can raise a question mark on the overall result.

There might be two or more publications on the same set of data with different focus or with different outcome variables. Each publication may be complete within itself for multiple comparisons but would be oblivious to the comparisons made for the other papers. The argument can be extended to Type I errors committed by other workers in similar studies and possibly in lifetime. For some researches, ignoring accumulation of Type I errors is one of the reasons that the statistically significant results fail to reproduce. Just be on guard for such fallacies.

1. Klockars AJ, Sax G. *Multiple Comparisons (Quantitative Applications in the Social Sciences)*. Sage, 2005.

## multiple correlation

This is the correlation between one variable $y$ and a linear combination of several other variables $(x_1, x_2, \ldots, x_K)$. Multiple correlation is easily computed when all these variables are quantitative but can be extended to include categories of **nominal** variables also when recoded by **indicator variables**. When nonlinear

combinations of $x$'s are considered, this becomes the **coefficient of determination**.

While considering multiple correlation coefficient as the Pearsonian **correlation coefficient** between one variable $y$ and the *best* linear combination of the variables ($x_1$, $x_2$,…, $x_K$), two words are important: best and linear combination. The linear combination of these variables is $b_1x_1 + b_2x_2 + \ldots + b_Kx_K$, but for multiple correlation coefficient, $b$'s are chosen in such a manner that the correlation between this combination and $y$ is maximum. This is what we mean by *best* linear combination. These $b$'s happen to be the same as the **regression coefficients** obtained by the **least squares method**. Thus, the best linear combination of the $x$'s is the same as the predicted value of $y$ by linear regression. This gives another definition of multiple correlation coefficient—this is the Pearsonian correlation coefficient between $y$ and its value **predicted** by the linear regression model. This coefficient is denoted by $R$, but for ease of interpretation, $R^2$ is used instead. The value of $R^2$ is the proportion of the total variance of $y$ explained by the linear regression. If $R^2 = 0.73$, this means that 73% of the variation in the values of $y$ can be explained by the differences in the values of the $x$'s from subject to subject. This, in a way, measures the utility of the regression in predicting the value of $y$. The other 27% is unexplained or is due to the factors not considered in this regression equation including the random fluctuations. In case of simple linear regression with one independent variable, $R^2 = r^2$, where $r$ is the notation for the usual correlation coefficient. The following example may help in fixing the ideas.

Consider the prediction of glomerular filtration rate (GFR) in kidney recipients 12 months after transplantation by graft kidney volume/recipient BSA ratio, donor age, and recipient gender [1]. A group of, say, 160 such patients were followed up for 12 months, and these measurements were recorded. A linear regression equation was obtained to predict the GFR at 12 months with the ratio, age, and gender as regressors. Now, plug in these values of the ratio, age, and gender in the regression equation and obtain the predicted GFR for these 160 patients. The Pearsonian correlation between these predicted values and the actual observed values will be the multiple correlation coefficient. If the square of this correlation coefficient $R^2 = 0.22$, you know that the three regressors (ratio, age, and gender) together, when their best linear combination is considered, are able to account for just 22% variation in the GFR at 12 months.

$R^2$ is also used to compare one model with another. If $R^2$ for one model is 0.63 and for the other model it is 0.76, then the model with the higher $R^2$ is considered a better fit. Increasingly better fit can be obtained by progressively adding new regressors, but a large number of regressors make the model difficult to interpret. Thus, a model is considered good when it contains a small number of regressors but gives a sufficiently large value of $R^2$. A value more than 0.70 is generally considered desirable, between 0.80 and 0.89 good, and 0.90 or more excellent. These are relative and not absolute since $R^2 < 0.50$ can also be useful in rare cases where almost nothing is known. The success in achieving large $R^2$ depends on the proper choice of the regressors and on the right specification of the model. In health and medicine, it is many times difficult to obtain a high value of $R^2$ because the appropriate regressors are not known for some situations, and they have to be assumed on the present knowledge or on available data. Both these sources may be inadequate.

Since $R^2$ continues to improve as more regressors are added, a realistic assessment is made when adjusted for the number of regressors. The details are given under the topic **adjusted $R^2$**.

A statistical procedure similar to the $F$-test for ANOVA is available to test the statistical significance of $R^2$ as well as of addition to $R^2$ made by any additional variable. Most statistical software packages do this easily. It is thus possible to specify a large number of regressors and ask the computer program to include only those that contribute significantly. Methods such as forward selection, backward elimination, or **stepwise** are used for this purpose. These methods tend to automatically exclude variables that have high **multicollinearity** with other regressors.

High $R^2$ for small $n$ is not much helpful because standard errors are still high and the reliability of the estimates is low. High $R^2$ can occur in a structurally inappropriate model. This can also occur in case of multicollinearity where many individual regression coefficients are not statistically significant.

### Coefficient of Determination

Overall statistical significance of a regression model is checked by an $F$-test similar to the one we use for ANOVA. This significance indicates only that the model is not useless, but it may or may not be a sufficiently good fit. A regression model (whether linear or nonlinear) is considered a good fit if all the **residuals** $e = (y - \hat{y})$ are small. Since residuals fluctuate around zero in any case, small residuals would necessarily yield a small **sum of squares** $\Sigma e^2$. This is the residual sum of squares, popularly called the sum of squares due to error (SSE). Its magnitude, when compared with the total sum of squares, $SST = \Sigma(y - \bar{y})^2$, provides a measure of "lack of fit" of the regression. The difference (SST – SSE) is called *regression sum of squares* (RegSS). Each of these sums of squares has associated degrees of freedom (df's) as in the case of ANOVA. RegSS measures how much of the total sum of squares the regression has been able to account for. This is the "goodness of fit" and its proportion of the SST is denoted by $\eta^2$, i.e.,

$$\eta^2 = \frac{\text{RegSS}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}.$$

The quantity $\eta^2$ is interpreted as the proportion sum of squares of $y$ explained by the regression and called the **coefficient of determination**. The larger the $\eta^2$, the better the fit. If the residual sum of squares is as small as, say, 10% of the total sum of squares, then $\eta^2 = 0.90$. The fit, then, is said to account for 90% variation in $y$. A fit with such high $\eta^2$ should be adequate in most cases, especially if $n$ is large.

Note that we have not restricted to linear regression in this case, and it can be nonlinear. When the regression is linear, $\eta^2$ is the same as $R^2$. To reiterate, coefficient of determination is the square of the multiple correlation coefficient when linear combinations are considered. The term *coefficient of determination* can be used for any regression but is more appropriate for nonlinear regressions.

Most researchers present results in term of $R^2$ and not $\eta^2$ because they generally consider only the linear regression. The notation $\eta^2$ for coefficient of determination is general and applies to all regression setups. However, in case of multiple or simple linear regression, this is denoted by $R^2$.

1. Lee CK, Yoon YE, Choi KH, Yang SC, Lee JS, Joo DJ, Huh KH, Kim YS, Han WK. Clinical implications for graft function of a new equation model for the ratio of living donor kidney volume to recipient body surface area. *Korean J Urol* 2013 Dec;54(12):870–5. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3866292/

## multiple imputations, *see* imputation for missing values

## multiple linear regression, *see also* linear regression, simple linear regression

A **linear regression** with more than one regressor is called multiple linear regression. When the regressor is just one, the term used is **simple linear regression**. If you are not familiar with this, review that topic first to understand what multiple linear regression is. A multiple linear regression is of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \varepsilon,$$

where the variable $y$ is called the dependent, whereas $x_1, x_2, \ldots x_K$ are the $K$ independent variables, and $\varepsilon$ is called the error term that is left over part of the variable $y$ not accounted by the $x$'s. The $x$'s can be nonlinear such as $x_3 = x_2^2$, but the $\beta$'s have to be linear for the regression to be called linear. If the $\beta$'s are nonlinear, this becomes a **nonlinear regression** and much more complex to obtain. The dependent is also called the response variable or the outcome variable depending on the context and the independents have names such as predictors, explanatory variables, and regressors. For details of these variables, see the topic **dependent and independent variables**. The values of the independent variables in this regression are considered fixed and not stochastic.

The $\beta$'s in the multiple linear regression equation are called the regression coefficients. See the topic **regression coefficients** to understand their interpretation, and for confidence intervals and tests of hypotheses on these coefficients. The notation $\beta$ is for the regression coefficients in the population from which the subjects have been selected and measured for running the regression. The corresponding sample estimates of $\beta$'s are denoted by $b$'s. When the values of $b$'s are substituted in the regression equation and the value of $y$ is computed for any set of values of $x$'s, this is called the predicted value of $y$ and denoted by $\hat{y}$. This predicted value, in fact, is the average of predicted $y$ for the fixed set of values of $x$'s. This implies that if several values of $y$ are obtained for the same fixed values of the $x$'s, $\hat{y}$ would be the average of the predicted values. This is explained a little later in this section by an example.

Two basic methods are used to estimate $\beta$'s. The first is to find those $b$'s that make the regression closest to the observed values in the sample in the sense that the sum of squares of the difference between the observed and predicted values of $y$ is minimum. In notations, this sum of squares is $\Sigma(y - \hat{y})^2$ and is called the residual sum of squares. The method that minimizes this sum of squares is called the method of **least squares**. The second method is to find those values of $b$'s that make the observed sample values most likely to occur. This is called the method of **maximum likelihood**. Quite commonly, particularly when the errors have Gaussian distribution, the two methods give the same estimates of the regression coefficients. For more details, see the topic **regression fitting (general method of)**. Statistical software packages give these estimates easily once you specify the regression and the dependent and independent variables.

Consider regression of birth weight of full-term healthy babies on the weight of the father and mother. Suppose the regression equation obtained on the basis of a random sample is

$$\text{BW} = 2.65 + 0.008 \times (\text{MW}) + 0.004 \times (\text{FW});$$
$$55 \le \text{MW} < 80;\ 60 \le \text{FW} < 90;$$

where BW is birth weight, MW is mother's weight, and FW is father's weight. All weights are in kilograms. Note that the regression is based on the mothers whose weight ranges from 55 to 80 kg and the fathers whose weight ranges from 60 to 90 kg. Thus, this regression is valid for these weight ranges only. Slight extrapolation on either side may be admissible.

Since there are two regressors, graphically a surface is obtained in place of a line. The response surface of this equation is shown in Figure M.8. If there are any square or any such term for weight of mother or father or both, the shape of the surface will not be a plane but curved. If there is a third regressor, the shape will be three-dimensional such as a pyramid or cuboid for linear regression and difficult to be represented graphically. Four or more regressors are even more difficult to visualize. Higher dimensions can only be conceptualized and not depicted.

The values of the regression coefficients show that for every kilogram weight of the mother, the birth weight increases on average by 8 g, and for every kilogram weight of the father, the birth weight increases by 4 g. The influence of the father's weight on birth weight is one-half of that of the mother in this example. When the mother's weight is 60 kg and the father's weight is 70 kg, the predicted birth weight is $2.65 + 0.008 \times 60 + 0.004 \times 70 = 3.41$ kg. This, in fact, is the predicted average birth weight of children of all those parents whose weights are 60 kg (mothers) and 70 kg (fathers). Note that this regression is for full-term healthy babies. Even in this group, it is known that the birth weight actually does not depend exclusively on the parental weight. Inclusion of other influential factors in the regression may improve the prediction.

As noted earlier, the prediction in this example is for the *average* weight of babies with specific weights of parents. The actual weight in individual births could be different. This difference would be due partly to deviation from the average and partly to the influence of factors other than parental weight.

In a classical multiple linear regression, all the $x$'s are quantitative and so, of course, is $y$. However, if some of the $x$'s are nominal, indicator variables can be created for each category of each nominal variable that would effectively give separate regression equations for each category of the nominal variables. For example, we can insert a variable gender of the child in our example with value gender = 1

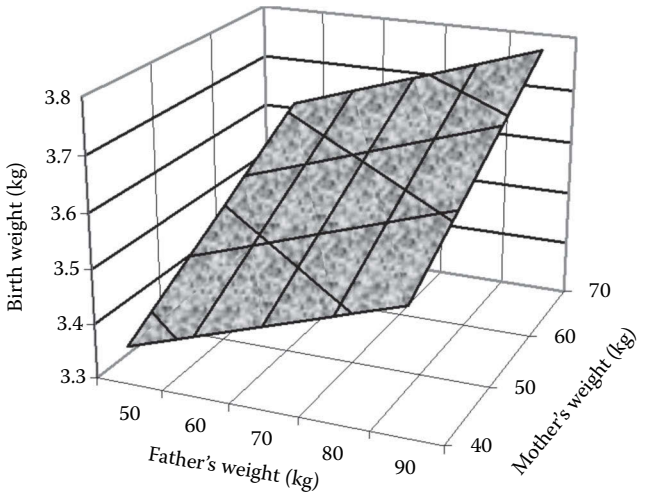



**FIGURE M.8** Response surface for a child's birth weight for different weights of the mother and father.

M

for male child and gender = 0 for female child. This is the indicator variable for gender. Insertion of this variable would most likely change all the regression coefficients. Suppose the new regression equation is

$$BW = 2.63 + 0.007 \times (MW) + 0.005 \times (FW) + 0.4 \times (\text{gender});$$
$$55 \le MW < 80;\ 60 \le FW < 90.$$

For gender = 1 (male child), this equation is BW(male child) = 2.63 + 0.007 × (MW) + 0.005 × (FW) + 0.4 × 1, or BW(male child) = 2.67 + 0.007 × (MW) + 0.005 × (FW). For gender = 0 (female child), this equation is BW(female child) = 2.63 + 0.007 × (MW) + 0.005 × (FW) + 0.4 × 0, or BW(female child) = 2.63 + 0.007 × (MW) + 0.005 × (FW). Because of the indicator variable, the same regression equation produces two equations—one for male children and the other for female children. The only difference we have represented in this example is only in the **intercept**, but regressions can be devised so that the regression coefficients for MW and FW are different for male children than for female children. This could be obtained by including **interaction** terms between the indicator variable and MW or FW or both.

As explained under the topic **models**, a multiple linear regression could be explanatory, predictive, or causal. **Adequacy of regression fit** is a separate topic but mostly, in case of multiple linear regression, is assessed by the value of the square of the **multiple correlation** coefficient $R^2$. If there are a large number of candidate variables for consideration in multiple linear regression, **variable selection** may become necessary. This also helps in obtaining a more parsimonious model. All regressions, including multiple linear regression, are valid under certain conditions as explained in the topic **regression requirements (validation of)**.



## multiple responses

You know that one person can have just one age and one hemoglobin level at a point in time but can have multiple symptoms or multiple diseases. The person may report that he/she is suffering from diabetes, hypertension, and vision loss at the same time. Such responses are called multiple responses. Data on these responses require special care in analysis and presentation.

Consider the data in Table M.14 on the number of cases of abdominal tuberculosis with major symptoms. This is a classification of the cases, but the symptoms are neither exclusive nor exhaustive. They are not exclusive because one patient can have two or more complaints: a patient may have vomiting as well as constipation. This is what we call multiple responses. The categories are not exhaustive because only major symptoms are listed: patients with other symptoms may be present but are not included in this table. When the categories are exhaustive and multiple responses are present, the sum total of frequencies would necessarily exceed the total number of subjects. In Table M.14, the sum is more even when the categories are not exhaustive. To avoid confusion, the total number of subjects in the case of multiple responses is called a base in place of total. Percentages are generally calculated using this base and not the total.

**TABLE M.14**
**Cases of Abdominal Tuberculosis with Major Symptoms**

| Symptom | Number of Cases | Percent |
|---|---|---|
| Pain in abdomen | 126 | 89.4 |
| Vomiting for a long time | 85 | 60.3 |
| Constipation for a long time | 57 | 40.4 |
| **Total cases (Base)** | **141** | **100.0** |

*Source:* Adapted from Das P et al., *Am J Proctol* 1975; 6:75–86. http://www.ncbi.nlm.nih.gov/pubmed/1119554#

**TABLE M.15**
**Conversion of Cases of Abdominal Tuberculosis with Major Symptoms into a Contingency Table**

| Group of Major Complaints | Number of Cases | Percent |
|---|---|---|
| Pain, vomiting, and constipation | 44 | 31.2 |
| Pain and vomiting, no constipation | 37 | 26.2 |
| Pain and constipation, no vomiting | 10 | 7.1 |
| Vomiting and constipation, no pain | 2 | 1.4 |
| Pain, no vomiting, no constipation | 35 | 24.8 |
| Vomiting, no pain, no constipation | 2 | 1.4 |
| Constipation, no pain, no vomiting | 1 | 0.7 |
| Other symptoms | 10 | 7.1 |
| **Total** | **141** | **100.0** |

Because of the multiple responses, Table M.14 is not a **contingency table** and cannot be analyzed by **chi-square**. Graphs that require a meaningful total such as **pie diagram** cannot be made on data with multiple responses. But the data can be converted to a contingency table when additional information is available. One way to do it is suggested in Table M.15. Note that the categories are now **mutually exclusive and exhaustive**. Perhaps this table provides more useful clinical information.

1. Das P, Kumar P, Gupta CK, Indrayan A. Clinical patterns of abdominal tuberculosis. *Am J Proctol* 1975; 6:75–86. http://www.ncbi.nlm.nih.gov/pubmed/1119554#

## multistage random sampling

In studies that involve a population of large size, it is sometimes helpful to draw samples in stages. If the subjects spread all over a state are the target, you may select a small number of districts or counties in the first stage (first stage units are called **primary sampling units**); then some blocks, colonies, or hospitals in the second stage from the selected districts or counties; and finally the subjects from the selected colonies/hospitals. Thus, there are sampling units of various sizes. When sampling is done in stages from bigger to smaller units within the units selected at the previous stage, it is called multistage sampling. When selection in each stage is random, this becomes multistage random sampling (MRS).

In a study to find the prevalence of smoking in females of age 20 years and above in a particular state with, say, a million families, you may, for example, first select 4 counties by the random method, then 12 census blocks within each selected county, and 50 families within each selected block. All females of age 20+ years in the selected families could be the unit of inquiry, although the sampling units successively are counties, blocks, and families. Some families may have two or more units of inquiry and some none at all but most may have just one. If there are many families with two or more eligible females, this can produce a **clustering effect**.

In the preceding example, a total of 4 × 12 × 50 = 2400 families could be in the sample. This looks like an extremely small number

compared with a total of a million families in the state. Yet this could provide a fairly precise estimate of the prevalence of smoking among females of age 20+ years in the state.

If a **simple random sampling** (SRS) of 2400 families out of a million is chosen instead of an MRS, the selected families may be scattered all over the state, say in 200 census blocks. A block may have to be visited for just one family in the sample. This could mean a substantially higher cost of travel by the survey team and loss of time. In the case of MRS in this example, only 4 counties need to be visited, and the survey work will concentrate in 12 blocks within each county. Thus, the major advantages of MRS are reduced costs and saving of time because of less travel. Another advantage is that a full **sampling frame** of the smaller units is not required. In this example, the frame required is the list of all counties in the state, the list of blocks in the *selected* counties, and the list of families in the *selected* blocks only. In the case of SRS, on the other hand, the frame will be the list of all families in the entire state. Preparation of this frame in itself could be a major exercise in some situations. Another advantage of MRS is that, in most practical situations, the smaller sample chosen by MRS may be sufficient to achieve good precision relative to SRS.

An example of MRS is provided by Milias et al. [1] who studied prevalence of self-reported hypercholesterolemia in Greek adults. In this nationwide survey, 5003 adults of age 18–74 years were enrolled by multistage sampling. Full details are not given, but they state that the multistage sampling was based on age–sex distribution of the Greek population according to the 2000 census. However, the stages are not identified. The accompanying table in their article on the age–sex distribution of the sample and population indicates that the age representation in the sample was nearly the same as in the population. The reporting made in this article has provided a few lessons. First, the sampling is stated as multistage, but stages are not identified. They may have selected a few districts in the first stage, a few counties in the second stage, and a few families in the third stage, but not stated. Second, the authors use the term "study population" instead of study sample for the selected 5003 adults. Such misuse of the term **population** is quite common in medical literature. Third, the proportionate representation of age indicates that possibly age stratification has been used but not explicitly stated.

1. Milias GA, Panagiotakos DB, Pitsavas C, Xenaki D, Panagopoulos G, Stefandis C. Prevalence of self-reported hypercholesterolemia and its relation to dietary habits in Greek adults: A national nutrition and health survey. *Lipids Health Dis* 2006;5:5. https://www.mysciencework.com/publication/read/4147242/prevalence-of-self-reported-hypercholesterolaemia-and-its-relation-to-dietary-habits-in-greek-adults-a-national-nutrition-health#page-null, last accessed March 20, 2015.

## multivariate analysis of variance (MANOVA)

Analysis of variance (ANOVA) is used where a quantitative variable is dependent on qualitative characteristics such as case and control group, or different types of treatment. This is called univariate setup since the dependent is just one variable. Now consider a setup where the dependent is not one variable but a set of quantitative variables, and the objective continues to be to study the difference in means in different groups.

### *Regular MANOVA*

Multivariate ANOVA (MANOVA) is used when a set of correlated quantitative variables is considered dependent on a set of qualitative variables. For example, the dependent could be kidney functions (creatinine clearance, urea clearance, and diotrast or *p*-aminohippurate clearance) in persons of age 50–59 years. These persons may be grouped into those taking different diets (vegetarian, meat based, fish based, etc.) and grades of physical activity (sedentary, mild, moderate, and heavy). They may also be categorized into male and female. The dependent variables in this case are various kidney functions. These are **multivariate data** and quantitative. The independents are gender and the category of diet and physical activity, which are qualitative. Just as in ANOVA, the primary purpose of MANOVA in this case would be to test the null hypothesis of equality of means of kidney function parameters in people in different categories of independent factors. Interaction among these independent factors can also be investigated. We illustrate this with the help of an example.

To evaluate the effectiveness of an exercise intervention for people with early and midstage Parkinson's disease, 51 men and women with the disease were randomly allocated to two groups [1]. One group received individual instructions for exercise three times a week for 10 weeks, and the other group remained in the usual care (control). In this study, 46 completed the trial. The outcome measures are functional axial rotation for spinal flexibility, functional reach, and the supine-to-standing time for measuring physical performance. Thus, there were three dependent variables. MANOVA performed for these three variables demonstrated a significant difference ($P < 0.05$) between the two groups. It was concluded that better improvement could be achieved by a 10-week exercise program for people in early and midstage Parkinson's disease.

Note that the conclusion in this example is based on simultaneous consideration of the three outcome variables. When each of them is considered individually in a univariate setup, functional axial rotation and functional reach showed a significant difference but not the supine-to-sitting time. It can be argued that univariate analyses allow a more focused conclusion because they tell which particular outcome is affected and which is not. However, combining the univariate conclusions on correlated variables can sometimes be erroneous because the joint conclusion is subject to a relatively high **Type I error** than is otherwise apparent. Also, in general, a combined conclusion cannot be drawn by univariate analyses when some outcomes show change and others do not. Multivariate analysis helps to draw a clear conclusion for all the variables together.

In the case of a univariate setup, the Student *t*-test is a special case of the ANOVA *F*-test when the number of groups is two. The corresponding analog of MANOVA for two groups is **Hotelling $T^2$**. In other words, MANOVA for two groups gives the same result as Hotelling $T^2$.

The following remarks about MANOVA may be helpful in clarifying certain issues:

- The underlying requirement for a MANOVA test is a **multivariate Gaussian** (normal) pattern of the observations. This is not easy to verify. However, each variable separately can be checked for a Gaussian pattern using the methods described in the topic **Gaussianity (how to check)**. When each is not far from Gaussian, there is a great likelihood that they are jointly multivariate Gaussian.
- The other requirement for a valid MANOVA test is homogeneity of the **dispersion matrices** of $y$'s in different groups. This matrix is the multivariate analog of the variance. Methods such as the **Box *M* test** are used for testing their homogeneity, keeping in mind that this heavily depends on multivariate Gaussianity of data. You can depend on a statistical software package for performing this test. When the same set of variables is measured in

two or more groups, the dispersion matrices should be comparable, and there would be rarely any need to worry on this account.
- The test criterion used for the MANOVA test is generally either Pillai trace or **Wilks Λ**. Their distribution in most cases can be transformed, at least approximately, to the usual *F* as applicable to ANOVA. Statistical software would do this and provide the ***P*-value**.
- Empty cells due to missing observations or otherwise are a more serious handicap in a multivariate setup than in a univariate setup. If information on just one variable is missing, the entire record is generally deleted. This may reduce the available sample size considerably when values for different variables are missing for different subjects.
- This method assumes that all dependent variables have equal importance. If homocysteine and insulin levels both appear as dependents in the data for coronary artery disease (CAD) cases, both are given the same importance unless a weighting system is applied in the analysis. For example, it is possible in the analysis of data of a study on CAD to incorporate six times more importance to homocysteine level compared with insulin level. This will modify the method to some extent, but the major problem for a clinician would be to determine these weights objectively.

### *MANOVA for Repeated Measures*

Another example where the dependent is a correlated quantitative response is drug concentration in the blood at different points of time (repeated measures) after its administration to two or more groups of patients (e.g., experimental/control or control/drug 1/drug 2). The qualitative independent in this case is the group. The objective is to find whether or not the mean response at different points of time is different in various groups, and not to explore the time trend of the response. The mean response in this case refers to the mean over the patients and not the mean over time. MANOVA would simultaneously compare several means, one at each time point, in one group with those in the other groups. Such simultaneous consideration obviates the need to examine univariate parameters such as time to reach the peak concentration ($T_{max}$) and the peak concentration ($C_{max}$) reached unless they are otherwise needed for evaluating pharmacological properties of the regimen. MANOVA in this case could be an alternative to the area under the (concentration) curve (AUC) used by many workers for this setup. As explained under the topic **area under the curve (AUC)**, this can lead to erroneous conclusions in some cases.

Repeated measures are naturally correlated and provide an apt situation for use of MANOVA whenever dependent is quantitative. Univariate **repeated measures ANOVA** is discussed separately, but MANOVA is considered better when the group sizes are equal (balanced design) because, then, MANOVA is quite robust to violation of **sphericity**. Sphericity is contrasts (differences at various time points with the previous value) being independent (covariance = 0) and having the same variance (homogeneity). This is tested by the **Mauchly test**. These are rather strong requirements for univariate repeated measures ANOVA but not as much for MANOVA. Univariate analysis is recommended for unbalanced design (unequal group sizes) where df's are corrected by **Huynt–Feldt** epsilon for *F*-test when sphericity is violated.

If you are using MANOVA because of balanced design in repeated measures, care is needed in specifying the design regarding what factors are between subjects and what are within subjects. In repeated measures, time will always be within subjects, but there might be other factors as well. Also, for MANOVA for repeated measures, prefer **Pillai trace** as the criterion instead of Wilks Λ for testing differences between groups, because Pillai trace is generally more robust to assumption violation. The interaction between groups and time will indicate whether different groups have the same time trend or not. Do not forget to test homogeneity of dispersion matrices between groups by the Box *M* test. Gaussian distribution of the errors is required for all these tests.

For further details of MANOVA, see Morrison [2].

1. Schenkman M, Cutson TM, Kuchibhatla M, Chandler J, Pieper CF, Ray L, Laub KC. Exercise to improve spinal flexibility and function for people with Parkinson's disease: A randomized controlled trial. *J Am Geriatr Soc* 1998;46:1207–16. http://www.ncbi.nlm.nih.gov/pubmed/9777901
2. Morrison DF. *Multivariate Statistical Methods*, Fourth Edition. Duxbury Press, 2004.

## multivariate data/methods

Simultaneous consideration of several measurements gives rise to multivariate data. The only condition is that they should be all **stochastic**—depend on chance and cannot be predicted with certainty, or are not fixed by design. In this setup, each individual measurement gives partial picture, and the complete picture is available when all these are considered together. For example, a child's weight does not tell you much about his or her growth unless you consider age, sex, and height also. Systolic level itself is not good enough for clinical evaluation, and diastolic level also should be simultaneously considered. For BP also, age and other comorbidities help in obtaining a complete picture. In fact, all health and disease conditions require that more than one measurement be considered in juxtaposition with one another to arrive at any unified conclusion. In this sense, all medical evaluations are intrinsically multivariate.

In view of what we have stated in the preceding paragraph, isn't it strange that most statistical analyses of data are done as though each variable is stand-alone? The Student *t*-test, *z*-test for proportions, binomial test, analysis of variance, and corresponding confidence intervals (CIs) are all univariate methods. Even ordinary regression and logistic regression are effectively univariate as only the dependent is considered stochastic—the independents in both these regressions are considered fixed. Whereas methods such as survival analysis are adequate as univariate methods because it can be considered by itself, most other analyses are oblivious of the correlations a variable has with other variables. For example, if there is any conclusion for HDL-cholesterol (HDL-C), it will not consider LDL-cholesterol (LDL-C) or triglyceride; if the study is on diabetic neuropathy alone, it will not consider age or body mass index. Such conclusions have risk of being compromised, even wrong, because other factors affecting it are not simultaneously considered. Even if you consider HDL-C and LDL-C both but independently and separately, the effect of their correlation is still ignored. A joint conclusion can be rarely drawn when correlated variables are separately studied. This underscores the need to be careful about conclusions regarding one variable at a time.

**Correlations** and **chi-square for *R* × *C*** contingency tables are examples of methods that consider two variables together—correlation where both the variables are quantitative and chi-square where both are qualitative. Both these are bivariate methods and good enough to assess the presence and degree of dependence. If one is quantitative and the other qualitative, **point-biserial correlation** can be used. But this kind of bivariate analysis does not

go beyond since, for example, the **simple regression** that finds the nature of relationship considers the independent variable on the right side of the equation as fixed and not stochastic. Thus, this is not a bivariate method.

As mentioned, the primary reason that several variables are considered together is that they depend on one another—they are correlated. Thus, the correlation structure among the variables is an important consideration in the multivariate setup. This structure can also be considered in the form of **dispersion matrix**. Even if the variables are not correlated, the results arrived at by individual variables could provide a different result than when considered simultaneously. Statistically, this happens because combination of individual ***P*-values** or the CIs is not the same as the *P*-value or the CI in the multivariate setup.

We have already given examples of multivariate data: child growth parameters where age, sex, height, weight, etc., are considered together; and lipid profile where total cholesterol and its components are considered together. Similarly, liver functions comprising albumin, bilirubin, AST, and ALP can be considered together; and kidney functions comprising creatinine, urea, uric acid, etc., can also be considered together. One example of actual multivariate data on lung functions in 70 adult males is in Table M.16, where multivariate lung functions comprising FVC, $FEV_1$, PEFR, and TLC can be investigated for dependence on age, height, and weight. We use these data to illustrate some multivariate methods in other sections, but the dataset here illustrates what multivariate means. There are a large number of such examples in health and medicine, but most publications consider the variables one at a time for ease of analysis interpretation and avoid the complexity of joint inference on a set of measurement. Wide availability of statistical software packages has made it easy to analyze multivariate data, but their joint interpretation still remains a concern for many of us. Thus, univariate methods still persist for good reason.

Among common multivariate statistical methods are **multivariate analysis of variance (MANOVA)** and **multivariate regression**. In both of these, the dependent is a set of quantitative variables and independents are qualitative in MANOVA and can be almost any in multivariate regression. Two-population analog of this is **Hotelling $T^2$**. If the dependent set is qualitative, multivariate logistic regression can be used. This is too complex for the level of this book, and we have not included this here. In all of these, the independents are considered fixed. If the dependent is one qualitative variable that depends on several stochastic variables, **discriminant functions** can be the right method where the independents are not considered fixed. Similarly in **factor analysis**, qualitative constructs are sought to be extracted on the basis of multivariate data. **Cluster analysis** is used to discover affinity structure among multivariate observations. Beyond these, not many multivariate methods are used in health and medicine.

One big difficulty with multivariate methods is that they ignore the entire set of values for one subject if just one measurement is missing. In the data in Table M.16, multivariate analysis will not consider the values for subject numbers 30 and 35 because of missing values and would be based on 68 subjects in place of 70 subjects in the sample.

## multivariate distributions

A multivariate distribution is the joint distribution of several variables. To get a flavor of multivariate distributions, see **bivariate distribution**, where only two variables are considered together. That also may tell you how difficult it does become to present a distribution of three or more variables. One example is in Table M.17 where **joint distribution** of 1544 adults is shown by sex, age group, and **smoking index**. Smoking index is zero for never smokers.

In the data in Table M.17, sex is qualitative and age and smoking index are quantitative. Actually these are exact values for each of the 1544 persons in the sample but are categorized for easy presentation in the table. This representation is still adequate to explain certain features of multivariate distributions. The top panel of the table is the age group–smoking index bivariate distribution of males, and the middle panel is for females. Thus, there are two bivariate distributions—one for males and the other for females. Their combination in the bottom panel of the table is a bivariate distribution of age group and smoking index. This is called the **marginal distribution** after collapsing sex into one group. Similarly, there are three bivariate distributions of age group and sex for each of the three categories of smoking index, and four bivariate distributions of sex and smoking index for each of the four categories of age. The marginal distributions are represented by the corresponding totals. Further collapsing also provides another level of marginal distributions. For example, the last row of the table is the marginal distribution of smoking index in these 1544 persons.

Imagine now that you have actual values of age and smoking index. The marginal distributions of these will be continuous. For example, the marginal distribution of smoking index (after excluding zero) would be like that shown in Figure M.9. The peak (mode) is around smoking index = 8 or 9 when the actual values are considered. Zero value of smoking index is not shown as the frequency 1223 for this value is too large to accommodate in this graph. If you look at the total frequencies in different age groups, it seems that the mode in marginal distributions of age would be around 40 years, i.e., the commonest age in people surveyed is around 40 years.

The shape of the joint and marginal distributions helps in deciding which statistical method would be appropriate for analyzing our data. For example, the distribution of smoking index in our example is far from **Gaussian**, particularly when zero value is also considered. Thus, conventional methods such as confidence intervals and tests of hypotheses for **regression coefficients** are not applicable if smoking index is to be considered as dependent on age and sex. The other important issue in multivariate distributions is the **dispersion matrix**. Whereas variances can be calculated of one variable at a time, covariances require that all pairs of the variables be considered. There will be one covariance between age and sex, another one between age and smoking index, and a third covariance between sex and smoking index. These covariances decide the correlations.

For simplicity, we have discussed multivariate distributions with the help of an example of a trivariate distribution. This could be of any dimension. We hope that our description will help you to conceptualize a *K*-dimensional distribution. A popular example of a discrete multivariate distribution is **multinomial distribution**.

## multivariate Gaussian distribution, *see also* bivariate Gaussian distribution

See the topic **multivariate distributions** to understand their essentials. A distribution is called multivariate Gaussian (Normal) when their joint distribution of many variables has a Gaussian pattern. For a bell-shaped picture of bivariate Gaussian distribution, see the figure in the topic **bivariate Gaussian distribution**. A distribution in more than two dimensions is difficult to depict on a paper, but that picture can give you an idea of what a multivariate Gaussian distribution would look like.

**TABLE M.16**
**Lung Functions in a Sample of 70 Adult Males of Age 20–49 Years**

| Subject Number | Age (Years) | Height (cm) | Weight (kg) | FVC (L) | FEV1 (L) | PEFR (L/s) | TLC (L) |
|---|---|---|---|---|---|---|---|
| 1 | 27 | 156 | 61 | 4.46 | 3.17 | 5.66 | 6.93 |
| 2 | 26 | 163 | 60 | 4.02 | 3.19 | 5.70 | 5.67 |
| 3 | 32 | 153 | 54 | 3.35 | 2.74 | 3.68 | 4.60 |
| 4 | 35 | 155 | 45 | 3.50 | 3.93 | 2.88 | 6.61 |
| 5 | 35 | 148 | 53 | 2.58 | 2.29 | 5.46 | 3.28 |
| 6 | 38 | 163 | 50 | 3.26 | 3.05 | 6.95 | 3.94 |
| 7 | 26 | 161 | 54 | 3.97 | 3.44 | 8.87 | 5.17 |
| 8 | 32 | 155 | 43 | 2.49 | 2.56 | 5.57 | 2.73 |
| 9 | 26 | 155 | 43 | 4.35 | 4.34 | 8.71 | 4.91 |
| 10 | 28 | 147 | 49 | 2.85 | 2.63 | 7.32 | 3.48 |
| 11 | 28 | 164 | 63 | 3.23 | 2.81 | 6.66 | 4.19 |
| 12 | 29 | 167 | 57 | 3.88 | 2.74 | 4.92 | 6.09 |
| 13 | 19 | 161 | 54 | 4.00 | 3.43 | 4.56 | 5.24 |
| 14 | 27 | 161 | 52 | 4.48 | 3.73 | 5.93 | 6.04 |
| 15 | 30 | 162 | 60 | 3.41 | 3.51 | 7.79 | 3.72 |
| 16 | 20 | 162 | 45 | 3.06 | 2.66 | 5.39 | 3.98 |
| 17 | 20 | 160 | 45 | 2.77 | 2.97 | 4.44 | 2.99 |
| 18 | 30 | 161 | 46 | 3.77 | 2.87 | 5.76 | 5.51 |
| 19 | 23 | 160 | 47 | 4.20 | 3.38 | 4.55 | 5.86 |
| 20 | 20 | 159 | 47 | 3.42 | 2.96 | 4.63 | 4.45 |
| 21 | 28 | 163 | 49 | 3.36 | 3.16 | 6.01 | 4.03 |
| 22 | 24 | 162 | 45 | 3.89 | 3.02 | 8.65 | 3.12 |
| 23 | 35 | 165 | 45 | 3.94 | 3.12 | 5.63 | 5.59 |
| 24 | 24 | 167 | 56 | 5.27 | 4.05 | 5.95 | 7.66 |
| 25 | 26 | 159 | 47 | 3.26 | 2.34 | 4.31 | 5.03 |
| 26 | 16 | 162 | 40 | 2.84 | 1.27 | 1.88 | 3.18 |
| 27 | 19 | 165 | 55 | 3.33 | 2.39 | 3.62 | 5.13 |
| 28 | 28 | 164 | 55 | 3.74 | 3.27 | 5.51 | 4.83 |
| 29 | 30 | 172 | 44 | 3.93 | 2.82 | 5.12 | 6.07 |
| 30 | 25 | 160 | 53 | – | – | 5.82 | 6.71 |
| 31 | 35 | 160 | 52 | 4.85 | 3.06 | 4.80 | 3.08 |
| 32 | 22 | 166 | 57 | 4.36 | 1.88 | 8.65 | 5.53 |
| 33 | 25 | 170 | 56 | 4.00 | 3.77 | 7.89 | 4.79 |
| 34 | 30 | 165 | 55 | 2.80 | 3.40 | 7.31 | 3.00 |
| 35 | 37 | 172 | 63 | 4.72 | 4.16 | 6.03 | – |
| 36 | 21 | 170 | 60 | 3.11 | 2.92 | 6.36 | 6.73 |
| 37 | 26 | 155 | 47 | 3.25 | 2.68 | 5.33 | 4.42 |
| 38 | 21 | 167 | 55 | 3.11 | 2.92 | 6.21 | 3.70 |
| 39 | 23 | 163 | 54 | 3.80 | 3.12 | 6.33 | 5.19 |
| 40 | 25 | 174 | 59 | 2.02 | 3.76 | 4.64 | 4.84 |
| 41 | 38 | 155 | 44 | 3.16 | 1.48 | 2.31 | 3.10 |
| 42 | 34 | 168 | 53 | 5.22 | 4.52 | 5.12 | 4.61 |
| 43 | 30 | 163 | 56 | 3.23 | 2.70 | 5.30 | 4.33 |
| 44 | 40 | 161 | 60 | 2.17 | 1.76 | 5.18 | 2.99 |
| 45 | 23 | 159 | 48 | 3.79 | 3.60 | 5.44 | 6.01 |
| 46 | 25 | 162 | 61 | 3.46 | 4.66 | 5.02 | 3.09 |
| 47 | 30 | 163 | 60 | 3.78 | 3.29 | 6.50 | 4.90 |
| 48 | 32 | 161 | 52 | 3.03 | 2.69 | 5.64 | 4.32 |
| 49 | 30 | 163 | 59 | 3.60 | 3.02 | 6.03 | 4.81 |
| 50 | 34 | 166 | 50 | 3.34 | 1.68 | 5.10 | 6.72 |
| 51 | 27 | 162 | 42 | 4.57 | 2.04 | 6.29 | 2.89 |
| 52 | 23 | 161 | 62 | 3.83 | 3.35 | 5.97 | 4.93 |
| 53 | 24 | 156 | 49 | 3.50 | 3.30 | 6.22 | 4.18 |

(*Continued*)

**TABLE M.16 (CONTINUED)**
**Lung Functions in a Sample of 70 Adult Males of Age 20–49 Years**

| Subject Number | Age (Years) | Height (cm) | Weight (kg) | FVC (L) | FEV1 (L) | PEFR (L/s) | TLC (L) |
|---|---|---|---|---|---|---|---|
| 54 | 40 | 158 | 54 | 3.50 | 3.01 | 8.01 | 4.57 |
| 55 | 24 | 157 | 52 | 3.84 | 3.15 | 4.74 | 5.26 |
| 56 | 25 | 146 | 45 | 3.48 | 3.12 | 5.96 | 4.39 |
| 57 | 22 | 169 | 77 | 4.26 | 4.68 | 6.11 | 4.81 |
| 58 | 27 | 170 | 60 | 4.54 | 1.28 | 5.91 | 5.42 |
| 59 | 40 | 156 | 55 | 3.22 | 2.77 | 8.63 | 4.22 |
| 60 | 30 | 170 | 58 | 4.38 | 3.51 | 4.64 | 6.12 |
| 61 | 41 | 169 | 55 | 4.18 | 2.14 | 3.15 | 8.32 |
| 62 | 26 | 161 | 64 | 3.45 | 4.43 | 5.17 | 4.33 |
| 63 | 50 | 157 | 53 | 3.60 | 2.89 | 5.72 | 5.02 |
| 64 | 25 | 160 | 60 | 3.80 | 3.30 | 8.20 | 5.45 |
| 65 | 52 | 162 | 76 | 3.62 | 2.99 | 6.78 | 4.92 |
| 66 | 25 | 175 | 68 | 2.77 | 3.99 | 8.25 | 3.41 |
| 67 | 33 | 174 | 82 | 3.85 | 3.31 | 8.37 | 5.05 |
| 68 | 25 | 163 | 47 | 3.20 | 2.61 | 3.82 | 4.41 |
| 69 | 34 | 162 | 50 | 3.85 | 3.31 | 8.32 | 6.12 |
| 70 | 28 | 159 | 55 | 4.84 | 3.51 | 6.78 | 3.81 |

*Note:* FEV1: forced expiratory volume in 1 s; FVC: forced vital capacity; PEFR: peak expiratory flow rate; TLC: total lung capacity.

**TABLE M.17**
**Trivariate Distribution of a Sample of 1544 Adults by Age Group, Sex, and Smoking Index in a Community Survey**

**Number of Subjects**

| Sex | Age Group (Years) | Smoking Index 0 | 0.1–9.9 | 10.0+ | Total |
|---|---|---|---|---|---|
| Male | 15–24 | 135 | 21 | 7 | **163** |
| | 25–44 | 202 | 51 | 33 | **286** |
| | 45–59 | 154 | 52 | 28 | **234** |
| | 60+ | 69 | 13 | 9 | **91** |
| | **Total** | **560** | **137** | **77** | **774** |
| Female | 15–24 | 176 | 13 | 2 | **191** |
| | 25–44 | 224 | 23 | 14 | **261** |
| | 45–59 | 187 | 35 | 13 | **235** |
| | 60+ | 76 | 5 | 2 | **83** |
| | **Total** | **663** | **76** | **31** | **770** |
| M + F | 15–24 | 311 | 34 | 9 | **354** |
| | 25–44 | 426 | 74 | 47 | **547** |
| | 45–59 | 341 | 87 | 41 | **469** |
| | 60+ | 145 | 18 | 11 | **174** |
| | **Total** | **1223** | **213** | **108** | **1544** |

*Note:* Smoking index = 0 for never smokers.

In a multivariate Gaussian distribution, marginal distribution of each variable is Gaussian. This, however, does not mean that Gaussian marginal distribution of each variable implies joint Gaussian. Examples can be constructed when this is not so, but that would be rare and hardly ever in practice. Thus, Gaussian pattern of marginal distribution of each variable is considered sufficient to assume that their joint distribution is also Gaussian.

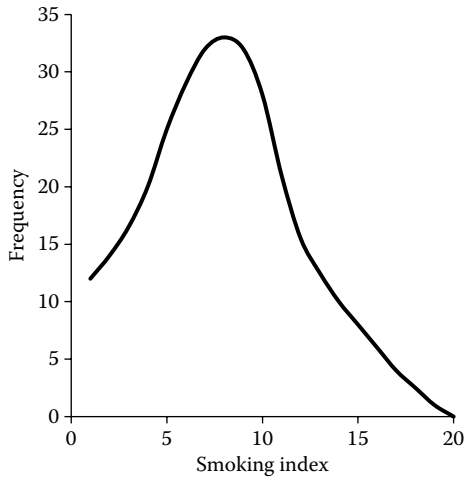



**FIGURE M.9** Marginal distribution of smoking index (excluding zero) based on exact values.

When the distribution of $K$ variables is jointly Gaussian, each variable can be expressed as a linear combination of the other $(K - 1)$ variables. This linearity is a big convenience, particularly in using Pearsonian correlation coefficient as a measure of the degree of dependence among variables. This correlation is able to capture only the linear dependence and not the nonlinear dependence as explained under the topic **correlation coefficient**, and that is all you need for multivariate Gaussian distribution. Such linear dependence also helps in using linear regression in this setup. The mean of the variables and the **dispersion matrix** completely specify a multivariate Gaussian distribution. The dispersion matrix when its elements are expressed as correlations tells us which pairs of variables have high dependence on each other and which have low dependence. They are all independent when the correlations for each pair are

zero. In samples, they may still exhibit correlations despite independence because of sampling fluctuations, but these should be low.

A multivariate Gaussian distribution is a requirement for many multivariate methods. For example, this is needed for testing the hypothesis of equality of group means by **Wilks Λ** in **MANOVA** and for repeated measures. For **multivariate regression** also, this is needed for obtaining confidence bands and tests of hypotheses, although not for obtaining the regression equation. For discriminant functions also, multivariate Gaussianity is helpful. However, this is not a requirement for multivariate methods of **cluster analysis** and **factor analysis**.

## multivariate logistic regression, *see also* logistic models/regression (multinomial, ordinal, and conditional)

This expresses a set of binary variables as a function of one or more regressors. Just as in the case of (univariate) logistic regression, the regressors can be qualitative or quantitative. Many times in the literature, univariate logistic regression when the regressors are more than one has been called multivariate logistic regression. This is a wrong usage as the right term for this setup is *multivariable* logistic regression and not *multivariate*.

Consider a set of data on patients where presence or absence of diabetes, hypertension, and arthritis are to be studied together and not one at a time. These are anticipated to be dependent on age, sex, body mass index, family history of diabetes, family history of hypertension, physical activity score, etc. The relationship of three binary variables together in this example with the regressors can be studied by multivariate logistic regression. Studying them together is all the more important in this case since diabetes, hypertension, and arthritis are interdependent conditions.

Although mathematically quite complex and not easy to implement by the existing software packages, the right approach in this situation is to consider all the binary variables in a multivariate setup and run the regression similar to what we do for univariate logistic regression. This would require complete specification of the joint distribution as discussed by Liang and Zeger [1]. An alternative approach generally advocated for this setup is **generalized estimating equations**. As discussed under that topic, this incorporates working correlations. Another approach is to run separate logistic for each dependent and make adjustment in the standard errors of the logistic coefficients for the possible correlation among the dependents. The third approach is to pool multiple binary outcomes into a single categorical outcome or some sort of score, and regress this on the regressors. For further details, see Lu and Yang [2]. The method of multivariate logistic regression is yet to evolve as a standard statistical tool and is rarely used.

1. Liang KY, Zeger SL. A class of logistic regression models for multivariate binary time series. *J Amer Stat Assoc* 1989;84:447–51. http://www.jstor.org/stable/2289928
2. Lu M, Yang W. Multivariate logistic regression analysis of complex survey data with application to BRFSS data. *J Data Science* 2012;10:157–73. http://www.jds-online.com/file_download/347/JDS-1040.pdf

## multivariate regression

In contrast to the ordinary (univariate) **regression**, multivariate regression has a set of quantitative dependent variables in place of just one variable and another set of regressors. Take the example of height, weight, head circumference, and chest circumference of a child dependent on parity, years of education of the mother, and per capita income of the family. Thus, there are four dependent and three independent variables in this case. All are quantitative. As an extension, now consider a general situation where there are $J$ independent variables denoted by $x_1, x_2, \ldots, x_J$ and $K$ dependent variables denoted by $y_1, y_2, \ldots, y_K$. In other words, the dependent vector has $K$ components. Let the $y$'s be correlated with one another so that they need to be considered simultaneously. The independent variables $x$'s may or may not be related to one another. In fact, it is desirable that the $x$'s are not highly correlated so that **multicollinearity** is not present. In any case, $y$'s are related to $x$'s; otherwise, there is no point in studying their relationship. The dependent set should preferably be the outcome or response of the independent set so that the relationship is plausible and interpretable. As before, the regression becomes multiple when the independent set contains more than one variable. Thus, a multivariate regression can be simple with one regressor or multiple with many regressors. This gives rise to the term *multivariate multiple regression*. This can be linear when all the **regression coefficients** are linear or could be nonlinear when these parameters are not linear. For simplicity, we restrict ourselves to multivariate multiple linear regression only in this section.

Multivariate multiple linear regression gives the same regression coefficients as the corresponding univariate regressions, and the regression equation does not change. Thus, there is no need for the multivariate method if the objective is to obtain regression equations only. However, if the objective is to check statistical significance of the regressions, then multivariate methods give the right answers for correlated dependents. This is explained next with the help of an example.

Consider the dependence of lung functions (forced vital capacity [FVC], forced expiratory volume in one second [$FEV_1$], peak expiratory flow rate [PEFR], and total lung capacity [TLC]) on age, height (Ht), and weight (Wt) in apparently healthy males of age 20–49 years. The data for a random sample of 70 subjects are given in the table in the topic **multivariate data/methods**. Part of the information is not available for two subjects. These two have been deleted casewise. The univariate and multivariate results for the remaining 68 subjects are as follows.

Regressions (univariate and multivariate are the same):

$$FVC = 0.67 - 0.0016(age) + 0.0182(Ht) + 0.0011(Wt)$$
$$FVV_1 = 3.07 - 0.0220(age) - 0.0073(Ht) + 0.0332(Wt)$$
$$PEFR = 5.04 + 0.0022(age) - 0.0173(Ht) + 0.0661(Wt)$$
$$TLC = -5.21 + 0.0122(age) + 0.0596(Ht) - 0.0006(Wt)$$

*Univariate tests of significance*: Univariate $P$-values (each lung function investigated separately to depend on three regressors, i.e., one dependent and three independent variables in the regression) are given in Table M.18.

**TABLE M.18**
***P*-Values for Univariate Regression in Our Example—Lung Functions Dependent on Age, Height, and Weight**

| Lung Function | *P*-Values | | | |
|---|---|---|---|---|
| | Age | Height | Weight | All Three Together |
| FVC | 0.899 | 0.250 | 0.924 | 0.580 |
| $FEV_1$ | 0.090 | 0.661 | 0.007 | 0.023 |
| PEFR | 0.939 | 0.632 | 0.014 | 0.071 |
| TLC | 0.576 | 0.037 | 0.975 | 0.128 |

The *scatter plot matrix*, which plots each dependent vs. each independent, is in Figure M.10. No trend is immediately discernible, although some *P*-values in Table M.18 are less than 0.05. The plot also indicates that the independents chosen in this study do not have much predictive power. A discerning eye can pick up subject number 65 with age 52 years in the data table, which should be between 20 and 49 years as stated for this example. This looks like a data entry error about which a caution should have been exercised.

First consider the univariate simple regression results when Ht, Wt, and age are the individual regressors on individual lung functions. At $\alpha = 0.05$, Ht has a statistically significant effect on TLC ($P = 0.037$) but not on any other lung function (Table M.18). On the other hand, Wt has a significant effect on $FEV_1$ ($P = 0.007$) and PEFR ($P = 0.014$), but not on FVC ($P = 0.924$) and TLC ($P = 0.975$). When Ht, Wt, and age are considered together, they have significant effect on $FEV_1$ ($P = 0.023$) and marginally on PEFR ($P = 0.071$), but not on FVC ($P = 0.580$) and TLC ($P = 0.128$). In view of these conflicting results, what sort of conclusion can be drawn about the effect of Ht, Wt, and age on lung functions as a whole? The answer is given by the multivariate results.

*P*-values under multivariate setup (four lung functions together investigated to depend on three regressors, i.e., four dependent and three independent variables in the regression) based on **Wilks Λ** are given in Table M.19.

Ht does not have a statistically significant ($P = 0.304$) influence on the "vector" of lung functions when assessed by these four measurements together, but Wt has a significant ($P = 0.024$) influence. Such a conclusion has less than 5% overall chance of being wrong. The multivariate method allows one to draw a conclusion on the basis of a joint Type I error for all four measurements of lung functions together. Multivariate multiple regression results for all the three independents together show that their joint effect on lung functions is not statistically significant ($P = 0.056$) if the maximum tolerable error of Type I is 0.05. If the level of significance is relaxed to 0.06, the conclusion in the multivariate setup is that the lung functions are

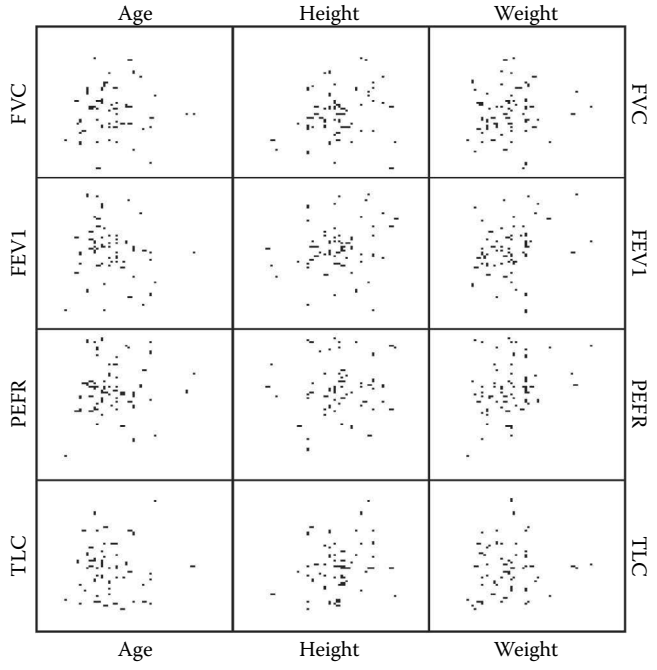



**FIGURE M.10** Scatter plot matrix of four dependent and three independent variables.

**TABLE M.19**
***P*-Values in Multivariate Setup for Our Example**

| | *P*-Values | | | |
|---|---|---|---|---|
| | **Age** | **Ht** | **Wt** | **All Three Together** |
| Lung functions (all four together) | 0.463 | 0.304 | 0.024 | 0.056 |

significantly affected. Therefore, it can be concluded with relaxed level of significance that age, Ht, and Wt together do influence the lung functions on the whole in the population from which this sample was drawn. If a threshold of 0.05 for Type I error is used strictly, the conclusion is that of nonsignificance. The cautious conclusion is that the effect of age, Ht, and Wt on lung functions is *marginally significant*. In any case, the conclusion is different from the one obtained by univariate regressions.

The requirement of external validation as for univariate methods applies to the multivariate methods as well. In the case of multivariate multiple regression, for example, the model may be an excellent fit to the data on which it is based, but the evidence of its utility on new subjects comes when the model is tested on a new set of data.

## mutually exclusive and exhaustive categories, *see also* multiple responses

Categories arise naturally for **nominal** and **ordinal** data but are many times also created for **continuous** data for convenience as explained under the topic **categories of data values**. Categories are called mutually exclusive when only one of them is applicable to one subject. While measuring body mass index (BMI), categories such as –l4, 15–24, 25–34, and 35+ $kg/m^2$ are mutually exclusive because a person's BMI can be in only one of these categories. These are exhaustive too because no BMI can be beyond these categories. In contrast to this, complaints in a patient are not mutually exclusive, nor are diseases. One person can have multiple diseases at the same point in time. Even if you list more than 100 diseases, they may still not be exhaustive as someone may have a disease not listed. A remainder category such as "others" can take care of this problem.

An important implication of mutually exclusive and exhaustive categories is that the probability of each category can be easily obtained and the sum of these probabilities is 1. This is the setup where multinomial distribution can be rightly used. Addition **law of probability** is applicable to mutually exclusive categories, although it does not require exhaustive categories. That is, in this case, the probability can be added.

Mutually exclusive and exhaustive categories have a special place in surveys. If the question is on use of tobacco, the person may say that he/she smokes a cigarette as well as chews a tobacco. These are not mutually exclusive and are not exhaustive either as the person smoking a cigar may also be using snuff, etc. Thus, the questionnaire should be structured to include all possible responses. It is customary in survey questionnaires to have the last category as "any other" so that nothing is left out.

Mutually exclusive and exhaustive categories are easy to analyze because the probabilities sum to 1. For regression analysis, such categories can be easily identified by indicator variables. This cannot be done for multiple response variables, and they need separate methods. Contingency tables and, consequently, chi-square are applicable to mutually exclusive and exhaustive categories, and not for multiple responses.

# N

## national well-being

National well-being is the average personal well-being of the population of a country. Well-being is a concept closely linked to health but goes beyond physical health and economic indicators. It considers health in its holistic form, which includes physical, mental, and social health. The mental health component dominates in this concept and includes aspects such as people's cognitive and affective evaluations of their lives. These valuations are mostly self-rated instead of being objectively measured and depend on how worthwhile or satisfied we rate our lives and our happiness, feelings of anxiety, and mental well-being [1]. The aim of this assessment is to provide a fuller picture of how society is doing by supplementing existing economic, social, and environmental measures [1]. A national index may help inform the society about balancing work with recreation and about nurturing relationships.

In its most simplistic form, one question, "How do you feel about your life as a whole?," with a rating on a 7-point response scale ranging from delighted to terrible can give you an answer about the self-rated feeling of well-being of the person [2]. Generally, though, a validated questionnaire consisting of several questions on physical, social, and mental aspects is administered to a random sample of adults. They can be asked to provide a rating of how yesterday was for them. The total response is scored on, say, a 0 to 100 scale, and the average provides an idea of the national well-being.

There are risks associated with the results of this kind of exercise. Too much satisfaction that will give a high score on well-being may leave people unmotivated, but the greater risk comes from its volatility as the score can quickly change with the mood of the person at the time of the interview. In addition, the score given by the young can be very different from the response given by the elderly because of their widely divergent experience. These problems occur with almost any subjective scoring system, but it still provides a baseline to work on and to compare different segments of the society and different time points.

The Office of National Statistics, United Kingdom, is trying to develop a measure of national well-being [1]. Efforts are made periodically to combine various indicators and come up with a comprehensive measure of well-being in its holistic sense. For example, there may be indexes that measure how people feel in control of themselves, make choices, and have a sense of purpose and belonging. However, an index that measures well-being in its entirety is not yet available. There is ongoing debate regarding what such an index should contain.

The **human development index** and the **quality of life index** are two of some of the currently more popular indexes of well-being. The former is computed each year for almost all nations of the world by the United Nations Development Programme after combining expectation of life, education, and income indicators. The quality of life index is popular for individual assessment, particularly for patients suffering from chronic ailments. In the United States, there is the Community Need Index that looks at economic and educational needs and identifies communities with barriers to access of resources [3]. A national well-being index is a work in progress and may produce worthwhile results shortly in the future.


1. Office of National Statistics. *Measuring National Well-being*. http://www.ons.gov.uk/ons/guide-method/user-guidance/well-being/index.html
2. Diener E. Subjective well-being: The science of happiness and a proposal for a national index. *Am Psychol* 2000;55(1):34–43. http://mina.education.ucsb.edu/janeconoley/ed197/documents/Dienersubjectivewell-being.pdf, last accessed March 23, 2015.
3. Econometrica, Inc., Eggers FJ. *Research to Develop a Community Needs Index*. U.S. Department of Housing and Urban Development, 2007. http://www.huduser.org/portal/publications/comm_index.pdf


## natural clusters, *see* cluster analysis

## natural experiments

When a rare or unique opportunity is available to observe or to study the effects of specific events as a result of naturally occurring changes, it is called a natural experiment. A natural experiment is a kind of **medical experiment** that occurs in nature without being planned for. To be called an experiment, you need a hypothesized cause that can be manipulated and an anticipated effect. The experiment is, by necessity, speculative (otherwise, there would be no need for an experiment in the first place) and you must be able to anticipate the outcomes that can be subsequently observed. All these conditions are fulfilled by a natural experiment although in a restricted sense. For example, the cause is not manipulated by humans but by nature. Statistically, these causes or precipitation factors are the independent variables and the outcome is the dependent variable. The group assignments in natural experiments are not strictly random but mimic random assignments in the sense that those receiving the intervention are not generally chosen on the basis of the population characteristics.

The tsunami of 2005 provided a rare opportunity to study the health consequences of such a disaster. The population affected was not much different from those unaffected in other coastal areas. The similar baseline in such comparisons helps in attributing the change to the intervening factor—in this case, the tsunami. Mining accidents, avalanche, and extreme weather conditions are also natural experiments. These help in studying the effect of such disasters. The effect of spraying insecticides on crops on human health or of genetically modified crops can be studied through natural experiments. The study of the effect of radiation therapy on the quality of life of cancer patients is also a natural experiment. To extend this further, the study of twins brought up in different environments can be a natural experiment. Whether or not there are different maternal complications in anemic or nonanemic women can be assessed by an experiment performed by nature since anemic and nonanemic women are already present and maternal complications among them can be observed. Similarly, the occurrence of goiter of various grades in areas with iodine deficiency in water where some are severely affected, some are mildly affected, and some not affected, is also a result of a natural experiment. John Snow's classical discovery that cholera is a waterborne disease was also the outcome of a natural experiment. None of these "interventions" is intentionally made by man.

Natural experiments have strength since they occur in natural settings, which makes the results more realistic. Participants do not know that they are being studied; thus, there is no **Hawthorne effect**. Such experiments provide an opportunity to study those interventions that are harmful and ethically not acceptable in human experiments. However, because of no control over extraneous factors, natural experiments may provide contaminated results. They may not provide a cause–effect relationship that a planned experiment would do. Nevertheless, they remain a useful tool to answer a research question that cannot be answered in any other way because of ethical problems in hazardous interventions.

Epidemiologically, natural experiments generally come under the domain of **observational studies**. Experimental evidence, even if from natural experiments, is generally more compelling than that available from the usual observational studies.

## negative binomial distribution

This is the statistical **distribution** of the number of independent "trials" needed to get $K$ "successes" and relates to **inverse sampling**. Suppose a hospital wants to set up an operation theater that can handle two open heart surgeries in a day. How many heart patients are needed on average per day to get two patients requiring open heart surgery? If this number is denoted by $x$, this would follow a negative binomial distribution. In this example, getting a patient requiring open heart surgery is "success" and the number of heart patients required to get two such patients is the number of "trials." Contrast this with **binomial distribution** where the number of trials $n$ is fixed and the interest is in the number of successes in those $n$ trials. In the negative binomial distribution, the number of trials is a variable and denoted by $x$ and the number of successes is fixed in advance. The distribution of $x$ obviously depends on the probability of success in any trial. If this probability is denoted by $\pi$, the probability that you need $x$ independent trials to get $K$ successes is given by

$$\text{Negative binomial distribution } (x) = \frac{(x-1)!}{(K-1)!(x-K)!}\pi^K(1-\pi)^{x-K},$$

$$x = K, K+1, \ldots,$$

where $x$ has no limit—it can theoretically reach infinity, which in our example means that you have to wait eternally to get two patients requiring heart surgery. However, the chance of this happening is almost zero. The factorial sign (such as $a!$) in this expression means the product of integers up to and including $a$, and $0! = 1$. The negative binomial distribution is sometimes expressed for the number of failures before $K$ successes, in which case $x$ is replaced with $y = x - K$, and now $y$ can be 0 as well.

In our example, if the probability of heart patients coming to that hospital requiring open heart surgery is $\pi = 0.06$ (i.e., on average, 6% of such patients require open heart surgery), the probability that you may have to wait for $x = 10$ patients to get $K = 2$ patients requiring open heart surgery is (using negative binomial distribution)

$$P(x = 10 \text{ for } K = 2) = \frac{(10-1)!}{(2-1)!(10-2)!})0.06^2(1-0.06)^{10-2};$$

$$= 0.02.$$

This shows that there is only 2% chance that you will have to wait for exactly 10 patients. This is nearly one-fifth of the binomial probability of getting 2 such patients out of a fixed 10, since this is $\frac{10!}{(10-2)!2!}\,0.06^2\,(1-0.06)^{(10-2)} = 0.099$. The difference is in the constant upfront in the two equations. The utility of a negative binomial is mostly in the cumulative probability; for example, the probability that you will have to wait for a maximum of 10 patients to get 2 successes is $P(x \le 10) = 0.12$. This chance is not high and says that most likely you may have to wait for more than 10 patients.

An estimate of the population proportion $\pi$ is $p = \frac{K-1}{n-1}$, where $K$ is the number of subjects with the characteristic you wish to have in the sample and $n$ is the size of the sample that had to be scanned to get $K$ subjects of the desired type. This is valid when the chance of getting a "success" remains the same with each subject and the sample subjects are independent of one another. The standard error (SE) of this estimate is given by

$$\text{SE } (p\text{: negative binomial distribution}) = \frac{\pi(1-\pi)}{n-2},$$

which can be estimated by replacing $\pi$ with its estimate.

A negative binomial distribution is sometimes used for regression models where the outcome is the count such as the relation between the number of teeth lost and the smoking intensity studied by Similä and Virtanen [1] for a Finnish birth cohort. Asiki et al. [2] explored the factors associated with the number of adverse pregnancy outcomes using negative binomial distribution.

When $K = 1$, this becomes, what is called, a *geometric distribution*. This gives the probability of waiting for $x$ trials to get one success. For its application to ocular chlamydial infection, see Lietman et al. [3].


1. Similä T, Virtanen JI. Association between smoking intensity and duration and tooth loss among Finnish middle-aged adults: The Northern Finland Birth Cohort 1966 Project. *BMC Public Health* 2015 Nov 17;15(1):1141. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4650303/
2. Asiki G, Baisley K, Newton R, Marions L, Seeley J, Kamali A, Smedman L. Adverse pregnancy outcomes in rural Uganda (1996-2013): Trends and associated factors from serial cross sectional surveys. *BMC Pregnancy Childbirth* 2015 Oct 29;15:27. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4627380/
3. Lietman TM, Gebre T, Abdou A, Alemayehu W, Emerson P, Blumberg S, Keenan JD, Porco TC. The distribution of the prevalence of ocular chlamydial infection in communities where trachoma is disappearing. *Epidemics* 2015 Jun;11:85–91. http://www.sciencedirect.com/science/article/pii/S1755436515000365


## negative trials

These are trials that do not substantiate the presence of an anticipated effect. This can happen for a variety of reasons: (i) the **effect size** is smaller than the study was planned for, or the effect is not present at all; (ii) the study is not sufficiently powered to detect the stipulated effect (the sample size is smaller than required for statistical significance); (iii) the study is not properly focused—there are too many variables in that study so that the combined conclusion is negative; (iv) the study is not properly planned and **bias** has occurred either at the planning, execution, analysis, or interpretation stage (which generally works in favor of a positive result but can inadvertently work against it).

**TABLE N.1**
**Duration of Taking Tranquilizer in the Support Group and the Conventional Group**

| | Tranquilizer Support Group | Conventional Management Group | Total |
|---|---|---|---|
| Still taking tranquilizer after 16 weeks | 5 | 10 | **15** |
| Stopped taking tranquilizer by 16 weeks | 10 | 5 | **15** |
| **Total** | **15** | **15** | **30** |

Freiman et al. [1] reexamined reports of 71 negative trials to determine whether a sufficiently large sample was studied. They concluded that 50 of these trials had a greater than 10% risk (**Type II error**) of missing a true 50% therapeutic improvement. This happened because the size of the sample was not sufficiently large. Dimmick et al. [2] reported similar findings for surgical trials. The sample size must be adequate to inspire confidence that a medically relevant difference would not go unnoticed. The following example explains this problem, although this is for the proportions and not for the means.

Table N.1 contains results of a trial in which patients receiving a regular tranquilizer were randomly assigned to continued conventional management and a tranquilizer support group. The null hypothesis is that the two groups are similar. Under this $H_0$, the expected frequency in each cell is $15 \times 15/30 = 7.5$. Since no expected frequency is less than 5, a chi-square analysis can be safely applied. When the **Yates correction for continuity** is applied, $\chi^2 = 2.13$ and $P > 0.05$ at 1 df. Note that the number of patients who stopped taking the tranquilizer in the support group is twice of that in the conventional group, yet the difference is not statistically significant. There is a clear case of a trial on a larger $n$. If the same type of result was obtained with $n = 30$ in each group, then the difference would be statistically significant. However, in many situations, the kind of effect needed to change our practice is just not present and, in this case, an increase in sample size will not help.

Many journal editors are much too keen to publish reports that give a positive result regarding the efficacy of a new regimen compared with the negative trials that did not find any difference. This is called the **publication bias**. In a straw poll of the published reports, those with positive results would hugely outscore those with negative results. Yet, the fact of the matter might be just the reverse. On the other hand, too many negative trials because of the reasons enumerated earlier can provide misleading results. Kurt et al. [3] have discussed how negative trials on lipoprotein(a) may have severely underestimated its potential role in atherosclerosis. Similar observations can be made for some other results.

BioMed Central publishes a large number of open-access journals. Among these is the *Journal of Negative Results in BioMedicine*. This is in realization of the need of professionals to know about negative trials as much as about positive trials.

1. Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: Survey of 71 "negative" trials. *N Engl J Med* 1978;299:690–4. http://www.ncbi.nlm.nih.gov/pubmed/0000355881
2. Dimmick JB, Diener-West M, Lipsett PA. Negative results of randomized clinical trials published in the surgical literature: Equivalency or error? *Arch Surg* 2001;136:796–800. http://archsurg.jamanetwork.com/article.aspx?articleid=391750
3. Kurt B, Soufi M, Sattler A, Schaefer JR. Lipoprotein(a)—Clinical aspects and future challenges. *Clin Res Cardiol Suppl* 2015 Apr;10(Suppl 1):26–32. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4361767/

## neonatal mortality rate, *see* mortality rates

## nested case–control studies, *see* retrospective studies

## nested designs, *see* hierarchical designs

## net reproduction rate, *see* fertility indicators

## NNT, *see* number needed to treat

## *n*-of-1 trials

An *n*-of-1 is a clinical trial with just one participant. The disease under investigation may be rare or unusual such as orthostatic hypotension or narcolepsy, for which enough cases are not available for conducting a full-blown **randomized controlled trial** (**RCT**). The treatment under investigation might work only in a very specific type of cases precluding its use in RCT groups. In such situations, a trial on a single patient can be considered for evaluating individual efficacy though not for generalized conclusion. Such trials can be conducted only for chronic stable conditions with reversible symptoms and quantifiable outcome.

An *n*-of-1 trial is a variation of a **crossover** strategy where several pairs of treatment periods are used. The test regimen is given in one period and the control treatment is given in the other period in a crossover trial so that the patient serves its own control. The order of these two treatments is randomized within each pair of periods separately in a crossover trial. *n*-of-1 is an extension of this methodology where there is just one patient but repeatedly administered a treatment, and is possible only when the patient is hugely willing to cooperate.

Obviously, a trial is done when efficacy is in doubt. This is always the case for a new regimen. If a patient is reluctant to comply with the existing regimen because it is not effective in his or her case or because of side effects, he/she might agree to participate in an *n*-of-1 trial. This trial provides a unique opportunity to test a new regimen for a particular patient.

As in the case of crossover trials, a suitable regimen for an *n*-of-1 trial is the one that can be rapidly started and stopped. There should not be any carryover effect, and the disease must come back to its original severity after the washout period. A minimum of three crossover pairs are advised. Blinding helps minimize patient and observer bias and should be used wherever feasible.

Nixdorf et al. [1] have used this kind of trial for stellate ganglion blocks on a patient with orofacial pain who was not responding to conventional treatment. Yuhong et al. [2] have described an *n*-of-1 trial of a Chinese medicine Liuwei Dihuang decoction versus placebo on 47 patients, each completing three pairs of periods. Since this is not restricted to one patient, the target is not one patient but to come up with a more general conclusion.

1. Nixdorf DR, Sobieh R, Gierthmühlen J. Using *N*-of-1 trials in clinical orofacial pain management. *J Am Dent Assoc* 2012 Mar;143(3):259–61. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3713805/
2. Yuhong H, Qian L, Yu L, Yingqiang Z, Yanfen L, Shujing Y, Shufang Q, Lanjun S, Shuxuan Z, Baohe W. An *n*-of-1 trial service in clinical practice: Testing the effectiveness of Liuwei Dihuang Decoction for kidney-Yin deficiency syndrome. *Evid Based Complement Alternat Med* 2013;2013:827915. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3794636/

## nominal categories, *see* categories of data values

## nominal scale, *see* scales of measurement (statistical)

## nomogram

A nomogram consists of a set of lines each representing a continuous variable plotted in such a way that the value of one variable can be read in relation to the values of two or more other variables (like a slide rule). The lines are rescaled so that corresponding values of the variables lie on a straight line and the values can be read using a simple ruler, thus obviating the need to do calculations.

For example, Figure N.1 shows a nomogram designed to find the number of clusters (C-lines) required in a **cluster sampling** for a survey to estimate the prevalence of a disease with specified precision ($L$ as percentage of $P$, where $P$ is the prevalence) and specified confidence level ($1 - \alpha$) when a guesstimate of prevalence ($P$-line) is used and the cluster size or the **design effect** is known [1]. The design effect ($D$) and the cluster size ($B$) is measured in terms of the ratio $D/B$. This nomogram was used for rapid surveys to assess the prevalence of senile cataract blindness in a community. As an illustration, a cross-sectional line is drawn in the figure, which shows that when the guesstimate of the prevalence of blindness attributed to senile cataract is $P = 0.04$, the ratio of design effect to cluster size ($D/B$) = 0.05, and then the number of clusters required is nearly 460 for 95% confidence for estimating prevalence within one-tenth of its value ($L = 10\%$ of $P$). The notation ($1 - \alpha$) is for the level of confidence.

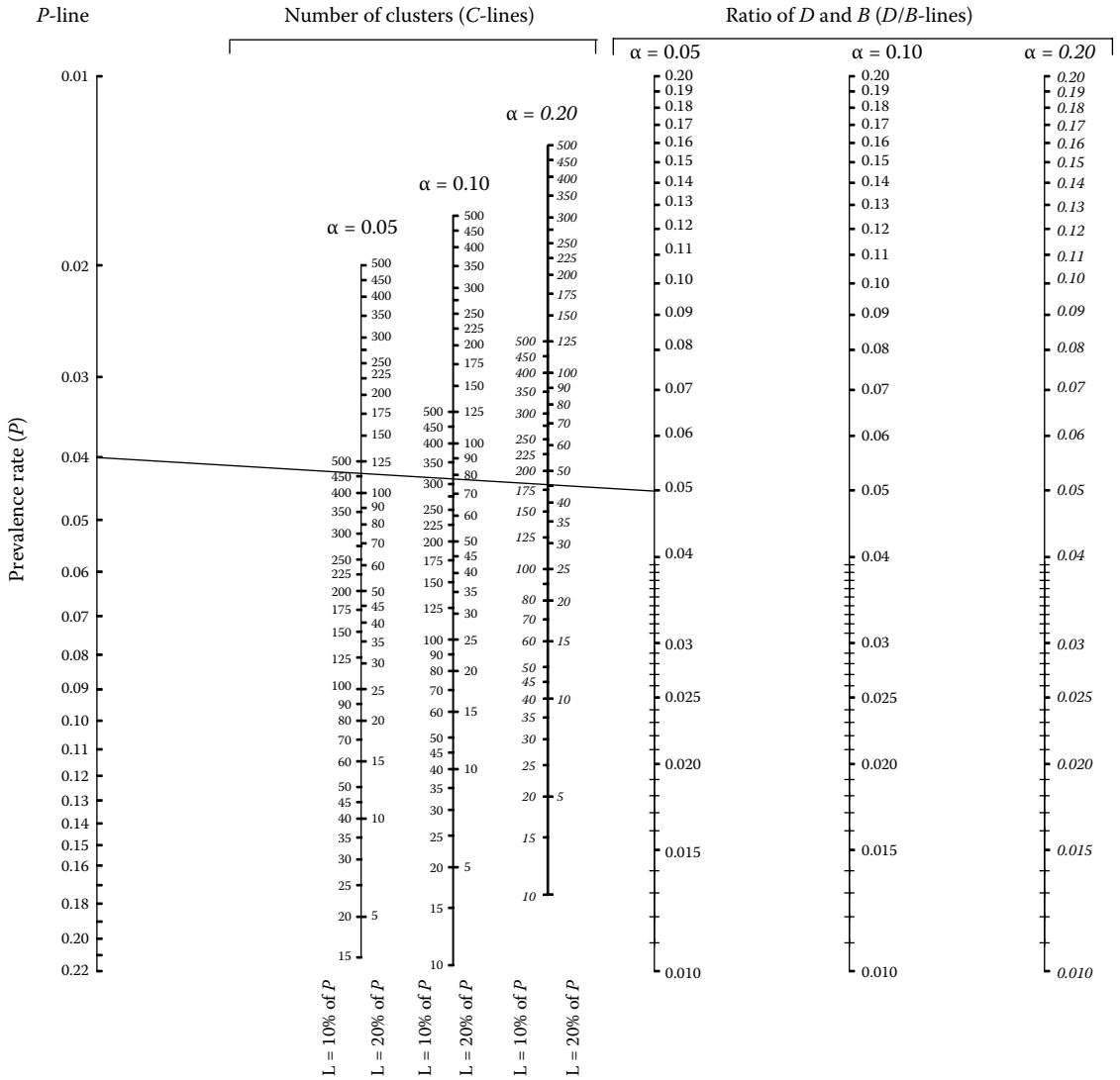



**FIGURE N.1** Nomogram: number of clusters required for rapid assessment of prevalence of a disease for different cluster sizes. (From Kumar A, Indrayan A. *Int J Epidemiol* 2002;31:463–7. http://ije.oxfordjournals.org/content/31/2/463.abstract.)

Such nomograms are used in several different applications. Lawoyin and Onadeko [2] developed a nomogram to predict the birth weight category of babies born at term on the basis of the maternal weight changes at different periods of gestation. Malhotra and Indrayan [3] constructed a nomogram for sample size for estimating sensitivity and specificity of medical tests with specified precision. Partin et al. [4] developed a nomogram based on clinical stage, Gleason score of the prostate needle biopsy, and serum prostate-specific antigen. This was designed to improve the ability to predict the pathologic stage of a prostate tumor.


1. Kumar A, Indrayan A. A nomogram for single-stage cluster sample surveys in a community for estimation of a prevalence rate. *Int J Epidemiol* 2002;31:463–7. http://ije.oxfordjournals.org/content/31/2/463.abstract
2. Lawoyin TO, Onadeko MO. A nomogram for screening low birth weight and large for gestational age babies for use in primary health care centres. *East Afr Med J* 1993;70:746–8. http://www.ncbi.nlm.nih.gov/pubmed/8026344
3. Malhotra RK, Indrayan A. A simple nomogram for sample size for estimating sensitivity and specificity of medical tests. *Indian J Ophthalmol* [serial online] 2010;58:519–22. http://www.ijo.in/article.asp?issn=0301-4738;year=2010;volume=58;issue=6;spage=519;epage=522;aulast=Malhotra
4. Partin, AW, Yoo J, Carter HB, Pearson JD, Chan DW, Epstein JI, Walsh PC. The use of prostate specific antigen, clinical stage and Gleason score to predict pathological stage in men with localized prostate cancer. *J Urol* 1993;150:110–4. http://www.ncbi.nlm.nih.gov/pubmed/7685418


## non-Gaussian distributions, *see* distributions (statistical)

## noninferiority test, *see* equivalence, superiority, and noninferiority tests

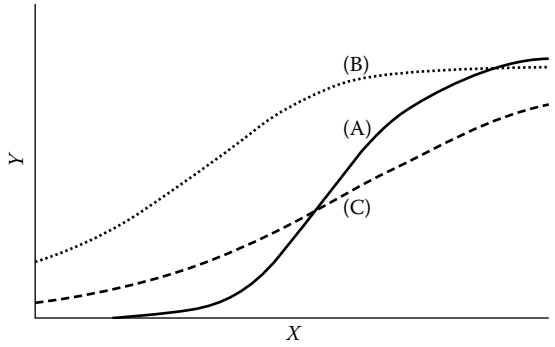
## noninferiority trials, *see* equivalence and noninferiority trials

## nonlinear regression, *see also* linear regression, curvilinear regression

To understand nonlinear regression, first refer to **linear regression** if you are not familiar with the basics. This is given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \varepsilon,$$

where $x_1$, $x_2$, …, $x_K$ are the regressors and β's are the **regression coefficients**. These are the unknown parameters of the regression model that we wish to estimate by the regression analysis. The term ε is the difference between the observed value of $y$ and its regression-predicted value. All the regression coefficients are linear—they would not be linear when, say, $\beta_2 = \beta_1^2$ or $\beta_2 = \log(\beta_1)$, or any other nonlinear function. Statistically, when any regression coefficient is nonlinear and cannot be converted to linear by any transformation, it is called nonlinear regression. We later give examples of regression that look nonlinear (graphically, they yield a curve instead of a line) but are intrinsically linear in the statistical sense. For some examples of nonlinear regression, see Figure N.2 in this section and the figures in the topic **regressions (types of)**.




(A) $\beta_1 = 1$, $\beta_2 = -3.22$, $\beta_3 = 8$, $\beta_4 = 3.33$
(B) $\beta_1 = 0.95$, $\beta_2 = -4.61$, $\beta_3 = 10$, $\beta_4 = 0.33$
(C) $\beta_1 = 1$, $\beta_2 = -2.3$, $\beta_3 = 4$, $\beta_4 = 1.25$

**FIGURE N.2** Concentration of drug at time $x$ after injection with different choices of the values of the parameters in nonlinear regression represented by Equation A. (From Gravbill FA, Iyer HK. Nonlinear regression, Chapter 9, in: *Regression Analysis: Concepts and Applications.* Colorado State University. http://www.stat.colostate.edu/regression_book/chapter9.pdf. With permission.)

Nonlinear models are difficult to fit—they are not fitted by the method of **least squares** that we commonly use for linear regressions but require an *iterative procedure*. The software asks you to specify the model and sometimes to supply your initial set of guess estimates of the parameters. It does calculations by a mathematical algorithm using guess estimates, calculates residual sum of squares, and presents adjusted estimates that are better fit. This is one iteration. Now, these adjusted estimates are inserted and another set of values that are better fit to the observed values is obtained. This completes the second iteration. These iterations go on till such time that the estimates obtained at any iteration and subsequent iteration are almost the same. This is called *convergence*. These are the best estimates from the available data though not necessarily good enough to provide a good fit to the data. It is a good idea to start with another set of initial guesses and see whether or not the same final estimates are obtained.

In some cases, convergence does not occur and the software stops after making reasonable attempts. In this case, the software will give you the message that the there is no convergence. This can happen either because of insufficient data or because of improper specification of the model.

Nonlinear models are complex and should be used only when linear models fail to provide a good explanation or a good fit to the data, particularly when stakes are high and it is necessary to get a more accurate model. Suppose the interest is in the 24-h pattern of concentration of a drug in blood after it is injected. This is expected to depend on a host of parameters such as the liver functions of the patient, kidney functions, age, and sex. One such model can be

$$\text{(A)} \quad \text{Concentration: } y = \frac{\beta_1}{1 + e^{-(\beta_2 + \beta_3 x)^{\beta_4}}} + \varepsilon,$$

where β's are the parameters such as various liver and kidney functions that determine the concentration of the drug in blood at time $x$ after injection. Note that the Equation A cannot be modeled as linear by any transformation and is a genuinely nonlinear regression. Depending on the values of these parameters, the drug concentration can follow one of many possible patterns; three of these for some specific values of β's are shown in Figure N.2. If the drug

is such that it starts at zero concentration, rapidly rises, and then stabilizes, curve (a) as per the values of the parameters shown at the bottom of the figure may be a good representation. If the drug is expected to have some low level immediately after injection and rising gradually, curve (c) can be chosen. If the drug is expected to have substantial level immediately on injection and rising gradually and then leveling off, curve (b) may be adequate. Other values of parameters will give other shapes, and the software packages can be asked to find estimates of the parameters that best fit the data.

The adequacy of a nonlinear fit can be assessed in some situations by the **coefficient of determination** ($\eta^2$), which is obtained as the regression sum of squares as a proportion of the total sum of squares. Square of the multiple correlation coefficient ($R^2$) is for linear regression and should not be used for nonlinear regression. A good statistical software package will give the value of $\eta^2$ and will also give the results of test of hypothesis regarding the parameters and the confidence intervals. In other setups, residual plots give a good idea of the adequacy of the regression. This plot should be random with small variance. For details, see Gravbill and Iyer [1].

Chen et al. [2] have discussed nonlinear growth trajectory of brain, and Chin et al. [3] have studied the relationship between aortic valve area and mean pressure gradient measurements with nonlinear regression with implications for patients with small-area low-gradient aortic stenosis.

Examples of the so-called nonlinear regressions that give a curved shape but are intrinsically linear are mostly curvilinear, as described under the topic **curvilinear regression**. All polynomial regressions are curvilinear. The exponential growth curve $y = ae^{bx}$, seen in the multiplication of organisms when not interrupted as in a laboratory, is apparently nonlinear but is also intrinsically linear since, in this case, $\ln(y) = \ln(a) + bx$, which takes the form $z = b_0 + bx$ with $z = \ln(y)$ and $b_0 = \ln(a)$. This is then a straight line but is between $\ln(y)$ and $x$. There are several other "nonlinear" forms of the equation that are intrinsically linear. They are not termed as nonlinear for statistical regression.


1. Gravbill FA, Iyer HK. Nonlinear regression, Chapter 9, in: *Regression Analysis: Concepts and Applications*. Colorado State University. http://www.stat.colostate.edu/regression_book/chapter9.pdf
2. Chen Y, An H, Shen D, Zhu H, Lin W. Tailor the longitudinal analysis for NIH longitudinal normal brain development study. *Proc IEEE Int Symp Biomed Imaging* 2014 May;2014:1206–9. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4232948/pdf/nihms583804.pdf
3. Chin CW, Khaw HJ, Luo E, Tan S, White AC, Newby DE, Dweck MR. Echocardiography underestimates stroke volume and aortic valve area: Implications for patients with small-area low-gradient aortic stenosis. *Can J Cardiol* 2014 Sep;30(9):1064–72. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4161727/


## nonparametric methods/tests (overview)

These methods are not concerned with **parameters** of the statistical **distributions** such as mean and variance but are concerned with other aspects such as shifts in the location of the distribution or association among the variables. The most commonly used nonparametric methods are the **chi-square test** and the **Wilcoxon tests**.

The term *nonparametric method* implies a method that is not for any specific **parameter**. The Student *t*-test, for example, is a parametric method because it is concerned with a parameter, namely, the mean. In the case of nonparametric methods, the hypothesis is concerned with the pattern of the distribution as in a **goodness-of-fit test** or with some characteristic of the distribution of the variable such as randomness and trend. More commonly, though, the interest would be in the location of the **distribution** without specifying the parameter. This is illustrated in Figure N.3, where the distributions are identical except for location. Location shift of the entire distribution is clear and it is not necessary to talk of mean or median or any such parameter in this case.

Nonparametric methods are sometimes also called **distribution-free methods**. These are methods that are based on functions of sample observations whose distribution does not depend on the form of the underlying distribution in the population from which the sample was drawn. The chi-square test is distribution free as it remains valid for any categorical data whether the underlying distribution of the variable is Gaussian (normal) or not. Although nonparametric methods and distribution-free methods are not synonymous, most nonparametric tests are also distribution free. Both categories of methods are generally considered as nonparametric methods. These methods derive their strength from transforming the values to ranks. Thus, they are applicable to ordinal data as well.

Nonparametric tests are preferred when the underlying distribution is far from Gaussian and, at the same time, $n$ is small. A small $n$ precludes the use of the **central limit theorem**, which allows the use of Gaussian distribution for many estimates. Nonparametric methods are especially suitable when outliers are present, which are genuine in the data set (not artifacts) and cannot be ignored. Outliers do not affect these methods because these methods mostly work with ranks. When Gaussian conditions are present, the performance of nonparametric methods is not as good as those of parametric methods such as the **Student *t*-test** and the **ANOVA *F*-test**, but they work well with non-Gaussian data.

Nonparametric methods are not as developed as parametric methods. If you have data more suitable for a nonparametric method for which such a method is not readily available, the recommended procedure is to use the parametric method on the actual data as well as on the rank-transformed data, which makes it nonparametric [1]. If the two methods give nearly identical results, you are done. If not, take a closer look at the data for outliers or for highly skewed distribution and use rank-transformed data for inference when such features are present.

Many researchers take Gaussianity for granted and use Gaussian-based parametric methods even for small samples. The tests for checking Gaussianity such as **Kolmogorov–Smirnov**, **Anderson–Darling**, and **Shapiro–Wilk** generally work well with large samples and not so much with small samples. Thus, small samples would not be able to provide evidence against Gaussianity because of lack of power. For small samples, try to have some extraneous evidence (from experience or literature) that the distribution is Gaussian or of any other form. Sometimes, plots give a good idea of the distribution pattern. If the plot suggests a non-Gaussian pattern, it might be safe to use nonparametric methods for small samples. On the other hand, some researchers are nonparametric enthusiasts and use these methods even when the sample size is large. This is not advisable as this could result in loss of power. Thus far, fortunately, the number of such enthusiasts is not high.

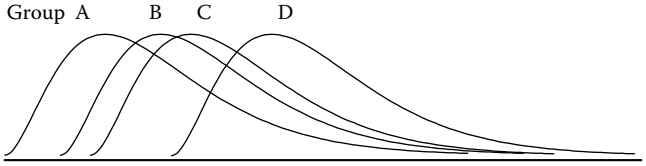



**FIGURE N.3** Distribution in four groups differing only in mean.

Besides the ones mentioned in this section so far, the other nonparametric methods discussed in this book are as follows:

- The **Sign test** for paired data to check if the two groups have the same location
- The **Wilcoxon signed-ranks test** for paired data when the magnitude and the direction of the differences are important
- The Mann–Whitney–Wilcoxon test for two independent samples (algebraically identical to the **Wilcoxon rank-sum test**)
- The **Kruskal-Wallis test** for the comparison of location of three or more groups
- The **Friedman test** for a two-way layout with one observation per cell

For further details of nonparametric methods, see Conover [1].

1. Conover WJ. *Practical Nonparametric Statistics*, Third Edition. Wiley, 1999.

## nonrandom sampling, *see* sampling techniques

## nonresponse

The inability to elicit full or partial information from the subject after inclusion in a study is termed nonresponse. This can happen as a result of the subject's noncooperation, relocation, injury, death, or any such reason. The opportunity for nonresponse is particularly present in follow-up studies, but even in the case of one-time evaluation, the subject may refuse to answer certain questions or may not agree to submit to a particular investigation or examination even when prior consent has been obtained. The problem of nonresponse can be particularly severe in mailed questionnaires and **telephone surveys**, primarily because of disinterest of the respondents.

Nonresponse has two types of adverse impacts on the results. The first is that the ultimate sample size available to draw conclusions reduces, and this affects the **reliability** of the estimates and the **power** to detect an effect. This deficiency can be remedied by increasing the sample size corresponding to the anticipated nonresponse. The second is more serious. If only 2200 respond in a sample of 3000 out of 1 million, the results could be severely biased. These responders could be conformers or those with strong views. If not biased, a sample of 2200 is not too bad to provide a reliable estimate or to test a hypothesis in most situations. Mostly, the nonresponding subjects are not random segments but are of specific type. For example, they may be seriously ill who do not want to continue in the study or very mild cases who opt out after feeling better, or some such peculiar cases. Their exclusion can severely bias the results. A way out is to take a subsample of the nonrespondents, undertake intensive efforts for their full participation, assess how these subjects in the subsample are different from the regular respondents, and adjust the results accordingly. A provision for such extra efforts to elicit responses from some nonrespondents should be made at the time of planning the study.

Experience suggests that some researchers fail to distinguish between nonresponse and zero value or the absence of a characteristic. Take care that this does not happen in the data you are examining as evidence for practice or in the data you are recording for research.

Elsewhere in this book, we have described methods such as **imputation for missing values** and **intention-to-treat analysis** that can partially address the problem arising from nonresponse, but no analysis, howsoever immaculate, can replace the actual observation. Thus, all efforts should be made to ensure that nonresponse is minimal if not altogether absent. Strategies for this should be devised at the time of planning the study, and all necessary steps should be taken to minimize nonresponse. A proper groundwork with the sample subjects such as explaining the benefits of the study may help in getting their cooperation. The investigations should be such that these cause least inconvenience or distress to the respondents and should have no adverse financial implications for them. In fact, it may be necessary in some cases to provide some sort of compensation to the subjects for their time and inconvenience so that they extend full cooperation.

## nonsampling errors

Nonsampling errors in the results are not due to not using sampling but are human errors that occur due to various types of biases in concepts, sample selection, data collection, analysis, and interpretation. These are the genuine methodological errors opposed to the **sampling errors** that in fact are not errors and occur due to variation among the subjects from sample to sample. Nonmethodological errors such as in diagnosis and medication are not included among nonsampling errors. Sampling errors are natural and random, whereas nonsampling errors are systematic and tend to destroy the study findings. Sampling errors reflect the **reliability** of the results, whereas nonsampling errors hit the **validity**—the study may not be answering the question you started with when substantial nonsampling errors are present.

Besides bias, which we will mention in the next paragraph, some prominent nonsampling errors are as follows:

- Unclear questions and not being able to communicate properly
- **Nonresponse**, **recall lapse**, and wrong response
- Wrong coding and wrong entry of data
- Inadequate analysis that does not consider all **confounders** and inappropriate use of statistical methods such as the **black-box approach**
- Interpretation based on partial analysis, or wrong interpretation of the computer output

Biases that result in nonsampling errors are a burning problem in most medical studies. For this reason, we have discussed these under several different topics in this book:

1. The topic **bias (overview)** enumerates and briefly describes more than 30 types of biases that can occur at different stages of study from the planning to arriving at the conclusion. Some of these such as **Berkson bias** and **Hawthorne effect** are discussed in detail as separate topics.
2. **Biased sample**.
3. Bias in **case–control studies, cross-sectional studies, and prospective studies**—discussed under these respective topics.
4. **Bias in literature review**.
5. **Bias in medical studies and their minimization**.
6. **Publication bias**.

Nonsampling errors are controlled by better thinking, better design, better methods, better training, and so on. It may be necessary to work intensively with the respondents to elicit correct answers. On the other hand, sampling errors are controlled mostly by increasing the size of the sample. A large sample with nonsampling errors tends

to aggravate bias instead of controlling, and a large sample in this case provides a false sense of security. Sampling errors in most cases are overt and can be easily handled with statistical methods but nonsampling errors can easily slip your attention. Often, they remain obscure for long till such time a new finding emerges that exposes previous wrong findings due to nonsampling errors.

For further details, see McNabb [1]. You may also like to explore the error matrix approach suggested by Keus et al. [2] for examining the validity of the available evidence including the nonsampling errors.

1. McNabb DE. *Nonsampling Error in Social Surveys.* Sage, 2013.
2. Keus F, Wetterslev J, Gluud C, van Laarhoven CJ. Evidence at a glance: Error matrix approach for overviewing available evidence. *BMC Med Res Methodol* 2010 Oct 1;10:90. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2959031/

## nonsense correlation, *see* correlation (the concept of)

## normal distribution, *see* Gaussian distribution

## normality (test of), *see* Gaussianity (how to check)

## normalization

The term *normalization* in statistics is used in two senses. First is normalization of values and the second is normalization of a database.



### *Normalization of Values*

Normalization of values is rescaling them on (0, 1) from minimum to maximum. All the values are rescaled accordingly. If you want to normalize values of a variable that ranges from $a$ to $b$, the value $a$ will be linearly rescaled to 0 and the value $b$ will be linearly rescaled to 1. Any value in between, say, $x$, will be rescaled as

$$\text{Normalized value of } x = \frac{x-a}{b-a}.$$

This will always be between 0 and 1. If you are measuring the serum urea of 230 persons and find that the minimum is 17 mg/dL and the maximum is 43 mg/dL, the normalized value of a person with a urea level of 24 mg/dL is $(24 - 17)/(43 - 17) = 0.27$. This gives an idea how far the value is from the minimum relative to the range. Normalized values are sometimes confused with **standardized values** that have a mean of 0 and a standard deviation (SD) of 1, and sometimes they are confused with normal values found in healthy subjects. Those are erroneous usages.

The purpose of normalization is to make values on different scales comparable. If the normalized value of the urea level of a person is 0.27 and the normalized value of the creatinine level of the same person is 0.65, you know that creatinine is higher relative to the urea level in this person. Both creatinine and urea are measured in milligrams per deciliter, but consider lymphocytes, which are measured in terms of percentage. If the normalized value of lymphocytes is 0.90, you know that it is relatively very high compared to the urea level. Normalized values do not have units, and that makes them comparable.

The second big advantage of normalized values, arising from the first, is that the values on different scales can be averaged. For some reason, if you are interested in finding the average level of serum urea, serum creatinine, and lymphocytes in the example just mentioned, the average normalized value is $(0.27 + 0.65 + 0.90)/3 = 0.61$. This may be absolutely meaningless for many health parameters but may be useful in isolated situations.

Kramer et al. [1] used normalized permutation entropy (PeEn) values to compare them in the sternal, mid-costal, and crural areas 1–81 days after phrenicotomy and concluded that electromyographic PeEn represents a new and distinctive assessment for characterizing intramuscular function after denervation and reinnervation. The United Nations Development Programme uses normalized values for indexes of education, life expectancy, and income and averages them to come up with the **human development index** for each country.

A big problem with normalization is its oversimplicity. A normalized value of 0.90 may look 20 points away from 0.70 but may not be so far away when mean, SD, or percentiles are considered. For this reason, they are seldom used for statistical conclusions.

### *Normalization of Database*

In the context of database, normalization is the process of efficiently organizing data in a database. It is a method to remove all the anomalies and bring database to a consistent state. This includes eliminating redundant data but preserving dependencies. When this is done, it can drastically enhance the capability to exploit the data without spending too much time and effort, besides saving the storage space.

Normalization of a database requires creating lots of tables and examining them for duplication and clarity. It involves hard work and some consider it a luxury. Perhaps the best strategy is to be careful at the designing stage itself so that no anomalies occur. Sometimes, they still occur as all of them can rarely be anticipated at the designing stage. Once the database starts to build up, some of these anomalies become apparent or can be perceived. It is a good idea to take preemptive action so that such instances do not recur. The design could be modified accordingly for future records and to change the existing records.

For details such as data normalization rules, see Ambler [2].

1. Kramer C, Jordan D, Kretschmer A, Lehmeyer V, Kellermann K, Schaller SJ, Blobner M, Kochs EF, Fink H. Electromyographic permutation entropy quantifies diaphragmatic denervation and reinnervation. *PLoS One* 2014 Dec 22;9(12):e115754. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4274091/
2. Ambler S. *Agile Database Techniques: Effective Strategies for the Agile Software Developer.* Wiley, 2003.

## normal probability plot, *see also* quantile-by-quantile (Q–Q) plot

Let us first explain the probability plot without involving the term *normal*. A probability plot is the graph of cumulative probabilities at different values of the variable. In the case of samples, the cumulative probabilities are replaced by cumulative relative frequencies. For example, if the variable is the body temperature of healthy subjects, the proportion of subjects with temperature less than 97.5°F, less than 98.0°F, less than 98.5°F, less than 99.0°F, and so on are plotted versus 97.5°F, 98.0°F, 98.5°F, and 99.0°F. Since these proportions are cumulative, subjects with a body temperature less than 98.5°F would contain those with temperature less than 98.0°F and that in turn would contain subjects with temperature less than 97.5°F and so on (Table N.2). The probability plot of the data in Table N.2 is in Figure N.4a although the data are based on cumulative relative

**TABLE N.2**
**Distribution of Body Temperature in 468 Healthy Subjects and the Cumulative Relative Frequencies**

| Body Temperature (°F) | Frequency | Cumulative Frequency | Cumulative Relative Frequency |
|---|---|---|---|
| <97.0 | 2 | 2 | 0.0043 |
| 97.0–97.4 | 17 | 19 | 0.0406 |
| 97.5–97.9 | 86 | 105 | 0.2244 |
| 98.0–98.4 | 152 | 257 | 0.5491 |
| 98.5–98.9 | 133 | 390 | 0.8333 |
| 99.0–99.4 | 65 | 455 | 0.9722 |
| 99.5–99.9 | 12 | 467 | 0.9979 |
| 100+ | 1 | 468 | 1.0000 |

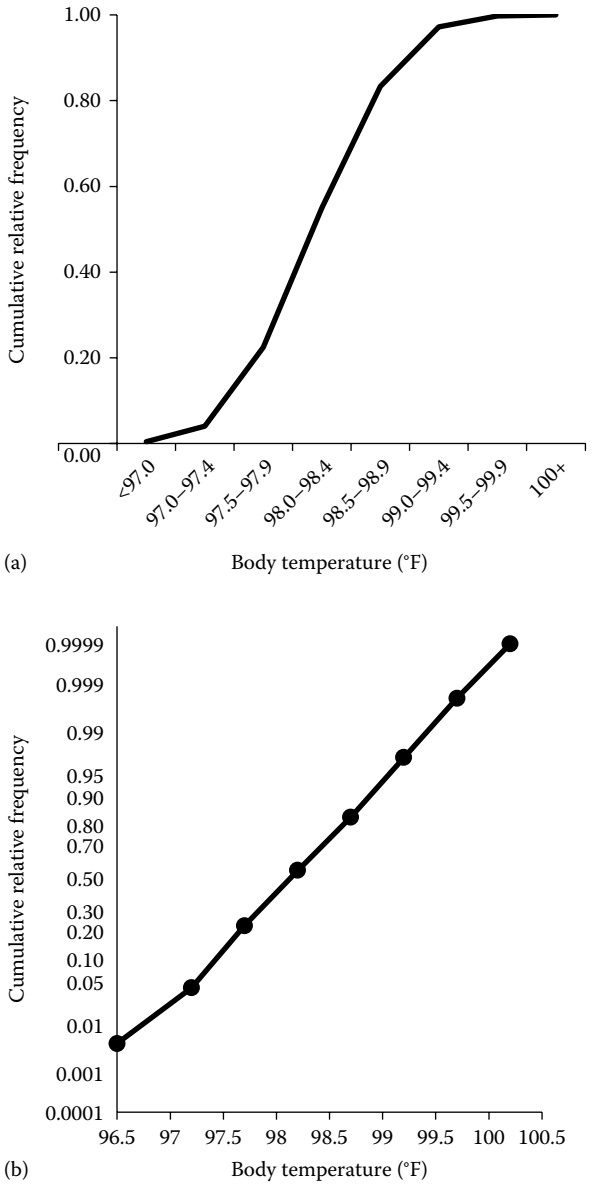



**FIGURE N.4** (a) Probability plot for the data in Table N.2. (b) Normal probability plot.

frequencies in place of the cumulative probabilities. In most cases, this will yield an S-shaped curve as shown in this figure.

Now, for a normal (Gaussian) probability plot, the *y* axis is compressed in the middle and stretched at both ends as per the probabilities in a Gaussian distribution (Figure N.4b). If the data follow a Gaussian distribution, the points will fall in a straight line when the scale on the *y* axis is the Gaussian probability. Sample values can show minor random variations. If the points do not fall in a straight line and follow some other systematic pattern, you have evidence that the distribution of values is not Gaussian. For example, you can have points at both ends away from the straight line. Thus, the normal probability plot is a graphical technique for assessing whether or not a data set has approximate Gaussian distribution, but this assessment could be very subjective in the absence of a test of significance. Tests such as **Kolmogorov–Smirnov** and **Anderson–Darling** are numerical but require a relatively larger sample.

We have illustrated the method by classifying the data into class intervals (Table N.2). The method can be used directly on the values without categorizing them into intervals. For this, the values are first rearranged in ascending order and **quantile** values are obtained using a normal inverse function that gives the Gaussian probability. See the topic **Q–Q plot** for details.

## normal range of medical parameters

The term *normal* is used in medicine with several different meanings. It is normal for a 70-year-old person to have myopia, and it is normal for women undergoing chemotherapy for breast cancer to lose hair. Murphy [1] has given a very interesting discussion of the different uses of this term in medical literature. A moot question is whether normal is the same as ideal or optimal. If yes, how does one define optimal? It is possible that a person with a 37.5°C body temperature and a 149/92 mmHg blood pressure (BP) does exceedingly well when accompanied by other corresponding physiologic changes. Who knows!

The normal level of a quantitative measurement in health can be defined in many ways. Most will agree that normal values are those that are generally seen in healthy subjects. However, each individual has his own normal (call it self-normal) in a healthy condition and, whenever possible, the evaluation of the current condition should be made against the value normally present in that person in a healthy state. If a person is known to have a BP of 110/70 mmHg when healthy, then a level of 130/80 mmHg would be considered a definite rise, sufficient to put the attending clinician on alert for contemplating some action despite the raised level being in the healthy range. For example, this can happen in pregnancy-induced hypertension.

The range of plasma levels of most hormones in healthy subjects is wide. As a consequence, the level of a hormone in an individual may be halved or doubled (and thus be grossly abnormal for that person) but still be within the so-called normal range. However, the healthy level of a patient coming to a clinic for the first time is seldom known. In such cases, the patient's level is evaluated against the levels generally seen in healthy subjects in that population. Call these population-normals that serve the purpose of reference values. Thus, **reference values** are the levels generally seen in healthy subjects in a population. Such references can also be used to delineate the maximum allowable variation in a subject even when his/her own healthy levels are known. In the example just cited, because a raised BP of 130/80 mmHg is still well within the normal limits for the other healthy subjects, therapy may not start unless the complaints are severe. Most likely, there would be no such complaints. Yet a big rise is surely enough to put you on alert.

### How to Establish Normal Range

Normal values or reference values are based on measurements of healthy subjects, preferably the healthiest segment of the population. The following are the criteria for normals to be reliable:

- They should be based on a sample that includes at least 200 individuals in each group (say, 200 males and 200 females) if the group-wise reference values are required. A smaller sample would be adequate only when the inter-individual variability is really small.
- There should not be any outliers in the data.
- The sampling procedure must be scientific so that the sample indeed represents the entire spectrum of values present in the group.
- Measurements should be carefully made by trained workers using adequately tested standardized instruments and methods. The data should be available to anyone for review.

When the inter-individual variation among healthy subjects is small, as in the case of body temperature, the normal level could actually be a single value (e.g., 98.6°F) instead of a range. This single value is a representative **central value** and would be a mean, median, or mode. When the distribution is **Gaussian** (normal), all three are the same and any one is nearly as good as another. However, the mean is preferred because of its good statistical properties, particularly its relative consistency from sample to sample. For distributions other than Gaussian, the choice again would mostly be the mean because of its easy understandability and better reliability. The median is used when the distribution is highly **skewed** or when extreme values or **outliers** are present in the data that cannot be excluded. When interest is specifically in the most common value for some reason, the choice naturally is the mode.

When the inter-individual variation in healthy subjects is large, a single normal value is not sufficient and we need a range of normal values. This is not as easy as it seems since there always are persons with very high or very low values who are still absolutely healthy. Thus, there could be considerable overlap between normal and abnormal values. Determining a threshold that works in all situations has remained an elusive objective. The following approaches are available.

*Disease Threshold:* The best method to delineate normal levels is to observe people with different levels for a sustained period and then to identify a threshold beyond which most people start feeling the burden in some sense—being unable to do work to one's full capacity or entailing a risk for an adverse condition later in life. This is an extremely complex procedure and requires consultation from experts, who in turn should have full evidence for the threshold they propose. The cutoff 140/90 mmHg for BP is such a threshold as experts have observed that a higher BP considerably increases the risk of coronary artery disease. Not many examples of this type of cutpoint are available, but there would be a caveat in this too as there would be people with level 145/92 who would be healthy and there would be people with level 136/88 and not healthy (with complaints such as headache and irritability). Thus, even this threshold does not rule out errors. This type of threshold is known as the disease threshold of normal level.

*Clinical Threshold:* The second alternative is to compare levels of those who are in perfect health with those who are not. Since each of these groups will have a distribution of its own, the situation typically will be as in Figure N.5. This figure has the following features:

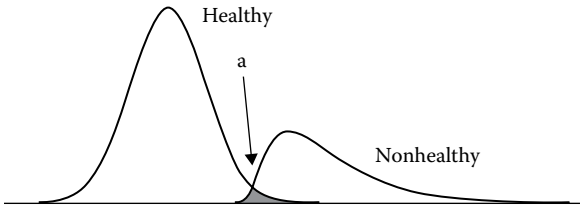



**FIGURE N.5** The pattern and overlap of measurement in healthy and nonhealthy subjects.

The number of healthy persons far exceeds the number of nonhealthy persons.

- The distribution among healthy subjects is Gaussian, whereas it is skewed in nonhealthy persons.
- The variation in the levels is smaller for healthy than for nonhealthy persons. Note how scatteredness in the levels of nonhealthy is relatively large.
- In this figure, nonhealthy subjects have higher levels. This is true for many measurements such as $T_3$, $T_4$, BP, blood glucose, and so on, but not for all. Higher levels of hemoglobin, peak expiratory flow rate, HDL cholesterol, and so on indicate good health. For such measurements, the curve for nonhealthy will be on the left side.
- There is some overlap between the levels seen in healthy subjects and the levels seen in nonhealthy subjects. This is shown as the shaded area in Figure N.5. If there is no overlap, the healthy levels and nonhealthy levels can be immediately defined. In practice, this overlap is substantial and causes problems in defining healthy levels. Whatever level is defined as the cutoff for healthy people, there would always be an error.

Statisticians have shown that the point where the two curves intersect provides the cutoff with the least number of misclassifications. This level is indicated as "a" in Figure N.5. This is the clinical threshold that could be used to define normal levels. Indeed, this is a very convincing approach but can be adopted only when the distribution in the healthy and nonhealthy groups is known and the overlap is minimal. The biggest problem in this approach is the choice of criterion to categorize a person as healthy or nonhealthy for drawing these curves. The threshold will not be known without the curves and the curves cannot be drawn without categorizing subjects as healthy and nonhealthy. Obviously, external criteria are needed, and those may or may not work. Nonetheless, such clinical threshold has inbuilt provision for tolerating error of misclassification as indicated by the shaded area. Errors are not ruled out by this method. The bigger the overlap, the larger the shaded area and the higher the chances of error.

### Statistical Threshold of Normal Values

When the distribution is Gaussian (mean − 2SD, mean + 2SD) are considered the normal limits, where SD is for the standard deviation. They exclude nearly 2.5% of healthy subjects with extreme measurements on either side. This is arbitrary but is now accepted around the world. The mean and SD are computed from measurements obtained on a large number of healthy subjects. These are statistical

thresholds and popularly known as ±2SD limits. Most of the normal ranges used in medical practice are obtained in this manner.

When the distributional shape is far from Gaussian, the range from 2.5th to 97.5th **percentile** points is considered normal instead of ±2SD limits. Note that the ±2SD limits for the Gaussian distribution are also from the 2.5th to 97.5th percentile. One can therefore forget about ±2SD limits and use the percentile-based range for all measurements irrespective of the shape of the distribution, but ±2SD limits are ingrained in the minds of many clinicians and statisticians alike. One reason for this is that the ±2SD limits fit well into the confidence interval and testing hypothesis strategy that are so common in statistical parlance. Note the following for such statistical thresholds:

- No matter how healthy the subjects are, there are always 2.5% healthy subjects at the lower end and another 2.5% at the upper end who will have levels outside such a normal range. This is an error but is tolerated because an error of this magnitude may always occur irrespective of the method used to establish normal limits. This error is at least quantitatively known for statistical thresholds but would not be easily known in other approaches.
- The ±2SD limits are purely statistical. A level beyond these limits is abnormal only in the sense that such an extreme level is rare in healthy subjects. Whether this translates to medical problems is not known. However, these limits seem to be working well as an aid in most situations encountered in medical practice.
- A measurement such as 106 mg/dL for fasting blood glucose level is not abnormal when the normal range is from 75 to 105; just that the chance of this value occurring in a healthy person is small—less than 2.5%. Gaussian theory stipulates that this chance reduces steeply as the measurement becomes farther and farther away from mean. A value of 104 mg/dL has nearly the same prognosis as the value 106 mg/dL. No miracle happens at the cutoff, such as 105, that would suddenly make a measurement abnormal. Nonetheless, such cutoff is needed somewhere as a guideline to start suspicion that the condition needs some kind of intervention and mean ±2SD provides such a cutoff, but it is applicable to one type of measurement at a time.
- If there are five different types of measurements such as different components of lipid profile, the chance of a healthy person labeled as healthy by such statistical criteria for all measurements together is not large because of accumulation of error of 2.5% multiple times.
- Some disease entities are based almost exclusively on a single parameter. Diagnosis of anemia caused by iron deficiency is based on hemoglobin level, that of hypertension is based on BP levels, diagnosis of diabetes mellitus is based on serum glucose level, and that of glaucoma is based on intraocular pressure. Other indications such as signs–symptoms play a minor role for classifying such diseases. There is evidence that persons with statistically abnormal levels do have an increased risk of the concerned morbidity and mortality. An intervention, such as therapy, to bring the level back to the normal range helps reduce this risk.
- The normal levels, whether statistical ±2SD limits or based on the healthy–sick dichotomy, should be determined by measuring a large number of subjects according to the guidelines given earlier. Only then do they command confidence. Normative data based on small samples can at best be indicative that they need confirmation in subsequent testing. Small-sample–based normals can seldom be used for diagnostic or prognostic purposes.

Reference values could be different for different segments of the population. A vital capacity normal for females would not be normal for males. A level of BP seen normally in adults would not be normal for children. The normal weight of 2-year-olds in Sudan may not be the same as the normal weight of 2-year-olds in Sweden. Normals may also change from time to time; for example, height seems to have increased all over the world during the past 50 years and lung functions also seem to be improving.

1. Murphy EA. The normal, and the perils of the sylleptic argument. *Perspect Biol Med* 1972;15:566–82. http://muse.jhu.edu/journals/perspectives_in_biology_and_medicine/summary/v015/15.4.murphy.html, last accessed March 30, 2015.

## notations (statistical)

Notations help make generalized statements. The following description of statistical notations may help you better understand the topics in this book, as well as to understand other statistical literature.

In statistics, a language that involves a number of conventions regarding notation with a mix of English (e.g., *P*-value) and Greek alphabets (e.g., alpha) has evolved. English (Roman) notations are used in italics (note *P* in *P*-value) whereas Greek notations are generally used straight. The convention in statistics is to use Roman notation for sample values and Greek for population **parameters**. Thus, $s$ is the notation for sample standard deviation (SD) and $\sigma$ is the notation for population SD; $\bar{x}$ for sample mean and $\mu$ for population mean; $r$ for sample correlation coefficient and $\rho$ for population correlation coefficient; $b$ for sample regression coefficient and $\beta$ for population regression coefficient, and so on.

It is customary in statistics to denote the variables by $x$ or $y$. Strictly speaking, capital letters such as $X$ and $Y$ are used for the name of the variables and small $x$ and $y$ are used for their realized values. For example, we may denote systolic blood pressure (BP) by $X$, and its value in a person such as 136 mmHg is denoted by $x$. We may have $x_1 = 136$ for the first person, $x_2 = 143$ for the second person, $x_3 = 125$ for the third person, and so on. These are all values of the variable $X$—in this example, the systolic BP. Such distinction between uppercase $X$ and lowercase $x$ is important for theoretical statistics but could be difficult for medical professionals. In this book, we have always used small $x$.

Universal notation for the sample size is $n$, which is generally indexed by $i$ and written as $i = 1, 2, \ldots, n$ to identify each of the $n$ subjects. Thus, in our BP example, we may have $x_n = 128$ that would tell that the last person in the sample has sysBP = 128 mmHg. Note that the subscripts 1, 2, and 3 of $x$ in the preceding paragraph are not in italics since they are not notation, but $n$ in the subscript is italic as this is a notation. The sample size $n$ could be 23 in one study and 386 in another study, or any number for that matter. For other numbers, such as the number of independent variables in a regression equation and the number of groups in a study, the notation we have used is $J$ or $K$. This is not standard but we find that convenient for nearly all setups. The indexing subscript for these are lowercase $j$ and $k$, respectively, written as $j = 1, 2, \ldots, J$, and $k = 1, 2, \ldots, K$.

You may find most books using lowercase $p$ for *P*-value. However, we have kept small $p$ to denote sample proportion. The notation for

corresponding population proportion, which generally is the same as the probability, is Greek π. The notation capital *P* is for probability as in *P*-value. This adaptation has allowed us to keep the distinction between a proportion and the probability.

Alpha (α) and beta (β) are the standard notations, respectively, for the probability of **Type I** and **Type II errors**. There are other standard notations such as chi-square ($\chi^2$), which is for a special statistic based on sample values, and similarly there are *F*, *t*, and *z* for statistical criteria used for testing of different hypotheses.

## nuisance parameters

A nuisance **parameter** is the one that we have to consider while drawing conclusions on the parameter of real interest. For example, while finding the confidence interval for the mean or for testing the hypothesis on the mean, we also need to consider the variance. Thus, variance is the nuisance parameter in this situation since the interest is only in mean. In a **multicentric trial**, the focus may be on estimating the effect size of the intervention but within- and between-centers variations are also considered in order to reach a valid conclusion. In clustered observations such as in longitudinal studies, within-subjects correlation is a nuisance parameter that cannot be ignored. The need to consider nuisance parameters makes the inference about the important parameters complicated: it will be good if a test statistic independent of the nuisance parameters can be devised, but that is not practically possible in many situations at the present state of our knowledge.

Cohen et al. [1] have discussed how age is a nuisance parameter for estimation of disease burden and its effect is sought to be neutralized by age-adjustment, **age-standardization**, and so on. Rubin and van der Laan [2] have discussed a situation of **structural estimating equations** that gives rise to a nuisance parameter while making adjustment for covariates.



1. Cohen SA, Chui KK, Naumova EN. Measuring disease burden in the older population using the slope-intercept method for population log-linear estimation (SIMPLE). *Stat Med* 2011 Feb 28;30(5):480–8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3043553/
2. Rubin DB, van der Laan MJ. Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *Int J Biostat* 2008;4(1):Article 5. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2669310/

## null and alternative hypotheses

A **hypothesis** is a supposition about the state of any phenomenon for starting an investigation about its truth or falsification. Depending on the quality of investigation, we may not be able to establish or deny a hypothesis, and may find that the evidence is not sufficient to reach any decision. Thus, there are three possibilities: convincing evidence that the hypothesis is true, clear evidence of it being false, and no sufficient evidence either way. Statistical testing of hypothesis also starts from setting up one, but this one is called null hypothesis as explained next, as well as a counter hypothesis, called the alternative. The evidence to confirm or refute them is in terms of the data obtained in the investigation, generally termed as a study.

### *Null Hypothesis*

In the case of statistical decisions, the supposition initially made generally is that there is no effect, called null hypothesis, which actually implies that the status of our knowledge about this phenomenon is the same as before the study. A clear understanding of this nature of the null is necessary to proceed further with the statistical aspects of a study. Data are collected with the aim of gathering evidence against the null and of rejecting the null if sufficient evidence against it is available. The study is done in the hope of overturning the null. Thus, a null hypothesis is equivalent to the presumption of innocence in the court setting where also evidence is presented against this presumption. The notation used for this is $H_0$. This is the hypothesis under scrutiny and is sought to be refuted by conducting a study on a sample of subjects just as criminal investigations are done to find evidence against a suspect. Thus, null is something you suspect or do not believe to be true.

Depending on the evidence available from the collected data, $H_0$ is either rejected or not rejected but is never accepted in empirical setup of the type we encounter in statistics. Not rejecting a null may mean two things: (i) carry out further investigations and collect more evidence so that a decision one way or the other can be taken, and (ii) continue to accept the present knowledge as though this investigation was never done. The "truth" remains unchanged in this case. This can be easily explained with the help of a telling example described in the next paragraph.

Ptolemy, in the second century AD, propounded that the sun revolves around the earth. It remained the "truth" for 14 centuries until Galileo came up with evidence against it in the 16th century and established a new truth that says that the earth revolves around the sun. Newton's laws of mechanics were accepted until Einstein discussed circumstances in which some of them did not hold. Peptic ulcer was believed to be caused by acidity until sometime ago when *Helicobacter pylori* was found to be a culprit in some cases. As of today, coronary heart disease is not considered to be caused by any infection, but who knows that a rebuttal will come soon.

Empirical strategy is to find evidence against a hypothesis that is stated in null form. Similar to a court judgment, if this evidence is sufficient, the null hypothesis is rejected; otherwise, it continues to be considered "truth" by default. Transplanted to the development of a new therapeutic regimen, it seems reasonable to demand evidence against the hypothesis that there is no effect at all. This process is difficult to implement without setting up a null hypothesis.

Since samples by their very nature are uncertain, the conclusion depends on what sort of data are obtained from the subjects. Statistical tests are precisely meant to deal with uncertainties arising from sampling fluctuations. They provide the answer to the question: What is the likelihood of sample values given that the null hypothesis is true? If this likelihood is exceedingly small, say, less than 5%, the null is considered implausible and rejected. At the same time, it would be ridiculous to accept a hypothesis whose likelihood of giving the obtained sample is only 15%. Thus, the only conclusion drawn in this case is that the sample fails to provide sufficient evidence to reject the null hypothesis. It is not accepted and the situation reverses back to what it was before that study. Many would consider it idiosyncratic that the sample values are searched for evidence against a null without recourse to finding what they are for—what they support. The alternative hypothesis explained next should clarify this dilemma, but the process of finding evidence against it is widely followed in empirical setup.

As another example, consider the claim of a manufacturer that his drug is superior to the existing angiotensin-converting enzyme inhibitors in improving insulin sensitivity in diabetic hypertensives. Suppose in a trial of matched cases, improvement was seen in 7 out of 10 patients who were given the new drug compared with 6 out of 10 who were given the existing drug. The methodology for testing the statistical **significance** of this difference is described separately, but you can see that the sample size $n = 10$ in each group is small

and the difference is too small to provide confidence to pronounce that the new drug is better. The difference could have arisen because of sampling fluctuation. If so, the claim of superiority is not tenable. The manufacturer needs to withdraw the claim forever or until such time that more evidence is available for scrutiny. The medical fraternity is expected to continue with their existing practice and not take cognizance of the claim until the claim is adequately substantiated.

The concept of null hypothesis is not restricted to the difference between two groups. It could concern difference between many groups, association or relationship between two or more characteristics or variables, or any other aspect of the problem under investigation.

### *Alternative Hypothesis*

If a null hypothesis is false, what alternative is true? The alternative hypothesis, denoted by $H_1$, is the opposite of $H_0$ that must be true when $H_0$ is rejected. If the null is that a drug is not effective, then the alternative has to be that it is effective. In the preceding example, the claim is that of superiority of the new drug, and this is the alternative hypothesis in this case. This is a **one-sided alternative** if inferiority is ruled out. Most often, it is not possible to claim that one group is better than the other, and the only claim is that they are different. In the case of peak expiratory flow rate (PEFR) in factory workers exposed to different pollutants, there may not be any a priori reason to assert that it would be affected more by one pollutant than another. Then, the alternative is that the mean PEFRs in workers exposed to different pollutants are unequal. This is called a two-sided alternative. The null is that they are equal in various groups. One-sided and two-sided $H_1$ are also sometimes called one-tailed and two-tailed $H_1$, although these terms should be used for the corresponding probabilities of **Type I error**.

Consider qualitative data where the interest is in proportion. Suppose that a pharmaceutical company claims that its particular drug has at least 72% efficacy in the long run. Thus, the null is $H_0$: $\pi = 0.72$. To test this claim, a researcher tries the drug on a random sample of $n = 40$ patients and 28 respond positively. Thus, the efficacy is 70%. Now, the question is: can population proportion still be at least 72%? If not, the inference would be that it is 72% or less. The null would be rejected in favor of this alternative. Thus, the alternative hypothesis is $H_1$: $\pi < 0.72$. Now, let us give it a slight twist. Suppose 32 out of 40 patients in the sample respond. Thus, $p = 0.80$. The question can now be whether the long run efficacy can still be less than the claimed 72%. The alternative hypothesis in this setup is $H_1$: $\pi > 0.72$. Thus, the alternative hypothesis can take either direction depending on what question is to be answered.

In a mathematical formulation of the null hypothesis, there will typically be an equality sign as illustrated in our example. The alternative hypothesis typically has inequality. This could have either a "less than" sign, a "more than" sign, or just a "not equal to sign."

## number needed to treat (NNT)

NNT measures the average number of patients that would need to be treated to prevent one additional adverse outcome or for one additional success compared with the standard or any other treatment (or even placebo). Statisticians have found that NNT = 1/ARR where ARR is **absolute risk reduction**; NNT converts the more difficult concept of ARR into a useful and readily understood quantity.

The concept of NNT has gained popularity because of its simplicity. It is especially useful in comparing the efficacy of two treatment regimens with binary outcome. For example, the NNT to cure one patient of atherosclerosis using a diet–exercise regimen may be 7 and the NNT using drug-X may be 5. Then, drug-X therapy would be considered more effective. This number also measures absolute efficacy in the sense that 5 patients on average are required to be treated by drug-X in this example to cure 1 patient, and thus this drug may not be considered a sufficiently effective strategy because the number 5 seems large for one positive response. Thus, NNT is a useful aid to the clinical decision process. Clinicians understand and appreciate NNT much better than probabilities underpinning the ARR. However, NNT does not work for one patient.

Also, if 32 patients of atherosclerosis are needed to be treated to prevent one death and 18 cases of diabetes are needed to prevent one death, you have some idea of where to put the better part of the resources, particularly if the financial and social cost of treatment of these types of cases is known. Also, due consideration should be given to the factors such as age, severity of disease, and duration of sickness while interpreting NNT.

Further caution is required when a regimen is to be used for mass adoption on the basis of NNT. Heller [1] gives an example of thrombolysis with NNT = 7 for prevention of one adverse outcome in secondary prevention after nonhemorrhagic stroke, and aspirin with NNT = 33. However, only 4% of the stroke patients are eligible for thrombolysis due to the time window and other factors, whereas 70% are eligible for aspirin. Despite better response, there may not be sufficient opportunity to use the regimen in practice.

The distributional properties of NNT are not fully known and a makeshift procedure is adopted to get its confidence interval (CI). This is easily implemented when the difference in efficacy of two treatments under comparison is statistically significant. In this case, the first step is to calculate CI for ARR. Call this ($ARR_L$ to $ARR_U$). Then, CI for NNT is $\left(\frac{1}{ARR_U} \text{ to } \frac{1}{ARR_L}\right)$. Note that L and U have switched their position (mathematically because we are now dealing with reciprocals). If the CI for ARR is from 0.04 (i.e., $ARR_L$) to 0.15 (i.e., $ARR_U$), the CI for ARR is from $\frac{1}{0.15} = 6.67$ to $\frac{1}{0.04} = 25$. For nonsignificant ARR, the situation is complex as explained next.

Generally, ARR would be small in a trial. If the efficacy of drug A is 70% and that of drug B is 75%, $(\pi_1 - \pi_2) = 0.05$. This has the same interpretation as ARR. Since $\frac{1}{\pi_1 - \pi_2} = 20$, NNT = 20 in this case. That is, 20 subjects are needed to be treated by drug B to get one extra success. The CI for $(\pi_1 - \pi_2)$ in this case can include negative values also such as −0.02 to +0.12. This interval includes zero where NNT is infinity. This causes problems. Zero within this CI indicates that ARR is not statistically significant. If nonsignificance is disregarded, the reciprocals give −50 to +18 as CI for NNT. This interval does not even include the original point estimate NNT = 20. Thus, this method of obtaining CI for NNT fails in this case. NNT is one example where sensible CI is obtained only when ARR is statistically significant.

For independent samples, NNT can also be obtained in terms of **relative risk** (RR) instead of attributable risk (AR):

$$\text{NNT} = \frac{1}{\text{AR}} = \frac{1}{R_2 - R_1} = \frac{1}{R_1(\text{RR} - 1)} = \frac{1}{p(\text{RR} - 1)},$$

where $p$ is the proportion affected in the control group. Note that $p$ and $R_1$ are the two notations for the same quantity. Just as RR can be approximated by **odds ratio** (OR) in special cases, so can NNT, and thus it is possible to calculate NNT through **case–control studies**

also in these conditions. In most situations, however, an immaculately carried out **randomized controlled trial (RCT)** is needed to correctly calculate NNT.

Beware of the anomalies that can occur in interpreting NNT. In a paper published in a reputed journal [2], the numbers show that NNT = 83 patients to be treated with statins to prevent one event of myocardial infarction (MI) and NNT = 142 patients to be treated with statins to prevent one event of stroke. These were for a mean follow-up of 3½ years. The authors wrongly interpreted and claimed that 24 patients needed to be treated for 1 year to prevent one MI and that 42 patients needed to be treated for 1 year to prevent one stroke, forgetting that NNT would be much higher for 1 year than for 3½ years.


1. Heller RF. Development of modern epidemiology: Clinical epidemiology, in: *The Development of Modern Epidemiology: Personal Reports from Those Who Were There*. Eds. Holland WW, Olsen J, du V Florey C. Oxford University Press, 2007:p. 269.
2. Savarese G, Gotto AM, Paolillo S, D'Amore C, Losco T, Musella F, Scala O et al. Benefits of statins in elderly subjects without established cardiovascular disease: A meta-analysis. *J Am Coll Cardiol* 2013;62(22):2090–9. http://content.onlinejacc.org/article.aspx?articleID=1732396


## numerical analysis/method

This is a strategy that tries to solve mathematical problems through various combinations of numbers in place of looking for solutions through notations. For example, those familiar with quadratic equations know that the solution to the equation $ax^2 + bx + c = 0$ is $x = [-b \pm (b^2 - 4ac)]/(2a)$. This is the algebraic solution and called the closed-form solution. Many mathematical problems can be directly solved in closed form with the help of other methods such as differentiation and integration. Do not worry if you are not aware of these mathematical terms. Important is that some problems cannot be explicitly solved and some may be intractable. For such problems, the numerical methods are many times used to arrive at a solution. The solution may be approximate but reasonable enough to go ahead.

Consider the **area under the concentration curve**. This plots the availability of the drug in an organ after it is ingested. Theoretically, this would be a curve as shown in Figure N.6. Without knowing the mathematical equation of this curve, it is not possible to correctly obtain the area under this curve. However, an approximate area can be obtained by the numerical method. For this, the area is divided into small segments and each segment is approximated by a trapezoid. We know how to calculate the area of a trapezoid, and their sum can be obtained to get the area under the curve as illustrated next.

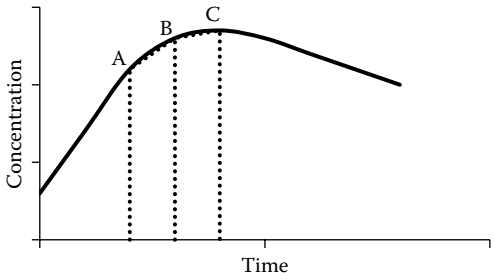



**FIGURE N.6** Numerical method for obtaining area under the concentration curve.

In Figure N.6, we show two trapezoids. Note that points A and B are connected in the trapezoid by a straight (dotted) line while they are actually connected by the (solid) curve. The curve is so close to the top of the trapezoid that the area under the curve between points A and B is nearly the same as the area of the trapezoid. Our trapezoids have big widths so as to illustrate the distinction between the curve and the top line but they can be made as small as we like. Then, the curve and the top line will be almost indistinguishable for small trapezoids. The area of each trapezoid can be calculated by using a computer algorithm. When these areas are added, we get the area under the curve. This is the numerical method of obtaining the area under the concentration curve and is followed by all statistical packages for this purpose.

For solving some mathematical problems, some guess values are tried in the first step and the error in the solution revealed by them is fed back to revise the solution. Sometimes, many iterations are needed. Quite often these days, a computer algorithm that does the iterations and finally comes up with a solution is used. A solution is considered to have been reached when the values with two successive iterations are sufficiently close—called *convergence* in mathematical terms. Such difficult and intractable problems are common in the engineering sciences but are now seen in medical sciences as well. This makes numerical analysis a broad discipline with close connections with computer science, mathematics, engineering, and the sciences.

Numerical analysis concerns all aspects of the numerical solution of a problem: from the understanding of numerical methods to their practical implementation in a reliable computer program. Most numerical analysts share some common concerns such as the size and the form of error. They try to develop algorithms that are efficient and give precise results. The algorithm must be sensitive to the changes in the data.

Medical applications of numerical analysis are wide and varied. Sometimes, the term has been used in a nontechnical sense for any method where numbers are used, such as for separating categories of patients. Anitha et al. [1] used the term *numerical analysis* to validate the findings regarding identifying the respective regions in the femoral neck and to prove that drug treatment elicits local changes in the mean outer radius and mean cortical thickness of femoral necks in postmenopausal women. Yu et al. [2] obtained the approximate function describing the relationship between the geometric parameters of the nasal airway and the nasal functions by what they called numerical methods.


1. Anitha D, Kim KJ, Lim SK, Lee T. Comparison of buckling ratio and finite element analysis of femoral necks in post-menopausal women. *J Menopausal Med* 2014 Aug;20(2):52–6. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4207002/
2. Yu S, Sun XZ, Liu YX. Numerical analysis of the relationship between nasal structure and its function. *Scientific World J* 2014 Feb 6;2014:581975. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3933016/

# O

## obesity (measures of), *see also* body mass index

Obesity is excess fat in the body than normally expected in healthy subjects. For assessing obesity by this definition, we need to measure body fat in a person and compare this with what is normally expected in healthy subjects. This requires that healthy subjects are defined and how much fat is expected in them is somehow obtained. This obviously would be different for different age and sex, and possibly also in different races and ethnicities.

Obesity has been found to be associated with risk of diseases such as hypertension, atherosclerosis, gallbladder disease, and diabetes. It is now standard practice in clinics to assess obesity and give advice accordingly. Since the amount of fat in the body is difficult to assess, a large number of surrogates are used. Many of them are variants of the **ponderal index**. A general form of this index is

$$\text{Ponderal index: weight/(height)}^b,$$

where $b$ is estimated from the **regression** of log(weight) on log(height) separately for each age, and weight is generally measured in kilograms and height is in meters. The coefficient $b$ varies from population to population. Freeman et al. [1], for example, found that $b = 2.08$ for boys of 7 years, $b = 2.20$ for girls of 7 years, $b = 2.44$ for boys of 16 years, and $b = 1.75$ for girls of this age in the United Kingdom. For neonates, $b = 3$ is considered more appropriate. The value of $b$ is large in children and progressively declines with age and settles to 2.0 in adults. Note that $b = 2$ makes it the popular **body mass index** (BMI).

A simple index of obesity is the **Broca index**, which compares weight (in kilograms) with (height in centimeters − 100). Among other measures of obesity are **waist–hip ratio** and waist–height ratio. They measure central or truncal obesity instead of the overall obesity. The latter is in consideration that waist measurement depends on height and cannot be considered as a stand-alone indicator. Whereas others are discussed as separate topics, we give details of two not described by us elsewhere in this book. Both focus on waist circumference (WC), which has been found to be a risk factor for early mortality in adults independent of BMI [2].

$$\text{Waist-height ratio} = \frac{\text{waist circumference}}{\text{height}}.$$

A value of more than 0.6 of this ratio is considered a health risk. This also has similar association with ischemic stroke as most other measures of obesity [3].

$$\text{Body shape index} = \frac{\text{waist circumference}}{\text{BMI}^{2/3} * \text{height}^{1/2}}.$$

This was obtained by regressing log of WC on log of height and log of weight. It has been found to have little correlation with weight, height, and BMI. Healthy body shape index is around 0.0800.


1. Freeman N, Power C, Rodgers B. Weight-for-height indices of adiposity: Relationships with height in childhood and early adult life. *Int J Epidemiol* 1995;24:970–6. http://ije.oxfordjournals.org/content/24/5/970.full.pdf
2. Krakauer NY, Krakauer JC. A new body shape index predicts mortality hazard independently of body mass index. *PLoS ONE* 7(7);2012; e39504. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0039504
3. Wang A, Wu J, Zhou Y, Guo X, Luo Y, Wu S, Zhao X. Measures of adiposity and risk of stroke in China: A result from the Kailuan study. *PLoS One* 2013 Apr 17;8(4):e61665. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3629147/


## O'Brien–Fleming procedure, *see also* Lan–deMets procedure

A medical experiment, particularly a clinical trial, is sometimes appraised one or more times while it is ongoing. The purpose of such interim reviews is to examine if the study has already provided believable evidence of the desired efficacy so that further investigations can be stopped, and cost and efforts are saved. This can also save exposure of the patients to inferior treatment. If no firm evidence is available, the trial runs its full course. The trial can also be stopped for futility if it turns out that there is practically little chance of reaching the minimum target effect.

When such multiple testing of hypothesis is done, the probability of **Type I error** is adjusted for each appraisal so that the total Type I error does not exceed the prefixed **significance level** and the statistical **power** of detecting a prespecified effect remains intact. This adjustment can be made in several ways, one of which is the O'Brien–Fleming procedure. The other popular method for this is the **Lan–deMets procedure**.

O'Brien and Fleming developed this procedure in 1979 [1]. This is restricted to comparing two treatments with **dichotomous** data. Under the usual setup, the result of such a trial would be in terms of a 2 × 2 table. The statistical test of significance for this is **chi-square** with 1 df. Multiarm trials and the trials with quantitative outcomes such as those aiming reduction in total cholesterol level are not covered by this procedure. The stages of appraisals and the number of appraisals are fixed in advance and would be part of the protocol. For example, the trialist may decide to do two interim appraisals—first after completing trial on one-third of the subjects and second after completing trial on two-thirds of the subjects. The last, of course, if needed, will be after completing trial on 100% of the subjects. When this scheme is followed, there are a total of $K = 3$ appraisals. They need to be equally spaced—a limitation later removed by the Lan–deMets procedure.

Suppose successive reviews are planned after every $n_1$ subjects in the treatment group and another $n_2$ in the control group. Let the number of planned appraisals be $K$—thus, the total subjects if the trial runs its full course is $K*n_1$ in the treatment group and $K*n_2$ in the control group. As mentioned earlier, all these are fixed in

**TABLE O.1**
**Value of O'Brien–Fleming $P(K,\alpha)$ for Selected $K$ and α**

| | Number of Stages ($K$) | | | | |
|---|---|---|---|---|---|
| α | 1 | 2 | 3 | 4 | 5 |
| 0.05 | 3.84 | 3.92 | 4.02 | 4.10 | 4.16 |
| 0.01 | 6.64 | 6.66 | 6.74 | 6.80 | 6.88 |

advance. For the hypothesis of no difference between the efficacy of the test and the control regimen at the $k$th appraisal,

O'Brien-Fleming procedure: Reject $H_0$ if $(k/K)\chi_k^2 \geq P(K,\alpha)$,

where $\chi_k^2$ is the usual chi-square calculated for all the samples till the $k$th stage and $P(K,\alpha)$ is the value derived by the O'Brien–Fleming procedure corresponding to the significance level α and the value of $K$. These values are derived in a manner that keeps the probability of Type I error under control. Some of these values are given in Table O.1. For $K = 1$, these values are the same as for the usual chi-square. If $\chi_K^2$ does not exceed $P(K,\alpha)$ after completing all $K$ tests, the conclusion is that the equality of the two regimens cannot be denied at the α level of significance.

For example, consider a trial to be appraised a total of $K = 3$ times, with $n_1 = 30$ and $n_2 = 30$. Thus, the total trial is planned to have 90 subjects in each group. After the results are available for the first 30 subjects in each group, suppose the data give $\chi_1^2 = 6.93$. This gives $(k/K)\ \chi_k^2 = (1/3) \times 6.93 = 2.31$. Since this is less than $P(3,0.05) = 4.02$, you are not able to reject the null. Had it been planned as a single stage trial with 30 subjects in each group with no appraisal, $\chi_1^2 = 6.93$ was enough to reject the null. In view of the nonsignificance, the trial adds another 30 subjects in each group and repeats this test at stage 2. For these 60 subjects in each group, let $\chi_2^2 = 10.87$. Now, $(k/K)\ \chi_k^2 = (2/3) \times 10.87 = 7.25$. Since this is more than 4.02, reject the null and conclude that the regimens are different in efficacy at the 5% level of significance. There is no need to go to stage 3—thus, the last 30 subjects are saved.

The O'Brien–Fleming procedure is applied to a **group sequential design**. In this design, the subjects are sequentially added if needed after testing at each stage. However, there are flaws in the O'Brien–Fleming procedure. For example, even a small **intraclass correlation coefficient** among responses of the subjects can substantially inflate the Type I error. Also, if the sample size is curtailed because of statistical significance at interim stages, consider whether the sample subjects covered so far are really representative or nonrepresentative of the population under target. A full sample is likely to be representative as usual, but a curtailed sample will not be if the first few subjects in sequence happen to be different, say, more severe than others. For details of such difficulties in this procedure, see Lui [3].


1. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979 Sep;35(3):549–56. https://www.musc.edu/psychiatry/research/cns/upadhyayareferences/OBrien_1979.pdf
2. Hammouri H. Review and extension for the O'Brien Fleming multiple testing procedure. *VCU Theses and Dissertations* 2013: Paper 3260. http://scholarscompass.vcu.edu/cgi/viewcontent.cgi?article=4259&context=etd
3. Lui K-J. The performance of O'Brien–Fleming procedure in the presence of intraclass correlation. *Biometrics* 1994;50:232–6. http://www.jstor.org/discover/10.2307/2533214?sid=21106346664793&ui=2&uid=4&uid=3738256


## observational studies, *see also* natural experiments

Observational studies are those that depend on just observations of what is naturally happening, in contrast to experimental studies and trials where a deliberate intervention is introduced to study its effect. However, observational studies also are **analytical studies** where the objective is to study the antecedent–outcome relationship such as between exposure and disease.

Observational studies exploit the premise that nature is a great experimenter. Many changes occur naturally, requiring no human intervention. Some people are exposed to iodine-deficient water because of environmental conditions, and some women have low hemoglobin levels because of their nutritional status. The study of such naturally occurring events can be invaluable in studying cause–effect relationship with conviction. Such a study can be based on records or on actual observations, or a combination of both. This type of study is also sometimes referred to as an **epidemiological study**. Since there is no deliberate human intervention (such as a drug) in this setup, such studies carry little risk of harm to the subjects or the society, although they are invasive with regard to time and privacy of the respondents. An observational study is generally conducted for specific groups such as those with high disease prevalence or those with high prevalence of a risk factor—thus, extrapolation to the general population is not immediate. Many observational studies are done in a hospital setup rather than in communities. For example, change in Oswestry disability index score 1 year after microdecompression and laminectomy in patients with central lumbar spinal stenosis, when based on records of previously treated patients, is an observational study [1]. This study had two types of surgery as interventions, but these were going on naturally without a design. As pointed later in this section, this can be a great limitation of observational studies in giving a valid conclusion.

An observational study can be done in one of many formats (Figure O.1): (i) **prospective studies** such as cohort, longitudinal, and repeated measures; (ii) **retrospective studies** such as case–control, nested case–control, and uncontrolled; and (iii) **cross-sectional studies**. For merits and demerits of each and their comparative performance, see the topic **retrospective studies**. **Ecological studies** and many **case studies/case series** are also observational.

Observational studies might look simple to carry out but they provide valid results only when carried out with full precaution. The antecedent under study and the outcome of interest must be sharply defined and all **confounders** must be under control. Such studies require representative and random samples so that there is no bias in the results, but this could be an uphill task in observational studies. In contrast to this, note that a random sample is not a strict requirement for clinical trials. Because of lack of randomization, the results of observational studies are seldom conclusive. Matching is adopted in some observational studies to address this limitation, but this also requires careful consideration of what should be matched and to what extent, and there is no **overmatching**. Some confounders may still escape attention and can contaminate the results. Thus, there is widespread perception that the results of observational studies are only indicative and not conclusive. On the contrary, in **randomized controlled trials (RCTs)**, the results are far more robust for the effect of intervention because of strict control of the conditions in which these are done. RCTs are insulated from extraneous factors that observational studies are not. Smith and Ebrahim [2] have cited an instance of the cardioprotective effect of hormone replacement therapy found in several observational studies overturned by subsequent RCTs. There are many such instances. For communication of the results of an observational study, the

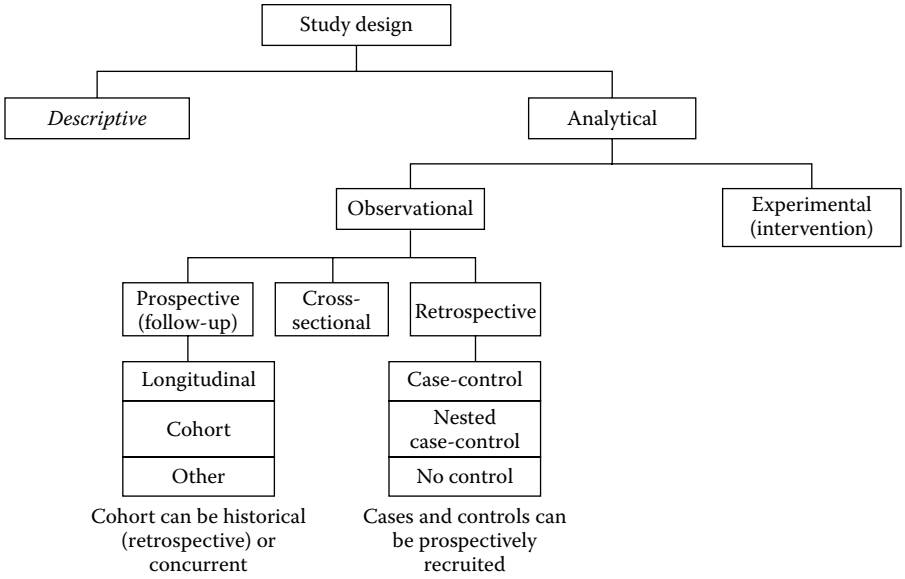



**FIGURE O.1** Types of observational studies.

**STROBE** format is recommended, which tends to take care of some of these discrepancies.


1. Nerland US, Jakola AS, Solheim O et al. Minimally invasive decompression versus open laminectomy for central stenosis of the lumbar spine: Pragmatic comparative effectiveness study. *BMJ* 2015 Apr 1;350:h1603. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4381635/
2. Smith GD, Ebrahim S. Data dredging, bias, or confounding: They can all get you into the BMJ and the Friday papers. *BMJ* 2002;325:1437–8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1124898/


## observations (statistical)

*Statistical observations* is the term used for realized values of a variable. The variable is the characteristic and the observations are the measurements of that characteristic in a set of subjects. For example, bone mineral density *T*-score is a variable, but its actual values in persons such as −1.52, −0.87, +1.20, and so on are the observations. This term is quite often used in statistical parlance, and generally these observations are denoted by $x_1$, $x_2$, $x_3$, and so on. The subscript identifies the subject number.

When we say mean = $(x_1 + x_2 + \ldots + x_n)/n$, $x$'s are the notations for the observations on $n$ subjects. In the regression equation $y_i = b_1x_{1i} + b_2x_{2i} + \ldots + b_Kx_{Ki} + e_i$, the observations are $(y_i, x_{1i}, x_{2i}, \ldots x_{Ki}, e_i)$ for the $i$th subject. Note that in this equation, there are $K$ independent and one dependent variable.

Statistical observations depend on their unit. We are not pointing to unit of measurement but to unit of observation. You can have family as a unit of observation for a variable such as family size or number of family members with similar history; you can have country as a unit for expectation of life or physician–population ratio; you can have a mouse as a unit in an experiment for measuring the duration of survival; and, of course, you can have individual human being as a unit for observations on pain score or creatinine level.

There are terms such as independent observations and dependent observations. The example of regression just mentioned is for independent and dependent variables and not for observations. The variable could be birth weight, but the observations are dependent when birth weight is measured for twins, or even when measured for siblings. Description of statistical methods tends to become confusing when the same variable measured in two groups of subjects is considered as two distinct variables. In our example, the birth weight of one sibling can be one variable and the birth weight of a second sibling can be another variable. Then, they become dependent variables instead of dependent observations. Many times, this kind of distinction is lost in statistical literature, and this does not seem to hurt either.

## observed zeroes, *see* contingency tables

## observer bias/errors/variation

Observer bias occurs when there is a systematic error in the values obtained by an observer and when the variation and errors occur because of lack of care. Bias is generally one sided (either positive or negative), whereas variation and error will be positive in some subjects and negative in others.

### *Observer Bias*

Observer bias occurs when wrong methods are used or inadequate assessments are done. It can arise if the subjects with disease are evaluated more intensively and more carefully than those without disease. **Blinding** is used as a means of safeguarding against this bias, but such blinding is rarely feasible in many clinical setups and in studies requiring long follow-up. Then, this bias can go on unnoticed.

There are many other examples. Some agencies deliberately underreport starvation deaths in their area as a face-saving device and overreport deaths caused by calamities such as flood, cyclone, and earthquake to attract funds and sympathy. Improvement in the condition of a patient for reasons other than therapy can be wrongly ascribed to the therapy. Tighe et al. [1] studied observer bias in the interpretation of dobutamine stress echocardiography. They concluded that the potential for observer bias exists because

of the influence of ancillary testing data such as angina pectoris and ST-segment changes. Lewis [2] found that psychiatric assessments of anxiety and depression requiring clinical judgment on the part of the interviewer are likely to suffer from observer bias. These examples illustrate some situations in which observer bias can occur.

Biases attributed to intentional abuses of methods and instruments, however, are nearly impossible to handle and can remain unknown until they expose themselves. One approach is to be vigilant regarding the possibility of such biases and deal sternly with them when they come to notice. Scientific journals can play a responsible role in this respect. If these biases are noticed before reaching the publication stage, steps can be sometimes taken to correct the data. If correction is not possible, the biased data may have to be excluded altogether from analysis and conclusion.

It is sometimes believed that bad data are better than none at all. This can be true if sufficient care is exercised in ensuring that the effect on the conclusion of bias in bad data has been minimized, if not eliminated. This is rarely possible if the sources of bias are too many. Also, care can be exercised only when the sources of bias are known or can be reasonably conjectured. Even the most meticulous statistical treatment of inherently bad data cannot lead to correct conclusions.

### *Observer Errors*

There are unintentional errors, and they will be positive in some cases and negative in others. Sometimes the exposure or the accompanying baseline information and sometimes the outcome are not correctly recorded. This can occur as a result of carelessness of the observer or of the recording clerk who may unwittingly classify a subject into a particular category. Incorrect information can also arise for the following reasons. These can occur in any setup but are typically more common in a prospective study setup.



- Human error in correctly assessing the condition of the patient. This can be attributed to either carelessness or lack of expertise of the observer. The physician may lack competence and the recording clerk may lack training or motivation. Assessment during the latter part of a longitudinal study may be less accurate as fatigue sets in or more accurate because of the learning effect.
- Inaccurate reports from the laboratory arising from use of nonstandardized techniques or chemicals, or from use of faulty instruments.
- Doubtful validity of the diagnostic or screening test.

There can be a lack of care in obtaining or recording information when, for example, sufficient attention is not paid to the appearance of Korotkoff sounds while measuring blood pressure by a sphygmomanometer or to the waves appearing on a monitor for a patient in critical condition. This can also happen when responses from patients are accepted without probing and some of them may not be consistent with the response obtained on other items. If reported gravidity in a woman does not equal the sum of parity, abortions, and stillbirths, then obviously some information is wrong. A person may say that he does not know anything about AIDS in the early part of an interview but states sexual intercourse as the mode of transmission in the latter part of the interview. The observer or the interviewer has to exercise sufficient care so that such inconsistencies do not arise.

All such errors in assessment can be reduced simply by being more careful and by using precise instruments, measurements, and classification criteria that have been pretested for their validity. Many inadvertent errors can be avoided by imparting adequate training to the observers in the standard methodology proposed to be followed for collection of data, and by adhering to the protocol as outlined in the instruction sheet. Many investigations do not even prepare an instruction sheet, let alone address adherence.

### *Observer Variation*

Barring some clear-cut cases, clinicians tend to differ in their assessment of the same subject. Interpretation of x-ray films is particularly notorious in this respect. Disagreement exists concerning simple tools such as a chart for assessing growth of children. Physicians tend to differ in grading a spleen enlargement. One physician may consider a fasting blood glucose level of 136 mg/dL in a male of age 60 years sufficient to warrant active intervention, but another might opt just to monitor. Variation in blood pressure readings due to differences in hearing acuity or in interpretation of Korotkoff sounds is on record. Some clinicians are more skillful than others in collating pieces of information into solid diagnostic evidence. Such variability on the part of the observer, researcher, or investigator is a fact of life and cannot be wished away. It is inherent in humans and represents a healthy feature rather than anything to be decried. Efforts are made from time to time to reconcile and come to a consensus, and indeed such a consensus is reached on many occasions. Yet, many issues remain unresolved and new ones keep cropping up, and they continue to contribute to the spectrum of uncertainty.

Quite often, an investigation is a collaborative effort involving several observers. Not all observers have the same competence or the same skill. Assuming that each observer works to his fullest capability, faithfully following the definitions and protocol, variation can still occur in measurement and in assessment of diagnosis and prognosis. This can happen because one observer may have a different acumen in collating the spectrum of available evidence than others.

### *Inability of the Observer to Get Confidence of the Respondent*

Patient–doctor equation plays an important role in extracting correct and full information. This inability can be attributed to language or intellectual barriers if the subject and observer come from widely different backgrounds. They may then not understand each other and generate wrong data. In addition, in some cases such as in sexually transmitted diseases (STDs), part of the information may be intentionally distorted because of the stigma or the inhibition attached to such diseases. An injury in a physical fight may be ascribed to something else to avoid legal wrangles. Some women hesitate to divulge their correct age. Some may refuse physical examination, forcing one to depend on less valid information. Correct information can be obtained only when the observer enjoys full confidence of the respondent.

1. Tighe JF Jr., Steiman DM, Vernalis MN, Taylor AJ. Observer bias in the interpretation of dobutamine stress echocardiography. *Clin Cardiol* 1997;20:449–54. http://onlinelibrary.wiley.com/doi/10.1002/clc.4960200509/pdf
2. Lewis G. Observer bias in the assessment of anxiety and depression. *Soc Psychiatry Psychiatr Epidemiol* 1991;26:265–72. http://link.springer.com/article/10.1007/BF00789218#page-1

**TABLE O.2**
**Odds for Some Probabilities**

| Probability ($\pi$) | 0.02 | 0.10 | 0.33 | 0.50 | 0.80 |
|---|---|---|---|---|---|
| Odds | 2:98 = 1/49 | 10:90 = 1/9 | 33:67 = 1/2 | 50:50 = 1/1 | 80:20 = 4/1 |

## odds, *see also* odds ratio

Odds of an event are defined as the ratio of the probability of its occurrence to the probability of nonoccurrence. If the probability of occurrence is $\pi$, the odds are $\pi/(1-\pi)$. Table O.2 gives the odds ratio for some values of $\pi$. For probability 0.02, the chance of nonoccurrence is 49 times of the chance of occurrence. For probability 0.5, the odds are even. For probability 0.80, the chance of nonoccurrence is one-fourth of the chance of occurrence.

Odds are commonly used in betting. For example, the odds of winning 1:3 imply that a loss is three times as likely as a win. In health and medicine, odds are used instead of probability for expressing the chances of presence of an antecedent. The chance of an outcome is expressed in terms of risk, whereas the chance of an antecedent is expressed as odds. This is because the probability is perceived for an event yet to occur, whereas the antecedent has already occurred. Odds are conventionally used in **case–control studies** that investigate the antecedents in the known cases and controls. If 67% of all cases of hypertension are obese in a particular segment of population, the odds of obesity are 67:33, or 2 to 1 in these subjects. That is, a subject with hypertension is twice as likely to be obese as being nonobese. This is a quirky measure but has been found extremely useful in medical research as explained next.

The most prominent statistical use of odds is in **logistic regression** where the dependent is **logit** of the probability of the event when observed as yes/no, present/absent, and so on. This is defined as log of odds: $\text{logit} = \ln[\pi/(1-\pi)]$. The main advantage of this transformation is that the probability with range (0, 1) is mapped onto $(-\infty, +\infty)$. This helps in mathematical manipulations and more useful interpretation. Note that for probability 0.5, the odds are 1 and its logit is 0. Odds are important measures by themselves, but their importance increases manifold when they are expressed as a ratio in one group relative to another. See the topic **odds ratio** for details.

## odds ratio (OR), *see also* odds

An odds ratio (OR) is the ratio of **odds** of an event in one group to that in another group, which itself is a ratio. The groups generally are cases with disease and controls without disease. Thus, OR is a ratio of ratios. This is commonly used in **case–control studies**. The OR represents the odds that an antecedent has a role in a given disease, compared to the odds of the antecedent's role in the absence of disease. This is treated differently in two independent samples compared to matched pairs.

### *OR in Independent Samples*

The data take the form as given in Table O.3 in an independent-samples setup. In this case, $\pi_{12} = 1 - \pi_{11}$ and $\pi_{22} = 1 - \pi_{21}$.

The odds of the presence of an antecedent among cases are $\pi_{11}/(1-\pi_{11})$ and those among controls are $\pi_{21}/(1-\pi_{21})$. Thus, the OR is

$$\text{OR} = \frac{\pi_{11}/(1-\pi_{11})}{\pi_{21}/(1-\pi_{21})} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

**TABLE O.3**
**Structure of Study for OR: Independent Samples**

| | Antecedent | | |
|---|---|---|---|
| **Outcome** | **Present** | **Absent** | **Total** |
| Present (cases) | $a$ ($\pi_{11}$) | $b$ ($\pi_{12}$) | $\boldsymbol{n_1}$ **(1)** |
| Absent | $c$ ($\pi_{21}$) | $d$ ($\pi_{22}$) | $\boldsymbol{n_2}$ **(1)** |
| **Total** | $\boldsymbol{O_{\cdot 1}}$ **(**$\boldsymbol{\pi_{\cdot 1}}$**)** | $\boldsymbol{O_{\cdot 2}}$ **(**$\boldsymbol{\pi_{\cdot 2}}$**)** | $\boldsymbol{n}$ |

If the sample estimate of $\pi_{11}$ is $p_1$ and that of $\pi_{21}$ is $p_2$, OR is estimated as

$$\text{OR} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{ad}{bc},$$

where $a$, $b$, $c$, and $d$ are as in Table O.3. Since the numerator is the product of the elements in the leading diagonal of this table and the denominator is that of the elements in the other diagonal, OR is also sometimes called the **cross-product ratio**. OR loses definition when any cell frequency is zero and becomes inflated if any observed frequency is exceedingly small. In that case,

$$\text{Modified estimate of OR: } \text{OR}_{\text{mod}} = \frac{(a+0.5)(d+0.5)}{(b+0.5)(c+0.5)}.$$

The interpretation of OR is similar to that of **relative risk (RR)**, although now for the antecedent rather than for the outcome. An OR = 2 means that the presence of the antecedent is twice as common among the cases with disease as in the controls. OR = 1/3 means that the presence of antecedent is one-third as common among the cases as in the controls. This could mean that the antecedent has a protective effect. Thus, OR is also a measure of the degree of **association** between two dichotomous outcomes, and OR = 1 implies that there is no association between the exposure and the outcome.

Cornfield [1] showed that the OR approximates the RR fairly well when the outcome of interest is rare, say, less than 5%, in the target population. Most outcomes of medical interest are rare. If the outcome is not rare, an OR of more than one leads to an overestimate of RR and an OR of less than one leads to an underestimate. This can be serious if the risk or the probability of outcome is more than 20%.

Consider the data in Table O.4 on the parity status of 210 anemic and 140 nonanemic women. Anemic and nonanemic women were assessed for parity so that the design is retrospective.

$$\text{OR} = \frac{98 \times 48}{92 \times 112} = 0.46$$

Thus, the likelihood of parity ≤ 2 among anemics is less than one-half of that among nonanemics. Since anemia in women is not rare, particularly in developing countries, this OR may fail to

**TABLE O.4**
**Parity Status in Anemic and Nonanemic Women**

| **Anemia** | **Parity ≤ 2** | **Parity ≥ 3** | **Total** |
|---|---|---|---|
| Present | 98 | 112 | **210** |
| Absent | 92 | 48 | **140** |

approximate RR. If it is really rare, then the conclusion could be that parity ≤ 2 has less than one-half the risk of anemia compared to the risk in parity ≥ 3, or the risk of anemia in women with parity ≥ 3 is more than twice that for women with parity ≤ 2.

Some features of OR are as follows:

- Interestingly, OR for the antecedent is the same as OR for the disease. For both, it is $(ad)/(bc)$. Thus, these terms can be interchangeably used.
- The sample OR is always a good estimate of the population OR irrespective of the disease being rare or common in the population.
- Besides, when the disease is rare, OR is a good estimate of the RR when the cases and controls are representative of the people with disease and without disease, respectively, in the population with respect to the history of exposure.
- In case–control studies, the response refers to the presence or absence of an antecedent characteristic. For this, it could be inappropriate to use the term *incidence*. Nor does the term *risk* seem appropriate to indicate the presence of an antecedent characteristic. The term *odds* is used, which appropriately describes the situation.
- Although OR = 0.5 and OR = 2 are half and double of OR = 1, respectively, their average is not 1. OR is a ratio and logarithm removes this discrepancy—ln(OR)s can be averaged as usual, but if you need to average ORs, use the **geometric mean**.
- RR can lead to a very different conclusion depending on whether a positive outcome or a negative outcome is being measured. OR is not affected by such a consideration. For instance, consider the data in Table O.5 from a cross-sectional study of breast and lung cancer cases.

  The estimated RR of death within 1 year among breast cancer versus lung cancer cases in these data is (2/50)/(1/100) = 4. The same data also lead to the OR (2/48)/(1/99) ≈ 4. Thus, both measures indicate nearly four times the risk of dying among breast cancer patients compared with lung cancer patients. Now, use the same data, this time with respect to survival instead of death. The RR of survival is (48/50)/(99/100) = 0.97, while the OR of survival is (48/2)/(99/1) = 0.24. The RR results are different depending on whether the study is summarized with respect to death or with respect to survival. For the OR, it really does not matter. Breast cancer has approximately four times the odds of death within 1 year compared with lung cancer or approximately one-fourth the odds of survival compared with lung cancer. Both convey the same result in case of OR.
- The OR from a **cross-sectional study** is computed the same way as that from a retrospective study. However, it is necessary that the sample is a true reflection of the proportion of subjects with and without outcome as well as of the proportion with and without antecedent. Thus, for valid results, a cross-sectional sample must be a truly random sample from the target population.
- Logistic regression can be used to adjust the OR to measure the association between an antecedent and an outcome after removing the effect of intervening factors and confounders. See the topic **adjusted odds ratio (OR)** for details.
- See the topic **pooled odds ratio (OR)** for calculating combined OR when the data are available for different strata.

Since OR is a ratio, it is best dealt with after taking logarithm. Note for estimated OR that $\ln(\text{OR}) = \ln a + \ln d - \ln b - \ln c$. This is linear and makes it eligible for **central limit theorem**. Thus, ln(OR) has an approximate **Gaussian distribution** for large $n$. With this, the **confidence interval (CI) for odds ratio** can be easily worked out as described under that topic. For test of hypothesis, note that OR = 1 is the null in this case as that says that there is no association. This can again be tested by the Gaussian ***z*-test** by using its standard error (SE) as mentioned for its CI, but it is more convenient to use **chi-square test for odds ratio**.

### OR in Matched Pairs

Consider Table O.6 on matched pairs. This is similar to Table O.3 but now uses $A$, $B$, $C$, and $D$ as notation for cell frequencies in place of $a$, $b$, $c$, and $d$ to distinguish the matched-pairs setup with the independent-samples setup. Note also that the labeling of the cells has now changed.

The total number of pairs is $A + B + C + D$. In this table, $A$ is the number of pairs with both case and control subjects found exposed and $D$ is the number of pairs with both found nonexposed. These two together are the concordant pairs. The OR is computed on the basis of the discordant pairs: $B$ is the number of pairs in which the case partner is exposed but the control partner is nonexposed, and $C$ is the number of pairs in which the case partner is nonexposed but the control partner is exposed. In case of a positive association between exposure and disease, clearly $B$ should be more than $C$.

$$\text{Odds ratio (matched pairs): } \text{OR}_\text{M} = \frac{B}{C}$$

The distribution of $\ln\text{OR}_\text{M}$ is nearly Gaussian for large $n$. Large $n$ also implies that no $B$ or $C$ is small, say, less than 5. For large $B$ and $C$,

$$\text{SE}(\ln\text{OR}_\text{M}) = \sqrt{\frac{1}{B} + \frac{1}{C}}.$$

**TABLE O.5**
**Deaths in Breast Cancer and Lung Cancer Cases**

| | Deaths within 1 Year | Survival after 1 Year | Total |
|---|---|---|---|
| Breast cancer | 2 | 48 | **50** |
| Lung cancer | 1 | 99 | **100** |

**TABLE O.6**
**Matched Pairs in a Case–Control Study with Dichotomous Antecedent**

| | Controls (Partner 1) | |
|---|---|---|
| **Cases (Partner 2)** | **Antecedent Present (Exposed)** | **Antecedent Not Present (Nonexposed)** |
| Antecedent present (exposed) | *A* | *B* |
| Antecedent not present (nonexposed) | *C* | *D* |

The 95% CI for log of OR, as usual, is $\ln OR_M \pm 1.96 * SE(\ln OR_M)$. In this case, this becomes

$$\ln \frac{B}{C} \pm 1.96 * \sqrt{\frac{1}{B} + \frac{1}{C}}.$$

Take the exponential of the limits and get

$$95\%\ \text{CI for OR in matched pairs: } \frac{B}{C} e^{\pm 1.96\sqrt{\frac{1}{B}+\frac{1}{C}}}.$$

While the CI for $\ln OR_M$ in symmetric, it is not symmetric for $OR_M$ itself.

The relevant null hypothesis in this case again is $H_0$: $OR_M = 1$. To test this against a one-sided alternative $H_1$: $OR_M > 1$ or $OR_M < 1$, calculate

$$z = \frac{B - C}{\sqrt{B + C}}.$$

For large $n$, refer it to the usual Gaussian distribution to find out whether the $P$-value is sufficiently small. For a two-tailed test, it may be easier to calculate

$$\text{McNemar } \chi^2_M = \frac{\left(|B - C| - 1\right)^2}{B + C}$$

and refer it to chi-square with 1 df. This incorporates the correction for continuity.

We illustrate the calculations with the help of an example. Consider a case–control study of births with multiple malformations. The malformations considered are cleft lip, cleft palate, anal atresia, heart defects, hypospadias, and so on. They are considered multiple when at least two are present. The controls were one-to-one matched for birth order, maternal age, socioeconomic status, and the place of delivery. The objective is to find any excess of one gender over the other in such births. Suppose the data obtained are as shown in Table O.7.

In these data, $B = 60$ and $C = 50$. Thus, the odds are 1.2 times in these subjects that the malformed child is male and not female. Now,

$$SE(\ln OR_M) = \sqrt{\frac{1}{60} + \frac{1}{50}} = 0.1915.$$

Thus, the 95% CI for the OR is

$$(1.2 \times e^{-1.96 \times 0.1915},\ 1.2 \times e^{1.96 \times 0.1915}),\ \text{or } (0.82,\ 1.75).$$

**TABLE O.7**
**Sex of Children with and without Multiple Malformations**

| | Matched Control | | |
|---|---|---|---|
| **Cases** | **Male** | **Female** | **Total** |
| Male | 70 | 60 | **130** |
| Female | 50 | 40 | **90** |
| **Total** | **120** | **100** | **220** |

Note that this interval contains $OR_M = 1$, which shows that this null hypothesis cannot be rejected. Otherwise, if the null hypothesis of $OR_M = 1$ is to be tested by the criterion just mentioned, then

$$z = \frac{60 - 50}{\sqrt{60 + 50}} = 0.95.$$

This gives $P = 0.3422$ for a two-sided alternative. Also, the McNemar chi-square in this case is

$$\chi^2_M = \frac{\left(|60 - 50| - 1\right)^2}{60 + 50} = 0.74.$$

This again gives $P > 0.05$.

None of the three procedures is able to reject $H_0$: $OR_M = 1$. Thus, these data do not allow the conclusion that either gender is more prone to multiple malformations at birth.

There are several points that you can note in this example: (i) The CI is not symmetric to $OR_M = 1.2$. It is skewed to the right in this case because $1.75 - 1.20 = 0.55$ is more than $1.20 - 0.82 = 0.38$. (ii) The antecedent characteristic is not conventional exposure but is the sex in this example. (iii) Multiple malformations are rare and thus case–control is a suitable design. Several confounders that could have influenced the outcome, namely, birth order, maternal age, socioeconomic status, and place of delivery, are controlled by matching. Perhaps there are no more confounders except those in the epistemic domain. Thus, any difference between male and female children can be ascribed to their gender and, of course, sampling fluctuations.

The results obtained by three methods for $H_0$: $OR_M = 1$ are consistent in our example as all three fail to reject this null, but you might see some variation in other situations. This variation tends to vanish as $B$ and $C$ increase. Use the McNemar test if the alternative is two sided and do not use the CI because the CI uses a greater degree of approximation.

1. Cornfield J. A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. *J Nat Cant Inst* 1951;11:1269–75. http://www.epidemiology.ch/history/PDF%20bg/Cornfield%20J%20(1951)%20A%20method%20of%20estimating%20comparative%20rates.pdf, last accessed April 10, 2015.

## ogive

Ogive is a **line diagram** that plots cumulative **frequency** (or cumulative percentage) at upper end points of the **class-intervals** of a quantitative data. This is the number of subjects with values less than or equal to the upper end of each interval. Thus, this is applicable where the data are quantitative and frequencies in different class-intervals are given or can be computed. This will not apply to contingency tables based on nominal data. For the age distribution of subjects in Table O.8, this plot is shown in Figure O.2. Note that the cumulative percentages for age 49, 59, 69, 79, and beyond 79 are 11, 32, 78, 95, and 100, respectively.

In the case of Gaussian distribution of values, the ogive takes a smooth **sigmoid** shape (S shape). It is this shape that has given rise to the term *ogive*. The shape tells us where the frequencies are steeply increasing and where they are gradually increasing. Whatever the distribution, ogive can be used for obtaining approximate **quantiles** in case of grouped data. To obtain a $p$th $s$-tile, draw a horizontal line at $100p/s\%$ and read the value on the $x$ axis where this horizontal

**TABLE O.8**
**Distribution of 1000 Subjects Coming to a Cataract Clinic by Age**

| Age-Group (Years) | Number of Subjects | Cumulative Frequency | Cumulative Percentage |
|---|---|---|---|
| 0–49 | 110 | 110 | 11 |
| 50–59 | 210 | 320 | 32 |
| 60–69 | 460 | 780 | 78 |
| 70–79 | 170 | 950 | 95 |
| 80+ | 50 | 1000 | 100 |
| **Total** | **1000** | | |

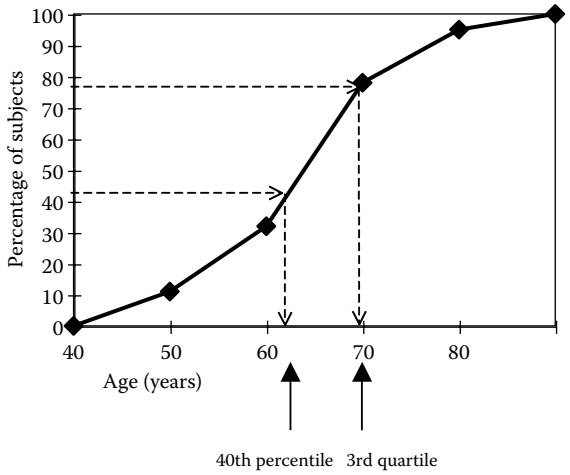



**FIGURE O.2** Ogive for the data in Table O.8 and its use for approximate calculation of some quantiles.

line intersects the percent-based ogive. Figure O.2 shows the 40th percentile and the third quartile for the data in Table O.8.

## one-sided bound, *see* confidence bounds

## one- and two-tailed alternatives/tests

You may be aware that statistical tests of hypothesis begin with setting up a null hypothesis and an alternative hypothesis. If not, see the topic **null and alternative hypotheses**. The nature of the alternative hypothesis determines that the statistical test is going to be one tailed or two tailed.

The alternative hypothesis, denoted by $H_1$, is the opposite of the null, denoted by $H_0$. The alternative must be true when $H_0$ is found false. If the null is that a drug is not effective, then the alternative has to be that it is effective. If the claim is that of superiority of the new drug over the existing one, this is the alternative hypothesis. This is a one-sided alternative if inferiority is ruled out. Mostly, it is not possible to claim that one group is better than the other, and the only claim is that they are different. In the case of peak expiratory flow rate (PEFR) in factory workers exposed to different pollutants, there may not be any a priori reason to assert that they would be affected more by one pollutant than another. Then, the alternative is that the mean PEFRs in workers exposed to different pollutants are unequal. It could be higher or lower. This is called a two-sided alternative. The null is that they are equal in various groups. One-sided and two-sided $H_1$ are also sometimes called one-tailed and two-tailed $H_1$.

It is sometimes asserted that there is no scientific reason to have a one-sided alternative—that the negative effect of a regimen can never be ruled out. This is too strong a statement, which may be mostly true but not always. Other things being equal, an exercise regimen among obese can only reduce or not reduce weight on average—it can never be suspected to increase the weight on average. Increase in weight can happen in isolated cases, but it does not seem possible when the average over the subjects is considered. Statistical tests are for groups and not for individuals. Thus, it would not be wise to set up a two-sided alternative in this case. If a regimen has passed previous two **phases of a clinical trial**, the effect can rarely be suspected to be even lower than that of placebo. Thus, in a phase III trial, if the comparison is with placebo, the alternative would be that its efficacy is better than that of placebo and the null hypothesis is that it is as good as placebo (no clinical effect). If this assertion is doubted, and lower than placebo efficacy cannot be ruled out, there must be something drastically wrong with the phase II of the trial that has established its minimum efficacy for proceeding to a phase III trial. If you want to see the effect of a hematinic, it would be absurd to suspect that it could even decrease the average hemoglobin level in a group of subjects. The only possible doubt is that it helps in specified cases or not. Note, again, that a statistical hypothesis concerns the average and not the individual subjects. Individuals can show decline, but that is not an issue in a statistical test of hypothesis setup.

One-sided alternative hypotheses could be that the parameter value is either more than a specified value or less than the specified value—one of the two. In our exercise–weight example, the hypothesis would be in terms of mean and would be $H_1$: $\mu_1 - \mu_0 < 0$, where $\mu_1$ is the mean weight after the exercise regimen and $\mu_0$ is the mean weight before the regimen. This is a left-sided alternative hypothesis. If you are testing a regimen that is expected to increase HDL cholesterol level, the alternative could be right sided, which says $\mu_1 - \mu_0 > 0$. If you are not fairly assured about the one-sided alternative, set up the two-sided alternative hypothesis. This would be $H_1$: $\mu_1 - \mu_0 \neq 0$. Often, this would be the case.

One-sided alternatives require a one-tailed test and two-sided alternatives require a two-tailed test. A one-tailed test could be right tailed or left tailed depending on the side of the alternative. A right-tailed test means that the null hypothesis is rejected when the value of the test criterion is *more* than the critical value at the specified level of significance. A left-tailed test means that the null hypothesis is rejected when the value of the test criterion is *less* than the critical value. For a Gaussian ***z*-test**, at 10% level of significance, the critical value is –1.28 for a left-tailed test. Recollect that level of significance is the threshold of the probability of a Type I error. If the sample values give a value of $z$ less than this value, the null is rejected in favor of the left-sided alternative. This is shown in Figure O.3b. For a two-sided alternative, the critical region would be both tails, mostly equally divided in the two tails. For a Gaussian $z$-test, this critical value at the 10% level of significance is either less than –1.645 or more than +1.645—each tail containing 5% probability (Figure O.3a).

You can see that it is easier to reject the null in favor of the one-sided alternative compared with the two-sided alternative. In our illustration, a left-sided alternative would be rejected at a 10% level of significance when the $z$-value from the sample is <–1.28 but if it is a two-sided alternative, you would reject if $z < -1.645$. The former is easier to achieve than the latter. The objective of most studies is to collect evidence to reject the null. Thus, it is preferable to set up a one-sided alternative rather than a two-sided alternative. However, caution remains that this should be done only if you are fairly assured that the other side of the alternative is not plausible for your study.

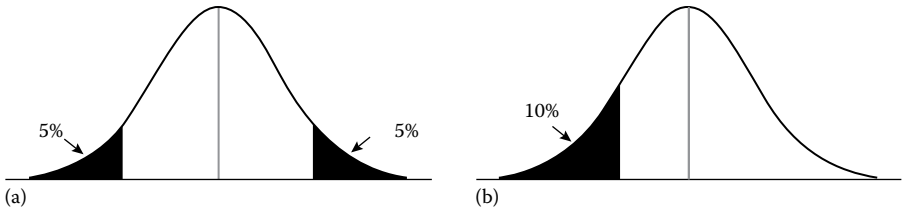



**FIGURE O.3** Two-tailed and one-tailed rejection regions for Gaussian distribution. (a) Two-tailed test. (b) One-tailed test (left tail).

Certain statistical tests are naturally two tailed, whereas some others can easily test one-sided alternatives. The chi-square test for association in a contingency table is the most prominent two-tailed test—first, because it is based on square of values (this also converts negative values to a positive number), and second, because it ordinarily considers categories as nominal. It can be devised as a one-tailed test but is seldom done. Tests for multiple groups such as analysis of variance *F*-test are also naturally two tailed—the alternative being unequal means across groups. On the other hand, Gaussian *z*-test for proportions and Student *t*-test for means in one or two groups can easily test one-sided as well as two-sided alternatives.

For another discussion on one- and two-tailed tests, see Peace [1].

1. Peace KE. The alternative hypothesis: One-sided or two-sided? *J Clin Epidemiol* 1981;42:473–6. http://www.ncbi.nlm.nih.gov/pubmed/2732775

## one-way ANOVA, *see also* analysis of variance (ANOVA)

One-way analysis of variance (ANOVA) is a method to find whether the mean in any of three or more groups is significantly different. The corresponding procedure for one and two groups is the **Student *t*-test**.

Consider a study in which the plasma amino acid (PAA) ratio for lysine is calculated in healthy children and in children with undernourishment of grades I, II, and III. This ratio is the difference in PAA concentration in blood before and after a meal, expressed as a percentage of the amino acid requirement. There are four groups in this study. The setup is called one way because no further classification of subjects, say, by age or gender, is sought in this case. Groups define the factor: in this case, grade of undernourishment. The response is a quantitative variable. When other factors are properly controlled, the difference in PAA ratio among subjects would be attributed to either the degree of undernourishment or the intrinsic interindividual variation in the subjects in different groups. The former is the between-groups variation and the latter is the within-groups variation. These variations are illustrated in Figure O.4 for a variable observed for four different groups. Group means are denoted by $\bar{y}_{\bullet j}$ ($j$ = 1, 2, 3, 4) and are represented by a circle. Within-groups variation is the difference between individual values and their respective group mean. This is summed over the groups. The overall mean is denoted by $\bar{y}_{\bullet\bullet}$ and is represented by a line in this figure. Between-groups variation is the difference between group means and the overall mean.

Note that part of the within-groups variation in the preceding example can be attributed to factors such as heredity, age, gender, height, and weight of the children. However, these are assumed under control and disregarded in this setup. All within-groups variation is considered intrinsic and random. In an ANOVA setup, this is generally called residual or error variance. This is measured by **mean square due to error**, popularly written as MSE. The term

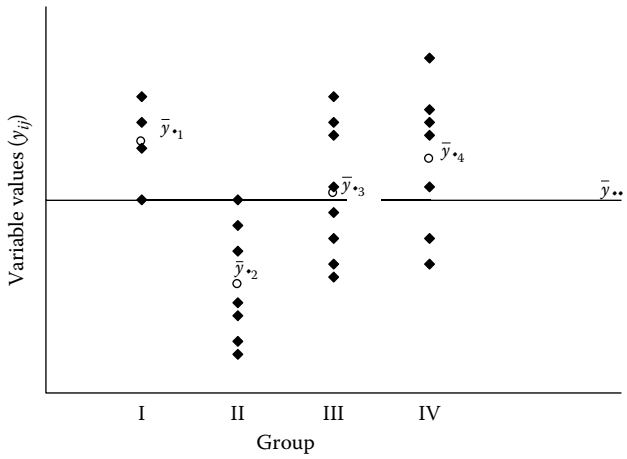



**FIGURE O.4** Graphical display of within-groups and between-groups variance.

*error* does not connote any mistake but stands only for random component, as for sampling error.

If the group differences were not really present, both the between-groups variance (call it mean square between groups and write it as MSB) and the within-groups variance would arise from intrinsic variation alone and will be nearly equal. The ratio of these two, with between-groups variance in the numerator, is the criterion *F* (for this reason, this is also called **variance ratio**). With MSB in the numerator, a value of *F* of substantially more than one implies that between-groups variation is large relative to within-groups variation. This is an indication that the groups are indeed different with respect to the mean of the variable under study relative to the intrinsic variation.

### *The Procedure to Test $H_0$*

The null hypothesis in this setup is that means in all groups are the same. A prerequisite for validity of the ANOVA procedure is that all groups have the same variance—called **homoscedasticity**. It is also stipulated in ANOVA that the pattern of distribution of the response variable is the same. This could be of almost any form if the number of subjects in each group is large. One possibility is shown in Figure O.5. The distribution in all four groups in this figure

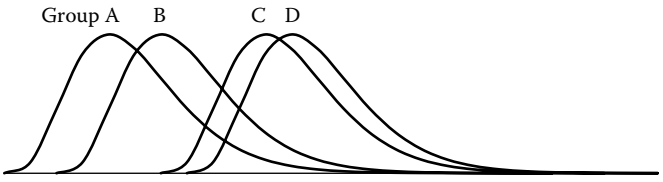



**FIGURE O.5** Distribution in four groups differing only in mean.

is identical except for a shift in location. Groups B and C are far apart from one another, but C and D are close. Only the means differ and other features of the distribution are exactly the same. If $n$ is small, the ANOVA procedure is valid only when this distribution is Gaussian. Thus, skewed distributions of the type shown in Figure O.5 are admissible for ANOVA only when $n$ is large.

In terms of notations, the null hypothesis in the case of one-way ANOVA is

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_J.$$

This says that there are $J$ groups and the means in all groups are the same. This hypothesis, in conjunction with the conditions mentioned in the preceding paragraph, implies that the response variable has the same distribution in different groups. The alternative hypothesis is that at least one mean is different.

When $H_0$ as just stated is true, mean square between groups (MSB) is also an estimate of the population variance $\sigma^2$. Mean square error (MSE) is an estimate of $\sigma^2$ whether or not $H_0$ is true. The ratio

$$\text{(A)} \quad F = \frac{\text{MSB}}{\text{MSE}}$$

is expected to be one under $H_0$. When $F \le 1$, it is surely an indication that group means can be equal. If group means are different, MSB would be large and $F \gg 1$ (substantially more than 1). Just like $\chi^2$ and $t$, the distribution of $F$ under $H_0$ is known. In place of a single df, the exact shape of $F$ depends on a pair of df, namely, $(J - 1)$, $J(n - 1)$ in case of one-way ANOVA. The first corresponds to the numerator of the criterion given in Equation A and the second corresponds to the denominator. In general, these df's are denoted by $(v_1, v_2)$, respectively. The probability $P$ of wrongly rejecting $H_0$ can be obtained corresponding to the value of $F$ calculated from the data. If $P < 0.05$, the evidence can be considered sufficient to reject $H_0$ at a 5% level of significance. Standard statistical packages provide the $P$-value so that a decision can be made immediately.

The following comments provide further information on one-way ANOVA:

- The primary purpose of ANOVA is to test the null hypothesis of equality of means. However, an estimate of the effect size of the groups can also be obtained as a by-product of this procedure. For this, see the topic **main effects and interaction in ANOVA**.
- ANOVA was basically developed for evaluating data arising from experiments. This requires **random allocation** of the subjects to different groups and then exposing these groups to various interventions. In the PAA example, the groups are preexisting and there is no intervention. The design is prospective because the undernourishment groups define the antecedent and the PAA ratio is investigated as an outcome. The subjects are not randomly allocated to the different undernourishment groups. However, it is important to ensure that the subjects in each group adequately represent their group and that there is no other factor, except the undernourishment grade, that separates the groups. Such precautions in selection of subjects are necessary for a valid conclusion from ANOVA.
- Although the ANOVA method has been described here for comparing means in three or more groups, the method is equally valid for $J = 2$ groups. In fact, there is a mathematical relationship that says that two-tailed $t^2$ for two independent samples of size $n$ is the same as $F$ with $(1, 2n - 2)$ df. Note that the df of Student $t$ for two independent samples, each of size $n$, is $(2n - 2)$. Both methods lead to the same conclusion, but $F$ is a two-tailed procedure, whereas $t$ can also be used for a one-sided alternative. This flexibility is not available with the $F$-test.

For illustration, consider the following example. Tokunaga et al. [1] conducted a study for effects of some $H_1$-antagonists on the sleep–wake cycle in sleep-disturbed rats. Among the response variables was rapid eye movement (REM) sleep time. Take a similar example of a sample of 20 rats, homogeneous for genetic stock, age, gender, and so on, which are randomly divided into $J = 4$ groups of $n = 5$ rats each. Let one group be the control and the others receive drug A (diphenhydramine), drug B (chlorpheniramine), and drug C (cyproheptadine). The REM sleep time was recorded for each rat from 10:00 to 16:00 h. Suppose the data shown in Table O.9 are obtained.

These data show a large difference in the mean sleep time in various groups. Another experiment on a new sample of rats may or may not give similar results. The likelihood of getting nearly equal means in the long run is extremely remote, as will be clear shortly.

To avoid the burden of calculations, given below are the values obtained from a statistical package for these data.

$$\text{SST} = 7369.8, \text{SSB} = 5882.4, \text{ and SSE} = 1487.4,$$

where SST is the total **sum of squares**, SSB is the between-groups sum of squares, and SSE is the error (within-groups) sum of squares. Mathematical expressions of these are being avoided because of complexity. However, note that SST = SSB + SSE in a one-way setup.

The degrees of freedom for SSB are $(4 - 1) = 3$ and those for SSE are $4(5 - 1) = 16$. Mean squares are obtained by dividing these sums of squares by the corresponding df's. These values, as well as the value of $F$, are conventionally shown in the form of a table, popularly called an *ANOVA table*. This is given in Table O.10 for this example. For these data,

$$F = \frac{5882.3/3}{1487.4/16} = 21.09.$$

**TABLE O.9**
**REM Sleep Time in Sleep-Disturbed Rats with Different Drugs**

| Drug | REM Sleep Time (min) | | | | | Mean (min) |
|---|---|---|---|---|---|---|
| 0 (control) | 88.6 | 73.2 | 91.4 | 68.0 | 75.2 | 79.28 |
| A | 63.0 | 53.9 | 69.2 | 50.1 | 71.5 | 61.54 |
| B | 44.9 | 59.5 | 40.2 | 56.3 | 38.7 | 47.92 |
| C | 31.0 | 39.6 | 45.3 | 25.2 | 22.7 | 32.76 |

**TABLE O.10**
**ANOVA Table for the Data in Table O.9**

| Source of Variation | df | Sum of Squares | Mean Squares | F |
|---|---|---|---|---|
| Drug | 3 | 5882.4 | 1960.8 | 21.09 |
| Error | 16 | 1487.4 | 93.0 | |
| **Total** | **19** | **7369.8** | | |

The null hypothesis is that there is no effect of dose on REM sleep time. This is the same as saying that means in all the four groups are equal.

A statistical package gives $P < 0.001$ under $H_0$ for $F = 21.09$ or higher. That is, there is less than one in a thousand chance that equal means in groups will give a value of $F$ this large. Therefore, these data must have come from populations with unequal means. In other words, another experiment on the same kind of animals is extremely unlikely to give equal means. The evidence is overwhelming against the null and it is rejected. The conclusion is that different drugs do affect the REM sleep time differentials.

In this example, mean REM sleep time in different drugs not only differs but also follows a trend. Had the groups represented increasing dose levels, it would have implied decline in sleep time as the dose level is increased. The conventional ANOVA just illustrated allows conclusion of different means in different groups but not of any trend. Evaluation of trend in means is discussed under the topic **regression**.

In our example, there are only five rats in each group. Statistically, this is an extremely small sample. The reason that such a small sample can still provide a reliable result is that the laboratory conditions can be standardized and most factors contributing to uncertainty can be controlled. The rats can be chosen to be homogeneous, as in this experiment, so that intrinsic factors such as genetic makeup, age, and gender do not influence the outcome. Random allocation tends to average out any effect of other factors that are not considered in choosing the animals. The influence of body weight, if any, is taken care of by adjusting dose for body weight. Thus, interindividual variation within groups is minimal. On the other hand, the variation between groups is very large in this case, as is evident from the large difference between the means. This provided clinching evidence in favor of the alternative hypothesis.

### *Cautions in Using ANOVA*

The following are cautions in using ANOVA:

1. The ANOVA is based on means. Any means-based procedure is severely disturbed when outliers are present. Thus, ensure before using ANOVA that there are no outliers in your data. If there are, examine whether they can be excluded without affecting the conclusion.
2. A problem in comparison of three or more groups by criterion $F$ is that its significance indicates only that a difference exists. It does not tell exactly which group or groups are different. Further analysis, called **multiple comparisons**, is required to identify the groups that have different means.
3. When no significant difference is found across groups, there is a tendency to locate a group or even a subgroup that exhibits benefit. This post hoc analysis is alright as long as it is exploratory in nature. In conclusion, a new study should be conducted on that group or subgroup.
4. The following are requirements for validity of the ANOVA $F$-test: (i) independence of the observations, (ii) homoscedasticity across groups, and (iii) Gaussian distribution of means in different groups. The last is generally fulfilled by the **central limit theorem** when the sample size in each group is large. The small numbers of subjects in different groups in our rat example should put you on alert regarding the pattern of distribution of the measurements and of their means. It should be Gaussian. See the topics **Guassianity (how to check)** for details. When the pattern is far from Gaussian, consider a **transformation** or nonparametric **Kruskal–Wallis test** for one-way ANOVA. For checking **homoscedasticity**, see that topic. If the variances are really unequal and the number of subjects in the groups is also widely different but large, the **Welch test** is used for testing the hypothesis of quality of means. For small samples and unequal variances, consider the **Brown–Forsythe test**. Independence is the most important requirement. This is checked by the **Durbin–Watson test** for autocorrelation and by **intraclass correlation** for clusters. If you are measuring electronic waves at six different sites of the brain in each subject, and consider each site as a group, the independence is lost since these measurements belong to the same person. Serial observations, such as in a time series or repeated measures, also violate the independence requirement. See **repeated measures ANOVA** for the analysis in this setup.
5. The procedure we have mentioned is for fixed effects. See the topic **fixed and random effects** for the distinction between the two. See the topic **random effects ANOVA** for analysis in case of random effects.

1. Tokunaga S, Takeda Y, Shinomiya K, Hirase M, Kamei C. Effect of some $H_1$-antagonists on the sleep–wake cycle in sleep-disturbed rats. *J Pharmacol Sci* 2007;103:201–6. https://www.jstage.jst.go.jp/article/jphs/103/2/103_2_201/_pdf

## one-way designs, *see also* two-way designs

Many times, the subjects of a study are classified into groups either to attain homogeneity within groups or to study the effect of the grouping factor. If the objective is to see how blood group affects a particular outcome such as blood glucose level, the subjects will be divided by blood groups. This is the grouping factor in this case. This is called one-way design when the division of subjects is by just one factor. Blood group is the factor and the actual values of the factor such as A, B, AB, and O are called the *levels of the factor*. Statistically, the factor is the independent variable in this setup and the blood glucose level is the dependent variable. Contrast it with **two-way** or multi-way designs where the subjects are divided according to the levels of two or more factors. In addition to the blood group, you can also divide the subjects by sex and by obesity. Then, it will be a three-way design. This will have three independent variables. In this setup, the independent variables are always **categorical** and cannot be continuous. Thus, these are better understood as factors and not as variables.

A one-way design can arise in three different ways. One is that you have a sample of subjects and you divide them into groups as per your choice. This happens when you have a group of tobacco users and you divide them by the age-groups you choose. You can have age-groups <15 years, 15–34 years, 35–54 years, and 55+ years, or if you want, you can divide groups into 10-year intervals. The choice is yours, although it may be guided by which groups you consider homogeneous with respect to tobacco use. Different groupings can give different results, and this should be considered when interpreting the results. However, a person of age 23 years has to be in his/her age-group. Second is that the groups are already there and you just classify them such as males and females. Nature forms these groups and not human beings. In both these setups, you cannot decide as to which subject will go to which group. Their age or sex will decide the group. Third is that you have one group that will receive treatment A, a second group receiving treatment B, and a third group receiving treatment C, and you can randomly allocate the subjects to these treatments after they are found eligible with inclusion and

exclusion criteria. This is done in most clinical trials, or you can decide on the basis of their clinical condition as to who will receive which treatment. In this situation, the choice is yours.

Consider the data in the table under the topic **one-way ANOVA**, where the rats are divided into four groups of five rats each and their REM sleep time is recorded. In this example, the data are quantitative and the number of subjects in each group is five. They may or may not be equal groups. If the data are the number of subjects (frequencies) in place of the quantitative values, this is called a one-way **contingency table**. If there is a sample of 100 subjects from a hypertension clinic and the number of subjects with normal blood pressure (BP), with mild hypertension, and with severe hypertension is observed, this would be a one-way classification. The focus is on the number of subjects in different groups, and the role of actual levels of BP is only in categorization. The data are not the quantitative values but are the number of subjects in different groups in this case.

The quantitative data in a one-way design is analyzed by one-way ANOVA where means are compared, and the data in a one-way classification is analyzed by chi-square **goodness-of-fit** where frequencies are compared. The null hypothesis in the former case is that the means of the quantitative values in different groups is the same, whereas in the second case, the hypothesis would be that the frequencies follow a particular pattern.

## open trial

In contrast to a **blinded trial**, an open trial (fully, an open-label trial) is one in which the subjects know which treatment he/she is receiving and the researcher also knows which treatment is being given to whom. This "openness" in such trials raises the specter of bias because of the distinct likelihood of prejudiced responses from the subjects on one hand and unfair assessment by the assessor on the other. Thus, open trials are not advised. They may still have to be used in situations where blinding is extremely difficult. For example, a trial on comparison of medical treatment with surgical treatment cannot be blinded.

Subjects in an open trial can still be randomized. In this case, the subjects are told after randomization which treatment they are receiving. Randomization helps in increasing the chances of baseline equivalence of the subjects in the different groups, particularly when the groups are large. One arm can be control, which can be either placebo or the existing regimen. Thus, a **randomized control trial (RCT)** can be an open trial or a blinded trial, and a double-blind RCT is considered the gold standard.

As mentioned, there are situations where blinding is not feasible. Gupta and Gupta [1] conducted an open trial on patients with plantar warts treated with adapalene gel 0.1% under occlusion in one group and with cryotherapy in the second group. Bhatia et al. [2] describe the protocol of an open trial to evaluate the impact of yoga on cognitive functions in schizophrenia. There are many situations where there is no choice. In all these trials, special care should be taken so that the responses are not biased and the observers carry out and report unbiased readings.


1. Gupta R, Gupta S. Topical adapalene in the treatment of plantar warts; randomized comparative open trial in comparison with cryotherapy. *Indian J Dermatol* 2015 Jan–Feb;60(1):102. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4318023/
2. Bhatia T, Mazumdar S, Mishra NN, Gur RE, Gur RC, Nimgaonkar VL, Deshpande SN. Protocol to evaluate the impact of yoga supplementation on cognitive function in schizophrenia: A randomised controlled trial. *Acta Neuropsychiatr* 2014 Oct;26(5):280–90. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4342363/


## operations research

Research focused in operations, in contrast to theories and statistical methodologies, is called operations research (OR) (or operational research). For example, observational studies or clinical trials are not OR but optimization of resources for controlling a swine flu epidemic is. OR has an overbearing approach to solve existing or upcoming problems instead of generalized approaches. What steps can help minimize the waiting time in the emergency department of hospitals and what can improve the turnaround time of patients admitted in wards are examples of OR. It examines trade-offs among various options and help us take a decision that meets a specific objective.

OR is a methodological driven discipline with focus on specific applications. It requires a variety of analytical methods, including statistical methods, to reach a conclusion. OR-based modeling can help in providing a structured framework that uses the best available evidence to capture relevant uncertainties, complexities, and interactions. It recognizes that most problems are multifaceted and draws its strength from an interdisciplinary approach to arrive at an ideal sort of solution within the specified constraints. However, mathematics plays a key role.

Two basic tools that OR uses are optimization and **simulation**. The first is generally calculation intensive and the second in any case requires a computer. Zai et al. [1] have used simulations for optimizing a population management system for cancer screening. Mistry et al. [2] proposed an optimization approach for controlling drug-resistant tuberculosis in Mumbai, India, that could include effects of newer technologies, changing the balance of ambulatory and inpatient care, the effects of initiatives to improve infection control, and so on. Bartsch et al. [3] used a simulation model to study the spread and control of norovirus outbreaks among hospitals in the United States.


1. Zai AH, Kim S, Kamis A, Hung K, Ronquillo JG, Chueh HC, Atlas SJ. Applying operations research to optimize a novel population management system for cancer screening. *J Am Med Inform Assoc* 2014 Feb;21(e1):e129–35. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3957383/
2. Mistry N, Tolani M, Osrin D. Drug-resistant tuberculosis in Mumbai, India: An agenda for operations research. *Oper Res Health Care* 2012 Jun;1(2–3):45–53. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3836418/
3. Bartsch SM, Huang SS, Wong KF, Avery TR, Lee BY. The spread and control of norovirus outbreaks among hospitals in a region: a simulation model. *Open Forum Infect Dis* 2014 Jul 2;1(2):ofu030. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4281820/


## optimistic index

This measures the extent of overfit (optimistic fit) of a **model** to a data set because it is based on the same data. A model based on any data set is likely to be a good fit because the calculations are based on the same data. This model may not apply with the same predictivity to a new data set. The optimistic index tries to correct the predictivity so that a realistic assessment is available regarding the practical utility of the model.

Among several methods, the most common one for finding the optimistic index for a model involves repeated fitting the model to a bootstrap subsample from the original sample and calculating the predictivity each time. This would invariably be lower than the predictivity you originally obtained based on the entire sample, since now only subsamples are being used. Average predictivity over, say,

100 bootstrap samples would be considered the actual predictivity. Difference between the original predictivity and the average predictivity based on bootstrap samples would be the optimistic index. This measures how much the model is overpredicting.

Suppose you fit a logistic model to predict mortality in acute necrotizing pancreatitis cases with a database of 400 cases. Let the classification accuracy of the model for these data be 86%. That is, the model is able to correctly predict survival or death in 86% of the cases but was wrong in the other 14% of the cases in the data set used for developing the model. In these 14%, the patient died when the model predicted survival or the patient survived when the model predicted death. Now, fit the same model to a random subsample of, say, 300 subjects from the available 400 and suppose it is found that the predictivity is 82%. Take another random subsample of 300 cases and find that the predictivity is, say, 75%. Do this 100 times. Suppose the average predictivity in these 100 random subsamples of 300 patients is 79%. Since the original predictivity is 86%, the optimistic index is 86% − 79% = 7%. The model seems to be overpredicting to an extent of 7%. You can see that the index may depend on the size of bootstrap samples. A smaller sample may have a lower predictivity not because of the small sample but because the smaller sample may not represent the full spectrum of subjects.

## ordered alternatives, *see also* one- and two-tailed alternatives/tests

These are extension of **one-tailed alternatives** hypotheses to more than two groups. In case of two groups with metric data, the one-tailed alternative takes the form $H_1$: $\mu_1 < \mu_2$ or $H_1$: $\mu_1 > \mu_2$. When the number of groups is more than two, the ordered alternative is $H_1$: $\mu_1 \le \mu_2 \le \mu_3 \le \ldots \le \mu_K$, or $H_1$: $\mu_1 \ge \mu_2 \ge \mu_3 \ge \ldots \ge \mu_K$, with at least one strict inequality, where $K$ is the number of groups. These are also called *directional alternatives*. The null is the same, that is, that means in all the groups are equal $H_0$: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_K$. These kinds of alternatives are possible only for ordered groups (such as none, mild, moderate, serious, and critical). For nominal groups, such as subjects belonging to blood groups, O, A, B, and AB, such ordered alternatives would not be applicable. For some other apparently nominal groups, such as cancer of blood cells (leukemia), lungs, prostate, and mouth, you may still be able to set up an ordered alternative for, say, average duration of survival if you believe that the cancers of different sites can be arranged by duration of survival.

Statistical methods for comparing more than two groups are in any case complex but they become even more intricate when the alternative hypothesis is ordered. Globally acceptable methods are yet to be developed for this setup. Among parametric tests for Gaussian distributed variables are the Bartholomew test and the Williams test as discussed by Alainentalo [1]. Among nonparametric tests based on ranks, the Jonckheere test (also called the Jonckheere–Terpstra test) for trend is meant for an ordered alternative [2]. Shan et al. [3] have proposed a new test, claimed to be more powerful, based on rank difference. Rice and Gaines [4] have proposed an ordered heterogeneity test for ordered variances. These are rarely used in health and medicine and we have not included these tests in this book.


1. Alainentalo L. *A Comparison of Tests for Ordered Alternatives with Application in Medicine*. Bachelor's Thesis, Department of Mathematical Statistics, Umeå University 1997. http://www.diva-portal.org/smash/get/diva2:479010/fulltext01.pdf, last accessed April 12, 2015.
2. Neuhauser M. *Nonparametric Statistical Tests: A Computational Approach*. Chapman & Hall/CRC, 2011.
3. Shan G, Young D, Kang L. A new powerful nonparametric rank test for ordered alternative problem. *PLoS One* 2014 Nov 18;9(11):e112924. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4236087/
4. Rice WR, Gaines SD. Extending nondirectional heterogeneity tests to evaluate simply ordered alternative hypotheses. *Proc Natl Acad Sci USA* 1994;91:225–6. http://www.pnas.org/content/91/1/225.full.pdf


## OR, *see* odds ratio (OR)

## order of a table, *see* contingency tables

## order statistics

When quantitative **observations** of a variable are ascendingly ordered, these are called order statistics. If total bilirubin level in milligrams per deciliter in six subjects is $x_1 = 1.14$, $x_2 = 0.97$, $x_3 = 1.20$, $x_4 = 1.28$, $x_5 = 0.93$, and $x_6 = 1.15$, the order statistics are $x_{[1]} = 0.93$, $x_{[2]} = 0.97$, $x_{[3]} = 1.14$, $x_{[4]} = 1.15$, $x_{[5]} = 1.20$, and $x_{[6]} = 1.28$ from minimum to maximum. The $i$th order statistics is denoted by $x_{[i]}$. It is customary to use square brackets as subscript for order statistics, and the value in these brackets is the rank of the value. Thus, the value 1.14 has rank 3 and the value 1.28 has rank 6. In general, for observations $(x_1, x_2, \ldots, x_n)$, the order statistics are $x_{[1]} \le x_{[2]} \le \ldots \le x_{[n]}$.

Some sample summaries are easily expressed in terms of order statistics. For example, the sample range is $x_{[n]} - x_{[1]}$ and the sample median is $x_{[(n+1)/2]}$ if $n$ is odd, and the average of $x_{[n/2]}$ and $x_{[n/2+1]}$ if $n$ is even. Similarly, all **quantiles** can be easily expressed in terms of order statistics, including the 2.5th and 97.5th percentiles we use to find the **nonparametric** 95% confidence interval for the median in case of highly **skewed** data. Order statistics are commonly used to devise nonparametric tests since these tests are based on ranks instead of the actual values.

**Sampling distribution** of all order statistics can be obtained. This can be used to find, for example, the distribution of the minimum value, which, in turn, can help in estimating the minimum possible value in the population of a variable with reasonable confidence. For example, if you have total bilirubin level in a sample of six subjects as in our example, the distribution of the minimum can tell you what minimum threshold becomes unlikely in the population from which this sample has been taken, provided the subjects are randomly chosen. The minimum in our sample is 0.93 mg/dL, but the statistical distribution of the minimum may tell you that the lower limit is 0.75 and the chance of anything lower than this in the population is extremely small, say, less than 5%. Similarly, if the minimum efficacy of a regimen in a series of trials is 56%, you can find how low it can go and decide if the regimen is worth pursuing. Similarly, the upper limit of the maximum can also be obtained with a reasonable confidence. This can also be used to detect **outliers** with unlikely high values. If the maximum value happens to be 2.38 in your sample and the next is 1.83, the question is that 2.38 is or is not likely to have come up from the population from which other values have come. We can find the probability that the largest value can be 2.38. If this probability is extremely small, this can be discarded as being wrongly reported or wrongly measured.

The order does not change under linear transformation of the type $y = a + bx$ when done for all the values of $x$. However, one big problem with order statistics is a tie between two or more values. If the fourth and fifth values in increasing order are both 1.23 mg/dL, what order do they belong? The answer is 4.5. Both are given the same rank. The next rank will be 6 as 4.5 takes away both ranks 4

and 5. Such ties are known to create problems in nonparametric tests and they require modification.

For further details of order statistics, see Arnold et al. [1].

1. Arnold BC, Balakrishnan N, Nagaraja HN. *A First Course in Order Statistics*. SIAM, 2008.

## ordinal association, *see* association between ordinal characteristics (degree of)

## outcome variables

The term *outcome* is used in several different contexts in medical research, and it is difficult to list all of them. The outcome of all medical research should be improved health care, but our concern here is with the outcome variables. The term is better contextualized in an antecedent–outcome setup where there are some inputs, some processes, and then some outputs. Antecedents could be an intervention, risk factors, or exposures, and outcome could be positive such as improvement in health or could be adverse such as side effects. In a clinical trial setup, the antecedents could be the regimen itself, its compliance, the preexisting disease severity, the quality of care, and so on, and the outcomes could be in terms of survival, duration of hospitalization, extent of relief, and so on.

Other terms for outcome are effect, response, and result. This always is a consequence and occurs after the antecedents. However, there are setups where the distinction between antecedent and outcome is blurred. Gender and blood group are determined simultaneously and none is the outcome of the other. They are not antecedent–outcome, but the association between them can still be investigated. Anxiety can cause disease and disease can cause anxiety. Cholesterol level is an antecedent for coronary outcomes but is an outcome for cholesterol-level–reducing agents. So long as you are clear in your mind and able to state it explicitly, such anomalies will not create problems.

In an **analytical study**—observational or experimental—where an antecedent–outcome framework is an essential ingredient, proper specification of the outcome variables is of paramount importance. Outcome variables depend on the main and secondary objectives of the study. For example, in patients undergoing mitral valve surgery, the outcome of interest could be sinus bradycardia, sinus arrest, atrial fibrillation, and prolonged asystolic period. It is important for research that each of them is precisely stated: when a sinus bradycardia would be called inappropriate, what would be considered prolonged asystolic period after tachycardia or otherwise, when will they be assessed—if repeatedly assessed, what time point will be considered most important for conclusion or whether trend would be considered, and so on. All these should be fully specified in the **protocol** itself so that there is no arbitrariness later on. Also, they must be carefully chosen so that they correspond with the objectives. They must be measurable and, as far as possible, should be on the metric scale because this is far more objective than ordinal and nominal scales.

In the statistical independent and dependent variable dichotomy, outcomes are dependents and the antecedents are independents. When the outcome is binary and the objective is to investigate the extent and form of relationship between the antecedents and outcome, **logistic regression** is the statistical method of choice. If the outcome is metric, ordinary least squares **regression** is used. In both these setups, the statistical requirements such as independence, homoscedasticity, and Gaussian distribution apply to the outcome variable and not to the antecedent variables. It is the nature of the outcome variable that primarily decides which statistical method is appropriate for a set of data. Thus, the outcome must be carefully chosen. Also, consider whether the outcome variable is the actual or a **surrogate**. Surrogate can be used when the actual is not measurable or extremely expensive to measure. If surrogate is used, be cautious in drawing conclusions. In this context, the article by Senn and Julious [1] may be of interest.

1. Senn S, Julious S. Measurements in clinical trials: A neglected issue for statisticians? *Stat Med* 2009;28:3189–209. http://onlinelibrary.wiley.com/doi/10.1002/sim.3603/pdf

## outliers

Outliers are those obstinate values that are located in isolation in the graph—not fitting with the others. Many of these can be attributed to errors such as misplaced decimal, wrong method of measurement, wrong reporting, misuse of instrument, and wrong record. Sometimes, rarely though, the outlier may be right and other values may be wrong, such as when the first few values are model-based estimates that are way off the actually observed most values. At other times, an outlier may be a genuine out-of-the-box value for a very unusual person or patient.

In Figure O.6a, the percentage of cholesterol esters is shown as dependent on total bilirubin in cases of cirrhosis, hepatitis, and common bile duct obstruction. In the presence of regurgitation jaundice, the flow of bile into the duodenum may decrease, causing a reduction in the esterified form of cholesterol. Note an outlier in this figure. Scatter diagrams easily provide the opportunity to spot such outliers.

Outliers are obviously defined in terms of data spread. Beside the scatter diagram as just mentioned, outliers can be spotted by a **box-and-whiskers plot**. In this plot, values more than 1.5*IQR (IQR is the

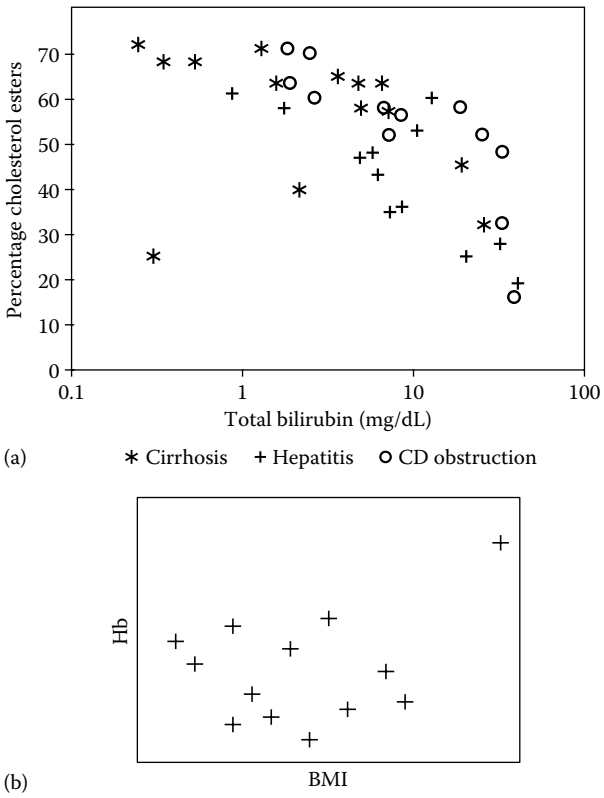



**FIGURE O.6** (a) Scatter diagram showing an outlier. (b) One outlier gives an impression of a trend.

**interquartile range**) away from the median are marked as marginal outliers and those more than 3*IQR away are marked as clear outliers. **Order statistics** can also be used to detect outliers with unlikely high values. If the maximum value happens to be 1.83 in your sample and the next is 2.38, the question is that 2.38 is or is not likely to have come up from the population from which other values have come. We can find the probability that the largest value can be 2.38 by using the distribution of order statistics. If this probability is extremely small, this can be discarded as being wrongly reported. A similar procedure can be used for unusually low values. A large absolute value of a **residual** for any particular $y$ in a **regression** is a definite indication that an outlier is present in the data, but that cannot be relied upon as a sole indicator. An outlier would most likely affect the entire regression, and thus many residuals could be moderately large and escape attention. Careful examination of residuals is required to find whether one or more outliers are present. Some software packages have an option to flag the outliers, which makes their spotting easier.

One or more outliers have the potential to distort the results—they have a disproportionate effect. The mean is the first one to be affected (the median is not affected). The mean of 5 values ranging from 7 to 20 is very different if one high value such as 36 is added. Because of this, results of all mean-based statistical tests such as Student *t*-test and ANOVA are vulnerable when one or more outliers are present. A large sample tends to moderate such distortion. Range is surely affected as everyone realizes but many do not realize that an outlier can substantially vitiate the standard deviation (SD) as well. This occurs because of the distorted value of the mean that is so intimately used in calculating the SD. Also, one outlier can present an artificial trend in regression analysis as in Figure O.6b where all other values are randomly located with no trend.

When one or more outliers are noticed in a data set, the first step is to examine whether they are genuine values or are due to assessment or data entry errors. If they are due to errors, they can be safely excluded provided they do not follow any pattern and are not too many. If there is a pattern or are too many, think of what may have gone wrong in terms of assessment or data entries. Genuine outliers are difficult to manage as their exclusion can distort the results. They can also arise as a result of missing important factors in the study that determine these values. If that is not the case, the best course would be to present results with and without these outliers and explain the difference. The other option is to use **nonparametric methods** since they are not much affected by outliers. They are based on ranks and the rank of the highest value is the same whether the highest value is 36 or 76.

On the positive side, genuine outliers can help you develop new **hypotheses**. Your subjects may suggest what possibly could have caused these values and one or more of these causes may not have been stipulated. In case the hypothesis looks plausible, you can plan a systematic study to examine this hypothesis, and present your results.

## overanalysis, *see also* data-dredging

Popularly termed as torturing the data until it confesses, data are sometimes overanalyzed, particularly in the form of post hoc analysis. A study may be designed to investigate the relationship between two specific measurements, but correlations between pairs of a large number of other variables, which happen to be available, are calculated and examined. This is easy these days because of the availability of computers.

When many correlations are examined and each is tested for statistical significance at a level of $\alpha = 0.05$, the total probability of **Type I error** increases enormously. Also, $\alpha = 0.05$ implies that 1 in 20 correlations can be concluded to be significant when actually it is not. If measurements on 16 variables are available, the total number of pairwise correlations is $16 \times 15/2 = 120$. At the error rate of 5%, 6 of these 120 can turn out to be falsely significant. Thus, there is clear risk of interpreting noise as signal. Hofacker [1] illustrated this problem with the help of randomly generated data. Any result from post hoc analysis should be considered indicative and not conclusive. Further study should be planned to confirm such results. This also applies to post hoc analysis of various subgroups that were not part of the original plan. There is always a tendency to try to find the age–sex or severity groups that benefited more from the treatment under review than the others. Numerous such analyses are sometimes done using a drop-down menu in the hope of finding some statistical significance somewhere. Again, such detailed analysis is fine for searching a ground to plan a further study but not for drawing a definitive conclusion. However, this is not to deny the existence of Americas because it was not in Columbus' plan. This is not empirical anyway. In most empirical cases though, it may be sufficient to acknowledge that the results are based on post hoc analysis and possibly need to be reconfirmed.

Another dimension of overanalysis is what has now come to be known as ***P*-hacking** as proposed by Simmons et al. [2]. This is chasing a small effect in the hope that it is hidden in the data. The data are reanalyzed in several different ways to find statistical significance somewhere. Head et al. [3] have also discussed this phenomenon, which they consider is widespread throughout science, but opine that this has not caused much damage to the science so far.


1. Hofacker CF. Abuse of statistical packages: The case of the general linear model. *Am J Physiol* 1983;245:R299–302. http://ajpregu.physiology.org/content/245/3/R299.full-text.pdf+html
2. Simmons JP, Nelson LD, Simonshohn U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 2011 Nov;22(11):1359–66. http://pss.sagepub.com/content/22/11/1359.long
3. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of P-hacking in science. *PLoS Biol* 2015;13(3):e1002106. http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002106




## overfit (of regression)

A regression model is said to be overfit when even the random fluctuations in the data are considered as trend. As an extreme example, if you have six data points and fit a regression with six parameters (regression coefficients), the regression will exactly go through all the points, ignoring that these data points have a random error component beside the trend. This can look like the curve in Figure O.7 between shock index and risk score for thrombolysis (thrombolysis in myocardial infarction [TIMI] score) in ST-segment elevated myocardial infarction cases. In this figure, the data are available for only five subjects, and the polynomial of degree 4 (this makes a total of five parameters when the intercept is also counted) is a perfect fit, passing through all the points. A linear fit is also shown for illustration.

As mentioned, this is an extreme example to illustrate the point, but an overfit can occur when the number of data points is not as many as needed for an adequate regression. If you have 10 data points, even a fourth-degree polynomial can be an overfit. It depends on what kind of trend you are expecting considering the biological process under study.

Note also that the regression analysis actually presumes that you have a series of values of $y$ for each value of $x$. In our example, regression would be far more satisfactory when you have a TIMI risk score for, say, 4 patients with shock index 0.5, 3 patients with shock index 0.6, 7 patients with shock index 0.7, and so on. Such

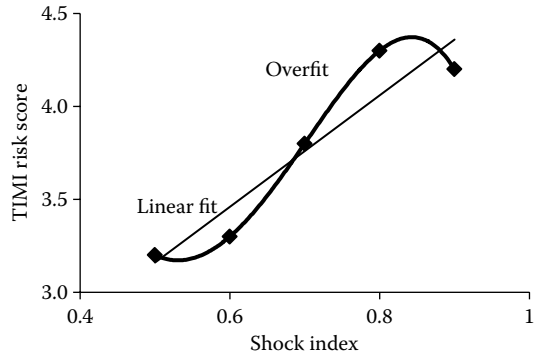



**FIGURE O.7** An overfit (thick curve) of regression passing through all the points and the linear fit (thin line).

multiple subjects for each value of $x$ help consider the average—thus ruling out, to some extent, the effect of random fluctuations. Remember that regression is for the *average* of $y$-values for each value of $x$. Many researchers tend to forget this aspect. The chance of an overfit is much more when you have one subject for each value of $x$. This is not a strict requirement for the set of $(x_1, x_2, \ldots, x_K)$ values in case of **multiple regression**. It could be an uphill task to include many subjects for each specific set of values of the $x$'s when $K$ is large. However, caution is also required in this case, particularly if you are fitting a complex regression. If you get an extremely high **coefficient of determination** (or $R^2$ in case of multiple linear regression), say, exceeding 0.90, suspect an overfit. Examine it carefully in the context of the sample size, the complexity of the model, and biological plausibility. If this coefficient is so high for a model that is not an overfit, you are in a win–win situation.

The best insulation against overfitting is a large sample. Another is to keep the model simple by judicious choice of the predictors. You should be able to justify why each of the regressors you have chosen needs to be in the regression. Also, it has now become a standard practice to **validate** the regression model. This can be external by using it on some other data set or internal by splitting the sample into training and validation data sets, or by **resampling methods**.

If you want to see overfit in action, Ryali et al. [1] have discussed the problem in the analysis of functional magnetic resonance imaging (fMRI) data because the number of regions considered is large compared to the number of participants.

1. Ryali S, Supekar K, Abrams DA, Menon V. Sparse logistic regression for whole-brain classification of fMRI data. *Neuroimage* 2010 Jun;51(2): 752–64. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2856747/

## overmatching, *see also* matching

**Matching** of subjects in the two groups under comparison is considered a good strategy in situations where randomization is not feasible or the sample size is small such as in some case–control studies. It helps rule out the effect of matched factors on the results. Proper matching enables one to validly estimate the effect without bias caused by the presence of extraneous factors. Although matching for all the factors that can affect the outcome except the ones under study is advocated, this operational clause is sometimes ignored. The real question is what to do with factors that affect the outcome as well as the risk factors under study—called **confounders**. Matching for these can sometimes result in overmatching, and biased results are obtained. Overmatching can also reduce statistical efficiency as well as cost efficiency. For example, in a study on the effect of postmenopausal estrogens on uterine cancer, matching on uterine bleeding can present invalid findings since uterine bleeding is also associated with uterine cancer. Any matching on such outcome-related variable can distort the results. The following example illustrates this further.

Marsh et al. [1] describe a study of relation between radiation exposure and mortality from leukemia in workers at a nuclear reprocessing plant. The matching factors in this study were site, sex, work status, age (within 2 years), and date of entry (within 2 years). Note that risk of leukemia varies with these factors and they looked like valid factors for matching. Date of entry was considered necessary as the risk of leukemia changes with calendar time. Examination of data since 1950 shows that the radiation dose steeply declined after 1980. Matching for date of entry unwittingly also matched for radiation exposure, which was the antecedent under study. Such overmatching obscured the relationship between radiation dose and risk of leukemia. Overmatching reduced the statistical significance of the effect.

Likewise, exposure-related variables that can make risk factor distribution similar in cases and controls should not be matched. Remember that the effect of matched factors cannot be studied as these have the same distribution in the two groups after matching. In most situations, matching on age and sex and possibly one or two more antecedents not under investigation would suffice. In any case, finding subjects that are matched with respect to several factors is difficult and can enormously increase the cost as many unmatched subjects have to be discarded. This is what we meant when we mentioned cost efficiency of matching.

1. Marsh JL, Hutton JL, Binks K. Removal of radiation dose response effects: An example of over-matching. *BMJ* 2002;325:327–30, http://www.bmj.com/content/325/7359/327

# P

## pack-years of smoking

Tobacco use in different forms is known to be a major killer in many countries and is gradually taking center stage in the rest of the world. Pack-years is a way of measuring the extent of a person's cigarette smoking. This is measured as

$$\text{pack-years} = m_1y_1 + m_2y_2 + \ldots + m_Ty_T,$$

where $m_t$ packs of cigarettes are smoked for $y_t$ years ($t = 1, 2, \ldots, T$). If 2 packs a day are smoked for 4 years and 3 packs for 10 years, the pack-years are $2 \times 4 + 3 \times 10 = 38$. As per this definition, one pack-year is smoking one pack a day for 1 year that actually means 365 packs a year.

An index such as pack-years suffers from at least one serious deficiency. According to this index, smoking 3 packs a day for 10 years is the same as smoking 1 pack a day for 30 years, although they may have very different implications. For lung cancer, duration may be more important, whereas for coronary artery disease, the intensity of smoking may be more important. Despite this deficiency, pack-years continues to be the most commonly used measure of extent of smoking.

Doctors studying the effect of smoking on patients use the concept of pack-years to determine who should be screened for lung and other cancers. Studies suggest that people who have a 30 pack-year history of smoking might be candidates for lung cancer screening. Factoring in additional criteria (age is old or young, continue to smoke or have quit in the past 15 years), studies have found that the mortality rate from lung cancer could be cut by 20% if people meeting these criteria underwent screening.

In general, the more pack-years one has smoked, the greater the chance of getting cancer. It is a dose–response and cause–effect type of relationship established after a large number of observational studies. That said, lung cancer can occur in people who have never smoked, and many people smoke heavily for decades without getting cancer. Statistically, according to Peto [1], pack-years is unhelpful as a correlate of lung cancer.

More precision can be obtained by considering the number of cigarettes rather than packs per day. To account for duration, a measure could be the total number of cigarettes smoked so far in life. This number is given by

$$S_1 = n_1x_1 + n_2x_2 + \cdots + n_Kx_K,$$

where $n_k$ ($k = 1, 2, \ldots, K$) cigarettes per day (intensity) are smoked for $x_k$ years (duration). This is more exact than the pack-years generally used for smoking. This is the beginning point of the **Indrayan smoking index**, which considers age at start of smoking, type of smoking, years since cessation by former smokers, etc., in addition to cigarette-years.

1. Peto J. That the effects of smoking should be measured in pack-years: Misconceptions 4. *Br J Cancer* 2012;107:406–7. http://www.nature.com/bjc/journal/v107/n3/full/bjc201297a.html

## paired samples, *see also* matched pairs

Paired samples in medical research arise primarily in three situations: (i) naturally matched such as in **before–after study** with no parallel controls: the same subjects are measured twice—first before the stimulus, and then after the intervention; (ii) in crossover trials where each subject receives two or more treatments on different occasions after a washout period; and (iii) in a deliberate and strict one-to-one matching where a parallel control group of different subjects is present, but each control subject is matched to a corresponding subject for nearly all the characteristics that can affect the outcome, except of course the regimen itself. Consideration of **matching** is important for statistical analysis since matched-pair analysis is different from independent groups analysis. Matching by one or two characteristics such as age and sex generally is not considered adequate for paired analysis—instead such partially matched control group is considered independent.

An advantage of using the paired samples approach is that the sample size can be smaller first because there is no separate control group and second since within-subjects (i.e., within pairs) variability is likely to be smaller than between-subjects variability. Also, since the sampled subjects are the same for intervention, there is less chance that some external factor (**confounding variable**) will influence the result. However, there is also a downside of paired samples. In the case of a before–after study, the **Hawthorne effect** and the placebo effect are confounded with the regimen effect—you would not know how much of the difference was because of the regimen and how much due to the hidden effect of these two factors. In the case of one-to-one matching, even when all known factors are matched, there might be unknown factors that can cause the difference, and this can give false assurance. Also, you would not be able to find the effect of matching characteristics—if the subjects are matched for age, you would not know the effect of age on the outcome. Moreover, there is always a risk of **overmatching** that could provide distorted results.

For statistical analysis, when the data are quantitative, the difference between each pair of values is obtained and the situation becomes of a one-sample scenario. Confidence intervals can be calculated, and hypothesis can be tested for this difference just as for one sample. Proportions in the case of paired samples are not straightforward: they may require the **McNemar** test.

## paired *t*-test, *see* Student *t*-tests

## pairwise deletion, *see* casewise, pairwise, and listwise deletion

## Palma measure of inequality, *see also* Gini coefficient, health inequality

This measures how unequal the distribution of a characteristic is across the population. Although primarily devised for assessing

income inequalities, it can also be used to measure inequalities in health parameters. The Palma measure tells us whether health is nearly the same for everyone or whether a few have excellent health and many have desperately poor health.


Jose Palma

The usual measure of health inequality is the **Gini coefficient**. However, it is nonspecific, does not reveal where the inequality exists, and has no physical interpretation. It is also oversensitive to the changes in the middle values and less sensitive to the changes at the top and bottom [1]. Jose Palma in 2011 [2] suggested his measure to remedy these deficiencies of the Gini coefficient. The Palma measure compares the highest 10% values with the lowest 40% in place of each with every other done by Gini. Among several conclusions, Palma observed that people in the 5th to 9th **deciles** of income acquire their rightful 50%, and this remains stable across populations; the other 50% income is mostly grabbed by the top decile leaving little for the bottom 4 deciles. Sharing between the top decile and the bottom 4 deciles varies widely from population to population, and this mostly defines the income inequalities. Thus,

$$\text{Palma measure of inequality: } P = \frac{\Sigma_{i=1}^{n/5} x_{[i]}}{\Sigma_{i=3n/5}^{n} x_{[i]}},$$

where $x[i]$ is the $i$th value after ordering from the lowest to the highest.

The Gini coefficient is mostly affected by the middle values, whereas the Palma measure ignores these middle values. A fair criticism of the Palma measure is that this is not based on all the values. Those interested in middle values as much as the tail values will continue to use the Gini coefficient.

In a medical context, the Palma measure of inequality is the ratio of aggregation of health (in whatever metric) of the top 10% and the aggregation of bottom 40% of values. No adequate metric is available for measuring comprehensive health, but consider for illustration purposes that we are able to measure it on a 0–100 scale for 200 persons. The top decile is the top 20 people, and let their health metric have an average of 97. Thus, the aggregation is 97 × 20 = 1940. If the health metric of the bottom 80 persons has an average of 14, their aggregate is 14 × 80 = 1120. This gives a Palma measure of 1940/1120 = 1.73. Palma stops here, saying that the health of the top 10% is 1.73 times the health of the bottom 40%. However, the interpretation is clearer if you realize that this would be ¼ = 0.25 if everybody has the same health. Thus, the inequality actually is 1.73/0.25 = 6.92. This is the same as the ratio of their averages (97/14 = 6.92). Now you can say that the health inequality between the top 10% persons and the bottom 40% is nearly 7 times. This value alone may not make much sense but would be extremely useful if you find varying values in different segments of population such as 5.81 for males and 6.56 for females, which reveals that females are more unequal with respect to health compared with males in this population.

1. Cobham A, Summer A. Is inequality all about the tails? The Palma measure of inequality. *Significance* Feb 2014;10–3. http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2014.00718.x/pdf
2. Palma JG. Homogeneous middles vs. heterogeneous tails, and the end of the 'Inverted-U': It's all about the share of the rich. *Dev Change* 2011;42(1):87–153. http://onlinelibrary.wiley.com/doi/10.1111/j.1467-7660.2011.01694.x/abstract

## parabolic curve, *see* curvilinear regression

## parallel controls, *see* controls

## parallel-line assays

These are situations in **bioassays** where the ratio of the response of the test preparation to the standard preparation is the same at each dose. For this purpose, response and/or dose may be in logarithmic term or any such transformation, called *metameters*, so that the **dose–response relationship** is linear. Parallel-line assays are most commonly found in the form of indirect analytical assays where the potency of the substance under test is observed in relation to the potency of a standard. Both test and standard preparations are used at several dose levels. If the number of dose levels of test preparation is $K_1$ and that of standard preparation is $K_2$, this is called $(K_1 + K_2)$-point assay and is symmetric if $K_1 = K_2$. The number of subjects for each dose can be any and can vary, but equal numbers help in easy interpretation. The underlying requirement is that the given doses of the test preparation will yield expected responses such as would be obtained from administering the standard preparations at doses that are ρ times the given test preparation's dose: ρ is known as the **relative potency**. When the expected response is assumed to be linearly related to the dose (or its transform), the graphs of expected response (or its transform) against dose metameter for the test and standard preparations will give parallel lines. In Figure P.1a, the relationship between response and dose is curvilinear but becomes linear (Figure P.1b) when dose is transformed to ln(dose).

With the help of the parallel-line model, the statistical validity of the following hypotheses can be tested:

- The dose–response (with or without transformations) relationship is linear for the standard and the test preparation.
- The dose–response line has a significant slope and not a flat line.
- The dose–response lines of the standard and test preparations are parallel. This means that the slopes of the dose–response relationship of both the preparations are the same.

For testing these hypotheses, the analysis of variance method is used to break between-doses **sum of squares** into components due to preparations, due to regression (slope), due to parallelism, and due to linearity. Each of these is compared with the within-doses sum of squares after dividing by the respective df's, and $F$-test is used for finding the statistical significance. For a proper parallel-line assay, the hypotheses of linearity and of parallelism must hold (not rejected), and the hypotheses of no slope and no difference in preparations must be rejected. When this happens, the relative potency

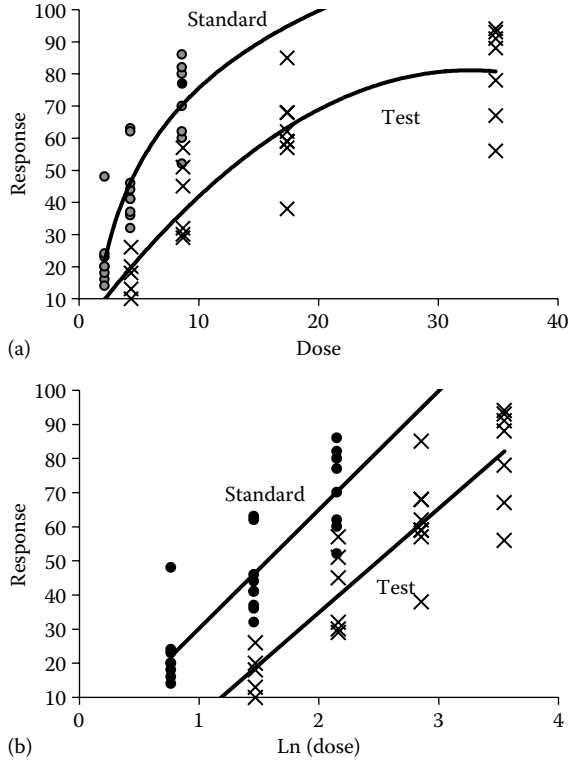



**FIGURE P.1** Graphical depiction of dose–response relationship: (a) curvilinear relationship becomes (b) parallel lines after logarithm of dose.

can be estimated as $\log\hat{\rho} = \left(\bar{x}_s - \bar{x}_t\right) - \left[\frac{\bar{y}_s - \bar{y}_t}{b}\right]$, where $\bar{x}_s$ and $\bar{x}_t$ are the average dose metameters of the standard and test preparations, respectively, $\bar{y}_s$ and $\bar{y}_t$ are the average response metameters in the standard and test groups, respectively, and $b$ is the common estimate of the slope of the dose–response relationship. The confidence interval for the **relative potency** ρ, which is a ratio, can be obtained by the **Fieller theorem**.

Deficiencies have been pointed out in the method we just outlined for detecting parallelism. Gottschalk and Dunn [1] have proposed a method based on chi-square, which is claimed to be more reliable and more appropriate. Parallelism required for the method just outlined may not hold in many situations because of effects such as partial agonism, differences in binding affinities, in vitro toxicokinetics, solubility, and sensitivity to environmental conditions as they can produce nonparallel dose–response curves [2]. To overcome such problems, Villeneuve et al. [2] proposed estimation of relative potency by multiple-point estimates over a range of responses in place of the single-point estimates.

Vieira et al. [3] report a (3 + 3) parallel-line assay for determining the potency of cefuroxime sodium in powder for dissolution for injection. See Collazo et al. [4] for an example of determination of potency of factor VIII in commercial concentrates.

1. Gottschalk PG, Dunn JR. Measuring parallelism, linearity, and relative potency in bioassay and immunoassay data. *J Biopharm Stat* 2005;15(3):437–63. http://www.ncbi.nlm.nih.gov/pubmed/15920890
2. Villeneuve DL, Blackenship AL, Giesy JP. Derivation and application of relative potency estimates based on in vitro bioassay results. *Env Toxicol Chem* 2000;19(11):2835–43. http://www.usask.ca/toxicology/jgiesy/pdf/publications/JA-253.pdf
3. Vieira DC, Fiuza TF, Salgado HR. Development and validation of a rapid turbidimetric assay to determine the potency of cefuroxime sodium in powder for dissolution for injection. *Pathogens* 2014 Jul 30;3(3):656–66. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4243433/
4. Collazo V, Alonso C, Frutos G. Validation of an automated chromogenic assay of potency of factor VIII in commercial concentrates. *Int J Lab Hematol* 2013;35(1):38–45. http://onlinelibrary.wiley.com/doi/10.1111/j.1751-553X.2012.01459.x/abstract

## parametric models

A parametric model in statistics is the one based on a family of statistical **distributions** that can be described using a small number of **parameters**. For example, an ordinary **regression** model requires that the errors follow a Gaussian (Normal) distribution with mean zero and some variance $\sigma^2$. The mean and variance are the parameters that describe the Gaussian distribution—thus, the errors in ordinary regression follow a parametric model. Gaussian distribution not only requires parameter values but is also bound by a specific shape, generally recognized as a bell shape.

Contrast parametric models with nonparametric models, which are based on distributions that are flexible and not based on specific parameters. They may or may not have a general requirement such as symmetric or asymmetric pattern, monotonic or cyclic, and smooth or curvy. For example, a spline regression is nonparametric as there are no fixed parameters, and no specific shape of the distribution of values is required. A **logistic regression** model is semiparametric as it requires parameters but does not require any specific shape of the distribution. Nonparametric models tend to be computationally intensive and not as powerful in detecting a trend as parametric models could when their conditions are satisfied.

No **model**, whether parametric or nonparametric, is perfect. It is important to realize though that parametric models (i) are critically dependent on the "size" parameter, which can be very difficult to determine accurately in the early stages of a project; and (ii) can be highly sensitive in that small changes to certain key parameters can result in substantial changes in output estimates (see **butterfly effect**). Regardless of which model is used in a research, understand that all models are based on conventional wisdom, but parametric models are especially susceptible to the input parameters that can substantially change the resulting estimates; this is affected by the range of uncertainty encompassed by model parameters, called the **parameter uncertainty**.

Parametric models work well when the parametric assumptions are adequately satisfied; otherwise, they may be misleading, and these assumptions could be fairly strong. Thus, use parametric models only after thorough checking of the distributions and after being satisfied with the validity and reliability of the estimates of the parameters proposed to be used. For example, if a Weibull distribution is proposed to be used for the survival pattern of a set of patients, make sure that the Weibull rightly depicts the pattern and the parameter estimates you will obtain will adequately fit to the data you have. When this framework is deficient for the data in hand, parametric models have the likelihood to give fallacious results. On the other hand, if this framework is right for the data, parametric models may give much more valid results than the corresponding nonparametric models. The greatest rescuer of parametric models is the **central limit theorem** that says that the distribution of a linear combination of a large number of values tends to become Gaussian (parametric) as $n$ increases.

## parameters

In medical terminology, a parameter sometimes refers to a measurement or characteristic of a person or patient such as his/her glucose level. For example, some disease entities are based almost exclusively on a single *parameter*: diagnosis of anemia caused by iron deficiency is based on hemoglobin level; hypertension on blood pressure levels; diabetes mellitus on serum glucose level; and glaucoma on intraocular pressure. Other indications such as signs–symptoms play almost no role in the diagnosis of such diseases. Evidence exists that persons with statistically abnormal levels of these *parameters* do have an increased risk of the concerned morbidity and the associated mortality. An intervention, such as therapy, to bring the level back to the reference range helps to reduce this risk. Thus, medical parameters have extensive role in health assessments.

In statistical parlance, a parameter can be defined as a constant in a **model**, or a constant that wholly or partially characterizes a function or a probability **distribution**. The behavior of a model is governed by its structure or functional form and unknown quantities or constants of nature called *parameters*. A statistical goal of running the model might be to estimate the unknown parameters so that the model can be fully specified. If the model has been constructed in such a way that the parameters correspond to clinically interpretable effects, the model can then be a way of estimating the influence of such factors on the outcome. In practice, statistical models usually provide a concise way of estimating parameters, obtaining confidence intervals (CI) on parameter values, and testing hypothesis on them.

The parameters in a Gaussian (normal) distribution are the mean and the standard deviation, denoted respectively by $\mu$ and $\sigma$. These are the only two parameters that completely specify a Gaussian distribution. Without the values of these parameters, you will not be able to use this distribution for any worthwhile inference. Many statistical methods are devoted to finding the best estimates of such parameters because they are mostly not known. The binomial distribution has just one parameter, namely, the probability of success $\pi$, although it needs the number of trials $n$ also for full specification. The Poisson distribution also has just one parameter, namely, the mean $\mu$. On the other side, ordinary multiple linear regression equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_K x_K + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$, has $(K + 2)$ parameters of which $(K + 1)$ are the regression coefficients $\beta$'s (including $\beta_0$) and one $\sigma$. The last ($\sigma$) is the **nuisance parameter**—of no intrinsic interest for any research, although this is needed to assess the goodness of the model. Statistical methods are used to estimate these parameters, to find the CI on them, and to test the hypothesis about them. The same can be stated for logistic regression. Depending on the situation, the parameter of interest may be the correlation coefficient, odds ratio, relative or attributable risk, etc. Sample values are used to make an inference on these parameters.

## parameter uncertainty, *see* sensitivity analysis, parametric models

## parsimonious models

A parsimonious model can be said to be the simplest model that adequately describes the observed data. A model invariably is a simple representation of a complex process so that some information is necessarily and knowingly ignored. It is thus right to suspect all the models and investigate their parsimony. The concept of parsimonious models derives from the parsimony principle: among competing models, all of which provide an adequate fit for a given set of data, the one with the fewest parameters (the simplest) is preferred. In statistical models, parameters are usually unknown and have to be estimated—thus, it is preferable to work with as few unknowns as possible without losing the vital information. This approach is pragmatic and the objective is simplicity.

Also recognized as *Occam's razor*: "entia non sunt multiplicanda praeter necessitate," i.e., "One should not increase, beyond what is necessary, the number of entities required to explain anything." Occam's razor is a logical principle attributed to the fourteenth century philosopher William of Occam [1]. The principle states that one should not make more requirements than the minimum needed, and underlies all scientific modeling and theory building.

In any given statistical model, we try to rid ourselves of those concepts, variables, or constructs that are not much helpful in developing the model of interest (see the topic **variable selection**). For example, one can always draw a straight line through two data points that can tempt us to conclude that all further observations would lie on that line. However, one could also draw an infinite variety of the most complicated curves passing through the same two points, and these curves would also fit the empirical data just as well if there are only two observations. Occam's razor would, in this case, guide us in choosing the *straight* (i.e., linear) relation as the best candidate model. A similar reasoning can be made for any kind of model. But such simplification can be far from reality, and thus, the concept of parsimony requires that not losing vital information is also an important consideration.

More specifically, in regression, when the number of possible explanatory variables is large, a computer program can be used to identify and include only the statistically significant variables in the regression model. This is one of those situations where the testing of the hypothesis strategy as opposed to confidence intervals is preferred. Many algorithms can do this, but the following are commonly used: **stepwise** that includes forward selection and backward elimination, and the **best subset**. These methods can be used to get a feeling of what predictors may be empirically important, but a more realistic solution is based on their real biological worth, if extraneously available, instead of purely numerical optimization procedures such as these. You must ensure that clinically important variables are not excluded in the final model. None of these algorithms may yield the *best* model, and different algorithms may lead to different models, yet it is a good idea to examine several possible models. The final choice is based on interpretability, parsimony, and convenience in obtaining the data.

1. Thorburn WM. Occam's razor. *Mind* 1915;24:287–8. http://mind.oxfordjournals.org/content/XXIV/2/287.extract

## partial correlation

In a multivariate setup, partial correlation is the measure of association between two quantitative variables, while controlling or adjusting for the effect of one or more of the other variables. When adjusted for just one variable, the partial correlation of $x$ and $y$ adjusting for $z$ can be calculated as follows:

$$r_{xyz} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{\left(1 - r_{xz}^2\right)\left(1 - r_{yz}^2\right)}},$$

where $r_{xy}$, $r_{xz}$, and $r_{yz}$ are the usual (Pearsonian) correlation coefficients between ($x$, $y$), ($x$, $z$), and ($y$, $z$), respectively; now better understood as *total correlations*, some call them simple correlations. For adjusting more than one variable, the formula is more complex. The variables used for adjustment are generally called covariates as these are not of primary interest. A partial correlation is also described as the correlation "after controlling for 'such and such' covariates." The controlled variables are stated after the dot in the subscript of the notation; for example, $r_{pq.rst}$ is the notation for the correlation between $x_p$ and $x_q$ after controlling $x_r$, $x_s$, and $x_t$. As for all Pearsonian correlations, partial correlation is valid for measuring the strength of only the linear relationship. Adjustment also is restricted to the linear effect. A partial correlation is generally contrasted with **multiple correlation**, which is the correlation between $y$ and its predicted value based on multiple linear regression on several regressors.

As with any correlation, there are hazards in interpreting partial correlations, especially when searching for causation. If changes in $x$ are associated with changes in a dependent variable, $y$, it is not necessarily clear why the changes in $y$ occur. Changes in $x$ may be a cause of changes in $y$, changes in $y$ may be a cause of changes in $x$, a third variable $z$ may be producing changes in both $x$ and $y$, or any combination of these possibilities may be true. One reason for this difficulty is the likely presence of **mediator** or **confounding** variables. A variable is confounding if it is related to both $x$ and $y$, making causal interpretations difficult. Mediating variables tend to enhance or reduce the correlation.

Consider the relation between body mass index (BMI) and ambulation time after a bariatric surgery. This may be affected by age, sex, duration of surgery, type of surgery, comorbidities, etc. Thus, the *net* correlation between BMI and ambulation time is obtained by removing the effect of these covariates. This is done by partial correlation. The total correlation (without any consideration of the covariates) may be 0.56 but could become 0.71 when the effect of the covariates is removed. The partial correlation can be more or can be less than the total correlation. However, as already cautioned, this adjustment is only for the linear effect. Let us also sound a cautionary note on the term *net* sometimes used for partial correlation: it is net only with respect to the covariates adjusted. Even when adjusted for all the covariates under study, it would still not be net in true sense since the adjustment can be made only for known covariates. Unknown covariates in epistemic domain continue to haunt the actual relationship.

Since the partial correlation $r_{xy.z}$ is a measure of the relationship between $x$ and $y$ keeping $z$ constant, if $r_{xy.z}$ is much smaller relative to $r_{xy}$, we can conclude that $z$ is a mediating variable. That is, $z$ may explain, at least in part, the observed relationship between $x$ and $y$. Although we talk about "explaining" the relationship based on correlations, we will not know what "causes" the relationship. Causal inference has much stringent requirements as discussed (see **cause–effect relationship**).

Partial correlation is encountered while using the statistical method of **factor analysis**. This analysis provides satisfactory results when multiple correlations of most variables are high, but partial correlations for most pairs of variables are low. This can be tested by using the **Kaiser–Meyer–Olkin measure**.

## partial least squares

The usual **least squares method** for estimating the coefficients in a **regression model** requires that none of the regressors has high correlation with any of the other regressors. When high correlations are present, called **multicollinearity**, this compromises the estimates of the regression coefficients in the sense that they tend to become less reliable. That is, the standard error of at least some of them would be unusually high. Partial least squares (PLS) method circumvents this problem and helps to obtain reliable estimates when multicollinearity exists. This also helps in a situation where the number of regressors is large relative to the sample size and the problem of overfitting occurs. The method has been seen to do well in **predictive models** but not necessarily in **explanatory models**. According to Tobias [1], the PLS method was developed by Herman Wold in the 1960s.

Consider an example of trying to predict the level of thyroid stimulating hormone (TSH) on the basis of signs and symptoms that are elicited anyway while examining a patient suspected to have hyperthyroidism or hypothyroidism. These signs and symptoms run into hundreds ranging from skin condition to trembling, brittle nails, and what not, and are the regressors. These also are highly collinear with each other. This exercise may be useful in a setup where laboratory facility for TSH is not available or is beyond reach. If there are only 80 cases, the usual multiple regression will not work since the number of regressors exceed 100 and they are collinear; the method of PLS may be helpful in this situation.

Similar to what is done in the principal components method of **factor analysis**, the PLS method tries to extract latent factors that account for the highest variation among the responses. These latent factors would be few but need to account for a large part of the variation for PLS to be successful. The extracted factors are used to predict factor scores, and these are, in turn, used to predict the response of our interest. The difference between the PLS and the principal components is that PLS uses the covariance structure between the response and the regressors, whereas the principal components method uses the covariance between the regressors alone. The PLS method is applied only after the regressors are standardized by subtracting the mean and dividing by the standard deviation (SD) that makes all the regressors on the same scale. Besides multiple linear regression, the method can also be used for **discriminant analysis** if similar constraints exist. An introduction to PLS is given by Tobias [1], and Naes and Martens [2] have provided further details.

Jiang et al. [3] used the PLS method for **age–period–cohort** analysis of body mass index in Ireland. You might be aware that there is redundancy in this analysis as two of age, period, and cohort determine the third. Thus, this is an appropriate setup for using the PLS method, although the number of regressors is not large in this case. Sousa et al. [4] used PLS-based discriminant analysis to demonstrate unsuitability of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry for bacterial typing for *Acinetobacter baumannii* clonal discrimination.


1. Tobias RD. An introduction to partial least squares regression. SUGI Proceedings, 1995. http://www.ats.ucla.edu/stat/sas/library/pls.pdf
2. Naes T, Martens H. Comparison of prediction methods for multicollinear data. *Commun Stat Simul Comput* 1985;14(3):545–76. http://www.tandfonline.com/doi/abs/10.1080/03610918508812458?journalCode=lssp20
3. Jiang T, Gilthorpe MS, Shiely F, Harrington JM, Perry IJ, Kelleher CC, Tu Y-K. Age–period–cohort analysis for trends in body mass index in Ireland. *BMC Public Health* 2013;13:889. http://www.biomedcentral.com/1471-2458/13/889
4. Sousa C, Botelho J, Grosso F, Silva L, Lopes J, Peixe L. Unsuitability of MALDI-TOF MS to discriminate *Acinetobacter baumannii* clones under routine experimental conditions. *Front Microbio* 2015 May 19;6:481. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4436932/

## partial likelihood

Cox [1] introduced the concept of partial likelihood in 1975 to obtain estimates of parameters of interest in survival analysis where an appropriate parametric form of the **hazard function** is often unknown and, in any case, is not of interest. The method of partial likelihood divides the **likelihood** into two parts—conditional and marginal. Only that part of the likelihood that contains the parameters of interest is maximized, and the part that contains the nuisance parameters is ignored. This does mean loss of information but becomes simple as the nuisance parameters are excluded. This also makes the method robust as the part of the likelihood does not have to be specified. Because of loss of information, the method does not provide as good results as the usual likelihood does, yet serves well in some specific situations.

A likelihood function usually depends on many parameters. Let the parameters of interest be denoted by θ and the parameters not of interest (nuisance parameter) by δ. For example, in a linear regression, interest typically lies in the regression coefficients and not in the error variance. The standard way to approach the estimation is to look for those values of both δ and θ that maximize the likelihood function. Since the primary interest lies in θ, the method of partial likelihood can be used to estimate θ without worrying about δ. A more technical explanation can be found in Armitage and Berry [2].

In survival analysis, the time to an event or failure (such as death) is recorded for each subject. If these are ordered as $t_1 \le t_2 \le \ldots \le t_n$ for $n$ subjects, then the partial likelihood function is derived by taking the product of the conditional probability of a failure at time $t_i$, given the number of cases that are at risk of failing at this time. In Cox regression, for example, where partial likelihood has the most pronounced application, the interest may be in this order, and the actual distribution of the survival times or of hazard can be ignored under some conditions. Since the hazard ratio is used in this regression, specification of the actual baseline hazard also is not needed.


1 Cox DR. Partial likelihood. *Biometrika* 1975;62(2):269–76. http://biomet.oxfordjournals.org/content/62/2/269.short

2. Armitage P, Berry G. *Statistical Methods in Medical Research*, Third Edition. Blackwell Science, 1994: pp. 484–5.


## partitioning of chi-square and of table,
*see also* **chi-square—overall**

After a general conclusion is drawn regarding the presence of association in a contingency table by using the usual chi-square test, partitioning is used to obtain a focused conclusion regarding which particular category (or a set of categories) is causing the association. It is also possible that the overall chi-square test indicates no association, but partitioning reveals some association somewhere. We use the data in Table P.1 for illustrating partitioning of chi-square that shows the observed number of AIDS subjects with different blood groups (O, A, B, AB) from a population where these groups are in the ratio 6:5:8:1. These ratios provide $H_0$: $\pi_1 = 6/20 = 0.30$, $\pi_2 = 5/20 = 0.25$. $\pi_3 = 8/20 = 0.40$, and $\pi_4 = 1/20 = 0.05$. Expected frequencies are based on these probabilities. The objective is to find whether or not the blood group pattern in AIDS cases is the same as in the population.

In this example, df = $K - 1 = 4 - 1 = 3$. Calculations are also presented in Table P.1, which show $\chi^2 = 4.91$. A relevant statistical software package automatically compares the calculated value of $\chi^2$ with its known distribution for 3 df and gives $P = 0.178$. Thus, the value $\chi^2 = 4.91$ obtained for these data is not all that unlikely when $H_0$ is true. That is, the frequencies observed in different blood groups in the example are not very inconsistent with $H_0$. The sample values do not provide sufficient evidence against $H_0$ and so cannot be rejected. A preponderance of any blood group in cases of AIDS cannot be concluded on the basis of this sample when this method is used.

**TABLE P.1**
**Calculation of Chi-Square**

| | Blood Group | | | | |
|---|---|---|---|---|---|
| | **O** | **A** | **B** | **AB** | **Total** |
| Observed frequency ($O_k$) | 57 | 36 | 51 | 6 | **150** |
| Expected frequency under $H_0$ ($E_k$) | 45.0 | 37.5 | 60.0 | 7.5 | **150.0** |
| $O_k - E_k$ | 12.0 | –1.5 | –9.0 | –1.5 | **0** |
| $(O_k - E_k)^2/E_k$ | 3.20 | 0.06 | 1.35 | 0.30 | **4.91 = $\chi^2$** |

Examination of the data in this example reveals that the observed frequency in blood group O is much higher than expected from the pattern in the general population (57 versus 45), and the other differences are not as large. To find out that this really is so, check whether the pattern in blood groups A, B, and AB (combined) is nearly the same as expected, and then check the difference in blood group O. The corresponding null hypotheses are

$H_{01}$: A, B, and AB are in the ratio 5:8:1, and
$H_{02}$: $\pi_1 = 6/20 = 0.3$ for blood group O, $\pi_2 + \pi_3 + \pi_4 = 14/20 = 0.7$ for the other three groups combined.

The former ratio is the same as in our example, and the latter ratio combines A, B, and AB. Not stating $H_{01}$ in terms of π is deliberate because that might give rise to confusion. The sum of π's in all contingency tables should be 1, and therefore the same π's cannot be used for the three cells covered by $H_{01}$.

For these two null hypotheses, calculations for $\chi^2$ are shown in Table P.2. The division of the earlier four-cell table into two tables as shown is called **partitioning**. The first partition gives $\chi^2_{\text{I}} = 0.38$ with 3 – 1 = 2 df, and from a software package, $P = 0.83$ for this value of $\chi^2$. Since it is more than 0.05, $H_{01}$ cannot be rejected at 5%

**TABLE P.2**
**Partitioning of Table P.1 and Calculation of Partitioned Chi-Square**

| | Blood Group | | | |
|---|---|---|---|---|
| **I.** | **A** | **B** | **AB** | **Total** |
| $O_k$ | 36 | 51 | 6 | **93** |
| $E_k$ | 33.2 | 53.1 | 6.6 | **93** |
| $(O_k - E_k)^2/E_k$ | 0.23 | 0.09 | 0.06 | **0:38 = $\chi^2_{\text{I}}$** |
| | **Blood Group** | | | |
| **II.** | **O** | | **Others** | **Total** |
| $O_k$ | 57 | | 93 | **150** |
| $E_k$ | 45.0 | | 105.0 | **150** |
| $(O_k - E_k)^2/E_k$ | 3.20 | | 1.37 | **4:57 = $\chi^2_{\text{II}}$** |

level of significance, i.e., the evidence is not sufficient to conclude that the pattern of blood groups A, B, and AB in AIDS cases is not the same as in the general population.

Part II of the table has only two cells, so $\chi^2_{II}$ has only one df. This gives $\chi^2 = 4.57$ and $P = 0.033$. Being less than 0.05, this is statistically significant, and it can be concluded without much chance of error that the pattern in part II is not the same as in the general population. Since the categories now are blood group O versus the others, it can be safely concluded that blood group O is *more* common (note that the observed frequency 57 is more than the expected 45) in AIDS cases. Nothing specific can be said about the other three blood groups.

The conclusion reached after partitioning is different from the one reached earlier when all the cells were considered together. This is because the lack of difference in A, B, and AB groups masked the difference in the O group also. Partitioning helped to uncover this difference.

Note the following: When $n$ is large, the values of $\chi^2_I$ and $\chi^2_{II}$ based on the partitioned table should add *approximately* to the overall $\chi^2$ based on all the cells. In this example, $\chi^2_I + \chi^2_{II} = 0.38 + 4.57 = 4.95$, which is only slightly different from 4.91 obtained earlier when all four cells were considered together. For this reason, this is also called partitioning of chi-square.

## partogram

A partogram is used by midwives in field conditions, especially in developing countries, for preventing prolonged labor in childbirth. Progress of labor is recorded in a graph in terms of cervical dilation against time. Observations are recorded at regular time intervals, say, every hour. The partogram contains an alert line and an action line (Figure P.2).

Crossing the alert line is associated with fetal distress. Neonatal resuscitation is more likely if the alert line is crossed. If the action line is also crossed, the chances of stillbirths are higher. The function of a partogram is to provide early warning for detection of abnormal progress of labor. Dujardin et al. [1] found the partogram useful and efficacious in a study in Senegal. Less postpartum sepsis was reported [2] after implementation of the partogram in a multicenter trial done in Indonesia, Malaysia, and Thailand.


1. Dujardin B, De Schampheleire I, Sene H, Ndiaye F. Value of the alert and action lines of the partogram. *Lancet* 1992;339:1336–8. http://www.thelancet.com/journals/lancet/article/PII0140-6736(92)91969-F/abstract
2. World Health Organization Maternal Health and Safe Motherhood Programme. World Health Organization partograph in management of labour. *Lancet* 1994;343:1399–404. http://www.sciencedirect.com/science/article/pii/S0140673694925283


## path analysis

Path analysis works on the correlations of each predictor with the outcome in observational studies and decomposes them into two components: (i) due to direct effect and (ii) due to indirect effect through another predictor, such as of body mass index (BMI) through diet. Thus, a path is traced through a diagram as illustrated in Figure P.3 for the size of prostate and other factors. This helps in investigating the interrelations of the variables and tries to segregate causation from mere correlation. However, a causation model must be postulated for path analysis to work. Although the target may be one particular variable, all other variables are also considered stochastic in this setup, as opposed to regression where only the target variable is stochastic and the regressors are considered fixed. The observed correlations are considered fixed in path analysis in the sense that whatever values of correlations are obtained, they are taken on face value without worrying about the measurement errors.

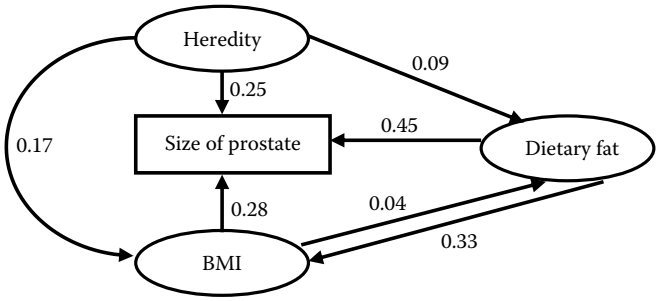



**FIGURE P.3** Example of path analysis of size of prostate affected by heredity, dietary fat, and BMI.

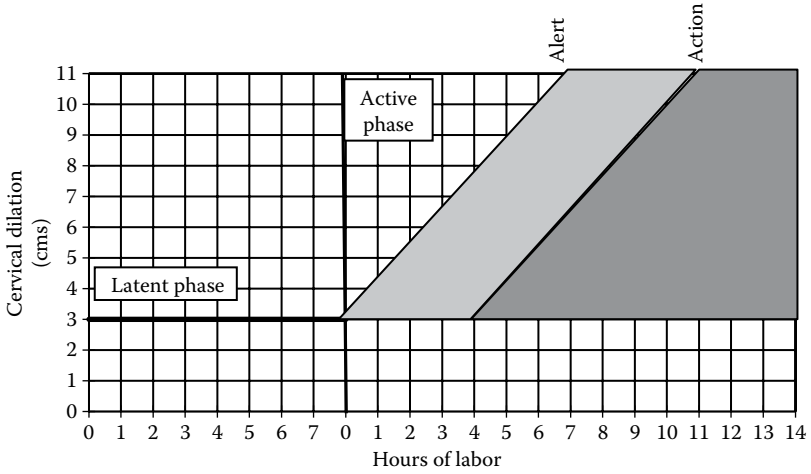



**FIGURE P.2** Partogram.

Despite its strength, path analysis cannot prove causation and its direction—it is just indicative. It provides a good graphical depiction of the extent and direction of the relationships on the basis of the correlations, and highlights the role of **mediators and moderators**. Note that this kind of exercise is not needed in experimental studies since there the design is such that causation comes up upfront anyway.

For example, in a study of the role of heredity, diet, and obesity on the size of the prostate gland, the first method of choice is ordinary regression, provided they are all quantitative. Suppose a heredity score can be developed on the basis of the proportion of male family members who have had prostate problem, diet measured for fat content, and obesity quantified by BMI. Ordinary regression may tell you, for example, that the net effect of heredity on determining the size of prostate is 12%, that of diet is 20%, and that of BMI is 15%, but the role of obesity through diet would remain obscure because diet and BMI are related. The results of path analysis can resolve this complexity as shown in Figure P.3 for this example.

Arrows and quantities in Figure P.3 indicate that heredity is not affected by any of the predictors considered in this figure, but it is affecting both fat intake and BMI—almost twice (0.17) as much BMI as fat intake (0.09). BMI has little effect on dietary fat (0.04), but fat has substantial (0.33) effect on BMI. Fat intake affects the size of prostate (0.45) more than BMI (0.28) does, whereas the effect of heredity (0.25) and BMI (0.28) is nearly the same. These values are path coefficients and interpreting them as "effects" is not necessarily appropriate, although that is how the literature describes the results of path analysis.

The path diagram is already looking complex with only three predictors in this example, and it can become difficult to comprehend if there are more predictors. It is for the researcher to postulate a right model for path analysis. In this example, the postulation is that the size of prostate is affected by dietary fat, BMI, and heredity, and nothing else. If the postulated model is wrong, correct results cannot be expected. Path analysis only provides an algorithm for decomposing the effect into direct and indirect within the postulated model, although one can test how well the postulated model fits to the correlation structure. As in the case of regression, a good fit does not preclude other models with better fit. Secondly, in Figure P.3, ovals are used for *regressor* variables and rectangle for the dependent variable. Sometimes ovals are reserved for latent traits when those also are considered.

The actual mathematics of path analysis are intricate and beyond the scope of this section, but they can be undertaken by a suitable statistical software package that uses various path models such as direct, independent, recursive, and nonrecursive. The above brief is just to apprise you of situations where path analysis can be tried, and the type of results expected from this analysis.

As mentioned earlier, path analysis requires each variable to be quantitative. The variables must be actually observed so that their measurements are available. If some variables are qualitative and others are unobserved underlying traits, also called latent traits of the type seen in factor analysis, **structural equation models** (SEMs) may be appropriate as they consider all these aspects and find out which variable should leave free and which of the others should be constrained or fixed. SEM assumes that the observed correlations are subject to measurement errors—path analysis does not. Details of path analysis and structural equation modeling are given in Loehlin [1]. A good example of where path analysis has been used is given by Tae et al. [2] on depression in Korean women with breast cancer–mediating effects of self-esteem and hope.

1. Loehlin JC. *Latent Variable Models: An Introduction to Factor, Path and Structural Equation Analysis*, Fourth Edition. Psychology Press, 2003.
2. Tae YS, Heitkemper M, Kim MY. A path analysis: A model of depression in Korean women with breast cancer-mediating effects of self-esteem and hope. *Oncol Nurs Forum* 2012;39(1);E49–57. https://onf.ons.org/onf/39/1/path-analysis-model-depression-korean-women-breast-cancer-mediating-effects-self-esteem-and

## pattern recognition

Pattern recognition is the area of research that deals with the operation and design of systems that recognize patterns within a set of data. Pattern-recognition technology is changing the way we do things as it is being used to evaluate, for example, speech, fingerprints, handwriting, and images, giving valuable insights with useful application to health and medicine as well.

While pattern recognition in its purest form is only possible using the human eye and a discerning mind, a computer system designed for the purpose is extremely useful and is very quick. When properly programmed, such a system can be relied on to provide consistent results. Validity, however, will depend on whether adequate dimensions of the problem have been visualized and properly accounted for in the program. This system is put through a training phase, and key known patterns are run through the program. The program initially guesses the answers; each guess is labeled correct or not correct, thus giving weights to the possible answers. The more weight assigned to an answer, the more the program will recognize the weighted pattern as the one it has been directed to find.

In clinics, physicians often learn to recognize disease patterns by knowledge and experience. A patient may present with a number of symptoms that all point to one disease, or they may not be so specific. Lacking expertise, the physician might create a differential diagnosis that is too large, or too small, and without assigning correct probabilities to the various potential diagnoses. Eventually, the physician learns to quickly determine the most likely diagnosis based on known disease patterns and as seen in a clinic by him or her. It is important, though, that the physician does not become sloppy and superficially misinterpret patterns by rushing to a diagnosis. A computer-aided diagnosis (CAD) can help in this respect.

CAD programs use pattern recognition methods to examine signs–symptoms, radiological images, and laboratory investigations for typical disease patterns. Each of these investigations should ideally have disease-specific definitive pattern for pattern recognition to be successful. The number of parameters for assessment may be huge, perhaps many more than a human mind can comprehend in one go, but that is not an issue with the present-day technology. In so doing, CAD allows one to spot possible patterns more quickly than by examining all the information manually.

Several types of reasoning are used by physicians during the diagnostic process: pattern recognition, algorithmic (using flow charts and algorithms) arguments, hypotheticodeductive (generating and rejecting hypotheses as more data are collected), exhaustive (gathering every possible piece of data to make the diagnosis), etc. Each is appropriate for certain situations and inappropriate for others. These are mostly done inadvertently in mind than on paper. Pattern recognition is an important ingredient and can be formally done using established methods. An example from the current literature is the article by Tiwari and Bhargava [1] for digital cancer diagnosis from chemical imaging data.

Statistically, **cluster analysis** is the method for pattern recognition in the data. This method groups subjects in clusters such that

the clusters are internally homogeneous and externally isolated. Several methods of **hierarchical clustering** are discussed in this volume such as average linkage, single linkage, and complete linkage. Farshad et al. [2] studied polymerase chain reaction–restriction fragment length polymorphism pattern for the gene of *Shigella sonnei* strains isolated from children with bloody diarrhea in Iran using the cluster analysis method.


1. Tiwari S, Bhargava R. Extracting knowledge from chemical imaging data using computational algorithms for digital cancer diagnosis. *Yale J Biol Med* 2015 Jun 1;88(2):131–43. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4445435/
2. Farshad S, Ranjbar R, Hosseini M. Molecular genotyping of *Shigella sonnei* strains isolated from children with bloody diarrhea using pulsed field gel electrophoresis on the total genome and PCR-RFLP of IpaH and IpaBCD genes. *Jundishapur J Microbiol* 2014 Dec 8;8(1):e14004. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4350046/


## Pearsonian correlation, *see* correlation coefficient (Pearsonian/product–moment)

## pedigree charts

Pedigree charts show ancestry, sibship, and progeny after arranging them in an organized manner. They are used by those interested in examining family tree to research the effect of heredity. This chart enables scientists to better understand the role of heredity and to make choices for manipulating future breeding that preserve or select specific strains. Information about the trait of each person of the selected families is charted for two, three, or more generations. Appearances (phenotypes) are used to study the genotypes. If the pedigree chart is about a particular individual patient, the chart will list that patient's parents, grandparents, and other known ancestors, as well as children and grandchildren if present, besides siblings in some situations.

Pedigree charts are commonly used for studying genetic diseases. One example is shown in Figure P.4 [1]. The pattern in such a chart helps to identify a trait as autosomal dominant or recessive since a dominant trait (e.g., familial hypercholesterolemia) shows a vertical pattern of inheritance (parents and children affected), whereas the recessive traits (e.g., beta thalassemia) show a horizontal pattern of inheritance (siblings affected). Males are represented by squares and females by circles, just in case the trait is sex-linked. Affected persons are represented by filled squares or circles and unaffected ones by hollow squares or circles. Carriers are represented by half-filled and half-hollow squares or circles. Pedigree charts can help to investigate X-linked disorders such as hemophilia A and color blindness.

Dwivedi and Aggarwal [2] studied pedigree charts of index cases of coronary artery disease (CAD) in India and forwarded the hypothesis that an individual's genetic profile functions as soil while various environmental factors such as physical inactivity, smoking, stress, etc. act as seeds in the etiopathogenesis of CAD. Another example is given by Pour-Jafari et al. [3] on oculocutaneous albinism Type1A (OCA1A) in an Iranian family.


1. Wilson JD, Braunwald E, lsselbacher KJ, Petersdorf RG, Martin JB, Fauci AS, Root RK (Eds.). *Harrison's Principles of Internal Medicine*, Twelfth Edition, Vol. I, International edition. McGraw Hill, 1991: p. 26.
2. Dwivedi S, Aggarwal A. Central obesity, hypertension and coronary artery disease: The seed and soil hypothesis. *World J Cardiol* 2011 Jan 26;3(1):40–2. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3030736/
3. Pour-Jafari H, Zamanian A, Pour-Jafari B. Genetic Analysis of oculocutaneous albinism Type1A (OCA1A) in an Iranian family. *Iran J Public Health* 2010;39(1):100–4. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3468964/


## penalized likelihood, *see* Akaike information criterion (AIC) and general AIC (GAIC)

## percentiles and percentile curves, *see also* growth charts

Percentiles are the values of the variable that divide the total number of subjects into ordered groups of 100 divisions. This is a form of **quantile**. See that topic for details of how it is calculated. For $n$ = 200 subjects, 35th percentile = (35 × 200/100) = 70th value after ordering from minimum to maximum.



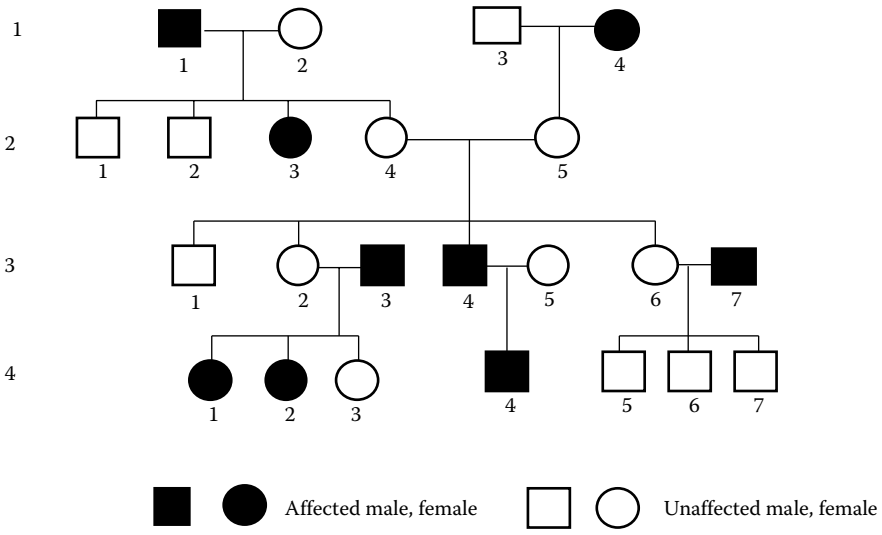



**FIGURE P.4** Pedigree chart.

Note that percentiles are very different from percentages. For example, the 90th percentile of esters in cholesterol could be 58%. This means that 90% of the people have esters less than or equal to 58%. Although theoretically we can have a 99.9th percentile, practically the highest is 99th percentile (better than 99%). No one is ever 100th percentile. Another feature worth noting for all quantiles, including percentiles, is that quantile of $(x + y) \neq$ quantile of $x +$ quantile of $y$. This needs to be understood in the context of mean since mean$(x + y)$ = mean$(x)$ + mean$(y)$.

A common use of percentiles is in growth charts of children for dimensions such as weight, height, and head circumference. Each can be assessed against age or against each other. See the topic **growth charts** for details and Figure P.5 for illustration. This shows various percentiles of height and weight of US girls at different ages from 0 to 36 months. Such charts use percentiles based on measurement of a large number of healthy children at each age and sex. The 50th percentile is like the median that can be used as a reference, and the 3rd and 97th percentile curves are often taken to define the lower and upper limits for healthy growth.

A good use of percentiles is in assessing the risk of heart disease in those who were thin at birth but gained weight fast—that is, initially at the lower percentile then became a part of the higher percentile in a matter of a few years—a phenomenon called *crossing the centiles*. There is a growing evidence that such children are at a greater risk of coronary heart disease [2].

Drawing percentile curves in these charts is not as simple as it looks. Intricate methods such as **LMS** and **BCPE** are used to calculate percentiles depending on the skewness and kurtosis in the data. Smoothing is done by **spline** functions, and measures such as **Akaike information criterion** are used to test the goodness of fit. The objective of all this complex methodology is to come up with valid percentile values and to obtain a trend that is close to the actual pattern of growth without missing any age of real slow or fast growth.

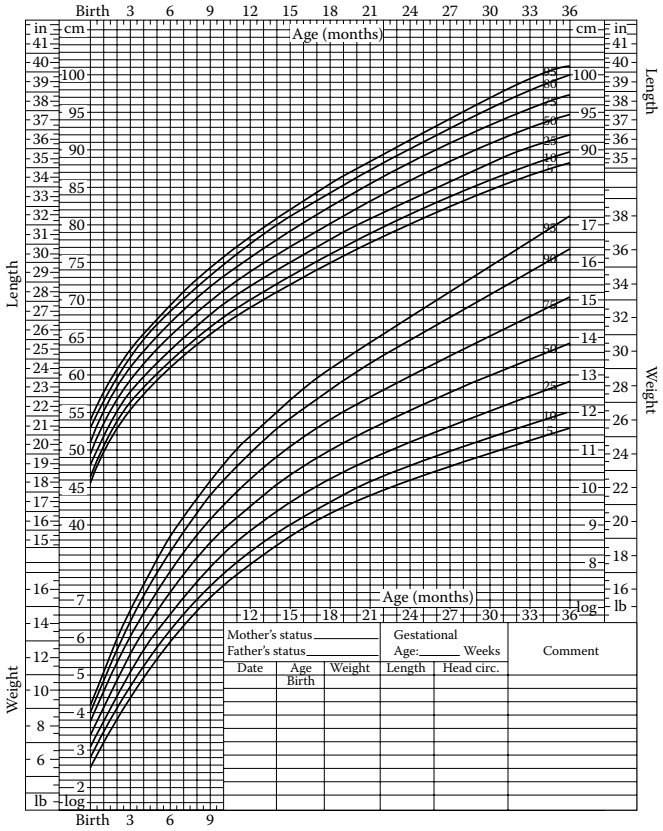



**FIGURE P.5** Height and weight charts of US girls from birth to 36 months. (From CDC. *Grow Charts*. http://www.cdc.gov/growthcharts.)

1. CDC. *Grow Charts*. http://www.cdc.gov/growthcharts
2. WHO. *Nutrition: Diet, Nutrition and Chronic Diseases in Context*. http://www.who.int/nutrition/topics/4_dietnutrition_prevention/en/index1.html

## periodogram

A periodogram is a plot of the behavior of a **time series** when broken down to different periods. The assumption behind this is that most time series tend to repeat the behavior of low and high values after certain periods of time, although there may be an increasing or decreasing secular (long-term) trend. Thus, a time series can be viewed as the sum of waves with different amplitudes and frequencies. This oscillatory behavior may occur once a day (circadian rhythms), once in four weeks (menstruation-related parameters), once a year (seasonal diseases), once in 6 years (influenza epidemics), etc. For some medical measurements, such periodicity is not known or is quite fuzzy. For example, this can happen in a patient with irregular heartbeats. A periodogram can help in extracting information on the dominant periodicity.

A periodogram estimates a particular behavior of time series, called spectral density, versus the frequency, where the latter is explored for all possible segments. It can identify the more common frequencies or the most common one. In simple terms, *spectral density* can be understood as proportional to the squared correlation between the observed series and waves with frequencies $r/n$ ($r = 1, 2, \ldots, n$). If there are $n$ values in the series, frequencies under examination would be $1/n$, $2/n$, $3/n$, etc. Generally, only the first few frequencies are examined. A peak in the periodogram indicates that an oscillation component exists near the frequency value corresponding to the peak. Some periodograms may exhibit multiple peaks—one dominant and the other recessives.

Consider brain cortex activity measured 128 times every 2 s for 256 s when a stimulus was applied for 16 time periods (of 2 s each) and not applied for another 16 time periods (of 2 s each). A time series plot of this (Figure P.6a) follows a regular pattern that seems to repeat about every 30 or so time periods in this example [1]. This is not surprising as the stimulus was applied in this fashion, and a repeating pattern every 16 + 16 = 32 time periods is expected. The periodogram (Figure P.6b) shows a dominant spike at a low frequency. It is hard to judge the exact location of the peak, but exact values reveal that the peak value of the periodogram corresponds to a frequency of 0.03125. The period for this value = 1/0.03125 = 32. That is, it takes 32 time periods for a complete cycle, as expected [1].

The example just mentioned is only to illustrate the features of periodogram, but the results were on expected lines because of the way the measurements were taken. In practice, the oscillations would not be easily discernible, and a periodogram can help in locating a common frequency if it exists.

1. The Pennsylvania State University, Eberly College of Science Resources for Online Courses 2015 STAT 516, Lesson 7: The periodogram. https://onlinecourses.science.psu.edu/stat510/?q=book/export/html/52, last accessed March 3, 2016.

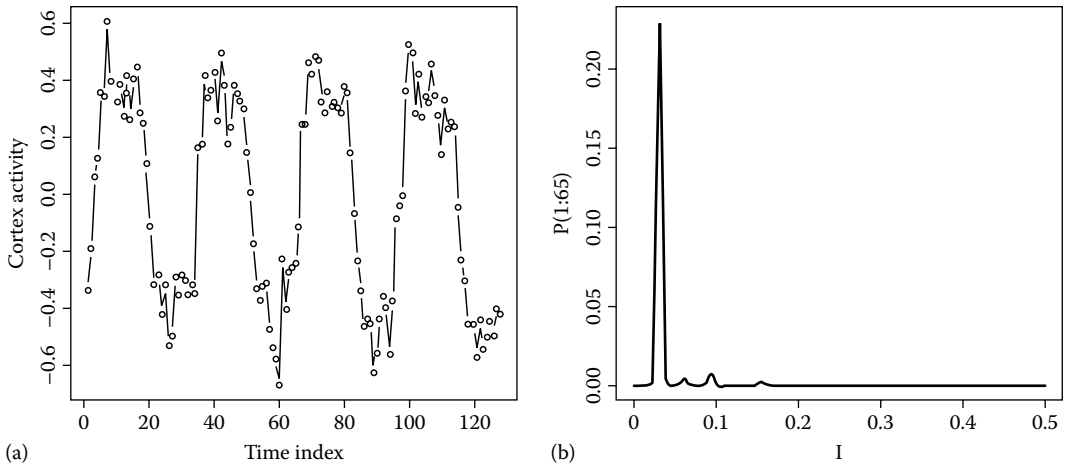



**FIGURE P.6** (a) Time series plot of cortex activity at different time points; and (b) periodogram of the series in (a). (From The Pennsylvania State University, Eberly College of Science Resources for Online Courses 2015 STAT 516, Lesson 7: The periodogram. https://onlinecourses.science.psu.edu/stat510/?q=book/export/html/52, last accessed March 3, 2016. With permission.)

## perinatal mortality rate/ratio, *see* mortality rates

## permutation tests

Permutation is the rearrangement of available values into distinct sequences. It is different from bootstrap because permutation is without replacement. Different permutations of the sample values to different groups allow us to compute the sampling distribution of a statistic in situations where it is too complex to find mathematically. This complexity occurs when the requirements of the regular sampling distributions are not met, such as when the values have not come from an established distribution (e.g., Gaussian or exponential), or the sampling scheme is too complex and does not fall into known models.

A permutation test is a statistical method of testing hypotheses based upon all potential permutations of subjects to groups. This test assesses whether, under the null hypothesis of no difference among groups, chance might account for the observed difference(s) in the samples. Say that we have a one-way analysis of variance (ANOVA) design with $K$ groups. If the null hypothesis is true, observations are just as likely to fall in one group as in the other. So we can permute the observations across the groups, calculating an $F$ value for each permutation. Repeat these steps approximately 5000 times, and calculate the percentage of repetitions in which the calculated $F$ values exceed the $F$ value obtained from the original data. This will give the $P$-value under the null hypothesis. This is a nonparametric procedure as this does not require any specific pattern of the distribution of values and is also called *randomization test*. A permutation test is exact when all possible permutations are considered.

There is a need to review the requirements underlying this approach. With observational studies, it is clear that we cannot randomly assign subjects across the groups. So we must assume that the observations are *exchangeable* in the sense that it is reasonable for a value to fall in either group under the null. This is the case if observations are statistically independent and if errors come from the same distribution, i.e., if they are identically and independently distributed.

This is straightforward with a one-way ANOVA as illustrated but not so with a two-way design because it is not always obvious what should be permuted and over what cells the permutation should take place. Also, there may be an interaction between the factors in a two-way experiment that could disturb in case of random permutation. In addition, permutation tests are computationally intensive, and they generally do not have as much statistical **power** as the regular tests when the conditions of the regular tests are fulfilled. Thus, permutation tests are recommended for only those setups that do not meet the established distributional requirements.

## personal probability, *see* probability

## person-time

It is sometimes not possible to observe each person in a **cohort** for the same duration. Also, the duration of exposure may vary from subject to subject. For example, persons in a stressful environment for different periods may be observed for incidence of peptic ulcer disease. One person may be under stress for 12 years, another person for 5 years, another one for 8 years, etc. These durations are totaled and called *person-years*. If the $i$th person is exposed for $x_i$ years, then

$$\text{total person-years of exposure for } n \text{ persons} = \Sigma x_i;\ i = 1, 2, \ldots, n.$$

This can be used as a base for the calculation of incidence per year of exposure or can be used as stand-alone for various indexes or scores, although there are limitations as mentioned later in this section. For example, for assessing smoking, cigarette-years are used as a stand-alone indicator as a measure of exposure to smoking, and for risk of diseases, it would be cases per 100 person-years of exposure. That is,

$$\text{incidence rate per 100 person-years} = \frac{\text{new cases occuring in the observed period}}{\text{person-years observed}} * 100.$$

In this setup, a uniform follow-up is not needed to calculate the incidence rate, although a follow-up is required in any case. The multiplier in the formula given in the above equation is not necessarily 100 and can be chosen per convenience.

Person-years is the most frequently used form of person-time, but this could also be calculated in terms of person-months, person-weeks, etc. For example, in the case of use of oral contraceptives, person-months are used, and incidence of a complication or of pregnancy is calculated per 100 person-months of use. For incidence of acute respiratory virus infections after an intense exposure, the follow-up may be a few weeks, and then the incidence will be worked out per 100 person-weeks.

The concept of person-time is valid only when the initial period is as important or as unimportant as the later period, and only if cumulative exposure is what matters and not the actual duration. In the case of a complex surgery, the risk of death in the first few days or weeks could be very different from the subsequent period when the patient stabilizes. In such a situation, person-time can lead to misleading results. A blatant and common misuse of the person-time tool is in calculating smoking exposure in terms of **pack-years**. Pack-years for a person smoking 2 packs of cigarettes a day for 5 years is the same as for a person smoking 1/2 pack a day for 20 years. Both are 10 pack-years. Intensity of smoking as measured by the number of cigarettes per day may have different implication as compared to duration for, say, incidence of cardiovascular disease. This is ignored when pack-years is used as done by Chatkin et al. [1] for assessing abdominal fat in smokers. Green et al. [2] reported incidence of primary outcome (cardiovascular death, nonfatal myocardial infarction, nonfatal stroke, or hospitalization for unstable angina) per 100 person-years of follow-up in patients of type-2 diabetes and cardiovascular disease to determine whether sitagliptin was noninferior to placebo when the patient follow-up was nonuniform with a median of 3 years. This is valid only if the incidence of outcome has nothing to do with the duration of follow-up.

1. Chatkin R, Chatkin JM, Spanemberg L, Casagrande D, Wagner M, Mottin C. Smoking is associated with more abdominal fat in morbidly obese patients. *PLoS One* 2015 May 15;10(5):e0126146. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4433108/
2. Green JB, Bethel MA, Armstrong PW, Buse JB, Engel SS, Garg J, Josse R et al. Effect of sitagliptin on cardiovascular outcomes in type 2 diabetes. *N Engl J Med* 2015 Jun 8. http://www.nejm.org/doi/full/10.1056/NEJMoa1501352



## pharmacokinetic parameters ($C_{max}$, $T_{max}$) and pharmacokinetic studies, *see also* area under the concentration curve (AUC curve), half-life of medications

Pharmacokinetics is the study of the movement of drugs in the body arising from absorption in the first phase, distribution in the second phase, and elimination in the last phase due to metabolism and excretion.

In order to undertake a pharmacokinetic study, the concentration of a drug in the body (mostly plasma or serum) is measured at a series of time points after ingestion to find how much is in the body and how much has been metabolized or excreted. Different drugs take different times to reach the peak concentration and its clearance. Thus, the primary presentation of pharmacokinetic data is the plot of concentration versus time—called the concentration curve. Sometimes the logarithm of concentration gives a more amiable picture. **Area under the concentration (AUC) curve** is just about the most important statistical parameter used to assess the

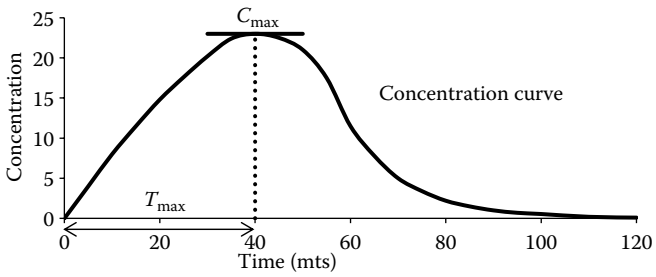



**FIGURE P.7** Concentration curve, peak concentration ($C_{max}$), and time to reach peak ($T_{max}$).

mechanical property of the drug that measures the total amount of drug that got into the body. Biological property is in terms of the efficacy and the side effects. Other pharmacokinetic parameters are peak concentration ($C_{max}$), time to reach peak concentration ($T_{max}$), and **half-life** that characterize the absorption phase. AUC curve and half-life are separately explained, and $C_{max}$ and $T_{max}$ are illustrated in Figure P.7. These parameters can be empirically estimated by drawing a concentration curve as in Figure P.7, where $C_{max}$ = 23 (in whatever units) and $T_{max}$ = 40 min. The elimination phase can take one of several forms, but generally, elimination is proportional to the existing level in the plasma/serum. This leads to exponential decline that converts to linear form in logarithmic units.

If there are 10 subjects for a study on pharmacokinetic properties of a drug, these parameters are obtained for each subject, and mean standard deviation are calculated as usual. The concentration curve can also be based on the average of these subjects, but that, as any average, can mask important age–sex and other variations that could tell a lot more about how drug works in different groups of subjects. Thus, analyze relevant groups separately if the number of subjects is adequate for this division. This may also be obtained separately for patients with mild disease than for patients with a severe form of disease, and may also be studied to depend on nutritional and other characteristics of the subject. If that is what you are interested in, a concentration curve should necessarily be drawn for each subject separately, and the area or any other parameter can be studied for the factors influencing these parameters by statistical methods such as **regression**.

Although the concentration curve empirically will be like a polygon with lines joining different points, it may be useful in some situations to model this as a smooth curve by means of an equation. That could be postulated as the population average. The population does not have to be the general population but can be restricted as always in statistics to a specific segment or specific type of cases.

An important use of pharmacokinetic parameters is in assessing **bioequivalence** of two or more drugs. You may be aware that this is generally assessed in terms of AUC curve and $C_{max}$. AUC can be misleading as two very different curves can give the same area.

For more details and pharmacokinetic models, see Ratain and Plunkett [1].

1. Ratain MJ, Plunkett WK. Principles of Pharmacokinetics, In: *Holland-Frei Cancer Medicine*. Sixth Edition (Kufe DW, Pollock RE, Weichselbaum RR et al., editors), Decker, 2003. http://www.ncbi.nlm.nih.gov/books/NBK12815/

## phases of (clinical) trials

Drug development is a complex process and requires a lot of research and extreme care. Before trying it on human subjects, the biochemical properties of a new drug are studied in a well-equipped laboratory, and then experiments are undertaken on a suitable animal model. Only after success at these two preclinical phases does the drug go into clinical trial. At clinical stage also, it is generally divided into phases. The purpose of phasing is to gradually increase the exposure and plough back the learnings while proceeding to advanced stages. This allows the developer and the regulatory authorities to supervise the development of the drug effectively. Details are provided later in this section, but in a nutshell,

- The focus of phase I is to assess the safety (tolerability) of the new drug by running a small clinical trial with graded doses on healthy volunteers (except when the nature of the drug precludes its administration to healthy subjects).
- Phase II determines the optimum dose and assesses the efficacy of the new drug in treatment of the target disease/condition.
- Phase III is a large comparative clinical trial to demonstrate the safety and efficacy of the new treatment with respect to the standard treatments available. This is needed to support product license applications.
- Phase IV is surveillance after the drug is released into the market.

The basic features of various phases of a clinical trial can be described as follows. Although these phases are used in some other setups, all these phases must be conducted with convincing success for developing a new therapeutic regimen. Such phasing is sometimes waived for a therapy or its variation that is already in use for some other condition, and the trial is conducted to examine its use in a new set of conditions.

### *Phase I Trial*

Phase I is usually undertaken on healthy human volunteers to study the pharmacologic properties of the regimen, such as concentration–time profile, food interaction, toxicity, and major side effects, and most of all to delineate the maximum tolerated dose. Thus, dose escalation may be done in this phase. It may not be easy to find volunteers for this phase of trial except courageous, healthy people who agree to participate for some compensation. The compensation should be proportional to the expected discomfort and not excessive that could be frowned upon as coercive or as unnecessary inducement. Except for regimens for diseases such as cancer that compromise tolerance, healthy subjects are preferred in this phase because therapeutic efficacy is not an issue at this stage. This phase generally needs less than 20 participants for each dose. A large number are not desirable since serious ethical issues arise due to the perils that volunteers may face. There is no control group in this phase.

If diseased cases are included for some specific reason, see that their varying severity does not affect the outcome. Since comorbidities can spoil this phase, it is necessary to rule out not only symptomatic diseases but also asymptomatic conditions such as low hemoglobin level and high triglyceride level so that the side effects are not unfairly attributed to the drug.

### *Phase II Trial*

Phase II of a trial is conducted on patients for whom the test regimen may be eventually indicated. The objectives of this phase are to (i) get an initial idea of potential clinical efficacy, (ii) assess short-term incidence of side effects, (iii) identify a dose schedule for various kinds of cases (such as for mild, moderate, severe, or for children and adults), (iv) investigate interaction with other drugs or effect of comorbidities, and (v) collect further pharmacologic data.

Phase II also establishes or refutes that the new regimen is likely to meet at least the minimum level of efficacy. If this level is not met, there is no use pursuing the regimen any further. This is a crucial phase that really establishes whether or not the regimen is likely to be clinically useful. Thus, it also provides the *proof of concept*. The number of participants in this phase is generally 100–200 and may be in a randomized trial mode with a control group on the pattern of a phase III trial.

Phase II can help in learning more about the treatment regimen and about the type of patients and kind of symptoms for which the treatment is beneficial. An appropriate dose and the appropriate subjects are identified for the phase III trial. Phase II may have to be stopped early if the regimen is found beyond tolerance in patients and serious side effects are seen. Failure of phase II may help to identify the problems with the regimen, which may indicate a need to go back to the basics for improving the formulation.

When interpreting the results of a phase II trial, keep in mind that the efficacy and toxicity might be interdependent. Thus, the error rate may be higher than that apparently obtained by considering them independently. In this phase, comorbidities are generally not excluded because applicability would suffer; sometimes subjects with potentially interacting comorbidities are intentionally included to avoid confusion. Comparison of efficacy and side effects in patients with and without comorbidities helps in defining the exclusion criteria for phase III trial.

### *Phases IIA and IIB*

There are times when there is a need for even more refinement in the development process by dividing phase II into phase IIA and phase IIB. For example, vaccine trials need even more precaution due to the applicability of vaccines to a large segment of populations who are not sick but are at risk, as opposed to therapeutics that is applied only to patients and administered under close supervision. A feature of vaccines is immunogenicity, which might be an important consideration in some diseases, in addition to protective efficacy. In other diseases, duration of protection may be important. Quality and quantity of immune responses required for protection against infection and against development of disease are scientific challenges. In the case of HIV, for example, there would be a vaccine that inhibits HIV infection, and there could be another that inhibits or retards development of disease—AIDS—in those already infected.

In view of the complexities involved in vaccine trials, an additional phase called phase IIB is sometimes advocated. This is also called the *test of concept* phase. The aim of phase IIA could be to establish the schedule of administration for different age groups as it would be most likely a factorial experiment with dose level as one factor and age group as the second factor. Thus, four phases are required for vaccine trials instead of the usual three for other regimens. The objective of phase IIA is to evaluate whether the vaccine has any efficacy at all, and in phase IIB, this objective shifts generally to at least 30% efficacy. The participants in phase IIB are not necessarily representative of the target population, whereas a representative sample is strongly indicated for phase III. Phase IIB also assesses the operational efficiency, whereas the objective of phase III is to produce compelling evidence of efficacy.

### *Phase III Trial*

The stage is set for a phase III trial once the early trials establish the overall safety of the regimen, its basic clinical pharmacology, its therapeutic properties, and its most important side effects. There must be a valid parallel control group and allocation of subjects fully randomized in this phase to various **arms of the trial**, including the control. For this reason, this is called a **randomized controlled trial** (RCT). The control arm should ideally get the best available treatment or, if no treatment is available, a placebo. Selection of the right cases and the appropriate controls, and randomization are important for this phase as it is through these that a phase III trial can provide compelling evidence of the efficacy and safety of the regimen or lack thereof. When benefits are explored, proper assessment of harm is equally crucial. In fact, these days, safety is an overriding consideration in many trials, particularly when efficacy of the existing regimen is already exceeding 80%. Phase III is also a prerequisite to meet the regulatory standards of license; for this reason, they are sometimes termed as *pivotal* studies.

For a vaccine, a phase III trial has to be on a larger scale so that adequate numbers developing the outcome, particularly in the control group, are available. The outcome may be in terms of immunogenicity and not necessarily the disease. Nevertheless, the total number of subjects may run into thousands, and the follow-up too may go for a long time. Since phase III is an expensive trial for vaccines, phase IIB becomes a highly desirable proposition for indicating whether or not to proceed to phase III. Phase IIB, however, increases the time frame because this too can take at least a couple of years.

### *Phase IV*

Phase IV are postmarketing studies after the drug has received the approval of the regulatory agency. For details, see **postmarketing surveillance**.



## phi coefficient, *see* association between polytomous characteristics (degree of)

## physical quality of life index, *see also* quality of life index

This is a composite index for measuring the quality of life at the *community level* by combining measures of child mortality, longevity, and education. These three are considered the cardinal components that govern the quality of life at the macro level. Note that these are the outcome indicators and not the inputs to the system. Thus, physical quality of life index (PQLI) measures performance or achievement rather than the resources. The components of PQLI reflect that the quality of life in less-developed countries is simply a "work-in-progress" version of that in industrialized countries. Comparison across countries or populations is considered valid since the same set of indicators is used. Morris [1] proposed PQLI in 1978.

At the individual level, the quality of life is mostly the subjective component of well-being. It can be defined as a composite measure of physical, mental, and social well-being as perceived by each individual or by a group of individuals. It includes aspects such as happiness and satisfaction as experienced in life, for example, in health, marriage, finances, education, and creativity. The quality of life can be evaluated by assessing a person's subjective feeling of happiness or unhappiness about various concerns of life. Attempts are made to reach one composite index by combining a number of health indicators.

The PQLI at the community level consolidates three community level indicators, namely, infant mortality, life expectancy at age 1 year, and literacy. Life expectancy considered for this index is at 1 year instead of the usual "at birth" since infant mortality is already included as a separate indicator. For each of these three components, the performance of individual countries or communities is placed on a scale of 0 to 100, where 0 represents an absolutely defined worst performance and 100 represents an absolutely defined best performance. This is similar to how the **human development index** (HDI) is derived and is called **normalization**. In the case of *positive* indicators of life expectancy and basic literacy, the best is shown by the maximum and the worst by the minimum. In the case of infant mortality, which is a negative indicator, the best is denoted by the minimum and the worst by the maximum. These are normalized as follows: achieved level = |minimum value – actual value|/(maximum value – minimum value). The composite index is calculated by averaging these three normalized indicators, giving equal weight to each of them so that each of these three receives equal importance. The resulting PQLI is also rescaled 0 to 100.

As just mentioned, PQLI can be compared over populations and over time to find the difference from one population to another and from time to time, but has been criticized since the three indicators are closely related to one another [2]. The index value has been found to be relatively high for some countries such as Sri Lanka that do not have high income—thus, it does provide an alternative way to look at development away from just income. The index is used more for assessment of social health than for physical health. An example of the use of PQLI in the Philippines is given by Disanayaka and Danuningrat [3].

1. Morris, MD. The physical quality of life index (PQLI). *Urban Ecol* 1978;3(3):225–40. http://www.sciencedirect.com/science/article/pii/0304400978900153
2. Larson DA, Wilford WT. The physical quality of life index: A useful social indicator? *World Dev* 1979 June; 7(6):581–4. http://www.popline.org/node/441921#sthash.G7RDq6pK.dpuf
3. Disanayaka M, Danuningrat MI. An assessment of welfare in the Philippines as measured by the physical quality of life index 1965–88. *DLSU Bus Econ Rev* 1991;4:25–34. http://ejournals.ph/index.php?journal=BER&page=article&op=view&path%5B%5D=8896

## PICO method, *see* population, intervention, comparison, and outcome (PICO) method

## pie diagram (exploded and wedged)

A pie diagram is a circular diagram divided into segments, each segment representing proportion in a category of the "whole." If the frequency in a category is $f_k$ out of $n$, the angle of the corresponding segment is $(360^* f_k/n)^\circ$. It is necessary that $\Sigma f_k = n$, that is, the categories must be mutually exclusive and exhaustive. A pie diagram cannot correctly display a **multiple response**.

As in the case of a histogram, percentages can replace frequencies in the case of a pie diagram also. In addition, the categories can be nominal or ordinal, which are not admissible for a histogram—thus increasing the scope of a pie diagram. Grouping of continuous data, as in the case of a histogram, is also required for a pie diagram if not already present, but discrete data may or may not be grouped. This means that a pie can be drawn for all datasets, which could be represented by a histogram, as well as in some other situations. When the metric data are in categories, the functional difference

between a pie and a histogram is that the pie is not adequate to depict the frequency distribution (the pattern) over various values. In addition, a pie does not provide a good representation when there are a large number of categories, although an adequate histogram can be drawn in that situation. However, a pie diagram is generally better in depicting the concentration of values in one category *relative* to the other categories.

Figure P.8a is a pie diagram for causes of death of children of age under 5 years in the world in the years 2000–2003 [1]. Acute respiratory diseases and diarrheal disease were the predominant causes. Figure P.8b is for Asia-Pacific region. Figure P.8c depicts the **disease spectrum** with the help of a series of pie diagrams that are linked to a segment of the previous pie [2]. In this diagram, manifestation of disease in mild, moderate, serious, and critical proportions is shown besides the proportion of survival and death in the case of critical manifestation of the disease.

As far as possible, each segment of pie must be labeled and must show the number or percentage of subjects it represents.

One useful application of a pie diagram is in comparing the relative distribution of two unequal groups of subjects divided into the same categories. Figure P.8a is based on 11 million deaths of children of age less than 5 years in the world per year during 2000–2003, whereas Figure P.8b is based on 4 million such deaths in Asia-Pacific. The higher number of death in the whole world is represented by a proportionate increase in the size of the pie. The segments are still comparable, and the greater predominance of malaria in the world relative to the Asia-Pacific is clear. Such comparability in groups of unequal size is rarely achieved by any other type of diagram.

Another useful feature of a pie diagram is wedging. When attention is to be specifically drawn to one particular category or a set of categories, the segments representing those categories are wedged out as shown for neonatal deaths in Figure P.8a and b. A pie can also be exploded so that all segments are wedged out. This is not shown in these figures.

1. WHO. *World Health Report 2005.* World Health Organization, 2005.
2. Indrayan A. *Basic Methods of Medical Research*, Third Edition. AITBS Publishers, 2008: p. 202.

## Pillai trace

Pillai trace is one of the multivariate test statistics utilized in **multivariate analysis of variance (MANOVA)** that is transformed into an *F*-ratio for assessing the significance level of the effect of the factors under study. This was proposed by Sreedharan Pillai in 1955 [1].

Sreedharan Pillai

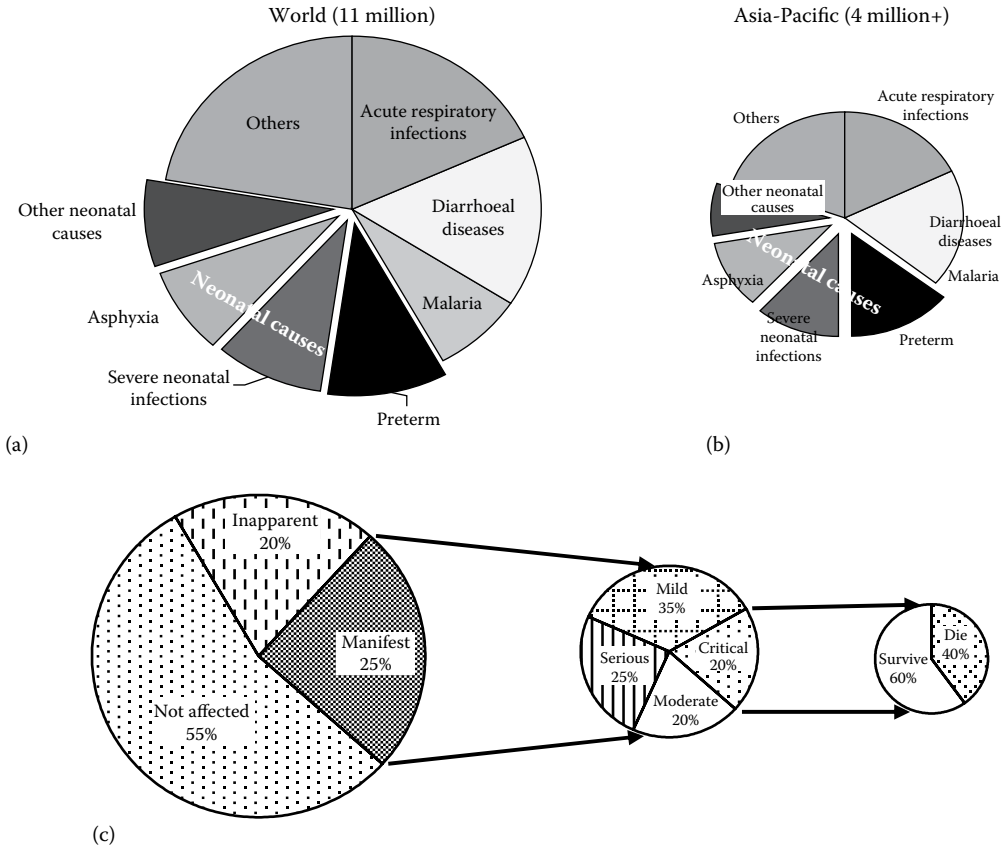



**FIGURE P.8** Pie diagrams: (a) causes of under-five mortality—world; (b) causes of under-five mortality—Asia-Pacific 2000–2003; ([a and b] From WHO. *World Health Report 2005.* World Health Organization, 2005.) (c) spectrum of disease by connected pies. (From Indrayan A. *Basic Methods of Medical Research*, Third Edition. AITBS Publishers, 2008: p. 202.)

MANOVA can be directly compared with the analysis of variance of univariate data, except that the groups are compared on *K* response variables simultaneously. In the univariate case, *F*-tests are commonly used to assess the hypotheses of interest. In the multivariate case, however, no one test statistic can be constructed to be optimal in every case. The most widely used of the available test statistics are Wilks lambda and Pillai trace. Both are obtained from eigenvalues of a matrix that comes up in case of multivariate analysis.

While Wilks lambda gives more accurate results when the requirements of multivariate Gaussian distribution and homogeneity of dispersion matrices are met, Pillai trace is generally more robust to the violation of these requirements. This works reasonably well for relatively small sample sizes also. It calculates the amount of variance in the dependent set of variables accounted for by the greatest separation of the independent variables, just as in discriminant functions. Most standard statistical software packages give a *P*-value based on Pillai trace that can be directly used to draw conclusions. As the sample size increases, the **central limit theorem** starts operating and the two tests tend to give the same result. Both can give the same result in some other specific situations also.

An example of the use of Pillai trace is given by Sánchez-Morla et al. [2]. They compared multivariate neuropsychological function between patients with schizophrenia and the control group using MANOVA and found Pillai trace = 0.493 that corresponds to F = 34.3 and yields P < 0.001.

1. Pillai KCS. Some new test criteria in multivariate analysis. *Ann Math Stat* 1955;26:117–21. http://www.jstor.org/stable/2236762
2. Sánchez-Morla EM, Santos JL, Aparicio A, García-Jiménez MÁ, Soria C, Arango C. Neuropsychological correlates of P50 sensory gating in patients with schizophrenia. *Schizophrenia Res* 2013 Jan;143(1):102–6. http://www.ncbi.nlm.nih.gov/pubmed/23148896

## pilot study and pretesting



It is generally considered highly desirable that all questionnaires, schedules, laboratory procedures, etc., are tested for their efficacy before they are finally used in the main study. This is called **pretesting**. Many unforeseen problems or gaps can be detected by such an exercise, and the tools can be accordingly adjusted and improved. This pretesting can reveal whether the items of information are adequate, feasible, and clear; that the space provided for recording is adequate; that the length of the interview is within limits; that the instructions are adequate; etc. This also serves as a rehearsal of the actual data collection process and helps to train the investigators. Sometimes pretesting is repeated to standardize the methodology for eliciting correct and valid information.

A study on a small number of subjects before the actual study is called a pilot study. This simulates the actual study and provides a preliminary estimate of the parameters under investigation. Such a preliminary estimate may be required for the calculation of sample size. If the phenomenon under study has never been investigated earlier, the pilot study is the only way to get a preliminary estimate of the required parameter. A pilot study may also provide information on the size of the clusters and sampling units at various stages that could be important in planning cluster or multistage sampling for large-scale surveys. In case of clinical trials, a pilot study can give an idea of the recruitment rate, duration of follow-up, and practical constraints on the blinding [1].

Sometimes observations on the first few cases of a study are used as constituting a pilot study, in which case this is called an *internal* pilot study. The objective in this setup is mostly to get a preliminary estimate of the parameter of interest so that the sample size can be projected. An internal pilot study is applicable only if the design of the study and instruments are final and need no modification. In other words, this presumes that pretesting has been done, and all changes as required have already been incorporated.

A pilot study is not undertaken to investigate statistical significance of the result, nor will the small scale of this study be adequate for estimation of the effect or to judge its practical utility. It only provides a preliminary estimate of the relevant factors for designing a full study when prior data are not available and helps to reduce the areas of uncertainty in the final investigation. In a rare case, when the effect size is really large and a pilot study carried out with proper selection of subjects, the results may even be conclusive, although many would question the conclusion based on a pilot study.

1. Birkett MA, Day SJ. Internal pilot studies for estimating sample size. *Stat Med* 1994; 13:2455–63. http://onlinelibrary.wiley.com/doi/10.1002/sim.4780132309/abstract

## placebo

A placebo is an inert substance or a procedure that is supposed to neither help nor harm the subjects. Saline injection, colored water, pills containing inactive substance, and sham surgery are examples. This is often used for comparison with the active regimen under test in a clinical trial. The objective of using a placebo is to simulate the conditions of the test regimen and make the patient believe that he/she is getting the treatment.

As in all medical experiments, clinical trials can have one or more treatment regimens, but a parallel or concurrent control group is almost invariably required except in crossover or before–after setups. Real controls are those that are similar but follow the natural course of disease without any intervention, but such subjects would be aware that they were receiving no treatment and that could influence the outcome. For this reason, a placebo is preferred over no treatment. In practice, thus, the reference control group is either treated with an existing regimen or administered a placebo.

The placebo is important because some subjects tend to behave or respond differently when no treatment is given compared with when a sham treatment is given. Although the phenomenon is well recognized and documented, the actual mechanism of how and why this happens is poorly understood. This is surmised to happen due to activation of mu-opioid reception in the brain by the *expectation* of relief [1]. As far as possible, a placebo should be indistinguishable from the active therapeutic agent under trial in terms of appearance, size, color, taste, smell, etc., but it is not always possible to make it absolutely identical. Also, it should be given in a parallel dose so that the **masking** is complete.

There is always an ethical question about using a placebo on patients who are known to have the disease because they need an active ingredient to cure their ailment. However, placebos can be justifiably used in the following situations, although using placebos may affect **randomization** in some of them.

- No standard treatment is available, i.e., the existing treatment modality has very doubtful results—perhaps no better than a placebo.
- New evidence has emerged regarding the doubtful efficacy of the standard therapy.

- Existing regimen is too costly or is rarely available to the patients at large.
- On patients who have already been given standard treatment and have not benefited, and no second line of treatment is available for them.
- Test regimen is an add-on to the existing regimen. This means that all patients in the trial, including those on placebo, would receive the normally prescribed therapy anyway.
- Patients refuse to accept existing therapy but are willing to be part of a trial where they know that they can receive a placebo.

In situations where these conditions are not met, and for procedures for which a placebo group is nearly impossible such as in renal dialysis and fitting of an artificial limb, a group on existing therapy can serve as control. But control is now widely considered as a scientific necessity. Whether existing treatment or placebo, the control group should undergo the same medicinal rituals, such as dietary regulations, as the treatment group. This is more easily said than done.

There is a debate whether surgical trials need a group with sham surgery as the placebo. Perhaps evidence is not enough that sham surgery has the same psychological benefits as a placebo in a drug trial. Because of the major ethical issues, studies in humans involving sham surgery are rarely performed. There is a debate as to whether a surgical trial needs a sham surgery group at all as there is a theoretical argument that the placebo effect of sham surgery may be even more operative than with placebo medications. Nevertheless, a sham surgery group can be adopted for a setup where it is not too expensive and is harmless to the participants. This can be easily adopted in animal experimentation as done by Losey et al. [2] on mice because of far fewer ethical concerns in this setup. For an example of sham surgery on patients, see Hunter et al. [3].


1. Zubieta JK, Bueller JA, Jackson LR, Scott DJ, Xu Y, Koeppe RA, Nichols TE, Stohler CS. Placebo effects mediated by endogenous opioid activity on mu-opioid receptors. *J Neurosci* 2005;25:7754–62. http://www.jneurosci.org/content/25/34/7754.full
2. Losey P, Ladds E, Laprais M, Geuvel B, Burns L, Bordet R, Anthony DC. The role of PPAR activation during the systemic response to brain injury. *J Neuroinflamm* 2015 May 22;12(1):99. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4450490/
3. Hunter JG, Kahrilas PJ, Bell RC, Wilson EB, Trad KS, Dolan JP, Perry KA et al. Efficacy of transoral fundoplication vs omeprazole for treatment of regurgitation in a randomized controlled trial. *Gastroenterology* 2015 Feb;148(2):324–33.e5. http://www.gastrojournal.org/article/S0016-5085(14)01208-6/fulltext


## point-biserial correlation

We try to address four questions in this section: (i) What is the point-biserial correlation coefficient? (ii) How is this coefficient calculated? (iii) How is this coefficient related to other **correlation coefficients**? (iv) How is this coefficient used in practice? Brief answers are as follows.

The point-biserial correlation coefficient ($r_{pb}$) is a measure of the degree of relationship between a naturally occurring **dichotomous** scale and a **metric scale**. For example, you might want to investigate the degree of relationship between gender (male/female) and hemoglobin level (a metric scale). This process is different from finding the mean difference between males and females. The dichotomous categories are coded as 0 and 1. Let the quantitative variable with which this correlation is to be computed be denoted by $x$.

$$\text{Point-biserial correlation: } r_{pb} = \frac{\bar{x}_1 - \bar{x}_0}{s}\sqrt{pq}$$

where

$\bar{x}_1$ = mean of $x$ for those with code 1
$\bar{x}_0$ = mean of $x$ for those with code 0
$s$ = standard deviation of $x$ for both the groups combined (with denominator $n$ and not $n-1$)
$p$ = proportion of those with code 1
$q$ = proportion of those with code 0 ($p + q = 1$)

Consider the shock index in male and female patients with ST-segment elevation myocardial infarction (STEMI). Suppose we have 28 male and 19 female patients. Let their shock index values be as follows:

mean shock index in 28 male patients = 0.83
mean shock index in 19 female patients = 0.72
SD of shock index in the total of 47 patients = 0.065

These give a point-biserial correlation coefficient between sex and shock index $= \frac{0.83-0.72}{0.065}\times\sqrt{\frac{28}{47}\times\frac{19}{47}} = +0.83$. Since the sign is positive, the higher the code (in this example, males), the higher the shock index; and the degree of relationship is an impressive 0.83 on the 0 to 1 scale. Figure P.9 supports this strong relationship. More compact values of the variable $x$ in each group give a higher value of the point-biserial correlation.

In fact, the point-biserial correlation is the same as the Pearsonian **correlation coefficient** ($r$) between $x$ and $y$, where $y$ is now recorded as 0 and 1. The Pearson formula reduces to what is just given for this setup. The interpretation of the coefficient is also similar. Like $r$, $r_{pb}$ can range from 0 to +1 if the two scales are related positively (i.e., in the same direction) and from 0 to –1 if the two scales are related negatively (i.e., in opposite directions). Confidence interval can be easily obtained, and test of significance can be easily done when $x$ follows a Gaussian distribution.

How is the point-biserial correlation coefficient used in practice? Say we are interested in the degree of relationship between being male or female, and the degree of anxiety as measured by the hospital anxiety and depression score (HADS score). The point-biserial correlation will help you to get the answer. Most

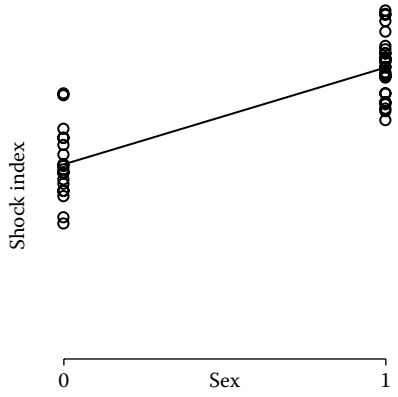



**FIGURE P.9** Relationship of shock with sex of STEMI cases.

commonly, however, researchers use $r_{pb}$ to calculate the item–total score correlation as another, more accurate, way of estimating item discrimination. The $r_{pb}$ shows the degree to which each item is separating the stronger subjects on the whole from the weaker ones. The higher the $r_{pb}$, the better the item in discriminating. Consider a setup where you are assessing some newly developed quality of life index of people with some chronic disease (code = 1) versus those without any disease (code = 0). If our index is good, those with code 1 (higher value) will have a substantially lower score than those with code 0 (lower value). Thus, the point-biserial correlation will be negative. If the correlation is not high on the minus side, you can conclude that the quality of life index used in this exercise fails to discriminate between cases with and without the disease under study. This way, point-biserial correlation can be used for item analysis.

If the dichotomous categories are artificial such as anemia <10 g/dL and nonanemia ≥10 g/dL, then this correlation is called just *biserial* instead of *point-biserial*. Such dichotomy of a metric scale, though, is not desirable. **Biserial correlation** may still fall in the comfort zone of some researchers. If this is so, realize that, in this case, you have ordered categories—you know that the category <10 g/dL is less than the category ≥10 g/dL. They can be legitimately coded as 0 and 1. This is not the case when the categories are male and female. The formula changes slightly in that the point-biserial correlation is further divided by the height of the standardized normal distribution at the point $d$, where $P(z < d) = q$ and $P(z > d) = p$ when the underlying distribution is Gaussian. Statisticians know that this height is $\frac{1}{\sqrt{2\pi}} e^{-d^2/2}$. This is the additional factor in the denominator of biserial correlation. The biserial correlation coefficient is always greater than the point-biserial correlation coefficient since the divisor is always less than 1.

## Poisson distribution

You may be aware that the **binomial** variable $x$ *counts* the number of successes out of $n$ independent trials. If $n$ becomes extremely large and the probability of success becomes extremely small, ultimately in the limit we get a Poisson distribution. Application wise, if migraine occurs 3 times in a year on average in established cases, the probability that it will occur 6 or more times in a year in a random case can be obtained by Poisson distribution. This distribution is given by

$$\text{Poisson distribution: } P(x|\mu) = \frac{e^{-\mu}\mu^x}{x!};\ x = 0, 1, 2, \ldots,$$

where μ is the mean. In the migraine example,

$P(x \geq 6|\mu = 3) = 1 - P(x \leq 5|\mu = 3)$ by the complementary rule of probability

$$= 1 - \left( \frac{e^{-3}3^0}{0!} + \frac{e^{-3}3^1}{1!} + \frac{e^{-3}3^2}{2!} + \frac{e^{-3}3^3}{3!} + \frac{e^{-3}3^4}{4!} + \frac{e^{-3}3^0}{5!} \right)$$

$$= 1 - e^{-3}\left(1 + 3 + \frac{9}{2} + \frac{27}{6} + \frac{81}{24} + \frac{243}{120}\right)$$

$$= 1 - (0.0498 \times 18.4)$$

$$= 0.084.$$

Thus, the chance is nearly 8.4% that a random migraine patient from this "population" will have 6 or more attacks during 1 year. Poisson tables are available in statistics books that give these probabilities, and statistical packages in any case have provision to give exact Poisson probabilities. The value of μ, being the average, does not have to be an integer. In most situations, this will be in decimals.

The mean μ of Poisson is estimated by the sample mean $\bar{x}$ as usual, and the variance of Poisson is also μ so that its estimate is also $\bar{x}$. Equality of mean and variance is considered a defining property of Poisson. In any set of data on rate where mean and variance are nearly equal, you should especially look for Poisson distribution. Examples of Poisson variables are (i) number of deaths occurring in a cardiac hospital per day; (ii) number of myocardial infarction cases arriving in a general hospital per day; (iii) number of measles cases occurring in a city per year; and (iv) number of handicap persons per 1000 population. The standard error of mean can be easily obtained as usual, and the exact confidence interval on population mean can be obtained by Poisson distribution.

In most practical situations where Poisson is applicable, the mean would be small, possibly less than 1. If it gets bigger, say more than 10, the Poisson too tends to behave in the same manner as Gaussian. For example, if μ = 20, then for Poisson $\sigma^2 = 20$, and by Gaussian approximation, the chance that this number is 15 or less in a specific case is

$$P(x \leq 15) = P(x < 15.5) \text{ with } \textbf{continuity correction}$$

$$= P\left(z < \frac{15.5 - 20}{\sqrt{20}}\right)$$

$$= P(z < -1.01)$$

$$= 0.156 \text{ from Gaussian distribution.}$$

The exact answer with Poisson distribution is 0.157. The Gaussian approximation is not far off.

Poisson distribution has been used extensively for medical conclusions. For example, Morris [1] has used this distribution for deleterious mutations in zygotes to study their genomic load in view of their relevance to death in infancy and childhood. A simple example is obtaining standard errors and 95% confidence limits of age-specific incidence rates of pediatric retinoblastoma in India by sex group in each population-based cancer registry using the Poisson distribution [2].


1. Morris JA. The genomic load of deleterious mutations: Relevance to death in infancy and childhood. *Front Immunol* 2015 Mar 16;6:105. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4360568/
2. Rangamani S, SathishKumar K, Julka PK et al. Paediatric retinoblastoma in India: Evidence from the national cancer registry programme. *Asian Pac J Cancer Prev* 2015;16(10):4193–8. http://www.apocpcontrol.org/page/apjcp_issues_view.php?sid=Entrez:PubMed&id=pmid:26028071&key=2015.16.10.4193


## Poisson regression

Refer to the topic **Poisson distribution** and the topic **regression** while reading this section. Poisson regression is a type of **generalized**

**linear model** (GLM) where the random component is specified by the Poisson distribution. A GLM has the following structure:

$$y = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \ldots + \beta_K x_K + \varepsilon.$$

The random component here is ε, which is ordinarily considered to have a Gaussian distribution with mean 0 and variance $\sigma^2$. However, in case the dependent variable $y$ is a count or a rate, such as the number of myocardial infarction cases arriving in a general hospital in a day, Gaussian distribution does not hold and Poisson distribution works better. When ε is considered to have a Poisson distribution, this is called Poisson regression.

You may be aware that all regression models are for the mean of $y$ than $y$ itself, and this mean could be a fraction even when individual values of $y$ are the counts. As with any kind of regression, a Poisson model is also fitted to the data values, the coefficients are estimated, and they are interpreted for conclusions. But the mathematics and the underlying theory are different from ordinary least squares regression, which makes Poisson regression a separate topic. It is generally observed that most counts of events have the feature of being mostly around the mode, but occasional value can be large and can also be exceedingly large on rare occasions. To correct this skewness, Poisson regression models the natural logarithm of the mean count instead of the count itself. After this transformation, negative values would occur corresponding to the mean counts between 0 and 1 (counts themselves cannot be negative). The logarithm is called the **link function** since this transforms counts to the values that can be considered to have Gaussian distribution. The logarithm of the response variable is linked to a linear function of regressor variables such that $\ln(y) = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \ldots + \beta_K x_K$. In other words, the typical Poisson regression model expresses the log outcome as a linear function of a set of regressors. However, this is valid in certain conditions: (i) as the regressor variables increase in equal increments, the logarithm of the dependent changes by the same amount; (ii) the combined effects of different regressors change the dependent in a multiplicative manner; (iii) at each level of the regressor, the dependent has variance nearly equal to the mean; and (iv) observations are independent.

Poisson regression can also be used to model the rates based on the person-time. This may be useful when all individuals are not followed for the same length of time. So, instead of having $\ln(y) = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \ldots + \beta_K x_K$, you can have $\ln(y/t) = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \ldots + \beta_K x_K$, where $t$ is the person-time such as 1000 for per thousand rate. This can be written as $\ln(y) = \ln(t) + \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \ldots + \beta_K x_K$— only the intercept changes.

Poisson-distributed data are intrinsically integer-valued, which makes sense for count data. That said, a Gaussian distribution can be a good approximation to a Poisson when the mean value is 10 or greater. This may be easier to fit and would actually be more general, since the Poisson regression requires that the mean and the variance are equal, while Gaussian-based usual regression can deal with unequal means and variances. For a count data model with means different from variances, one could also use a **negative binomial distribution**, for instance.

An example of the use of Poisson regression is given by Pocock et al. [1]. They used it for predicting the 3-year mortality rate in heart failure patients on the basis of the data in 30 studies. The predictors in this model were age, lower ejection fraction, New York Heart Association class, serum creatinine, diabetes, etc. Pereira et al. [2] used Poisson regression to find the factors influencing prevalence of dyslipidemia in a section of Brazilian population.

1. Pocock SJ, Ariti CA, McMurray JJV, Maggioni A, Køber L, Squire IB, Swedberg K et al. Predicting survival in heart failure: A risk score based on 39372 patients from 30 studies. *Eur Heart J* 2012;34:1404–13. http://eurheartj.oxfordjournals.org/content/early/2012/10/23/eurheartj.ehs337
2. Pereira LP, Sichieri R, Segri NJ, Silva RM, Ferreira MG. Self-reported dyslipidemia in central-west Brazil: Prevalence and associated factors. *Cien Saude Colet* 2015 Jun;20(6):1815–24. http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-81232015000601815&lng=en&nrm=iso&tlng=en

## polynomial regression, *see* curvilinear regression

## polytomous categories, *see* categories of data values

## ponderal index, *see also* body mass index (BMI)

Ponderosity is about the body heaviness and so is directly related to weight. Ponderal index measures the weight in relation to the size of the body, particularly the height. This term is primarily used for neonates. The corresponding term for adults is obesity.

Although ultrasonographic measurements can be used to assess the growth of the fetus, the first measurement of physical health of a child after birth is the weight. This generally declines for a few days after birth and then regained. If the weight immediately after birth cannot be recorded, as in some underdeveloped countries, the weight on the seventh day in many cases would be a good approximation. The normal range is 3.2–3.7 kg. A birth weight less than 2.5 kg is conventionally considered low in many countries. A low birth weight not only has been found to be associated with increased risk of early mortality but also is surmised to affect growth and development during adolescence and trigger diseases in adulthood such as coronary artery disease and diabetes. Similarly, high weight can be precursor to obesity in adulthood with all its adverse consequences. Thus, it is important to assess ponderosity in neonates. A useful index for neonatal weight is the ponderal index:

ponderal index for neonates:

$$\text{PI} = \text{weight in kg}/(\text{crown–heel length in m})^3$$

A child with a ponderal index of 30 or greater can be considered overweight, but in some conditions, such as in maternal smoking, reduced length may also be implicated. An index between 25 and 30 is considered normal, and that between 20 and 25 is considered marginal; a child with an index less than 20 is classified as *small for gestational age* (SGA). The index can be used as a prognostic indicator and to advise the family to take special care. When weight and length are both low, this index may not reveal the deficiency, but the prognosis is poorer. Such a symmetric SGA child is generally classified as *intrauterine growth retarded* (IUGR). This is identified by a very low weight but a nearly normal ponderal index. Some organizations do not distinguish between SGA and IUGR children, and both are identified only by low weight, generally below the 10th percentile point for the gestational age, irrespective of the length of the child.

The Ponderal index as just defined can also be used for adults, but the normal range will be different. Moreover, for this age, body mass index is a well-established index of obesity.

A general form of ponderal index is weight/$(\text{height})^b$, where $b$ is estimated from the simple linear regression of ln(weight) on ln(height) separately for each age. This can be used for any

age—only the value of $b$ changes from age to age. Freeman et al. [1], for example, found that $b = 2.08$ for boys of 7 years, $b = 2.20$ for girls of 7 years, $b = 2.44$ for boys of 16 years, and $b = 1.75$ for girls of this age in the United Kingdom. This varies from one population to another. Generally speaking, the value of $b$ declines from 3.0 for neonates to 2.0 for adults as we move up for age.

1. Freeman N, Power C, Rodgers B. Weight-for-height indices of adiposity: Relationships with height in childhood and early adult life. *Int J Epidemiol* 1995;24:970–6. http://www.ncbi.nlm.nih.gov/pubmed/8557455

## pooled chi-square, *see* Mantel–Haenszel procedure

## pooled OR, *see* Mantel–Haenszel procedure

## pooled RR, *see* Mantel–Haenszel procedure

## pooled variance

This is the variance in two or more groups combined and can be obtained on the basis of the variance in individual groups. The objective in case of samples is to get a better estimate of the variance because of a large combined sample size when pooling is permissible. The only requirement is that the sample variances in different groups do not differ significantly. Pooling of widely dispersed variances may lead to wrong results.

Consider the scenario where you have two sample means from two independent samples. The standard error (SE) of the difference between the two sample means is given by

$$\mathrm{SE}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

where $s_1^2$ is the sample variance in the first sample with size $n_1$, and $s_2^2$ is the sample variance in the second sample with size $n_2$. If the variances are nearly equal (see the **Levene test**), you can pool the variances of the two samples and get a better estimate. Then

$$\mathrm{SE}(\bar{x}_1 - \bar{x}_2) = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ is the pooled variance. This is better in the sense that this is now based on $n_1 + n_2$ values. The pooled variance (two independent samples) can also be written as

$$s_p^2 = \frac{\Sigma_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + \Sigma_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}.$$

Such pooled variance is commonly used in the two-sample **Student *t*-test** for significance of difference in means of two independent groups. The test changes to the Welch test in case the variances are unequal and cannot be pooled. This is also used for finding the confidence interval for the mean difference in the population when the variances are equal.

The method can be easily extended to many samples. In this case, the pooled variance for $K$ independent samples is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_K - 1)s_K^2}{n_1 + n_2 + \ldots + n_K - K}.$$

This may look like a new expression, but those familiar with one-way analysis of variance (ANOVA) know that this is the same as **mean square error (MSE)**, alternatively written as MSE = $\dfrac{\Sigma_{k=1}^{K}\Sigma_{i=1}^{n_k}(x_{ki} - \bar{x}_k)}{n - K}$, where $n = n_1 + n_2 + \ldots + n_K$, although in the case of ANOVA, slightly different notations are used. This is valid only if the sample variances are not significantly different across groups—commonly understood as the **homoscedasticity**. In this case, the pooled variance is the estimate of the variance of the errors in the ANOVA model. As abundant caution, note that this is not the same as the variance of all $n$ values combined for all the groups because that uses overall mean whereas MSE uses group means.

## population (the concept of)

Population in common parlance is the collection of people residing in a defined area, but it has a special meaning in statistics. It can be understood as the target group to which the findings of a study would apply. This is primarily understood as the **target population**. This is not necessarily a population of human beings—it could be a population of blood samples, of animals, of cases of kidney diseases, of healthy children in a school, of hospitals, etc.

In a descriptive study of acute respiratory infection (ARI) in a country, the target population could be all existing cases of ARI in that country. For a cervical cancer control program, the target population could be all married women of age more than 40 years. For studying risk factors of enlarged prostate, the population of interest could be all men of age 50 years and above in an area. In a laboratory experiment, this could be the population of mice or the population of histological specimens. In a clinical trial setup, the inclusion and exclusion criteria define the population of interest.

For most medical studies, the number of subjects in the population, called the size of the population, is large. Cost and logistic considerations seldom allow the study of all the subjects in a population. Therefore, sampling becomes a natural choice. However, sampling should be such as to provide representative cross-section of the population so that the results can indeed be generalized to the entire target population. This is what makes the **sampling techniques** so important.

In rare situations, all subjects in the target population can also be investigated. Then such a study is called complete enumeration or **census**. However, even if all existing cases are surveyed, there is no guarantee that the results would apply to *future* cases. Thus, the concept of population in the context of medicine is more hypothetical than real. When all existing cases are indeed included, it still remains a sample considering that future cases as well as those who have died are not included. Medical empiricism implies that the findings on the existing cases are used for future cases, and sampling is a prerequisite for this paradigm. Nonetheless, if the objective is to find the prevalence of diabetes mellitus in 2016 among females of age 40 years and above residing in a particular city, complete enumeration is possible. Similarly, if a complete registry of cancer cases in a defined population is available, perhaps sampling is not needed for assessing the existing situation.

A futuristic perspective brings in the concept of statistical **universe**. A universe could be larger than the target population that has

implications for the result. Future cases are part of the universe but not of the population.

## population attributable risk and population attributable fraction, *see also* attributable risk

**Attributable risk (AR)** due to a particular factor is the difference in risk of disease in subjects with that factor and the subjects without that factor when other factors remain constant. This factor is generally termed as exposure, although this can be any, such as low hemoglobin level, high obesity, and adverse family history. Attributable risk can be expressed as

$$\text{AR} = R_1 - R_0,$$

where $R_1$ is the risk (or incidence) in the exposed group and $R_0$ is the nonexposed group. This measures how much risk increases due to the exposure.

In view of the public health importance of AR, it is sometimes of interest to estimate the excess rate of disease attributable to the exposure in the *total population* under study. This excess is called the population attributable risk (PAR) and is calculated as

PAR = **incidence rate** in the total population (including exposed subjects) – incidence rate in the nonexposed subjects.

This can also be called population risk difference. The first component of this equation can be estimated only when the sample contains the exposed and the nonexposed subjects in the same proportion as in the total population. Otherwise, an extraneous estimate would be needed. Note that PAR is the rate of disease in the population minus the rate in the unexposed group. This is different from AR because the population comprises both the exposed and the nonexposed groups of people. In fact, it can be shown that PAR = AR*$p$, where $p$ is the **prevalence** of exposure. This measures the impact of eliminating the exposure from the entire population—and thus helps in developing strategies for control.

The PAR fraction measures the proportion of disease in the population that is attributable to the exposure and is the proportion of incidence that could be eliminated if the exposure were eliminated. This can be directly obtained from the **relative risk** (RR) as follows when the proportion of persons with the given risk factor is known:

$$\text{PAR fraction} = \frac{p(\text{RR}-1)}{p(\text{RR}-1)+1},$$

where $p$ = the proportion of persons having the given risk factor and is the same as the prevalence of exposure. The PAR fraction is also called the **etiologic fraction**. Whenever RR can be approximated by odds ratio (OR) (this requires low prevalence of disease), PAR can be estimated on the basis of case–control data. For a method of doing so with multiple risk factors in case–control studies, see Bruzzi et al. [1].

As an example, suppose oral cancer has a 10-year risk of 0.0002 among nonchewers of tobacco and 0.015 among the chewers. If only 2% are chewers, there are 2000 chewers and 98,000 nonchewers in a population of 100,000. At the risk presumed above, 20 among nonchewers and 30 among chewers are expected to develop oral cancer in a 10-year period. These two together amount to 50 per 100,000 or 0.0005 as the risk in the total population. Thus, PAR = 0.0005 – 0.0002 = 0.0003. If chewing of tobacco is completely eliminated from the population through educational campaigns or otherwise, 3 cases per 10,000 population would be saved at the end of a 10-year period. A health administrator may not view this as a substantial improvement considering the cost involved in the educational campaign and the fact that 2 cases per 10,000 would in any case occur among nonchewers.

The result may look dramatically different if PAR *fraction* is calculated. Using the formula given earlier, since RR = 0.015/0.0002 = 75 in this example,

$$\text{PAR fraction} = \frac{0.02(75-1)}{0.02(75-1)+1} = 0.60.$$

This means 60% of the risk of oral cancer in the population is attributable to chewing of tobacco. This is the same as PAR = 0.0003 out of population risk of 0.0005.

PAR is extensively used in epidemiological studies. Suhrke and Zahl [2] estimated the population attributable fraction of 8.2% of breast cancer incidence in 2006 due to use of hormonal therapy in Norway. Azimi et al. [3] have highlighted how sum of PAR fractions for different exposures can exceed 100% when sufficient care is not taken in its calculation.


1. Bruzzi P, Green SB, Byar DP, Brinton LA, Schairer C. Estimating the population attributable risk for multiple risk factors using case-control data. *Am J Epidemiol* 1985;122:904–14. http://www.ncbi.nlm.nih.gov/pubmed/4050778
2. Suhrke P, Zahl PH. Breast cancer incidence and menopausal hormone therapy in Norway from 2004 to 2009: A register-based cohort study. *Cancer Med* 2015;8(4):1303–8. http://onlinelibrary.wiley.com/doi/10.1002/cam4.474/full
3. Azimi SS, Khalili D, Hadaegh F, Yavari P, Mehrabi Y, Azizi F. Calculating population attributable fraction for cardiovascular risk factors using different methods in a population based cohort study. *J Res Health Sci* 2015 Winter;15(1):22–7. http://jrhs.umsha.ac.ir/index.php/JRHS/article/view/1761/




## population growth model and curve

Population growth refers to the changes in population in a particular area brought about by births, deaths, and residents moving into and out of an area. This can be modeled to produce a population growth curve. Knowing the size of the population and its projection allows the governments and businesses to make decisions to meet the expectations. Besides humans, these models are also used for bacterial colonies and viruses.

According to the World Population Clock [1], at the dawn of agriculture, about 8000 BC, the human population of the world was approximately 5 million. Over the 8000-year period up to 1 AD, it grew to 200 million with a growth rate of under 0.05% per year. A tremendous change occurred with the industrial revolution: whereas it had taken all of human history until around 1800 for the world population to reach 1 billion, the second billion was added in only 125 years (1927), the third billion in less than 35 years (1960), and the fourth billion in 14 years (1974). It has slowed down a bit since then and it takes nearly 12 years each time to add a billion. The world population in 2016 is estimated as 7.5 billion.

The preceding data may have convinced you that all populations, in general, tend to grow in exponential fashion—multiplying per unit of time instead of adding. That is, they double—the doubling time could be a day, a year, or 100 years depending upon the organism. This is particularly true for bacteria and viruses when allowed to grow unhindered, and for them, the doubling

time is short. The corresponding statistical model that depicts this pattern is **exponential curve**. This model is applicable when the population grows as a proportion of itself, such as 8% each day of what it was at the beginning of the day. If the growth rate is 0.01% per day, it will take 6930 days for population to double; if the growth rate is 1%, it will take 70 days to double; and if the growth rate is 3%, it will take only 23 days to double. If you have some data on a population for some time points, exponential curve (Figure P.10a) can be easily obtained as a regression on time after taking the logarithm of the population. As opposed to an exponential pattern, the populations that tend to limit themselves for whatever reason will automatically relent and stabilize. This gives a logistic type of curve (Figure P.10b). An important feature of this curve is that it shows slow initial growth, then increasing growth rate, rapid growth during the middle period, declining growth rate a little later, and plateauing thereafter. The third pattern could rise fast to a point after which the population tends to plateau (Figure P.10c).

Human populations have recently seen widespread interventions such as contraception. Thus, this requires a relatively complex model of the following type:

$$\text{human population growth model: } y_t = \frac{a}{1+be^{-ct}},$$

where $y_t$ is the population at time $t$, and $a$, $b$, and $c$ are the parameters in this model that would vary from population to population. No transformation of variables $y$ and $t$ can reduce this equation to a linear form. Thus, this model is genuinely nonlinear in parameters. The parameter $a$ mostly defines the baseline, $b$ is the growth rate, and $c$ is the limiting parameter. This is shown in Figure P.10c for specific values of the parameters, which we call classical.

Growth in many populations across the world is primarily due to high birth rate relative to death rate such as in India, but in countries such as the United States, this would be mostly due to immigrants and their descendants [2]. Some countries such as Russia and Belarus have negative population growth as the death rate exceeds the birth rate [3]. This is rare in natural course but can happen in specific situations.


1. World Population Clock. http://www.worldometers.info/world-population/, last accessed June 13, 2015.
2. SUSPS. *Population Numbers, Projections, Graphs and Data*, http://www.susps.org/overview/numbers.html, last accessed June 13, 2015.
3. About Education. *Negative Population Growth*. http://geography.about.com/od/populationgeography/a/zero.htm, last accessed June 13, 2015.


## population, intervention, comparison, and outcome (PICO) method

When researching medical literature, questions often come to mind that makes finding answers a real challenge. With regard to each of these questions, the way forward is to dissect the question into its component parts and then to restructure it in such a way that it is easy to find answers. For example, the most common clinical questions in **evidence-based medicine** are about how to treat a disease or condition. Many such questions can be divided into the following four components, summarized by the acronym PICO: P—population, patient; I—intervention, indicator; C—comparator, control; and O—outcome. The details are as follows.

P: What is the population, patients, or participants for whom the answers are sought?
I: What is the intervention (diagnostic test, exposure, or the treatment regimen) that you are interested in?
C: Is there a control or alternative test or exposure for comparison, also known as the comparator?
O: What are the patient-relevant outcomes of interest of the intervention?

Restructuring the research question as above makes way for a lucid answer. In medical research, this is used for selecting relevant articles for a **systematic review**. These reviews are considered the backbone of evidence-based medicine. For example, Patel et al. [1] studied cervical spine collar clearance in the obtunded adult blunt trauma patient by using PICO as follows: population—obtunded adult blunt trauma patient; intervention—cervical collar removal with adjunct imaging; comparator—cervical collar removal without adjunct imaging; and outcome—peri-clearance events. This helped them to select appropriate studies for a systematic review. Zou et al. [2] used PICO structuring of a clinical question regarding astragalus in the prevention of upper respiratory tract infection in children with nephrotic syndrome, and searched several medical literature databases for relevant papers. PICO seems to be gaining wide acceptance for choosing studies for systematic reviews.


1. Patel MB, Humble SS, Cullinane DC, Day MA, Jawa RS, Devin CJ, Delozier MS. Cervical spine collar clearance in the obtunded adult blunt trauma patient: A systematic review and practice management guideline from the Eastern Association for the Surgery of Trauma. *J Trauma Acute Care Surg* 2015 Feb;78(2):430–41. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4409130/

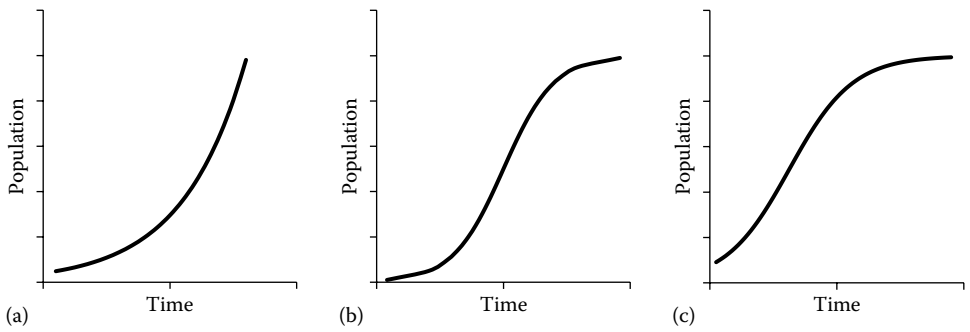




**FIGURE P.10** Population growth models: (a) exponential; (b) logistic; and (c) classical (see text).

2. Zou C, Su G, Wu Y, Lu F, Mao W, Liu X. Astragalus in the prevention of upper respiratory tract infection in children with nephrotic syndrome: Evidence-based clinical practice. *Evid Based Complement Alternat Med* 2013;2013:352130. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3638577/

## population pyramid

This is a special bar diagram for showing age–sex distribution in a population. Population pyramid is commonly used in demography but can also be used in a large-scale study to depict the age–sex distribution of the subjects of the study. This diagram shows the number or percentage of males and females in different age groups in a population. Figure P.11 shows the changing age–sex distribution of the population in a developing country such as Bangladesh over a long period. Such pyramids can be used to study the population dynamics. The details are as follows.

The vertical axis running down the center is age, and horizontal bars on the left side are the percentage or number of females and those on the right side are the percentage or number of males. *The age groups must be equal for this kind of depiction*. Note the striking difference between the shapes of the population pyramids for the different periods. The shape initially in 1975 is triangular with a broad base, showing the predominance of the child population. The decline in the population for each age group shows that deaths occurred at almost every age in 1975. Fifty years later (2025), due to slow transition, the pyramid may take a conic shape. After that, the transition is expected to be fast and the projected distribution of the population up to the age of 60 years is nearly uniform for the year 2050. Practically, no deaths occur in age groups up to 60 years because almost everybody tends to live longer. After that age, the deaths are fast and frequent. As the population evolves, the women of old age groups increase: the width of the right-hand side (females) is broader than that of the left-hand side (males). As of now, men die earlier than women, and this is reflected in the projections into 2050 and 2075: it is clearly visible in the pyramid projected for the year 2075 that there will be many more females of old age than males.

Population pyramid can be an effective tool in studying epidemiological impact of diseases. Belle et al. [1] presented population pyramids of Lesotho in the years 1976, 1986, 1996, and 2006 to show impact of HIV/AIDS on the demographic profile of that country. Iani et al. [2] used this to draw a proportionate stratified sample according to the age–sex structure in Italy.

1. Belle JA, Ferriera SB, Jordaan A. Attitude of Lesotho health care workers towards HIV/AIDS and impact of HIV/AIDS on the population structure. *Afr Health Sci* 2013 Dec;13(4):1117–25. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4056471/

2. Iani L, Lauriola M, Costantini M. A confirmatory bifactor analysis of the Hospital Anxiety and Depression Scale in an Italian community sample. *Health Qual Life Outcomes* 2014 Jun 5;12:84. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4054905/

## positive health

Positive health can be understood as the ability to cope with physical, biological, psychological, and social stress [1]. This can become far too abstract. Yet, measurements such as hemoglobin (Hb) level, high-density lipoprotein (HDL) level, immunity level, vital capacity, and pain-bearing capacity can be possibly used to assess positive health. The same is inversely true for low cholesterol, low sedimentation rate, and low bleeding time. Thus, an index of positive health can indeed be developed. Measurement of positive health has remained unexplored and requires the attention of researchers. The profile of persons who rarely fall sick and are able to do more work than others while leading an enjoyable life can be studied to identify factors that contribute to positive health. It is possible, though, that psychological factors such as personality profile, absence of stress, and carefree attitude contribute more to positive health than physiological parameters. This needs to be explored. This is a flagship concept mooted by Indrayan and Sarmukaddam in 2001 [1], and has been picked up by others for exploration such as by Seligman [2].

Assessment of the level of health and disease in individuals can be very subjective. Perfect health can seldom be defined, and the meaning of well-being changes from person to person. The subject's own perception also changes from time to time, and health becomes an unattainable ideal in a true sense. This, however, does not distract scientists, and they always make efforts to measure the level of health. The focus so far, though, remained on lack of health rather than on its presence. Thus, there is a need to define and pursue the idea of positive health.

Note that the way positive health is explained in this section by us is very different from positive health outcomes, positive health behaviors, positive health benefits, etc. These terms have been in use for a long time. Our concept of positive health is the state of health of an individual for a sustained period of time in the sense of capacities and abilities albeit restricted to psychosomatic health. This needs to be explored for proper assessment.

1. Indrayan A, Sarmukaddam S. *Medical Biostatistics*. Chapman & Hall/CRC, 2001.
2. Seligman MEP. Positive health. *Applied Psychology* 2008;57:3–8. http://onlinelibrary.wiley.com/doi/10.1111/j.1464-0597.2008.00351.x/full

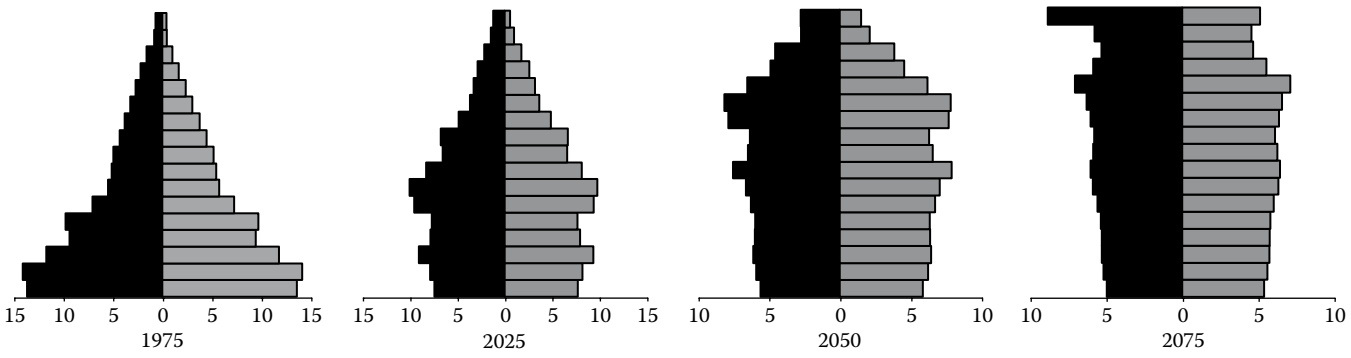



**FIGURE P.11** Population pyramids (projections for a developing country—1975, 2025, 2050, and 2075).

## posterior probability, *see* prior and posterior probability

## post-hoc comparisons, *see* multiple comparisons

## postmarketing surveillance, *see also* phases of (clinical) trials

A regimen, a drug, or a device can be released for marketing after getting the license from the regulatory authorities on the basis of successful completion of the phase III trial. However, the monitoring continues for the effects and side effects. This is called postmarketing surveillance, and many consider it the phase IV of a clinical trial.

Phase IV is generally carried out in an observational studies mode than in a trial mode. The drug is used in a routine setup and not intentionally ingested by the experimenter. The efficacy and adverse reactions are observed as they occur, and the conditions are not controlled. Under this surveillance, all adverse reactions and other events attributable to regimen are monitored, and the effectiveness is also evaluated. Patient preferences are studied, and health professionals and the public are encouraged to voluntarily share their experiences of any unexpected occurrence. This helps to uncover side effects not seen in a controlled trial and those that require sustained exposure to the regimen. Findings about tamoxifen carrying a risk of endometrial cancer [1] and arthroscopic surgery not beneficial for osteoarthritis of knee [2] are results partially attributable to such surveillance. Appropriate actions are taken to address the problems including risk management of the cases who unwittingly suffer. Investigations are launched to find if the manufacturers adhere to the terms and conditions of approval and that the drug or device is produced in a consistent and controlled manner as stipulated. Sometimes unannounced inspections are also carried out. Companies in any case are required to report any such events.

Any evidence of compromise on safety is a sensitive issue that can be easily hijacked by the media without gathering sufficient evidence. Pharmaceutical companies may have to react swiftly to investigate any such report because its damages can be irreversible.

### *Pharmacoepidemiology*

A whole new science of pharmacoepidemiology has emerged to study postmarketing issues of medical products, including drugs. A good reference on this subject is the book by Strom [3]. Pharmacoepidemiologic research is the study of the use and effects of health care products (e.g., pharmaceuticals, devices, and vaccines) after they are in practice for a while. It has now expanded to include clinical, economic, and other health outcomes, requiring new study methods. Pharmacoepidemiology is being used increasingly to evaluate health care systems, interventions, and health-related behaviors when actually used in practice. This now is considered the core science of therapeutic risk evaluation and management since these risks many times emerge after the product is marketed and used for a sufficiently long time.

The Guidelines for Good Pharmacoepidemiology Practices (GPP) [4] are intended to address these activities and other pharmacoepidemiologic studies. The GPP are intended to apply broadly to all types of pharmacoepidemiologic research, including feasibility studies, validation studies, descriptive studies, and etiologic investigations, as well as all related activities beginning from design through publication. The GPP also support risk management activities.


1. Fisher B, Costantino JP, Redmond CK, Fisher ER, Wickerham DL, Cronin WM. Endometrial cancer in tamoxifen-treated breast cancer patients: Findings from the National Surgical Adjuvant Breast and Bowel Project (NSABP) B-14. *J Natl Cancer Inst* 1994 Apr 6;86(7):527–37. http://www.ncbi.nlm.nih.gov/pubmed/8133536?report=abstract
2. Mounsey A, Ewigman B. Arthroscopic surgery for knee osteoarthritis? Just say no. *J Fam Pract* 2009;58:143–5. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3183924/
3. Strom BL (Ed.). *Pharmacoepidemiology*, Fourth Edition. Wiley, 2005.
4. International Society for Pharmacoepidemiology. *Guidelines for Good Pharmacoepidemiology Practices (GPP)*, 2015. https://www.pharmacoepi.org/resources/guidelines_08027.cfm


## poststratification

This is stratification of the sample after the selection is complete and all the data have already been collected. This is used when strata are not envisaged before the sampling but become clear after the data collection. Thus, the strata sizes here are random variables, as opposed to conventional stratified random sampling where strata and strata sizes are generally known before sampling is done.

Say that you have collected all your data but then find that unforeseen confounding has occurred. Confounding is said to exist when a factor affects the outcome as well as the antecedent under study. Confounding factors compete with the hypothesized risk factor as an explanation for the observed response. Frequent egg eaters generally have raised lipid levels. If these subjects are found to be at higher risk of coronary artery disease, it is difficult to say that this is due to eating eggs per se or is contributed, at least partially, by the raised lipid level that can occur for other reasons as well. To evaluate their separate effects, special care should be taken to include egg eaters with normal lipid profile and noneaters with raised lipid levels. When the stratification is known, this can be done beforehand. Unforeseen confounding can also occur or relevant strata might not have been visualized before the study. If one or both of these occur, a poststratification can be tried after the data are collected. This is also done if a particular group turns out to be of some special interest because of exceedingly good or exceedingly poor outcome. A sufficient number of cases may or may not be available in each group after this kind of stratification. If there were very few cases in one or more groups, the validity of the results would suffer.

It is quite common for researchers to use poststratification techniques in survey data analysis when it is discovered that some group or groups are grossly overrepresented or underrepresented. In this case, the estimate can be accordingly adjusted, and a relatively unbiased estimate can be obtained. The adjustment weight is often called the poststratification weight. This is obtained as $\pi_k/p_k$ for the $k$th stratum, where $\pi_k$ is the proportion in the population and $p_k$ is the proportion in the sample. If the weight for any stratum is 2, it really means that each sample value in this stratum will count as two values. These can be calculated only if the strata sizes (or the proportions) in the population are fairly well known. All these weights are always positive and nonzero.

The method to obtain these weights when population percentages are known is illustrated in Table P.3. In this example, strata A and B were underrepresented in the sample compared with the

**TABLE P.3**
**Calculation of Poststratification Weights**

| Stratum | Percentage in the Population | Percentage in the Sample | Poststratification Weight |
|---|---|---|---|
| A | 32 | 21 | 32/21 = 1.5238 |
| B | 45 | 33 | 45/33 = 1.3636 |
| C | 13 | 37 | 13/37 = 0.3514 |
| D | 10 | 9 | 10/9 = 1.1111 |
| **Total** | **100** | **100** | |

population, and stratum C was overrepresented. Stratum D was just about right. Poststratification weights correct this imbalance. When these weights are used for calculating something like mean (i.e., obtain the **weighted mean**), the value obtained will be a much better estimate of the population mean. The strata could be age–sex, blood group, disease severity, or any other. They must, however, be **mutually exclusive and exhaustive**.

Poststratification is also used by epidemiologists while analyzing health survey data. For example, certain diseases, e.g., cancer, are more common among older populations. When comparing the prevalence rates among geographic regions with different age structure, it is necessary to make adjustments according to such demographic categories and to compute relative prevalence rates of the diseases.

Poststratification is like data dredging since this is done after seeing the sample. Thus, this should be done only when it becomes necessary. Any modification done after seeing the data has potential to introduce bias. For example, imbalance in the sample could be due to differential nonresponse that may affect the outcome. Try to obtain the self-weighted data right at the outset. Poststratification weights work well for mean and proportion but not for something like median and quartiles. Extreme weights, which can arise due to gross underrepresentation or overrepresentation of strata, can produce unstable results. In this case, consider if collapsing the categories can help. You may have to exercise caution and make provisions for weights in analyses such as regressions and analysis of variance. Also, different characteristics can give very different weights, and obtaining a combined weight could be an uphill task. In this case, obtain the cross-classification of the population and think of considering each category as a stratum. This is possible only when the distribution of the population subjects is available by all combinations of these characteristics and will give reasonable results only when the sample size is large.

## potential-years of life lost (PYLL), *see also* disability-adjusted life years (DALYs)

Also called *person years of life lost*, potential-years of life lost (PYLL) is defined as the sum total of the years of life lost due to "premature" deaths in a population. Years of life lost for one person is the **expectation of life** at the age at death. This is the additional years he/she might have lived on average if he/she does not die at that age. If a person dies at age 74 years and the expectation of life at age 74 is 16 years, years of life lost is 16 for that person. For this calculation, expectation of life can be for the country of that person but is mostly taken for the population with the highest expectancy. Japan is the country where the expectation of life is the highest. In Slovakia, their expectation of life at 74 may be 7 years, but if expectation of life in Japan at age 74 is 16 years, the years of life lost will be considered 16 and not 7. This helps in obtaining the years of life lost on comparable scale across countries. PYLL is estimated by drawing up a **life table**: by linking life table data to each death of a person of a given age, sex, and race.

PYLL is generally calculated per year. If there are n deaths in a population in one particular year, the years of life lost in the population = x1 + x2 + … + xn, where xi is the years of life lost of the ith person. For comparability, this is also calculated per 1000 population. If PYLL in India is 182 years per 1000 population in 2016, and in 2011 it was 196 years, you know that improvement has taken place. If this is 75 years in Germany in 2016, you know how much ahead of India this country is with respect to this indicator.

What has just been mentioned is for all causes of death, but PYLL can also be calculated for any particular disease. For example, one can conclude from PYLL that cancer causes more person-years lost than do all the other diseases combined. Understanding the concept is also quite exacting: we can draw upon an example contained in a cancer trends progress report—2015 update [1]. PYLL caused by cancer helps to describe the extent to which life is cut short by cancer. On average, in the United States, each person who died from cancer in 2012 lost an estimated 15.7 years of life. This loss has been calculated per person who died of cancer and not per 1000 population. Average years lost due to a disease is the PYLL divided by the number of deaths by that cause. This is potentially a better method, or at least a complementary method, of measuring the burden of lives lost than death rates alone. Soneji et al. [2] were able to conclude by using PYLL that cancer burden increased as a result of decline in cardiovascular mortality, and observed that prior assessments have underestimated the impact of cancer interventions.

PYLL can be a useful metric for comparison of diseases also. If PYLL from cancer is 42 per 1000 population in a country and PYLL from coronary diseases is 38, you know that both are afflicting nearly the same loss of life in that country on average. (These values are illustrative and not real.) In another country, the affliction by coronary disease may be early in life, which would result in higher years of life lost by this disease in this country. You can also calculate this separately for men and women and get the idea which disease causes more early deaths in females than males. For example, when deaths are restricted to age 0–69 years, PYLL in 2012 from all causes was 2.048 in the females of Austria per person but was 3.578 in Hungary, whereas these values in males were 3.835 and 7.507, respectively [3]. Note the disparity between males and females. Among others, this provides another evidence that males are dying early in both these countries.

PYLL can be used for morbidity also when equated to death by some weightage as mentioned for **disability weight**.

1. National Cancer Institute. *Cancer Trends Progress Report: Person-Years of Life Lost*. (November 2015). http://progressreport.cancer.gov/end/life_lost
2. Soneji S, Beltrán-Sánchez H, Sox HC. Assessing progress in reducing the burden of cancer mortality, 1985–2005. *J Clin Oncol* 2014;32:444–8. http://jco.ascopubs.org/content/early/2014/01/13/JCO.2013.50.8952.abstract
3. OECD.StatExtracts. *Health Status: Potential Years of Life Lost*. Organization for Economic Co-operation and Development. http://stats.oecd.org/index.aspx?DataSetCode=HEALTH_STAT, last accessed September 11, 2015.

## poverty index

Poverty is generally equated with lack of income, but Oxford Poverty & Human Development Initiative (OPHI) [1] uses a multidimensional

**TABLE P.4**
**Components of Multidimensional Poverty Index**

| Indicator | Weight of the Major Component | Division of Weight into Subcomponents |
|---|---|---|
| I. Education | 1/3 | |
| Ia. No one has completed five years of schooling | | 1/6 |
| Ib. At least one school-age child not enrolled in school | | 1/6 |
| II. Health | 1/3 | |
| IIa. At least one member is malnourished | | 1/6 |
| IIb. One or more children have died | | 1/6 |
| III. Living conditions | 1/3 | |
| IIIa. No electricity | | 1/18 |
| IIIb. No access to clean drinking water | | 1/18 |
| IIIc. No access to adequate sanitation | | 1/18 |
| IIId. House has dirt floor | | 1/18 |
| IIIe. Household uses "dirty" (dung, firewood, charcoal) cooking fuel | | 1/18 |
| IIIf. Household has no car and owns at most one of: bicycle, motorcycle, radio, refrigerator, telephone, or television | | 1/18 |

approach to measure it. They define it as experience of deprivation and take it to constitute poor health, lack of education, inadequate living standard, lack of income, disempowerment, poor quality of work, and threat from violence. Under this formulation, lack of income is one of the many constituents of poverty.

The United Nations Development Program (UNDP) [2] has adopted a multidimensional poverty index (MPI) as per the details provided in Table P.4. This includes measures of health, education, and living standard—each making up one-third of the index. The yes/no type of information, with a score of 1 and 0 for yes and no, respectively, is compiled for the households but applied to each individual of the household. For example, health component has nutrition measured by "at least one member of the household is malnourished" and "one or more children have died." If one of these occurs in a household, all members of the household get the score of 1. For this, information on household size is also recorded.

Note that living conditions have six indicators, and all get the same weight. MPI is the weighted average of the number of the "yes" answers. This index can be considered as an inverse of what **human development index (HDI)** measures. The higher the MPI, the greater the poverty. The UNDP divides it also into headcount ratio and intensity of poverty. For details, see their latest HDI report.

1. OPHI. *Policy—A Multidimensional Approach*. http://www.ophi.org.uk/policy/multidimensional-poverty-index
2. UNDP. *Human Development Reports*. http://hdr.undp.org/en/2014-report

## power (statistical) and power analysis

The following discussion on power presumes that you are familiar with the terms **Type I error**, **Type II error**, and **level of significance**. If not, you may like to review these topics.

The complementary of the probability of Type II error is called statistical power and is denoted by $(1 - \beta)$. Thus, the power of a statistical test is the probability of correctly rejecting a null hypothesis $H_0$ when it is false. In other words, this is the proportion of times that repeated samples of similar nature would give *P*-values less than $\alpha$ when the effect specified by the alternative hypothesis $H_1$ is present. Thus, this is the ability of a test to detect the specified difference or any other measure of effect. The power of a test is high if it is able to detect a small difference and can easily reject $H_0$. Suppose the mean peak exploratory flow rate (PEFR) in workers in a tire manufacturing industry is 296 L/min and that in workers in a paint varnish industry is 307 L/min so that the mean difference is 11 L/min. This difference seems small relative to the PEFR values. A test with high power is needed to detect this difference when present and to call it statistically significant. A test with low power will not be able to reject the $H_0$ of equality in this case and will lead to the conclusion that the difference is likely to have arisen by chance in the samples studied. Power measures the degree of assurance that the specified difference will not be missed even when clouded by high variability in the measurements.

Statistical power becomes an especially important consideration when the investigator does not want to miss a *specified* difference. For example, an antihypertensive drug may be considered useful if it reduces diastolic blood pressure (BP) in certain type of cases by an average of at least 5 mmHg after use for, say, 1 week. A sufficiently powerful statistical test would be needed to detect this difference with high probability. Thus, the magnitude of $(1 - \beta)$ is an important consideration in this setup. However, one would like that the minimum **medically important difference** (5 mmHg in this case) is chosen on some objective basis. The choice may not be easy in some situations since this is determined on *clinical* considerations based on benefit to the subjects. Do not confuse it with the actual effect size, which could be more or less.

The number of subjects in the study that determines the standard error is the most important consideration for power when all sources of bias and uncertainties, such as lack of knowledge and inadequate design, are in control. The best approach to achieve good power for detecting a minimum medically relevant difference is to increase the number of subjects in the study. Formulas are available that can give this number for different settings (see the **sample size** formulas in this volume). Power calculation depends on whether the characteristic under assessment is quantitative or qualitative, the form of its statistical distribution in the target population, the variance across subjects, the minimum difference between groups that can be considered medically relevant, and the chosen level of significance. A power of 0.80 or 0.90 seems to have become the norm for medical studies. For analogy with statistical significance, it is customary to call successful detection of a medically relevant difference as **medical significance** of results.

There is a trade-off between Type I and Type II errors in all statistical tests of hypothesis setups. In the court analogy also, which is popularly quoted in this context, eagerness to convict a guilty person has a direct bearing on the risk of an innocent being convicted. A Type II error (not being able to punish a guilty person because of lack of evidence) is preferred over a Type I error in a court to uphold civil rights if the evidence is not sufficiently convincing. Thus, power has a kind of direct relationship with the **level of significance**. The lower the level of significance, the lower the power. Using 1% level

of significance instead of the conventional 5% will make it difficult to reject a false null, and the power will be reduced. The relationship between significance level α and power (1 – β) can be obtained in terms of a receiver operating characteristic (ROC) curve similar to the one between sensitivity and (1 – specificity). However, utility of such ROC is limited since α is mostly fixed in advance.

Also, the power is low if the variance is high. You can intuitively realize that when values are highly variable from subject to subject, the difference between two or more groups will be difficult to detect. In addition, irrespective of α and variance, a smaller difference is difficult to detect, and the corresponding power will be low. The larger the minimum medically relevant difference, the greater the power. For example, in the case of the difference between capillary (by stick) and venous random blood sugar, the difference of possibly as much as 20% could be tolerated. In this case, medically important difference is large. The basic statistical consideration in determining the power is the sample size. An adequately large sample can override the limitation of low level of significance, high variance, and small medically relevant difference for detection.

In the case of large *n*, when **Gaussian conditions** prevail, it can be shown for one-tailed test that

$$\text{power} = P(Z \geq z_\alpha \text{ when the specified medically important effect is present}),$$

where $Z$ is the criterion under the null, assumed Gaussian here, and $z_\alpha$ is the value of $z$ corresponding to the α level of significance. The sample size $n$ would occur on the right side of the above equation. This can be solved to obtain $n$ when everything else is specified. However, the criterion may not be $Z$ and would depend on the kind of data you have. If the power is to be 99% for detecting a specified difference, obviously a larger sample size is required than for a power of 80%. Power and $n$ have a direct relationship for any fixed level of significance. The following example illustrates the power calculation in a simple situation.

The ratio of heart rate and the systolic BP is called the shock index. Suppose you measure this for 100 patients of ST-segment elevation myocardial infarction at the time of admission to test the null hypothesis that the mean is $\mu_0 = 0.70$ and the alternative is $\mu_0 > 0.70$. If the sample mean is 0.73 and $\sigma = 0.16$, what is the power of the study to find statistical significance when actually the mean is $\mu_1 = 0.75$ and the level of significance is 5% for one-tailed test? For many tests, such a calculation is difficult, but it is easy in this case since the sample size is large and the mean can be assumed to follow a Gaussian pattern even when the original values do not. Statistical significance in this case will be achieved when $Z \geq 1.645$ under the null hypothesis, where 1.645 is the Gaussian value for one-tailed probability 0.05. Under the null, $Z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq 1.645$ gives $\bar{x} \geq 1.645 \times 0.16/10 + 0.70$, i.e., $\bar{x} \geq 0.7263$. Power is the probability of this when mean $\mu_1 = 0.75$. That is,

$$\begin{aligned}
\text{power} &= P(\bar{x} \geq 0.7263 \text{ when the actual mean is } 0.75, \\
&\quad \sigma = 0.16, \text{ and } n = 100) \\
&= P\left(\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \geq \frac{0.7263 - 0.75}{\frac{0.16}{10}}\right) \\
&= P(Z \geq -1.48) \\
&= 0.93.
\end{aligned}$$

In this example, there is a 93% chance that the null of $\mu_0 = 0.70$ will be rejected when the actual mean is 0.75. In practice, we rarely calculate power; we calculate $n$ for the desired power so that the study can be planned accordingly. This exercise is done at the planning stage before the data are collected and the required sample size is mentioned in the protocol on the basis of the power calculations. For this, see the following and the topic **sample size**.

### *Power Analysis*

Good researchers determine the sample size on the basis of power required to detect a specified minimum difference, and try to do their study on the size so determined. Power is specific for the difference or effect size we wish to detect. Because of some exigencies, sometimes it becomes difficult to study as many subjects as calculated. This reduces the power. Sometimes a study is done on specific $n$ without worrying about power.

There is a considerable overlap in the literature about the term **power analysis**. It is used for the calculation of power for a given $n$ as well as for the calculation of $n$ for a specified power. Both calculations require prior specifications of the effect size to be detected based on literature or previous experience. The real utility of power analysis is in designing a study and not so much in interpreting nonsignificance of the result once obtained. Nonetheless, midcourse power calculation (**interim analysis**) can be done on the basis of the observed data and can tell you how much more $n$ you need. Post-hoc calculation can tell you what might have gone wrong. Use this experience for planning a better study next time. Post-hoc power can also be used to assess the ability of the study to discern a false hypothesis.

There is another exception, though. If the study throws up very different values of the parameter estimates such as the proportion $p$ and the standard deviation (SD) $s$ than assumed earlier for calculation of the sample size, and the effect size is found statistically not significant, recalculation of power with new $p$ or SD is justified and can give you better leads about the strength of your result. This can lead to **sample size re-estimation**. See also **adaptive clinical trials**.



## power transformation, *see also* Box–Cox power transformation

A number of statistical analysis methods for testing of hypothesis, such as the Student *t*-test, regression, and analysis of variance (ANOVA), require that the data have a Gaussian distribution. This distribution also helps in obtaining the correct confidence intervals easily. When the data are not distributed in a Gaussian manner, specific features of non-Gaussianity are explored, and appropriate remedial actions are taken. Data transformation is one of those remedial actions that may help to make the data Gaussian. An understanding of the concept of transformation may prepare you better to work with non-Gaussian data.

Transforming data means performing the same mathematical operation on each value in the original data. An example from daily life is converting temperature from degrees Celsius to degrees Fahrenheit. This transformation is called a linear transformation because the original data are simply multiplied or divided by a particular coefficient, or a constant is subtracted or added. But linear transformation does not change the shape of the data distribution, and it remains as non-Gaussian as before the transformation. Thus, nonlinear transformations are examined for achieving Gaussianity.

**TABLE P.5**
**Common Power Transformations**

| Power | –2 | –1 | –0.5 | 0 | 0.5 | 1 | 2 |
|---|---|---|---|---|---|---|---|
| Transformation | $1/y^2$ | $1/y$ | $1/\sqrt{y}$ | $\log(y)$ | $\sqrt{y}$ | $y$ | $y^2$ |

These transformations can also stabilize the variance in case that is needed. In many situations, both non-Gaussianity and variance heterogeneity go hand in hand. One set of these transformations is $y^{\lambda}$, where the $\lambda$ value is the power to which all the values should be raised. This is the reason that it is called power transformation. The value of $\lambda$ varies from one situation to another and depends on the nature of departure from Gaussianity. For example, $\lambda < 1$ is for correcting right skewness and $\lambda > 1$ is for correcting left skewness. The higher the skewness, the greater the power of $\lambda$. For aberrations other than skewness, such as sharp or flat peakedness, this transformation may not work. Also note that these apply only to nonnegative values. If there are negative values in your data, add slightly more than the maximum negative value to all the values so that none is negative.

See Table P.5 for some common transformations. The user may explore the data using several of these and other transformations and then select the most appropriate one.

Box and Cox [1] proposed a family of slightly different transformations that could correct non-Gaussianity in a variety of situations. These are given by

$$\text{Box–Cox power transformation:}$$
$$W(\lambda) = (y^{\lambda} - 1)/\lambda, \text{ for } \lambda \neq 0; \quad \text{and} \quad \ln y, \text{ for } \lambda = 0;$$

where $W(\lambda)$ is a continuous function of $y$. The value of $\lambda$ is chosen in a manner in which the transformed values have Gaussian distribution. A modification of the method of maximum likelihood is used for estimating the value of $\lambda$ for any given dataset. See the topic **Box–Cox power transformation** and Sakia [2] for further details.

Using a power transformation is not in itself a guarantee for achieving Gaussianity. You must check Gaussianity by using one or more of the methods listed under the topic **Gaussianity (how to check)** after the data are transformed.


1. Box GEP, Cox DR. An analysis of transformations. *J Royal Stat Soc, Series B* 1964;26(2):211–52. http://www.jstor.org/stable/2984418
2. Sakia RM. The Box–Cox transformation technique: A review. *The Statistician* 1992;41:169–78. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.469.7176&rep=rep1&type=pdf


## P-P plot, *see* proportion-by-probability (P-P) plot

## PPS sampling, *see* probability proportional to size (PPS) sampling

## pragmatic trials

These are those trials that give high importance to the practical feasibility of the regimen. A regimen may have 90% efficacy, but what is it worth if it is extremely difficult to implement? Effectiveness under practical conditions has brought pragmatic trials into focus. The patients recruited for this kind of trial are not homogeneous as in a regular clinical trial but reflect variations that occur in real clinical practice. Strategies such as **randomization** and **control** are also not structured in this trial. Instead, patients receiving existing regimen or not receiving any treatment serve as control. Because of a large number of intervening factors in this setup, the interpretation could be difficult. Statistically, the standard deviation could be relatively large. For details of pragmatic trials, see Roland and Torgerson [1].

**Clinical trials** are generally done in ideal conditions that do not exist in practice. The subjects are carefully chosen with strict inclusion and exclusion criteria, administration is done in standard conditions, efforts are made for full compliance, patients get full attention, the results are adjusted for dropouts and other missing observations, and the response is carefully assessed by experts. These steps help to draw a causal inference and establish the efficacy of the regimen under trials. However, the actual performance of the regimen in practice may differ. Efficacy of a treatment is what is achieved in a trial that simulates optimal conditions, and effectiveness is what is achieved in practical conditions when the treatment is actually prescribed. For clarity, the latter is sometimes called use-effectiveness. Effectiveness could be lower than efficacy because of lack of compliance of the regimen due to cost or inconvenience, inadequate care, nonavailability of the drugs, etc. These deficiencies do not occur in a trial. Experience suggests that nearly three-fourths of the patients, in general practice, do not adhere to or persist with prescriptions. Thus, patients and maneuvers adopted during a trial do not translate their results for patients at large. Generally, such external validity of the trial results is not high. But clinical trials do establish the potency of a regimen to effect a change. Effectiveness, on the other hand, is a suitable indicator to decide whether or not to adopt that regimen in practice, or what to expect. The concept of effectiveness in contrast to efficacy has given rise to pragmatic trials.

In the absence of blinding and placebos in a pragmatic trial, the results could be biased because of the **Hawthorne effect**. The expectation of participants could be favorable or unfavorable, and that will determine the actual bias. If patients are allowed to choose a treatment as can occur in practice, a further bias may creep in. Causal inference, which says that the effect is due to certain regimen, suffers. Thus, a better approach would be to do a pragmatic trial for assessing usefulness in a real-life situation *after* efficacy in ideal conditions has been established by regular clinical trials.


1. Roland M, Torgerson DJ. Understanding controlled trials: What are pragmatic trials? *BMJ* 1998;316:285. http://www.bmj.com/content/316/7127/285.short


## precision, *see also* reliability

Imagine you are shooting a gun at a target and that the aim of your gun causes shots to be displaced to one side or the other from the bull's eye. As your shooting proceeds, the shots form a group. The dispersion of this group is the precision. How close the mean of these shots is to the bull's eye is the validity, and not the precision.

It is easy to translate this into the realm of clinical measurements. When repeated measurements of the same characteristic under same conditions give the same value, you know that it is precise. Precision is the same as repeatability or reliability. Measurement of body temperature is quite precise, but that cannot be said about alkaline phosphatase level. Even if you split the sample and analyze separately in the same laboratory in a blind manner, the phosphatase level may differ. This measurement is not as much precise.

Statistically, the term *precision* is used for estimates we generate on the basis of sample values. High precision of estimates (by which

it is meant that the estimates from sample to sample are more tightly grouped) mainly results from a good sample size. When the sample size is large, summaries such as mean and proportion in different samples from the same population would be close to one another. You can intuitively imagine that the means based on two samples of size 100 are likely to be close to one another than the means based on a sample size of 10 each. In clinical trials, the underlying theme with regard to sample size considerations is precision. Imprecision of an estimated effect such as a mean, proportion, and difference is a consequence of measurement error, person-to-person variability, etc. These are factors that are mostly beyond human control; however, the sample size is within our control. By specifying quantitatively the precision of measurement needed in an experiment, the investigator is implicitly outlining the sample size.

Population parameters are seldom known and cannot be calculated (if they were known, there would be no need of a sample). They are almost invariably *estimated* from the corresponding values in the sample. The **standard error** (SE) of these estimates is the inverse measure of precision. This measures sampling error. You may have noted that $\sqrt{n}$ appears in the denominator of most SEs. The higher the *n,* the lower the SE and the higher the precision. This underscores the importance of sample size and explains why statisticians are so particular about the sample size. Precision frequently defines the credibility of the research you do and report.

However, caution is needed in interpretation of the SEs, particularly for proportions. For example, consider $\mathrm{SE}(p) = \sqrt{\frac{\pi(1-\pi)}{n}}$. You may like to check that SE($p$) is maximum when $\pi = 0.5$ and smaller when $\pi$ is either small or large. It would appear that the precision is lowest at $\pi = 0.5$. But its interpretation requires additional care. For $n = 100$ and $\pi = 0.5$, SE($p$) = 0.05, but for same $n$ and $\pi = 0.25$, SE($p$) = 0.043. In absolute terms, the SE in the second case is less than what it is in the first case, but, in relative terms, the second SE is almost one-fifth of $\pi$ while the first is one-tenth of $\pi$. Thus, the precision in the second case is lower in a relative sense. These are referred to as the *absolute precision* and *relative precision*, respectively, of *p.* The difference is more apparent while considering the confidence intervals (CIs). A CI from 0.02 to 0.12 around 0.07 may have smaller width than 0.42 to 0.58 around 0.50; however, the former provides more imprecise information about the value of $\pi$, whereas the latter is not so imprecise.

Distinction between precision and validity is important just as between reliability and validity. In terms of shooting, you can be far away from the target, yet very precise. Shooting is precise if all shots are very close to each other even if all of them are far away from the target. When this happens, you know how to adjust your angle of shooting so that the shots hit the target. This has a direct relation with statistical precision. The estimate may be very precise but highly biased. Bias in statistics implies that the average is far away from what you aimed. Bias can be corrected by adding or subtracting a constant from each value—precision cannot be corrected this way. For medical studies, bias is corrected by framing correct questions, by standardized laboratory results, by randomization or matching, etc.

## predicted value and prediction interval in regression

In 2012, statistician Nate Silver created a furor by predicting the correct results of all 50 states for the presidential election in the United States [1]. Nobody believed that this could be done. Such is the power of prediction. If you can predict the years of life remaining for a patient of cancer, and your prediction turns out correct for many such patients, you will be idolized. However, medical predictions are fraught with enormous uncertainties, surpassed only by attitude and behaviors among human endeavors. The mortals among us use empiricism to predict the future. They are guided by what has actually happened in the past or what is happening in the present and what can possibly affect the outcome. Many researchers take recourse to statistical models for prediction. **Regression** models are just about the most common statistical tools for prediction of medical outcomes. The following discussion is restricted to prediction based on these models.

A regression model can be written as

$$\hat{y} = f(x_1, x_2, \ldots, x_K),$$

where function $f$ is a rule that relates the values of ($x_1$, $x_2$,…, $x_K$) to a variable $y$, and $\hat{y}$ is the predicted value. In case the purpose of the regression is prediction, the $x$ variables are called the predictors of $y$. Their number is $K$ in this model. The function $f$ is chosen in a manner that specific values of the $x$'s uniquely predict the value of $y$. Ordinarily, each value of $y$ must be independent of other values of $y$ for valid regression; for example, they cannot belong to different sites of the same body.

The relationship $\hat{y} = f(x_1, x_2, \ldots, x_K)$ may take any of several different forms. One that is relatively simple to comprehend and most commonly studied is the **linear** form. For $K$ predictors, this is expressed as

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_Kx_K.$$

Any of the $x$ variables can be a function of another $x$ variable. For example, $x_3$ could be $x_2^2$, or $x_4$ could be $\ln(x_1)$—thus, this model includes curvilinear regressions. The regression coefficients ($b_1$, $b_2$, …, $b_K$) are obtained by the method of **least squares** or by the method of **maximum likelihood**. When these coefficients are known, the values of the $x$'s can be substituted in the model to predict the value of $y$. Remember though that prediction models are not necessarily the same as explanatory models. For prediction, choice of the $x$ variables is not important—one set of $x$'s may be as good as another set of $x$'s for predicting $y$, but in explanatory models, $x$'s must be those that are biologically or functionally related to $y$.

One of the important functions of regression is to provide an estimate or to predict the value of $y$ for a given set of $x$'s. There is a fine statistical distinction between an estimated value and a predicted value. The former term is generally used for the *mean* of $y$ for a group of individuals with a specific value of ($x_1$, $x_2$, …, $x_K$), and the latter is used for the value of $y$ for a single individual. The former has less variance than the latter, although both have the same value. Because the estimated and predicted values are the same, these terms are sometimes interchangeably used in the literature.

The predicted value of $y$ will be the best that the regression model can predict, but it may still be far from reality. Correct prediction depends on the proper choice of predictors. These should include all those that are associated with the outcome $y$, although they may or may not be causally associated with $y$ as this is not important for prediction models. The difficulty is that for many medical outcomes, all associated factors are rarely known—nobody knows, and we work within the limitation of our knowledge. Such epistemic gaps are the most pronounced stumbling blocks for correct prediction. In addition, a linear model is not necessarily the most appropriate even when it is extended to include curvilinear forms. Most medical phenomena are not as simple, and many times nonlinear models provide better prediction. Linear models were

a favorite because of the ease of computation and understanding, but now computation is not a restrictive feature—understanding remains a block that is still to be overcome by many researchers.

Consider the problem of predicting body surface area (BSA) (in $cm^2$) by weight (in g) of well-proportioned children using simple linear regression [2]. If the regression is BSA = 1321 + 0.3433(Wt), it tells what BSA to expect in children of different weights. If this regression is good, the predicted BSA of a child of 20 kg is (1321 + 0.3433 × 20,000 =) 8187 $cm^2$. The actual value would most likely differ from this predicted value. If the actual BSA of a child of 20 kg is found to be 8093 $cm^2$, then the difference from the expected is –94 $cm^2$, and if another child of 20 kg has BSA 8214 $cm^2$, then the difference is +27 $cm^2$. Thus, prediction is never perfect. Statistically, the error in the long run depends on the standard error (SE) of the estimate.

The SE of $\hat{y}$ is different when it is for the mean of $y$, earlier called as the *estimated value* of $y$, and when it is for individual values of $\hat{y}$, earlier called as the *predicted value* of $y$. Prediction of individual values has a much larger SE. These SEs have complex forms for multiple regression, but they can be easily stated for a simple linear regression. Under certain general conditions, for a simple linear regression of $y$ on $x$,

$$\text{SE (predicted individual value of } y_x) = \sqrt{\text{MSE}\left[1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\Sigma(x-\bar{x})^2}\right]},$$

and

$$\text{SE (predicted mean value of } y_x) = \sqrt{\text{MSE}\left[\frac{1}{n}+\frac{(x-\bar{x})^2}{\Sigma(x-\bar{x})^2}\right]},$$

where $y_x$ is the value of $y$ at a given $x$. MSE is the mean square error and is always generated by the regression package. If the dependent of interest is duration of analgesia ($y$) in min induced by different doses of a drug ($x$) in μg, and $n$ = 15, MSE = 6.50, $\bar{x} = 11.2$ μg, and $\Sigma(x-\bar{x})^2 = 18.07$, then for $x$ = 8 μg, SE(predicted individual value of $y$ at $x$ = 8) $= \sqrt{[6.50\{1+1/15+(8-11.2)^2/18.07\}]} = 3.26$ min by the above equation. The SE of the estimated mean is 2.03. The SE of the predicted value is much higher than the SE of the estimated value of $y$. Prediction of the individual value of $y$ is always less precise than the estimation of the mean of $y$. This has direct bearing on the **prediction interval**, which is the same as the confidence interval for the predicted value.


1. Champkin, J. Timeline of statistics pull out. *Significance* Dec 2013;10;23–6. http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2013.00707.x/abstract
2. Current JD. A linear equation for estimating the body surface area in infants and children. *Internet J Anesthesiol* 1997;2(2). https://ispub.com/IJA/2/2/10302


## prediction interval, *see* predicted value and prediction interval

## predictive analytics, *see* data analytics

## predictive models, *see* explanatory and predictive models

## predictive validity, *see* predictivities (of medical tests)

## predictivities (of medical tests)

Predictivities of medical tests are their ability to correctly identify presence or absence of a disease or any health condition. Positive predictivity refers to the ability to detect presence of disease, and negative predictivity is the ability to detect absence of disease. These are also called predictive value of a positive test and of a negative test, respectively. These are kind of the opposite of what **sensitivity and specificity** are. Sensitivity and specificity are also indicators of the validity of a test, but they do not measure the diagnostic value of the test. Sensitivity and specificity are based on the correct identification of the cases when it is known that the groups of subjects have disease or do not have disease, i.e., the disease status is already known. Diagnostic value of a test is obtained in terms of predictivities.

### *Positive and Negative Predictivity*

The actual problem in practice is detecting the presence or absence of a suspected disease by using a test. The diagnostic value of a test is measured by the probability of actual presence of disease among those who are test positives and the probability of actual absence of disease among those who are test negatives. Besides the names just mentioned, these can also be understood as posttest probabilities and measure the utility of a test in correctly identifying or correctly excluding the disease. Positive predictivity can also be understood to measure how good the test is as a marker of the disease. In terms of notations,

$$\text{positive predictivity } P(+) \text{ or } P(D+\,|\,T+) = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

and

$$\text{negative predictivity } P(-) \text{ or } P(D-\,|\,T-) = \frac{\text{TN}}{\text{TN}+\text{FN}},$$

where $D$+ is for presence of disease, $D$– for its absence, $T$+ is for positive test and $T$– for negative. TP, TN, FP, and FN are given as in Table P.6.

Suppose that 4700 patients coming to a cardiac clinic with some complaints had ECG done. Out of these, 425 had positive ECG and 275 negative ECG. All these were further investigated, and 416 of 425 ECG positives were found to have myocardial infarction (MI) (Table P.6), and 104 of 275 ECG negatives were found to have MI.

**TABLE P.6**
**Myocardial Infarction (MI) and ECG in Cases of Acute Chest Pain**

| | MI | | |
|---|---|---|---|
| **ECG** | **Present** | **Absent** | **Total** |
| Positive | 416 (TP) | 9 (FP) | **425** |
| Negative | 104 (FN) | 171 (TN) | **275** |
| **Total** | **520** | **180** | **700** |

Then, the predictivity of a positive ECG for the presence of MI in such cases is

$$P(+) = \frac{416}{425} \times 100 = 98\%,$$

and the predictivity of a negative ECG for the absence of MI is

$$P(-) = \frac{171}{275} \times 100 = 62\%.$$

The diagnostic value of a test is not reflected by sensitivity and specificity but is reflected by predictivities. These also are indicators of the validity of a test. If these 700 subjects in this example are considered good representatives of such cases, the specificity is high at 95% (171/180), yet the test is poor in excluding MI—only 62% are correctly ruled out. This happens in this case because many subjects with negative ECG have MI (i.e., 104 out of 275). What it tells us is that the ECG can be safely used to detect the presence but not to detect the absence of MI, as far as this example is concerned.

A screening test should have high negative predictivity, whereas a confirmatory test should have high positive predictivity. When both are equally important, use

$$\text{predictive validity} = \frac{\text{TP} + \text{TN}}{\text{total subjects tested}}.$$

**Predictive validity** combines the two predictivities assuming that both are equally important. If they are not, an index can be devised that gives differential weight as needed. This equation is similar to

$$\text{inherent validity of a test} = (\text{TP} + \text{TN})/n,$$

but the interpretation is different since the denominator in the latter equation is the total subjects with and without disease and in the former equation is the total subjects who are test negative and test positive.

### *Predictivity and Prevalence*

You can see that the concepts of sensitivity and specificity work in an inverse direction. They move from disease to the test. The predictivities do provide assessment in the right direction, but they are severely affected by the prevalence of disease among those tested. This is illustrated in the example below. The advantage of sensitivity and specificity is that they are absolute and do not depend on prevalence.

Out of the 700 tested, let the number with MI in Table P.6 be changed to 300 rather than 520 (the prevalence is now 43%, i.e., 300 × 100/700). If the sensitivity and specificity remain as before at 80% and 95%, respectively, then the different numbers would be as shown in Table P.7. Now,

$$\text{positive predictivity} = \frac{240}{260} \times 100 = 92\% \text{ and negative predictivity}$$

$$= \frac{380}{440} \times 100 = 86\%.$$

These values are very different from the ones obtained earlier because of the entirely different prevalence rate. The previous prevalence rate was 520/700 = 0.74 or 74%, and now it is 43%.

**TABLE P.7**
**Increased Prevalence of Disease Than in Table P.6**

| | *D*+ | *D*– | Total |
|---|---|---|---|
| *T*+ | 240 | 20 | **260** |
| *T*– | 60 | 380 | **440** |
| **Total** | **300** | **400** | **700** |

When the prevalence rate is available, denoted by $P(D+)$, we can calculate predictivities from sensitivity–specificity as follows. From **Bayes rule**, positive predictivity is given by

$$P(+) \text{ or } P(D+ \mid T+) = \frac{P(T+ \mid D+)P(D+)}{P(T+ \mid D+)P(D+) + P(T+ \mid D-)P(D-)}$$
$$= \frac{S(+) * p}{S(+) * p + [1 - S(-)] * (1 - p)},$$

where $p$ is the prevalence rate per unit, and $S(+)$ is the sensitivity and $S(-)$ the specificity. When sensitivity and specificity are constants, it is easy to see from the above equation that positive predictivity increases as prevalence increases. Similarly, it can also be shown for negative predictivity that

$$P(-) \text{ or } P(D- \mid T-) = \frac{S(-) * (1 - p)}{S(-) * (1 - p) + [1 - S(+)] * p}.$$

As prevalence increases, the negative predictivity decreases. Dependence of predictivity on prevalence arises from putting the information in its proper context. A patient with high fever and shivering might be diagnosed with influenza in Europe but malaria in West Africa [1].

The predictivities for some specific values of sensitivity and specificity and for different prevalences are shown in Table P.8. As

**TABLE P.8**
**Predictivities for Some Specific Values of Sensitivity, Specificity, and Prevalence**

| Sensitivity *S*(+) | Specificity *S*(–) | Prevalence | Positive Predictivity *P*(+) (%) | Negative Predictivity *P*(–) (%) |
|---|---|---|---|---|
| 0.20 | 0.20 | 0.10 | 3 | 69 |
| | | 0.50 | 20 | 20 |
| | | 0.90 | 69 | 3 |
| 0.20 | 0.90 | 0.10 | 18 | 91 |
| | | 0.50 | 67 | 53 |
| | | 0.90 | 95 | 11 |
| 0.90 | 0.20 | 0.10 | 11 | 95 |
| | | 0.50 | 53 | 67 |
| | | 0.90 | 91 | 18 |
| 0.90 | 0.90 | 0.10 | 50 | 99 |
| | | 0.50 | 90 | 90 |
| | | 0.90 | 99 | 50 |

P

the prevalence increases, the positive predictivity also increases, and this increase is more pronounced when the specificity is low. Higher prevalence leads to less negative predictivity, more so when sensitivity is low.

The two equations just mentioned also express the relationship between sensitivity–specificity and predictivities. If prevalence is known, predictivities can be obtained by using sensitivity and specificity. Herein lies the importance of these two somewhat reversed indices. Based on confirmed cases, sensitivity and specificity are easy to obtain. Use them to calculate diagnostically important positive and negative predictivities with the help of these equations. Direct calculation of predictivities requires follow-up studies to find the confirmed cases with and without disease among those that have shown positive and negative tests. Such follow-up studies can be expensive and can be avoided for predictivities when prevalence is known and sensitivity–specificity easily obtained.

1. Chatfield C. Confessions of a pragmatic statistician. *Statistician* 2002;51 (Part-1):1–20. http://www.jstor.org/stable/3650386

## predictors

A predictor variable is the one that is useful for predicting the value of the response variable. It may not fully predict the value but helps at least partially when chosen with care. Obviously, this would be used in a setup where response is not known or is difficult to observe or measure. Otherwise, why would one want to have a predictor? Predictors need not be biologically or causally related to the response, and they may have an indirect relationship. The stronger this relationship is, the better the predictor becomes. Among statistical methods, predictors are used predominantly in a **regression** setup that exploits the relationship between these so-called predictors and the response. In this setup, variables with high correlation with the response would be good predictors.

The general form of regression is $\hat{y} = f(x_1, x_2, \ldots x_K)$, which expresses a variable $y$ in terms of another set of variables $x_1, x_2, \ldots, x_K$. Whether or not this relationship is adequate is another question. When the objective of the regression is prediction of $y$ based on observed values of $x's$ (see **explanatory and predictive models**), then $\hat{y}$ is the predicted value of $y$ and $x's$ are the predictors. In case of multiple linear regression, this takes the form $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_Kx_K$, where $x's$ are the $K$ predictors; and in case of logistic regression, this becomes $\hat{\lambda} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_Kx_K$, where $\hat{\lambda} = \ln\frac{p}{1-p}$ and $p$ is the proportion of subjects with positive response. The response variable in this case is $\hat{\lambda}$, which is called the logit of $p$. In both ordinary and logistic regressions, the predictors could be dichotomous, polytomous, continuous, standardized, etc., and the interpretation of the regression changes accordingly. The details are as follows. Also see the topic **regression coefficients**.

### *Types of Predictors*

Continuous predictors are those that can have any value within the specified range. Age, blood pressure, bilirubin level, and creatinine level are the examples. In the case of linear regression, which we are considering in this section, the effect of continuous predictors on the response is that it increases by $b_k$ per unit increase in the value of $x_k$. Thus, the predicted value of $y$ depends on whether $b_k$ is small or large, positive or negative. This is adjusted for other predictors in the model. If you want to compare the effect of two predictors to find which of them has more influence on the response, it is necessary to standardize the predictors. A predictor that has a value in hundreds such as cholesterol level may have small $b_k$ but will still have large effect compared with, say, creatinine level, which is generally less than 1 when measured in mg/dL units. Thus, it also depends on the units. Weight (of, say, an infant) measured in kg will have a different meaning compared to weight in g.

Response will be in logit units in case of logistic regression, and the effect of a continuous predictor has to be assessed with a degree of caution. If a predictor is continuous such as age in years, a logistic coefficient of 0.15 would imply odds ratio (OR) = $e^{0.15}$ = 1.16, indicating that each year of increase in age increases the OR by a factor of 1.16. A 10-year increase in age would increase the OR by a factor of $(1.16)^{10}$ = 4.41. You can say that a 10-year increase in age increases the OR by 341%. In this case, the OR may look only slightly more than 1.0, but it can translate into an enormous effect. The argument can be turned on its head for another predictor such as waist–hip ratio that can rarely increase by 1 unit.

Interpretation of dichotomous and polytomous predictors is not so straight. Dichotomous predictors are those that have just two categories such as male or female, healthy or sick, prescription followed or not followed, discharged or died, etc. These categories are customarily coded as 0 and 1. Consider a simple example of linear regression for predicting systolic level of blood pressure (sysBP) in healthy adults with age and sex as the predictors. This regression could be sysBP = 112 + ⅓*Age – 4*Sex, where Sex is coded as 1 for males and 0 for females. Sex is a dichotomous predictor in this example. When these codes are substituted in the equation, you get sysBP = 108 + ⅓*Age for males and sysBP = 112 + ⅓*Age for females. Coding in this case gives two equations—one for males and the other for females. The case of polytomous predictors is more complex. First, they could be nominal or ordinal, and these two types require separate treatment. For nominal predictors, **indicator variables** are used as predictors as explained for that topic. For ordinal predictors, see if you can convert them into linear or nonlinear **scores**. Whereas the interpretation of polytomous predictors is simple in case of ordinary regression, logistic regression requires some explanation.

Interpretation of logistic coefficients for predictors with three or more categories depends on the coding system adopted. If a particular predictor is polytomous ordinal such as disease severity categorized as none, mild, moderate, serious, or critical, many types of coding can be done. The simplest can be 0, 1, 2, 3, 4 when the number of categories is five. This kind of coding quantifies predictor values and works like a score. In this case, the corresponding $e^b$ is the factor by which OR multiplies when the category moves one up in terms of this score. If $e^b$ = 1.15, this means that the odds for category $x$ = 1 (mild disease) are 1.15 times the odds for category with $x$ = 0 (no disease), and the odds for category $x$ = 2 (moderate disease) are 1.15 times the odds for category $x$ = 1 and 1.15 × 1.15 = 1.32 times the odds for $x$ = 0 (no disease).

This interpretation of the logistic coefficient is valid for ordinal predictors. For a predictor in real nominal categories, suitable **contrasts** of interest should be defined. They are generally defined in terms of the difference of one group from one or more of the others. If the objective is to examine 5-year survival with site of malignancy, you can have one particular site, say lung, as the reference category and compare this with oral, prostate, esophagus, etc., by forming contrasts such as (lung – oral), (lung – prostate), etc. Each of these contrasts would appear in the predictor set of the logistic model. If you consider it more appropriate, you can have "prostate cancer" as the reference category. In this case, all other sites will be compared with this cancer. This requires coding accordingly. The advantage with these kinds of contrasts is that each contrast will have its own logistic coefficient, and the coefficient for one contrast can differ from that of another contrast. Thus, differential effects of the categories on the

response, if present, will emerge. It is possible to find out if one particular category, say category 2, is a significantly higher contributor to the response than category 3 or category 1. The serial coding 0, 1, 2, 3, 4 mentioned in the preceding paragraph does not have this feature.

For interpretability, it is necessary that each category of each predictor is explicitly defined. For example, the "others" category that includes leftovers as in a kitchen sink is not admissible. In cases where reference category is used for all comparisons, conclusions are more reliable when the number of subjects in the reference category is reasonably large.

## pretesting, *see* pilot study and pretesting

## prevalence and prevalence rates

In our context, prevalence is the presence of morbidity, and incidence is its fresh occurrence. Thus, prevalence is computed on the basis of the existing cases and incidence on the basis of the new cases. They can be obtained either by counting the subjects affected or by counting the episodes that have occurred. Prevalence is considered to measure the load on health care services, and incidence is considered to measure the risk of getting the disease. In its simplest form, prevalence is governed by incidence (how much new cases are added), duration of disease (due to which accumulation occurs), **case fatality** (how many people with the disease are dying), and recovery (what percentage of cases recover) (Figure P.12).

Depiction of prevalence in Figure P.12 is rather too simplistic. In fact, the prevalence increases when

- Incidence increases: This can happen either because people become more susceptible due to change in lifestyle, because of the increase in expectation of life that can increase the proportion of the vulnerable section, or because the causative agent such as bacteria has increased. If people become healthier and more resistant, the incidence will decrease.
- Duration of disease increases: See Figure P.13 where prevalence on day 7 is only 1 case when the duration of disease is short and is 9 when the duration of disease is long (for both the diseases in this figure, the onset [incidence] is the same). This will happen when better knowledge and technology are available to save lives—the patients remain sick for longer time but are not allowed to die.
- Recovery rate reduces: This is rare but can happen when the disease starts afflicting mostly old-age people who may have a slower rate of recovery. Fast recovery due to early detection and availability of better treatment regimens can decrease prevalence.
- Detection improves either because of better surveillance or because of the availability of more sensitive tools.
- Better reporting occurs due to more awareness.
- There is immigration of cases.

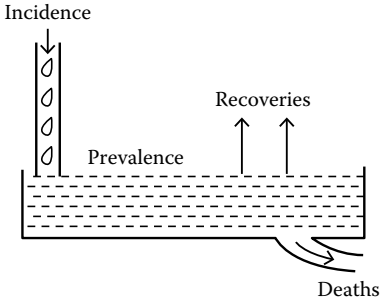



**FIGURE P.12** Accumulation of prevalence.

The last three indicate artificial increase and not the real increase. Cancer is the disease where most of these factors are operating almost all around the world.

By its nature, incidence is related to a period such as a week, a month, or a year, but prevalence is related to a point of time. Yet, there are concepts of point prevalence and period prevalence.

### *Point Prevalence*

Point prevalence is the number of cases existing at a specific point in time. Some cases may be preexisting, and some may have occurred on the day of inquiry. Thus, many texts define this as old and new cases. A survey to identify affected cases in a population generally takes weeks or months, but the count obtained is a point prevalence when the inquiry is with regard to the presence or absence of morbidity at the time of contact or a particular reference time point. This can be obtained by a cross-sectional survey. Point prevalence for the episodes would be the same as for the subjects because, at any point of time, a person cannot have two episodes of the same illness. In any case, point prevalence is generally obtained for chronic conditions rather than for acute illnesses.

For the purpose of comparison between groups, areas, diseases, etc., it is customary to calculate the prevalence rate (percent, per thousand, or per million persons) at a particular point in time. *It is actually a proportion but is conventionally called a rate.* It does not measure the speed of occurrence that a rate should. Persons counted

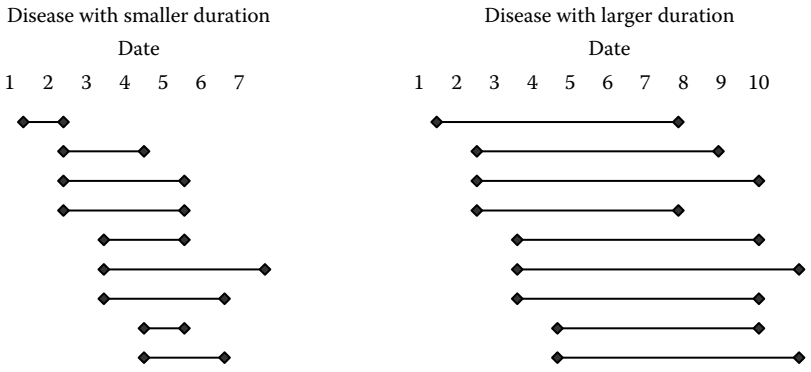



**FIGURE P.13** Effect of duration of morbidity on prevalence.

for the denominator are those who are exposed or at risk for the disease in question; others are excluded. For example, for smoking, children below the age of 12 years can be excluded. A prevalence rate can be calculated for a specific age, gender, occupation, etc. The prevalence rate estimates the probability of the presence of morbidity in a randomly selected person from that group. This has also been called a pretest probability in the context of **predictivities** of medical tests, although there prevalence is calculated among those suspected. If peptic ulcer is found in 5% of the cases of hypertension, this is the prevalence rate of peptic ulcer among cases with hypertension.

### *Period Prevalence*

For calculating period prevalence, the number of persons affected or the number of episodes of illness present during a specific period such as 1 week or 1 month in a defined population is counted. This includes the number of cases arising before but extending into or through the period as well as those arising during that period. The information sought from the respondent in this case is whether he/she is suffering from the disease at the time of inquiry or had suffered at any time during the last 1 week, 1 month, etc. For measuring period prevalence, generally only a short duration is considered, and it is mostly obtained for acute conditions.

### *Prevalence Rate Ratio*

If prevalence of a disease in males is 17% and in females 24%, it could be useful to say that the prevalence in females is nearly 1 ½ times of that in males. The ratio of prevalence in two mutually exclusive groups is called the prevalence rate ratio (PRR). Thus,

$$\text{PRR} = \frac{\text{prevalence rate in one group}}{\text{prevalence rate in another nonoveralapping group}}.$$

If $n_1$ is the number of subjects with diabetes and $a$ of them are also hypertensive, the prevalence of hypertension among diabetics is $a/n_1$. If $n_2$ is the number of subjects without diabetes and $b$ of them are hypertensives, the prevalence in this group is $b/n_2$. In this case,

$$\text{PRR} = \frac{a/n_1}{b/n_2}.$$

This is useful in measuring how commonly a condition is present in one group relative to the other.

## prevalence studies, *see* descriptive studies

## prevented fraction *see* attributable risk (AR) fraction

## primary data, *see* data sources (primary and secondary)

## primary sampling unit

In studies that involve a population of largesize, it is sometimes helpful to draw samples in stages (see **multistage random sampling**). If the subjects spread all over a state are the target, you may select a small number of districts or counties in the first stage; then some blocks, colonies, or hospitals in the second stage from the selected districts or counties; and finally the subjects from the selected colonies/hospitals. Thus, there are sampling units of various sizes. The large unit used in the first stage is called the primary sampling unit (PSU). For example, the National Health and Nutrition Examination Surveys in the United States use counties as the PSUs. From each selected PSU, they select segments, households, families, and individuals in stages [1].

In a study to find the prevalence of smoking in females of age 20 years and above in a particular state with, say, a million families, you may, for example, first select 4 counties by the random method, then 12 census blocks within each selected county, and 50 families within each selected block. All females of age 20+ years in the selected families could be the unit of inquiry, although the sampling units are counties, blocks, and families. Some families may have two or more units of inquiry and some none at all, but most may have just one. If there are many families with two or more eligible females, this can produce a **clustering effect**.

When the size of the PSUs is not large, i.e., when they generally contain a small number of subjects, then it is sometimes advisable that these units are not sampled further. All the elements in the selected primary units are then surveyed. This tends to increase the total number of subjects in the sample without a corresponding increase in the cost. Since many subjects in close proximity are included in this kind of sample, the travel time and the cost are saved. When this is done, it is convenient to understand a primary unit as a **cluster**. The units within each cluster are likely to be similar in some sense. The **design effect** arising from including clusters of units in the sample will depend on the size of the clusters and the homogeneity of units within clusters. The estimates and particularly their variance will be affected accordingly.

1. NHANES. *Task 1: Key Concepts About the NHANES Survey Design*. http://www.cdc.gov/nchs/tutorials/dietary/SurveyOrientation/SurveyDesign/Info1.htm

## primordial factors

According to conventional wisdom, a primordial factor is the one that is used at the very beginning of your study: its inclusion marks a beginning to that process. In our context, primordial factors are those underlying factors that start serial changes in health. The sequence generally is that primordial factors prepare a fertile ground for generation of risk factors, and these risk factors trigger the disease process when a suitable environment is available. Primordial factors have an important place in the etiology of disease; when they are controlled, the disease would not occur—this is called primordial prevention.

For most health conditions, social and environmental conditions are considered primordial. In the case of coronary diseases, obesity and hypertension are risk factors, but their origin may lie in improper diet and physical inactivity. They may have started to make changes at the beginning of life, perhaps even during pregnancy, because of imprudence, lack of awareness, traditions, attitudes, and practices. Thus, awareness could be the main primordial factor. Some of these may lie within—attitudes, personality traits, beliefs, and behaviors may be notable contributors. We can go further down and note that lack of awareness may be rooted in lack of education of parents, awareness in people around, universal lack of knowledge regarding the effect of these factors, and the like. For some cancers, primordial

factors may be environmental pollution, pesticides and chemicals in the food we eat, carcinogens in beauty products we regularly use, etc. There is a growing realization that control of risk factors is difficult unless primordial factors are targeted. For many diseases, these are not fully known, and the search for such factors has given rise to the science of primordialism in health that was earlier mostly restricted to philosophy.

Statistical methods that can investigate primordial factors may require complex modeling such as multilevel models because the effect of primordial factors is many times indirect and mediated through other cofactors that work at different levels. For example, environmental pollution is a macro variable, whereas the use of cosmetics is a micro variable. These can be studied together by multilevel models. Moreover, the statistical properties of most such variables would rarely be Gaussian, and methods such as generalized linear models with suitable link function may be needed to study them.

## principal components

If you are dealing with a really large number of variables that need to be considered together, a statistical method called principal components (PCs) can be used to reduce their number in many situations without losing much information. That can make them manageable and possibly also more meaningful for subsequent analysis. The aim of PC analysis is to reduce the dimensionality of the data.

It should be clear that PC analysis is used for multivariate data and not for univariate data. The original variables are transformed into new ones that are uncorrelated and that account for major proportions of the variance in the data. The new variables are the PCs and are defined as linear functions of the original variables. These are obtained in stages. For $K$ variables, the first PC may be $y_1 = a_0 + a_1x_1 + a_2x_2 + \ldots + a_Kx_K$. This will be the combination that can capture higher variation among $x$'s than any other linear combination. The second PC would be the one that is independent of the first (called *orthogonal*) and captures more of the remaining variation than any other combination. This can be written as $y_2 = b_0 + b_1x_1 + b_2x_2 + \ldots + b_Kx_K$. Similarly, other PCs are obtained. A complex method based on matrices and eigenvalues is used to obtain these PCs. Theoretically, one can obtain as many PCs as the number of variables, but generally only the first few are enough. If the first few PCs account for a large percentage of the variance of the observations (say above 80%), they can be used to simplify subsequent analyses and to display and summarize the data in a more parsimonious manner.

The loss of information in extracting just a few PCs in most cases would be small. But the PCs so derived are purely statistical and would not generally have any biological meaning. These are artificial variables and can make interpretation very difficult. They are sometimes "rotated" (see **varimax rotation**) to be able to attach meaning. PCs are commonly used for psychometric evaluations of questionnaires (see, e.g., Bian et al. [1]) but rarely for medical investigations. If you are looking for meaningful extracts in medical context, use the method of **factor analysis**. This also is based on PCs. That possibly is the most pronounced use of PCs in medical studies.

Among sparing uses of PCs in medical research, one is by Caine et al. [2] on the study of cognitive profile of prion disease. Their PC analysis showed a major axis of frontoparietal dysfunction that accounted for approximately half of the variance observed. This correlated strongly with volume reduction in frontal and parietal gray matter on magnetic resonance images. Tareque et al. [3] developed a wealth index on the basis of PC analysis of socioeconomic data to study the association of diabetes and hypertension with socioeconomic status.

A tutorial on PCs is given by Shlens [4] for those who want to know more.


1. Bian W, Li M, Wang Z, Wang X, Liu Y, Wu Y. Psychometric properties of the Chinese version of the Amblyopia and Strabismus Questionnaire (ASQE). *Health Qual Life Outcomes* 2015 Jun 12;13(1):81. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4465144/
2. Caine D, Tinelli RJ, Hyare H, De Vita E, Lowe J, Lukic A, Thompson A. The cognitive profile of prion disease: A prospective clinical and imaging study. *Ann Clin Transl Neurol* 2015 May;2(5):548–58. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4435708/
3. Tareque MI, Koshio A, Tiedt AD, Hasegawa T. Are the rates of hypertension and diabetes higher in people from lower socioeconomic status in Bangladesh? Results from a nationally representative survey. *PLoS One* 2015 May 27;10(5):e0127954. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4446365/
4. Shlens J, *A Tutorial on Principal Components Analysis*. Cornell University Library 2014. http://arxiv.org/abs/1404.1100, last accessed June 20, 2015.


## principles of experimentation, *see* experimentation (statistical principles of)

## prior probability and posterior probability

The probability of any event without availability of any particular information is called the **prior probability**, and the probability after that information is available is called the **posterior probability**. The latter obviously depends on the kind of information available to alter the prior probability.

When a patient comes with a certain presentation, several possible diagnoses may flash into mind. The clinician mentally works out the chances of each specific diagnostic category and assigns the patient to the most likely category. The diagnosis, thus, is essentially a probabilistic entity. Until sufficient information is available to objectively evaluate probability, personal probabilities can be used. **Personal probability** is a measure of one's belief in a statement. If you believe on the basis of experience or otherwise that among the patients who present with long-standing complaints of abdominal pain, vomiting, and constipation simultaneously, only 16% are cases of abdominal tuberculosis, your personal probability of abdominal tuberculosis given these three complaints (and nothing else) is 0.16. This is the prior probability. It alters as and when additional information on the patient becomes available such as the laboratory test results and x-ray findings.

Posterior probability is the chance of occurrence of an event when part of the information governing its occurrence is known; for example, high fever, rigors, splenomegaly, and presence of the malarial parasite in the blood are the stages that progressively confirm malaria. As the information increases, the diagnosis of malaria becomes firm, and the probability of absence or presence of the disease becomes concrete. This probability depends on what information is already available. The chance part is restricted to the uncovered information.

In an antecedent–outcome setup, if outcome is denoted by $O$ and antecedent state by $A$, the probability to be calculated is $P(O|A)$. This is conditioned on $A$ being present and is called the posttest probability with the same meaning as the posterior probability. $P(O)$ by itself without knowledge of the presence or absence of $A$ is the prior probability. The other prominent use of prior probability is in **discriminant analysis**—to run this analysis, we need to assign group probabilities upfront.

Prior and posterior probabilities are the backbone of **Bayesian inference**. See this topic for details. **Bayes rule** gives the method to convert posterior probability of one kind to another kind. For example, you can convert $P(A|O)$ to the more useful $P(O|A)$. The first is the probability of particular signs–symptoms in a disease, whereas the second is the probability of disease for given signs–symptoms. The first is how our books generally describe the disease, and the second is the way you use for establishing a diagnosis in clinics.

Brase and Hill [1] give a review of this topic.

1. Brase GL, Hill WT. Good fences make for good neighbours but bad science: A review of what improves Bayesian reasoning and why. *Front Psychol* 2015;31:340. http://www.ncbi.nlm.nih.gov/pubmed/25873904

## PRISMA Statement

PRISMA stands for preferred reporting items for systematic reviews and meta-analyses. It is an evidence-based minimum set of items for reporting in **systematic reviews** and **meta-analyses**. This is an effort to improve the conduct and report medical research so that more objective decisions can be made. The details are available at the PRISMA website [1]. A brief is given as follows.

You may be aware that diverse medical literature is regularly systematically reviewed to extract a common theme so that a more plausible conclusion can be reached regarding efficacy of a regimen or its side effects, its effective implementation, etc. Meta-analyses are their numerical counterparts for combining the evidence. Many times, both go hand in hand. These were being reported in the literature with varying format, sometimes missing the vital information. The aim of the PRISMA statement is to help authors improve the reporting of systematic reviews and meta-analyses. This is primarily designed for reviews of randomized trials, but PRISMA can also be used as a basis for reporting systematic reviews of other types of research, particularly those evaluating interventions.

The PRISMA Statement consists of a 27-item checklist and a four-phase flow diagram. The checklist is divided into Title, Abstract, Introduction, Methods, Results, and Discussion as is usual in the IMRAD format, and states what each of these should contain. For example, the checklist wants that full electronic search strategy for literature should be described, including any limits used, in a manner that it could be repeated. It also wants that results of any assessment of risk of bias across studies should be presented, and limitations fully discussed. There are several others. All these are pretty much the same as known to the researchers, but availability of checklist helps in ensuring that nothing is missed. More important from the statistical viewpoint is the flow diagram (Figure P.14). This shows how many articles have been identified, removed, screened, assessed, and finally included.

PRISMA Statement is an evolving document that is subject to change periodically as new evidence emerges. In fact, the present PRISMA Statement is an update and expansion of the now outdated QUOROM Statement.

1. PRISMA. *Transparent Reporting of Systematic Reviews and Meta-Analyses*. http://www.prisma-statement.org/
2. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses: The PRISMA Statement. *PLoS Med* 2009;6(6):e1000097. http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000097

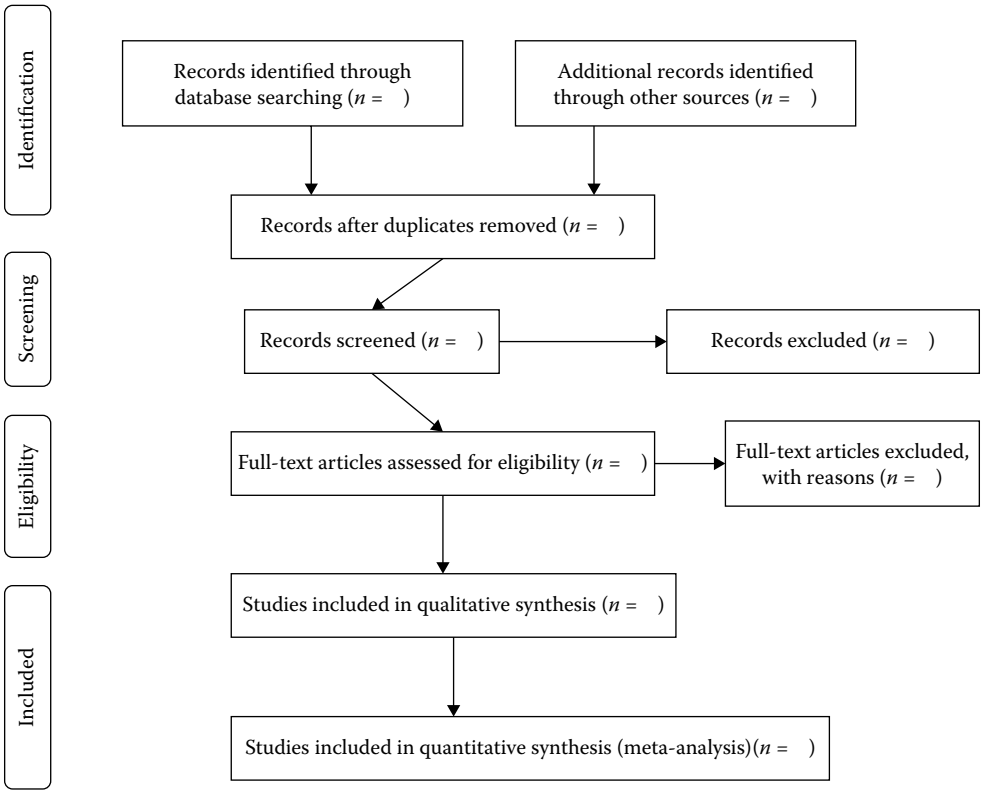



**FIGURE P.14** PRISMA flow diagram. (From Moher D, Liberati A, Tetzlaff J, Altman DG, *PLoS Med* 2009;6(6):e1000097. http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000097.)

## probability, *see also* laws of probability (addition and multiplication)

Uncertainties in medical decisions are profound. These can be minimized but not eliminated, and we must learn to live with them. Measurement of their magnitude is the first step in this direction. An accepted measure of uncertainty is probability. The term has everyday meaning, but its computation can be nerve-wrecking in some intricate cases. Mathematically speaking, an event that cannot occur, such as a human male giving birth to a child, has probability 0. This is as low as it can get. An event that is certain to occur, such as death, has probability 1. No probability can be negative, nor can it exceed 1.

As opposed to mathematics, statistical definition of probability is based primarily on empiricism and thus is milder. If a woman of age 58 years has never conceived in the history of a community, the statistical probability of occurrence of such an event in that community is 0. It does not necessarily imply that the event is impossible. Probability may seem ludicrously abstruse to some clinicians but is pervasive in all medical decision processes. See the two topics on **medical uncertainties** in this volume to be convinced that this really is so.

This section is concerned with statistical probabilities. In simple terms, if oral cancer is seen to occur in 8% of a *large number* of habitual tobacco chewers for more than 10 years, the probability *P* that a randomly picked tobacco chewer of this type will get oral cancer is 0.08. An interpretation of probability is the relative frequency in a large number of cases. Thus, it measures the likelihood of an event and is complementary to uncertainty. Their calculation is illustrated in an example later in this section. In an etiologic investigation, this probability is referred to as risk—a term extensively used to delineate the hazards of a disease on exposure to an unfavorable factor. If the risk of oral cancer among nonchewers is 0.005, then the risk in chewers in our example is 16 times of that in nonchewers.

Beside empirical approach to measure probability as just mentioned, statistical probabilities are also obtained on the basis of mathematical formulation of the statistical **distributions**. These are obtained as area under the curve in the specified range. Thus, for example, we can obtain the probability that a Student *t* statistic with 12 df is more than, say, 4.53, or a chi-square statistic with 28 df is less than 7.81. Such probabilities are routinely obtained for testing a statistical hypothesis. We have separately explained the method for this for the popular Gaussian distribution in the topic **Gaussian probability (how to obtain)**. Similar method is used for all other distributions.

Although probabilities can be exactly obtained in many situations, they are many times interpreted in a subjective manner. One such overlapping classification is as follows: virtually certain 99–100% probability, very likely 90–100%, likely 66–100%, about as likely as not 33–66%, unlikely 0–33%, very unlikely 0–10%, exceptionally unlikely 0–1%. Additional terms (extremely likely: 95–100%, more likely than not >50–100%, and extremely unlikely 0–5%) may also be used when appropriate [1]. In statistics, though, we like to use exact values in place of such subjective interpretations.

### *Personal and Imprecise Probabilities*

In many medical setups, precise probabilities are difficult to obtain. Many clinicians work with what is called **personal probability**. This is based on personal perception of chance of occurring of an event—the rate at which a person is willing to bet. Thus, this is also known as subjective probability. Ramsey [2] published an article in 1926 in which he interprets probability as related to individual knowledge that formulates personal beliefs. This led to the notion of personal probability. Savage [3] has also discussed these probabilities in some details.

In medicine, personal probability is the belief of the doctor in the occurrence of an event—a patient having a particular disease, efficacy of a treatment regimen, a patient getting fully relieved, not surviving for more than a year, etc. This is based on personal experience with the type of case in hand as evaluated from signs, symptoms, investigation results, etc. As more and more information becomes available by way of additional results, response to the treatment, etc., the personal probability changes. The difficulty with personal probabilities is that these are not objectively determined and may substantially vary from one clinician to another. Thus, these are not used for research.

Coined by Peter Walley in 1991 [4], the term **imprecise probability** signifies the absence of exact information that one needs to calculate precise probability. When precise probability cannot be obtained, imprecise probability, marked by lower and upper limits, can help in some cases. Another term used for the same is *interval probability*. Long-term chance of occurrence of cancer in a person with a specific trait would be rarely known. Imprecise probability is a term in relation to this kind of occurrence. Such a probability could be based on personal belief or other considerations that are fuzzy and have an element of gambling. Probability is imprecise when stated in interval such as between 10% and 15%. Such probability arises when the information is scarce, vague, or conflicting, and where the preferences are of the "may be" type. There is no bar in using imprecise probability where precise probability is not available, but note that a decision based on imprecise probability can only be tentative. Remember also that the output of any model cannot be more precise than the inputs. If at all, the output will be less precise. Thus, do not expect a model to work wonders when it is based on imprecise probability. It is better to use precise probabilities wherever possible.

To illustrate it further, suppose a patient comes to you with longstanding complaints of pain in abdomen, vomiting, and constipation. These are considered cardinal complaints, but how confidently can you say that the person has abdominal tuberculosis or, for that matter, any other related disease? Incomplete information propels the use of imprecise probability. Such instances are galore in critical care setup, where the patient requires immediate help and is not in a position to provide history. Even when full information is available, precise probability of a disease or of prognosis is difficult due to omnipresence of **epistemic uncertainties** in health and medicine. In addition, *risk* and *safety* are terms with inherent uncertainty, which are so commonly used in health and medicine setups. It is easier in these setups to think of a range of values of probability that allow greater flexibility for uncertainty quantification to take care of the missing information. One can even postulate that precise probability framework is more of a mirage than reality. It is apparent that weaker expertise is required for specifying imprecise probability compared with the expertise required for precise probability. However, full range of imprecise probability models has not been developed yet, and it is in the pipeline.

Not many researches that have used imprecise probability are available in medical literature. One is by Zaffalon et al. [5] that has successfully used imprecise probabilities arising from incomplete data for diagnosis of dementia and discriminating it with Alzheimer disease. Gurrin et al. [6] have discussed use of imprecise probabilities in randomized clinical trials.

### *Conditional, Marginal, and Complementary Probabilities*

The probability we assign to occurrence of an event A depends on what is known about the situation at that time. If we come to know more subsequently, this probability is revised. This revised probability is statistically called the conditional probability and is denoted as $P(A|B)$, where B is the additional information now available. The original probability is called the unconditional probability. These can also be understood as **prior and posterior probabilities**. This happens all the time in any medical setup such as revising the chance of disease when the test results are available, radiological findings are known, the patient reports improvement or deterioration, or new knowledge becomes available. The conditional probability is defined as

$$\textbf{conditional probability}: P(A|B) = \frac{P(A \cap B)}{P(B)},$$

where ∩ refers to simultaneous occurrence and is called the intersection. This probability really means that out of those with event B, what is the chance that A will occur? This can be easily explained with the help of an example of a cross-sectional study on disease and test results (Table P.9).

For the data in Table P.9, $P(D+) = 160/400 = 0.40$, $P(D–) = 240/400 = 0.60$, $P(T+) = 310/400 = 0.775$, and $P(T–) = 90/400 = 0.225$, where the + sign is for presence and the – sign for absence. Because of their division in this table, these will now be called *marginal probabilities* instead of original probabilities. The intersection probabilities are as follows: $P(D+ \cap T+) = 120/400 = 0.30$, $P(D+ \cap T–) = 40/400 = 0.10$, $P(D– \cap T+) = 190/400 = 0.475$, and $P(D– \cap T–) = 50/400 = 0.125$. Note that the sum of all these intersection probabilities is 1 since these are **mutually exclusive and exhaustive categories** in this example. Now the conditional probabilities can be obtained. For example, $P(D+ | T–) = P(D+ \cap T–)/P(T–) = 0.10/0.225 = 0.444$. This is the same as 40 disease positives among 90 test negatives (40/90 = 0.444). This particular conditional probability is the chance that a test negative patient has the disease.

Note that the conditional probability can be very different from the unconditional probability. For example, probabilities of death at different ages will add to 1. Conditional probabilities do not have this feature. The conditional probability of death between 80 and 84 years may be 0.6, and the conditional probability of death between 85 and 89 years may be 0.7. These are conditioned on the person being alive at that age. These probabilities add to more than 1. This cannot happen with unconditional probabilities.

The scheme of medical knowledge is sometimes to provide probabilities of the form $P(\text{complaints}|\text{disease})$, whereas the probabilities actually required in practice are of the form $P(\text{disease}|\text{complaints})$. For simplicity and for generalizability, denote the set of complaints and investigation results by $C$ and the particular disease by $D$. When some additional information is available, $P(D|C)$ can be obtained from $P(C|D)$ by using **Bayes rule**.

A word about *complementary probability*: When probability of occurrence of an event A is $P(A)$, the complement is the probability of nonoccurrence of A. This is $1 – P(A)$. In our example, $P(T–)$ is the complementary probability of $P(T+)$ assuming that if test is not negative, it has to be positive, when the third possibility of indeterminate is ruled out.

**TABLE P.9**
**Test Results and Disease Status in 400 Subjects Attending a Clinic**

| | Test Positive | Test Negative | Total |
|---|---|---|---|
| Disease | 120 | 40 | **160** |
| No disease | 190 | 50 | **240** |
| **Total** | **310** | **90** | **400** |

### *Further on Probabilities*

Probabilities are the foundation of statistics. You must have a good grasp of its different features to be able to correctly interpret the statistical results. This includes the **laws of probability** such as law of multiplication, which helps to calculate the probability of joint occurrence of two or more events, and the law of addition, which helps to calculate the probability of one or the other event. Refer to that topic for more details. Probabilities play an important role in wide-ranging clinical activities as described for **probabilities in clinical assessment**.


1. Climate News Network. *The IPCC's Fifth Assessment Report*. http://www.climatenewsnetwork.net/2013/09/the-ipccs-fifth-assessment-report/
2. Ramsey FP. Foundations of mathematics. *Proc London Math Soc* 1926;25:338–84. http://www.jstor.org/stable/2249944
3. Savage LJ. Implications of personal probability for induction. *J Philos* 1967;44:593–607. http://www.jstor.org/stable/2024536
4. Walley P. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, 1991.
5. Zaffalon M, Wesnes K, Petrini O. Reliable diagnoses of dementia by the naive credal classifier inferred from incomplete cognitive data. *Artif Intel Med* 2003;29(1–2):61–79. http://www.aiimjournal.com/article/S0933-3657(03)00046-0/pdf
6. Gurrin LC, Sly PD, Burton PR. Using imprecise probabilities to address the questions of inference and decision in randomized clinical trials. *J Eval Clin Pract* 2002 May;8(2):255–68. http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2753.2002.00358.x/abstract;jsessionid=3F0C1769A18201093F175263ACD5EE36.f03t03


## probabilities in clinical assessment

**Probabilities** play an important role in wide-ranging clinical activities. When a diagnosis is reached on the basis of the complaints and physical examination, this is generally only the most *likely* diagnosis. As the investigation reports become available or the response to a therapy is known, the probability changes and sometimes even the most likely diagnosis also changes. *If they do not change the probability, then there is no use of these investigations and they are a waste of effort.*

Prospects of recovery after a therapeutic intervention are almost always stated in terms of probability. When the first heart transplantation was done, the chances of success were rated as 80%. The probability of recovery of a patient from tetanus after clinical manifestation is often considered less than 60%. The probability of survival for more than 5 years after detection of acute lymphoblastic leukemia is generally assessed as less than 30% despite therapy. Thus, probabilities have extensive usage in clinical assessments as illustrated next.

### *Probabilities in Diagnosis*

Consider the following for the process of diagnosis. A set of signs, symptoms, and other evidence, occurring together more often than expected by chance and generally found useful in treatment and

prognosis, is given a name, and the process is called diagnosis. It requires discovering clusters such that similar patients fall within the same cluster, but the clusters themselves remain relatively distinct from one another.

There are two kinds of diagnostic entities. First are those that are based mainly on the value of a set of particular measurements. Diagnosis of essential hypertension depends almost exclusively on levels of systolic and diastolic blood pressures, diagnosis of glaucoma on intraocular pressure, and diagnosis of diabetes mellitus on serum glucose level. Once a defined cutoff point is available, the diagnosis is immediately obtained in these cases. Nevertheless, because of the wide-ranging variability even among healthy subjects and the distinct possibility of overlap between measurements found in healthy and ill subjects, a diagnosis based on such clear-cut definitions does not always remain above board. Probabilities play a preeminent role.

The second type of diagnostic entities is essentially multifactorial. They depend on signs and symptoms, and on findings of laboratory and radiological investigations. Most cardiopulmonary disorders and hepatic dysfunctions fall in this group. Even though gold standards are available in some cases, such as positive cerebral angiography in embolism, the facility to carry out such an investigation may not be immediately available. In some cases, the gold standard itself can be in error. Positive histological evidence in carcinomas is considered confirmatory, but negative findings sometimes fail to correctly exclude the disease. When diagnosis depends on a multitude of factors, the clinician's personal judgment becomes important. Childhood leukemia and abdominal tuberculosis are examples of such diagnoses. When a patient presents with complaints of pain in the right hypochondrium, anorexia, and dyspepsia, liver disease is suspected. Liver function tests are then performed, and various enzymes in the blood are estimated. They too, in many cases, fail to provide a differential diagnosis on cirrhosis, malignancy, or hepatitis. Looking at the totality of the presentation, only the most probable diagnosis is made.

Suppose the analysis of records shows that 70% of patients with abdominal tuberculosis (abdTB) present with long-term complaints of abdominal pain, vomiting, and constipation. Then, $P$(pain, vomiting, constipation|abdTB) = 0.70. Because of the restriction to a specific group, which is mentioned after the vertical slash (|) sign, namely, the cases of abdTB in the above equation, such a probability is called the **conditional probability**. Note that the probability herein is of very little value to a clinician. It tells what is seen in the patients but does not tell about the presence or absence of the disease when these complaints are reported. The inverse probability $P$(abdTB|pain, vomiting, constipation) is useful to a clinician because it gives the diagnostic value of the complaints. The difficulty, however, is that the hospitals maintain records disease-wise and not complaint-wise. Thus, the records to compute the probability in the first equation above are easily located. This requires only screening the records of cases of abdTB with regard to those complaints. But to compute the probability in the second equation, records of all the cases need to be screened irrespective of the disease, and the cases separated according to the reporting of these three complaints. Among these, the cases with abdTB need to be counted. This is relatively a big exercise. One way to get around the problem is to use **Bayes rule** that converts $P$(A|B) into $P$(B|A).

### *Probability in Treatment*

Choice of treatment is a complex process and includes single or multiple therapy, dosage and duration, surgery, etc. It is known that the progress of disease and its severity differ from patient to patient and that a patient's response to a treatment always remains uncertain. When a choice of treatment modalities is available, that modality is prescribed, which is *most likely* to result in best relief for the patient according to the assessment of the treating clinician. Thus, the choice of the treatment too is a matter of probability, so is its effectiveness. This probability again is conditional on a priori information regarding patient characteristics, disease prognosis, facilities available, cost, etc. Although the efficiency, efficacy, and effectiveness of various treatment regimens are established with the help of clinical trials, in many cases, a physician is again guided by his/her or others' experience and uses **personal probabilities** to choose a particular treatment.

### *Assessment of Prognosis*

In prognosis, the effort is to correlate the outcome (survival with or without residual effect, or death) with the antecedent state (disease severity with reference to agent, host, and environment factors). The relationship depends, among other things, on correct diagnosis, promptness of treatment, type of treatment, facilities available, cooperation of the patient, attention of medical personnel, their expertise, etc. Because of the variability in all these factors, not all patients with a given similar antecedent state have the same outcome. Again, the probabilities are helpful in making a decision. As in the case of diagnosis, the probability here is also conditional on the antecedent state.

If the outcome is denoted by $O$ and the antecedent state by $A$, the probability to be calculated is $P(O|A)$. In fact, long-term follow-up studies of patients of different types and with different severities are needed before $P(O|A)$ can be objectively evaluated. Books and other literature may carry information on such probability of a specific outcome for a particular antecedent state. In the absence of such studies, the clinician's own experience of the percentages of patients who in the past have recovered, been disabled, or died can be used. So far, the prognosis lacks much of such quantification.

## probability proportional to size (PPS) sampling



This sampling method applies to situations where blocks of units are to be selected and where larger blocks are given higher chance of selection. This presumes that the blocks vary greatly in their size. Blocks can be villages, schools, or hospitals, whereas the units can be persons, students, and patients, respectively. PPS is best understood using the following example. This example has a feature of systematic sampling woven into it.

Consider **cluster random sampling** (CRS) for assessing prevalence of poor vision in old age. For a survey on prevalence of poor vision (visual acuity <6/36 in the better eye with corrective glasses if any) in persons of age 50 years and above (50+) in a district with half a million population divided into census blocks, suppose 20 clusters (one cluster per block) of size 30 each are selected as per the following scheme:

(a) A list of census blocks is prepared along with the population of each, and this is cumulatively added.
(b) A sample of 20 clusters in half a million population means that the sampling fraction is one cluster per 25,000 population. One number less than or equal to 25,000 is randomly selected. Then 25,000 is sequentially added every time in a systematic fashion, and thus a sample of 20 numbers is obtained. Twenty blocks containing the chosen 20 numbers are selected from the list made in (a). These blocks are now in the sample.

**TABLE P.10**
**Illustration of PPS Sampling**

| Block Number | Block Size (Population) | Cumulative Block Size | Selected Blocks by PPS (Random Number) |
|---|---|---|---|
| 1 | 4387 | 4387 | |
| 2 | 12,275 | 16,662 | |
| 3 | 8931 | 25,593 | 3 (18,789) |
| 4 | 5365 | 30,958 | |
| 5 | 7987 | 38,945 | |
| 6 | 2361 | 41,306 | |
| 7 | 4580 | 45,886 | 7 (43,789) |
| 8 | 15,439 | 61,325 | |
| 9 | 9438 | 70,763 | |
| 10 | 6753 | 77,516 | 10 (68,789) |

Let us illustrate this procedure for selection of initial three blocks where the block sizes are as in Table P.10. The first random number selected is 18,789. Since all numbers from 16,663 to 25,593 are in block number 3, this would be selected. Now add 25,000 to the first random number and get 43,789. This number falls into block number 7. This is the second block selected. Again add 25,000 and get 68,789. This falls into block number 10, and this is selected. This process goes on till 20 blocks are selected.

(c) Home visits are made from a geographically random point in each of the selected blocks, and the first 30 persons of age 50+ residing in contiguous houses are listed and examined for visual acuity. This gives one cluster of 30 subjects from each selected block.

The scheme in this example of CRS is similar to the one recommended by WHO for surveys to assess immunization coverage in developing countries, but is not exactly the same. Among several features of the CRS, note the following in this example.

The selection of blocks is based on the size of the population in these blocks. This is inherent in step (b). Blocks with a larger population have a higher chance of being included in the sample. This is what is called sampling with probability proportional to size (PPS). The size in this case is the population in the census blocks, and the subjects are the persons of age 50+. It is reasonable to expect that this age group would have nearly the same proportion in each census block. The technique was suggested by Hansen and Hurwitz [1] in 1943.

This sampling becomes self-weighting because of PPS, and this makes estimation easier. Weighted calculations, as done for **stratified random sampling**, are not required when this kind of PPS sampling is done. The sampling strategy we have given in our example has the feature that you can choose the size of clusters as per your convenience, such as being able to manage the cluster within a day. In this example, the cluster size is 30, which could be covered by one team in 1 day. This would make it easier to complete the survey in a more efficient manner. If you think that a size of 40 could be covered by one team in a day, you can choose clusters of size 40. However, there are limitations as well. First, PPS can be adopted only when the **primary sampling unit** is a block containing several units of inquiry. Second, it requires that the information on size of all blocks in the population is available. In many situations, this would be a challenge, and a surrogate may have to be used. In our example also, we have used population as the size, whereas the actual requirement was people of age 50+. Similarly, for example, when the number of patients admitted in a hospital is not directly available, the bed strength can be used as a surrogate for the number of patients in the hospital. Using estimated size in place of the actual size introduces some complication in the estimates, but this is surmountable. For details, see Cochran [2].

PPS sampling is quite common in surveys. For example, O'toole et al. [3] report data on two cities in the United States from a community-based, probability populations-proportionate sample of randomly selected homeless adults to compare substance-abusing to non-substance-abusing respondents with respect to not just sex and education but also characteristics such as stealing and the need to learn how to manage money.


1. Hansen MH, Hurwitz WN. On the theory of sampling from finite populations. *Ann Math Stat* 1943;14:333–62. https://projecteuclid.org/download/pdf_1/euclid.aoms/1177731356
2. Cochran WG. *Sampling Techniques*, Third Edition. Wiley, 1977.
3. O'toole TP, Conde-Martel A, Gibbon JL, Hanusa BH, Freyder PJ, Fine MJ. Substance-abusing urban homeless in the late 1990s: How do they differ from non-substance-abusing homeless persons? *J Urban Healt*. 2004 Dec;81(4):606–17. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3455928/


## probability sample, *see* sampling techniques

## proband

A proband is a person who is the first affected within a family and through whom attention is first drawn to a pedigree of particular interest to human genetics. Here *pedigree* is a term used for any grouping of related individuals, for example, pairs such as twins and spouses, and larger groups such as nuclear or extended families. Probands help in the general understanding of genetic disorders—how they pass on to the siblings and progeny.

Proband studies are quite common in human genetics. Sun et al. [1] studied a proband in China who had typical homozygous phenotype of familial hypercholesterolemia to identify the gene defect. They did not study a series as is generally done in this setup. Messinger et al. [2] examined sex differences in autism spectrum disorder (ASD) outcome and in the development of ASD symptoms and cognitive functioning among the high-risk younger siblings of such probands and low-risk children, in view of female protective effect hypothesis for this disease.

The concept of proband is also loosely used in calculating the secondary **attack rate** of a communicable disease, although a more appropriate term is *index case*. This rate measures how many people are affected by the probands on average. When a proband affects others, those become proband for others. Secondary attack rate measures the intensity of spread of infection or risk among the susceptible contacts after exposure to an infective case. When the primary case (proband) is infective for a long period as in tuberculosis, the duration of exposure becomes important. Secondary attack rate then is computed per 100 person-weeks, person-months, or person-years of exposure.

Study of probands and their progeny poses a statistical challenge because of the clustering effect. For something like regression in such a study, you may have to use the method of **generalized estimating equations** because of possible clustering. Luckily, genetic studies seldom require study of the factors contributing to the outcome (as in regression) because genetic factors can be directly studied in many situations.

1. Sun LY, Zhang YB, Jiang L, Wan N, Wu WF, Pan XD, Yu J, Zhang F, Wang LY. Identification of the gene defect responsible for severe hypercholesterolaemia using whole-exome sequencing. *Sci Rep* 2015 Jun 16;5:11380. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4468422/
2. Messinger DS, Young GS, Webb SJ, Ozonoff S, Bryson SE, Carter A, Carver L. Early sex differences are not autism-specific: A Baby Siblings Research Consortium (BSRC) study. *Mol Autism* 2015 Jun 4;6:32. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4455973/

## probit transformation

The acronym for probability integral transformation, probit is the value of standard **Gaussian** (normal) $z$ corresponding to the specified probability. For example, probit = 0 for $P = 0.5$ since $P(z < 0) = 0.5$ for this distribution. The expression $z < 0$ indicates that $P$ is the cumulative probability. As another example, it is well known that for $P = 0.975$, $z = 1.96$, and for $P = 0.025$, $z = -1.96$. These are the probits. Sometimes, 5 is added to the values of $z$ so that none is negative. When we add 5 to such value of $z$, the probit of 0.975 is $5 + 1.96 = 6.96$ and probit of 0.025 is $5 - 1.96 = 3.04$. The negative value $z = -1.96$ for $P = 0.025$ explains how adding 5 is helpful. This addition of 5 is needed when doing calculations by hand and not needed when working with computers. Computer software packages on probit generally do not add 5.

Probit transformation is helpful when the category for which the probability $\pi$ is obtained is based on a continuous variable such as for systolic blood pressure <140 and ≥140 mmHg, and this continuum, in this example blood pressure, follows a Gaussian distribution. When this transformation is used in the case of simple linear regression, the target regression is

$$\text{probit of } \pi = \alpha + \Sigma\beta_k x_k \, (k = 1, 2, \ldots, K)$$

for $K$ regressors. This is an alternative to logistic regression and uses the Gaussian property of the underlying variable. The logistic does not use this property—and thus has applicability to non-Gaussian setups as well. Sometimes, probits are considered more relevant for problems of estimation of relative potency in bioassays. Methods are available to find the estimate of $\alpha$ and $\beta$ for standard and test preparations—thereby obtain the relative potency. However, the method is complex. See Finney [1] for details.

1. Finney DJ. *Probit Analysis*, Reissue Edition. Cambridge University, 2009.

## product limit estimator, *see* Kaplan–Meier method

## product–moment correlation, *see* correlation coefficient (Pearsonian/product–moment)

## profile analysis

This is the term used for analysis of the characteristics of the subjects to find what kind of subjects are more commonly affected and what kind are not affected. It can also tell which type of cases get severe form of disease and which ones get away with mild form only, or which type of cases suffer for longer duration and which ones do not.

The easiest and most common is the profile of cases in a study that gives their age–sex distribution, their socioeconomic status, their distribution by severity of disease, from where they came (geographical area, the institutions, or the like), etc. This is a **descriptive analysis** just to apprise the reader about the basic information of the subjects. This contextualizes the study and helps to assess where the results could be applicable. Profile analysis can go deeper into cross-classifications so that predominance or underrepresentation of particular type or types of cases can be highlighted. However, for this, the comparison should be with a valid group. For example, if you find that the age group 30–39 years is dominant among accidental deaths you are studying, compare it with the age distribution in the population from which these cases are coming.

This kind of analysis also helps in confirming or denying that the groups in clinical trial or any other experiment under study were similar to begin with. If the profile of subjects in different groups under study is the same and significant differences in outcome after an intervention are present, you would have more confidence in assigning these differences to the intervention. This may require using the Student $t$-test for two groups or analysis of variance (ANOVA) for multiple groups in case of quantitative data, and chi-square for qualitative data. Help from graphical displays is also taken to compare profiles of two or more groups.

Medical uses of the term *profile analysis* seem to be greatly varied. Gribskov et al. [1] used the term profile analysis for a method of detecting distantly related proteins by sequence comparison. Shen and Richards [2] mentioned profile analysis in the context of spectral processing of two harmonic complexes. Ghosh and Chinnaiyan [3] talked of genomic outer profile analysis. Whereas the last one is statistical, the previous two are not.

There is another term called *latent profile analysis*. This is a pattern statistical technique discovered in quantitative data, similar to the cluster analysis. (For qualitative data, this is called *latent class analysis*.) Cluster analysis is a nonparametric technique, but latent profile analysis uses distributional properties (such as multivariate Gaussian distribution) to apportion data values into natural categories such that the values within categories are similar and the values in other categories are dissimilar. Thus, latent profile analysis is very different from profile analysis we mentioned earlier. For details of this method, see Hagenaars and McCutcheon [4].

1. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: Detection of distantly related proteins. *Proc Natl Acad Sci U S A* 1987 Jul;84(13):4355–8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC305087/
2. Shen Y, Richards VM. Spectral processing of two concurrent harmonic complexes. *J Acoust Soc Am* 2012 January;131(1):386–97. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3272713/
3. Ghosh D, Chinnaiyan AM. Genomic outlier profile analysis: Mixture models, null hypotheses, and nonparametric estimation. *Biostatistics* 2009 January;10(1):60–9. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2605210/
4. Hagenaars JA, McCutcheon AL. *Applied Latent Class Analysis*. Cambridge University Press, 2009.

## proforma, *see* questionnaire, schedule, and proforma

## propensity score approach

The interest in many regression situations is in finding the effect of one exposure or an antecedent on one outcome. The exposure could be an intervention such as a treatment, or a preexisting risk factor such as obesity, or any other of interest. However, several covariates or confounders are also considered in this setup so that the net

relationship can be identified when other factors are under control. When the covariates are almost equally divided among the comparison groups, such as by randomization in clinical trials, these do not cause much of a problem. But in observational studies, they are generally not balanced. This hinders causal inference. Propensity score approach helps in adjustment of these covariates particularly when the number of covariates is large and subjects with exposure (or without exposure) are few.

Under this approach, all covariates are converted into a single value, called propensity score. Let the exposure of interest be denoted by $x_1$ and covariates by $x_2, x_3, \ldots, x_K$. It is sometimes possible to find a combination of multiple covariates that together reasonably determine the exposure level, or the probability of exposure if it is of the yes/no type. That is, $x_1$ can be determined mostly by $x_2, x_3, \ldots, x_K$. This would be the propensity score, which summarizes the covariate information into a single value.

The propensity score is obtained by running a regression of the actual variable of interest $x_1$ on the covariates $x_2, x_3, \ldots, x_K$. If $x_1$ is binary (exposure status: exposed/not exposed), logistic regression is obtained, and if $x_1$ is quantitative (exposure level: quantitative such as the degree of exposure), quantitative regression is obtained. If this scoring is successful in terms of high $R^2$ or percentage correctly predicted, you can use this score as a regressor in place of $x_2, x_3, \ldots, x_K$. Now a regression with only two regressors is required—$x_1$ and the propensity score. The method assumes that you have sufficient, valid, and reliable data on these covariates, and these covariates are able to reasonably predict the exposure level or exposure status. Note that the exposure level (or exposure status) continues to be in the regression. In the case of binary exposure, for example, the exposure is adjusted for the propensity, and that can be a big help in causal inference. However, the propensity score method is purely statistical and may not have much biological interpretation. The method was first proposed by Rosenbaum and Rubin [1] in 1983.

In most experiments, the subjects are randomly and equally allocated to the test and control group so that they have equal "propensity" to be in any group. Because of random allocation, all covariates and confounders (known and unknown) are expected to be equally divided. This is not so in observational studies. You cannot allocate who will smoke and who will not—thus, there is no random allocation. Covariates are not equalized. Here you can match them for propensity score. In a case–control setup, generally an equal number of subjects are selected by the outcome; but cases may have 70% exposed, while control may have 40% exposed. For example, 70% of patients undergoing kidney transplant (cases) may be relatively young and nutritionally better, and only 40% of those on conventional treatment (controls) may be so. The success rate (say, living for at least 10 years after the treatment) in these two groups depends not just on receiving or not receiving transplant but also on their age and nutrition status, among other factors. These are the covariates in this setup. The outcome is the success rate. The conventional method to adjust for such covariate differential is by including these in the regression (in this case, logistic since the outcome is binary). One can also think of matching for the covariates. Both these options are practical for a small number of covariates. What if the number of covariates is large, say 10, even 50? Matching is not feasible for so many covariates, and logistic would require an extremely large sample at the rate of at least 8 events (not subjects) per covariate [2]. The propensity score approach can be helpful in this situation as this does not require such a big sample. If you have a sufficient number of positive outcomes, do the analysis by both approaches: (i) by using $x_1$ and all the covariates as such and (ii) by using $x_1$ and the propensity score. If the propensity scoring is successful, you will get nearly the same result. In this case, you may like to prefer propensity score as this makes the regression so simple.

The propensity scores can also be used for matching or for stratification of the subjects. Matching is easier with these scores compared to matching on 10 covariates because now the matching is to be done for just one variable. Subjects with the same propensity score are likely to have similar effect of the set of covariates. Stratification can be done on subjects with high and low scores (or by **quantiles**) that really are now a surrogate for the level of exposure. Matching, as always, is able to take care of the known factors whereas randomization takes care of the unknown factors in the epistemic domain also. As for any regression model, the following precautions may be helpful in using propensity scores. These are mentioned for the setups where there are many regressors of interest besides a large number of covariates that are reduced to propensity score.

- Use a parsimonious approach as is always the case with regression.
- Test all the regressors for statistical significance.
- Examine all regression coefficients carefully for their proper interpretation.
- Perform regression diagnostics regarding its applicability and adequacy.
- Examine the residuals with care.
- Hold out a part of your data for cross-validation, performing external validation on a new sample of data.
- Ensure that the propensity model has not missed any important factor determining the exposure status.

For further details, see Guo and Fraser [3].


1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70(1):41–55. http://biomet.oxfordjournals.org/content/70/1/41.full.pdf+html
2. Cepeda MS, Boston ER, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol* 2003;158(3):280–7. http://aje.oxfordjournals.org/content/158/3/280
3. Guo S, Fraser MW. *Propensity Score Analysis: Statistical Methods and Applications*. Sage, 2010.




## prophylactic trials, *see* clinical trials

## proportional hazards, *see also* Cox regression

This section presumes that you are familiar with the concept of hazards. If not, review the topic **hazard rate**. This is the same as risk but has a time dimension. Since measured per unit of time, this becomes a rate.

Proportional hazards imply that the ratio of hazards in the test and the control group remains the same over the entire survival time. For example, if a person of age 60 years with hypertension has hazard of myocardial infraction (MI) 1.25 times (i.e., 25% higher) that of the person without hypertension, this ratio should continue to be 1.25 at 70 years and at age 80 years if these ages are in your data. This means that the hazard of MI increases with age in the same proportion in hypertensives as in nonhypertensives. This may not hold, for example, for intricate surgery (compared with, say, medical treatment) because of the high risk of death at early stages due to perioperative causes and relatively lower risk of mortality at later stages of recovery when the surgery is seen as successful. When proportionality holds, no correction is required for the varying length

of follow-up of different subjects. This is the basic advantage of proportional hazards. Also, this makes the analysis independent of the statistical distribution of survival as shortly explained.

In terms of notations, consider two groups defined by indicator variable $x_1 = 0$ for control and $x_1 = 1$ for the treatment group. For simplicity, suppose this is the only variable under consideration. In this case, this gives $h(t) = h_0(t)$ for $x_1 = 0$ and $h(t) = h_0(t)\, e^b$ for $x_1 = 1$, where $h(t)$ is the hazard at time $t$ when covariates are present and $h_0(t)$ is the hazard again at time $t$ when no covariate is in operation. Thus, the ratio of these two hazards is $e^b$ for all $t$. It does not depend on time. The form of $h_0(t)$ does not matter for this ratio. This is what proportional hazards imply. This simple and useful interpretation has made the hazard ratio so popular. It is because of this proportional property that the method of analysis is independent of the pattern of the survival curve. Proportional hazards in the two groups would mean the plot of logarithm of hazards versus survival time would be parallel (Figure P.15). This is an important requirement for validity of **Cox regression**, where the logarithm of the hazard ratio is modeled to depend on a linear combination of the regressors. This essentially implies that the effect of covariates on the hazard rate is multiplicative instead of the usual additive.

There could be situations where proportionality does not hold. While comparing a short-acting regimen with the one with continued action for a long time, the hazards will not be proportional (Figure P.16) because hazard rises fast after the active phase of the short-acting regimen is over, whereas it remains low for substantial period with the long-acting regimen.

How does one check that the hazards are indeed proportional? For this, divide the total time period of follow-up into two or three plausible segments, and run the usual Cox regression after including an interaction term between time segments and other risk factors in the model. If any interaction term is statistically significant, conclude that the requirement of proportional hazards is violated. As for any test, detecting significance will also depend on the size of the sample—thus, this must be adequate for this procedure.

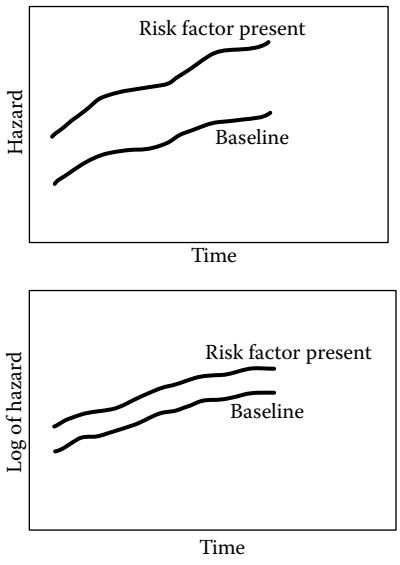



**FIGURE P.15** Hazards and proportional hazards in terms of logarithm.

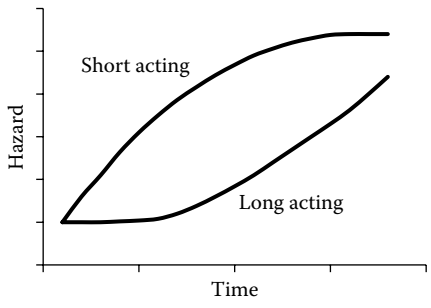



**FIGURE P.16** Nonproportional hazards.

## proportional reduction in error (PRE)

PRE is one of the measures of strength of relationship between two qualitative characteristics, but it has more useful interpretation in some specific situations. The primary purpose of PRE is to measure the utility of one characteristic in predicting the other: the higher the PRE, the more useful the predictor, and it is directly dependent on the strength of the relationship. The PRE ranges from 0 to 1 and can be interpreted from no association to perfect association. But there is no negative association. The basic principle underlying PRE is easily explained with the help of an example.

Table P.11 contains data on age and visual acuity (VA) of 1000 patients coming to a cataract clinic. Based on these data, if a guess is to be made about the visual acuity of a random person coming to the same cataract clinic, the best guess (with least error) is (6/60 > VA ≥ 1/60) because this is the most commonly (657/1000 = 65.7%) occurring acuity in such subjects. This guess can be wrong in the other 100 – 65.7 = 34.3% of the cases. This is the error of prediction. If we know that the age of the patient is between 60 and 69 years, then this guess is strengthened further because 325 out of 460 (70.7%) is a higher proportion in this acuity category than for any other age group. The error is now reduced to 29.3%. The PRE by knowing the age in this case is (34.3 – 29.3)/34.3 = 0.15 or 15%. This measures the utility of age in predicting VA category. However, knowing age on the whole is not helpful in predicting the VA category in the data of Table P.11 because the most common VA category is always (1/60, 6/60) whatever the age may be.

A study was carried out on 80 subfertile men with varicocele on spermatozoal morphology with the objective of finding whether head abnormalities can be used to predict neck abnormalities in spermatozoa. Suppose the data obtained are as shown in Table P.12.

If head abnormality is present, the best prediction is that neck abnormality is doubtful because this is the most commonly observed (24 cases) category in the head abnormality group. This prediction will be wrong in 44 – 24 = 20 cases. Table P.13 has been constructed

**TABLE P.11**
**Age and Visual Acuity (VA) in Patients Coming to a Cataract Clinic**

| Age-Group (Years) | Visual Acuity (VA) | | | Total |
|---|---|---|---|---|
| | ≥6/60 | 6/60–1/60 | <1/60 | |
| –49 | 19 | 69 | 22 | **110** |
| 50–59 | 39 | 142 | 29 | **210** |
| 60–69 | 46 | 325 | 89 | **460** |
| 70–79 | 21 | 98 | 51 | **170** |
| 80+ | 7 | 23 | 20 | **50** |
| **Total** | **132** | **657** | **211** | **1000** |

**TABLE P.12**
**Head and Neck Abnormality in Spermatozoa in Subfertile Men**

| | Neck Abnormality | | | |
|---|---|---|---|---|
| **Head Abnormality** | **Present** | **Doubtful** | **Absent** | **Total** |
| Present | 11 | 24 | 9 | **44** |
| Absent | 3 | 7 | 26 | **36** |
| **Total** | **14** | **31** | **35** | **80** |

**TABLE P.13**
**Calculation for PRE in Table P.12**

| **Head Abnormality** | **Best Prediction for Neck Abnormality** | **Extent of Error** |
|---|---|---|
| Present (44 cases) | Doubtful (24 cases) | 44 – 24 = 20 cases |
| Absent (36 cases) | Absent (26 cases) | 36 – 26 = 10 cases |
| If not known (total 80 cases) | Absent (35 cases) | 80 – 35 = 45 cases |

with similar arguments. The extent of error stated in the last column is obtained by subtracting the maximum in the row from the corresponding total. By knowing that the head abnormality is present or absent, the error is reduced from a total of 45 cases to a total of 20 + 10 = 30 cases. Thus,

$$\text{PRE (by knowing the head abnormality)} = \frac{(80-35)-[(44-24)+(36-26)]}{80-35} = 0.33.$$

Knowledge about the presence or absence of head abnormality reduces the error in predicting neck abnormality by 33%. This reduction is not really high in this example and shows that the association of head abnormality with neck abnormality in spermatozoa is not really strong.

In terms of notations, the PRE in predicting the column category when the row category is known can be written as

$$\text{PRE} = \frac{(n - \overset{max}{C}\, O_{\bullet c}) - \Sigma_r (O_{r\bullet} - \overset{max}{C}\, O_{rc})}{n - \overset{max}{C}\, O_{\bullet c}}, r = 1,2,\ldots,R,\ c = 1,2,\ \ldots, C,$$

where
the order of the table is $R \times C$.
$O_{rc}$ is the frequency in the cell in the $r$th row and the $c$th column.
$O_{r\bullet}$ is the marginal total in the $r$th row.
$O_{\bullet c}$ is the marginal total in the $c$th column.
$n$ is the total number of subjects.

The notation $\overset{max}{C}$ is for the maximum value among the columns.

For PRE in the formula given in the above equation to be interpretable, it is necessary that one variable is considered dependent on the other in the sense that one can be predicted by the other. In this example, neck abnormality is treated as dependent because it is sought to be predicted from head abnormality.

The formula given in the above equation assumes that the column category is to be predicted by the row category. The notations will change if the row category is to be predicted by the column category. The PRE considers all categories nominal. If categories are ordinal or metric, and if the order is important, then other measures of association should be used. These are discussed by Freeman [1]. Yamamoto et al. [2] have discussed PRE for ordered categories.

1. Freeman DH. *Applied Categorical Data Analysis*. Marcel Dekker, 1987.
2. Yamamoto K, Yoshida E, Tomizawa. Harmonic, geometric and arithmetic means type measures of proportional reduction in error for ordered two-way contingency tables. *J Appl Math Stat* 2014;1:1–8. http://paper.uscip.us/jams/JAMS.2014.1001.pdf

## proportionate sample

This is used in stratified sampling. A sample is called proportionate when the number of subjects from each stratum is in proportion to the size of the stratum. If the $k$th stratum has $N_k$ subjects in the entire target population, the sample size from the $k$th stratum should be $n_k \propto N_k/N$ for it to be proportionate. In exact terms, this is $n_k = n^*(N_k/N)$, where $n$ is the total sample size. Consider the following example.

Evidence has accumulated that a large waist–hip ratio (WHR) may be a health risk. In a study of 100 hypertensive males, the subjects are divided into thin, normal, and obese according to WHR ≤ 0.89, 0.90 ≤ WHR ≤ 1.09, and WHR ≥ 1.10. With this categorization, the **simple random sampling** (SRS) of 16 subjects happens to contain 2 thin, 11 normal, and 3 obese subjects. However, the actual numbers of thin, normal, and obese in the population according to this categorization are 14, 38, and 48, respectively. Thus, the obese are clearly underrepresented in the sample. If you calculate any summary such as mean cholesterol level in these 16 subjects, the sample mean would be biased because possibly high values of cholesterol levels of obese subjects would not be as many as they should due to their underrepresentation in the sample. If you divide the population in the first instance and then take a **stratified random sample**, the result could be very different. A commonly adopted strategy in this case is to take a proportionate sample from each stratum. When this is done, the sample mean, sample proportion, etc., would be unbiased without any correction.

If the sample sizes in different strata are not proportionate, an adjustment would be needed to get an unbiased estimate. In our example, the strata sizes are 14, 38, and 48, respectively. For obtaining the right sample mean, these are multiplied by the means in the respective strata and divided by the population size, i.e.,

$$\bar{x}_{st} = \sum N_k \bar{x}_k / N,$$

where $N_k$ is the size of the $k$th stratum and $\bar{x}_k$ is the mean obtained for the $k$th stratum. The stratum size is now the "weight" for calculation of the mean. Such weighting is necessary to get a valid estimate and is regularly done for stratified samples when not proportionately chosen. The same weighting procedure is used for obtaining the sample proportion, sample standard deviation, etc.

In the case of proportionate samples, the probability of selection of each subject is the same in different strata. The advantage with this is that the sample becomes self-weighting. This can be explained as follows. In our example, 16 out of 100 gives a sampling fraction of 0.16. Applying this to the strata sizes gives a sample of 2 from the first stratum, 6 from the second stratum, and 8 from the third stratum. Proportionate sample implies that the sample from each stratum is in the same proportion as in the population. This means

$$\frac{N_k}{N} = \frac{n_k}{n} \text{ for all the } k \text{ (i.e., for each stratum).}$$

When this is substituted in the formula earlier given for the sample mean, we get

$$\bar{x}_{st} = \sum n_k \bar{x}_k / n.$$

This can be easily seen to be the same as the usual sample mean without any weights. Thus, in the case of proportionate samples, the mean for the stratified sample can be calculated just as a usual unweighted mean is calculated.

## proportion-by-probability (P-P) plot

This is the plot of empirical cumulative proportion versus cumulative probabilities based on any hypothesized distribution and is used to check whether or not they agree. If they agree, the plot would be a straight line. If the plot is not a straight line, consider that the data do not follow the hypothesized distribution. This check is visual and not fully scientific because minor variations would always occur.

For a P-P plot, the observed values of the variable are first sorted into ascending order from minimum to maximum as $x_{[1]}, x_{[2]}, \ldots, x_{[n]}$. The $i$th observation is plotted as $i/n$ (i.e., the observed cumulative proportion) against the other axis as $F(x_{[i]})$, where $F(x_{[i]})$ stands for the value of the theoretical cumulative probability for the respective observation $x_{[i]}$. If they approximate well, all points in this plot should fall onto the diagonal line. Table P.14 contains survival years of 10 patients after the detection of a cancer. Does it follow a Gaussian distribution? The mean of the survival years is 12.3, and the sample standard deviation (SD) is 8.29 years. We used these to calculate the standardized deviate $z$. This is shown in the third column of Table P.14. The Gaussian probabilities obtained from software for these values of $z$ are in the last column. The plot of the probability versus proportion is in Figure P.17. Diagonal line is also shown for comparison. The P-P plot is away from the line—remains below the line all the time. Thus, the distribution of survival years is not Gaussian.

The P-P plot considers all aspects of Gaussianity including **kurtosis**. In place of P-P, you can also try the **quantile-by-quantile (Q-Q) plot**. However, both these methods are approximate. More exact methods require calculations and checking statistical significance of the departure from Gaussian. Such significance tests based on

**TABLE P.14**
**Calculations for P-P Plot**

| *i/n* | Survival Years (*x*) | *z* = (*x* – Mean)/ SD | Gaussian Probability |
|---|---|---|---|
| 0.1 | 2 | –1.24288 | 0.106956 |
| 0.2 | 5 | –0.88088 | 0.189192 |
| 0.3 | 7 | –0.63954 | 0.261236 |
| 0.4 | 8 | –0.51887 | 0.301925 |
| 0.5 | 10 | –0.27754 | 0.390684 |
| 0.6 | 11 | –0.15687 | 0.437674 |
| 0.7 | 12 | –0.03620 | 0.485561 |
| 0.8 | 18 | 0.68781 | 0.754213 |
| 0.9 | 20 | 0.92914 | 0.823593 |
| 1 | 30 | 2.13582 | 0.983653 |
| **Mean** | **12.3** | | |
| **SD** | **8.28721** | | |

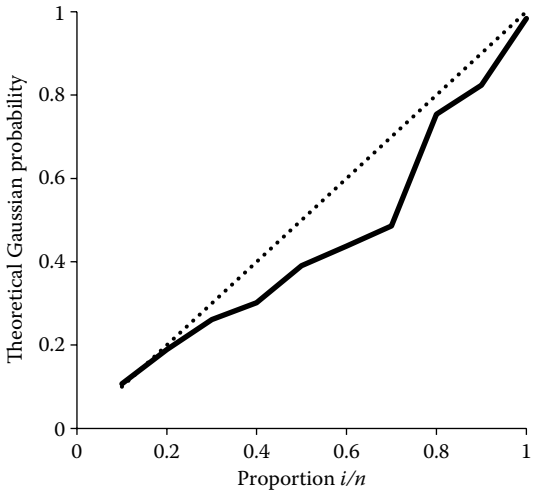



**FIGURE P.17** P-P plot for the data in Table P.14.

**Anderson–Darling**, **Shapiro–Wilk**, and **Kolmogorov–Smirnov test** are discussed in their respective topics.

## proportions

Proportion is the fraction of the whole. When we say that 30% of those affected had mild form of disease, the proportion is 0.30. Proportion is always out of 1, but many times, it is expressed in terms of percentage.

Proportion is best understood when the categories are **mutually exclusive and exhaustive** so that there is no overlap and nothing is left out when all categories are considered. Proportion of cases with mild, moderate, serous, and critical form of a disease is a good example. In this case, proportion tells how big or small the slice of a pie is. Best graphical representation of a proportion is indeed a **pie diagram**. You can show many proportions of the same pie by this diagram, although it becomes too cluttered if the number of categories is large. Since something like site of injury after an accident does not form mutually exclusive categories since one person can have injury at multiple sites, the pie diagram is not appropriate. Nonetheless, it is perfectly valid to say that 22% of all fatal accidents have head injury. This example may have also alerted you to distinguish between these two kinds of proportions—the overlapping and nonoverlapping. Nonoverlapping proportions give you a total equal to 1 when all possible categories are considered. Overlapping categories will give a total of more than 1.

The most prominent clinical application of proportion is in expressing the efficacy of a treatment—generally expressed as, say, 85%—and in expressing how common side effects are. Many times, probabilities of disease, or of recovery, or of death are also expressed in terms of proportions. For example, you may tell the patient that 4 out of 5 such patients recover within 7 days. This is just another method to say that the proportion recovering within 7 days is 0.80. However, proportions are different from odds. In our example of 4 patients recovering within 7 days out of 5, the odds of recovering within 7 days are 4:1. But odds can be converted to proportion: odds $a_1$:$a_2$ are the same as probability $\pi = a_1/(a_1 + a_2)$. Some statisticians and researchers also make a distinction between a proportion and a probability. Proportion is what you observe in a sample, whereas probability is what is seen in the corresponding population. The term *population* here is the statistical **population**

and not the general population. Probability is the proportion in the population and is applicable to individuals when randomly chosen or randomly arriving. In other words, probability is the long-term manifestation of the proportion: the expectation when the proportion is based on an exceedingly large sample, theoretically an infinite sample. Customarily, probability is denoted by $\pi$ and the sample proportion by $p$.

Probability of a sample proportion exceeding any given value, limiting (less than) to a given value, or lying between a specified range, is generally obtained by binomial distribution. The only condition is the probability in one case should be known and it should remain constant from person to person. Restricting to fatal accidental injuries, for example, we can find the probability that the number of persons with head injury will be between 20 and 30 out of 120 fatal accidents in a year in a city when the probability of this in one person is $\pi = 0.22$. Since $20/120 = 0.1667$ and $30/120 = 0.2500$, this translates to $P(0.1667 \le p \le 0.2500)$, where $p$ denotes the proportion of head injuries in fatal accidents in the sample subjects. The method to obtain this probability is given under the topic **binomial distribution**. When **Gaussian conditions** prevail, this probability is easily obtained by Gaussian approximation rather than by using the tedious method of binomial distribution.

In most situations, probability or the population proportion $\pi$ is not known and has to be estimated: the best estimate of $\pi$ is the sample proportion $p$. See the topic **confidence interval (CI) for proportion** and exact **confidence intervals (CIs)** for the methods to obtain interval estimates of $\pi$. We have also presented the method for obtaining the **confidence interval (CI) for difference between proportions**. See **$z$-test** for test of hypothesis on $\pi$ for large samples. For small samples, binomial distribution is used for this purpose as already indicated.

For comparing proportions in two groups, the groups are classified either as independent or matched pairs. Independent groups imply that the two samples taken are independent, that is, sample values selected from one population are not related in any way to sample values selected from the other population. Males and females are a common example of independent samples. Matched pairs consist of two samples that are dependent. Proportion of patients with a particular symptom before the treatment and after the treatment is an example of matched pairs. The method of choice for testing equality of two proportions in independent samples is **chi-square**, but that is applicable for large samples only. For small samples, the **Fisher exact** test is used. Exact methods are available in many statistical software packages for CI also for the difference between two proportions in small samples. If the samples are matched pairs, a large sample test is **McNemar**, and a small sample **exact test** is based on binomial distribution.

For proportions in multiple categories (polytomous), see **goodness of fit** test and **chi-square for trend in proportions** for large samples, and **multinomial distribution/test** for small samples.



## prospective studies, *see also* retrospective studies, case–control studies

Statistically, a study is called *prospective* when it investigates the outcome for a given set of antecedents. Antecedents could be exposure, intervention, or risk factors. Since outcome takes time to appear after the exposure, this generally requires a follow-up of cases forward in time. This could be a few minutes, a few hours, a few days, or a few years depending on whether the outcome is quick or slow to appear. There is a convention not to consider studies with a short follow-up as prospective. Thus, a study assessing the effect of anesthesia is not prospective in that sense. As we would shortly explain, the follow-up does not have to be in the future—it can also be in the past, called the *retrospective follow-up*. The follow-up can be just at one time point or at multiple time points. We also cite examples later of rare prospective studies that do not require actual follow-up in time frame. The defining feature remains that prospective studies are those that move from the antecedent to the outcome. Recruitment of subjects in the future as they come to a clinic or otherwise does not make it a prospective study since this refers to the recruitment only and not for the design.

Prospective studies can be contrasted with **retrospective studies** that move from the outcome to the antecedent, and **cross-sectional studies** that elicit both the antecedent and the outcome together.

*Prospective* is an umbrella term that includes various types of studies based on follow-up of the subjects: a cohort study with concurrent or historical cohort, repeated measures, and longitudinal study are all prospective studies. **Repeated measures studies** are separately described because of their special importance in medical and health setup; the others are explained in this section. **Before–after studies** too are technically prospective studies, and these also are separately discussed as they are commonly used. Some researchers use the term *prospective study* as a synonym of cohort study, but this is not entirely correct. A study of antenatal women for birth outcome, coming to a clinic staggered over a period of time at different gestations, is a prospective study but would not be called a cohort study. A cohort is a predefined group of subjects followed up for one or more outcomes. In view of its special importance in health and medical investigations, **cohort studies** are discussed as a separate topic in this volume.

It may be clear that prospective, cohort, longitudinal, and repeated measures are not mutually exclusive terms. A prospective study can be **longitudinal** comprising observations of each subject at several points of time or can have only two assessments: one at the beginning and the other at the end. In most longitudinal studies, measurements are taken at different points in time such as the patients are investigated before surgery, during surgery, and 1, 2, and 3 h after surgery. Another patient could be measured before, immediately after, and 90 min after surgery. The same characteristics such as extent of pain can be measured at these points of time. The objective is to assess change over time, and often the outcome of interest is the time taken to reach a particular end point such as the time when pain score is 1 or less on the scale of 0 to 10. In some situations, the interest may be in the pattern of change such as whether pain is high initially and then declines either rapidly or slowly, or declines initially and increases thereafter. In this case, the interest may also be in time at which the pain is the highest or lowest.

The primary objective of a longitudinal study is to track the trend over time, generally of a quantitative measurement. This means that the time points are an important consideration. For example, a **pharmacokinetic study** that evaluates peak concentration of a drug and time to reach the peak would require a longitudinal study since observations at several time points are needed for this kind of study. Similarly, a study on growth of children would need a longitudinal study to track their trajectory. In both these setups, the outcome is quantitative, but that is not a prerequisite for a study to be longitudinal. Time-invariant risk factors such as sex and family history may be measured only at the baseline, and other associated risk factors may be measured repeatedly over time.

Whereas longitudinal studies are mostly forward in time requiring follow-up, they could also be backward in the sense that the cases are investigated regarding their history at different points in time in the past. Smoking history of cancer cases that includes how many cigarettes were smoked for what duration in the past is the

classic example. Prospective studies could be based on past cases also. For example, Sokal et al. [1] carried out a cancer risk study in 1992 on the basis of the records of women sterilized with transcervical quinacrine hydrochloride pellets in Chile between 1977 and 1991. Traceable women were also interviewed. Despite being based on past records, it is not a retrospective study since the direction of investigation is from the antecedent to the outcome. Terms such as *retrospective follow-up* and *historical prospective* are also used for this kind of methods. This requires that past records are fully available.

Exceptional prospective studies can be cited that do not require any follow-up in actual time frame. In a study on the effect of profession on smoking habits, a cohort of people joining different professions in one particular year can be followed up for a 10-year period. This would be a standard prospective study. But the effect can also be studied by selecting people who have been in different professions for nearly 10 years and noting their present smoking habit at the end of the 10-year period in the profession. This also is a prospective study since the direction of the study is from the antecedent (in this case, profession) to the outcome (in this case, smoking), but there is no follow-up of the subjects in a conventional sense.

Among the most popular longitudinal studies is the **Framingham Heart Study** in the United States that has repeatedly measured initial cohort for their lifetime. Recent publication on this by Pencina et al. [2] asserts that investigation of apolipoprotein B improves risk assessment of future heart disease beyond LDL-C and non-HDL-C. Tarnanas et al. [3] carried out a longitudinal study on ecological validity of virtual reality daily living activities screening for early dementia.

Analysis of data from prospective studies is mostly done in terms of **relative risk (RR)** and **attributable risk (AR)**. These refer to the occurrence or nonoccurrence of an event of interest such as recovery, a medical parameter reaching to a threshold, and death. The outcome must be qualitative for these measures to be applicable. If the outcome of interest is quantitative such as actual creatinine level, and when the average over subjects at different time points is sensible, think of regression with time as a factor of interest for analyzing such data. You can have other covariates also in this model. All limitations of **regression models** apply. When the data at different time points are required to be considered together, **generalized estimating equations** may be a better method, particularly when the objective is to study the contribution of various factors to the outcome. This allows correlated values (in this case, values at different time points are correlated) that most other methods prohibit. When the duration of appearance or occurrence of an event at different points in time is the variable of interest, including whether the event occurred within the follow-up period or not, the data are analyzed by **survival analysis** methods. This considers survival (or event-occurrence rate) at different points in time.

### *Selection of Subjects for a Prospective Study*

An important consideration in the selection of subjects for any study is the feasibility of obtaining data on them. This means that the subjects must be approachable and cooperative. For a prospective study, especially, accurate and complete information must be available on them at baseline so that they can be correctly classified into exposed and nonexposed groups, and the effect of other characteristics on the outcome can be properly assessed. It is natural to expect that the subjects included in the study truly represent the target population. Thus, the target population must be clearly defined. It could be, for example, patients attending a particular group of diabetes clinics who are observed for development of retinopathy, or those exposed to a carcinogen in a particular district. Note the geographic limitation associated with the definition of a population as is nearly always done.

In a prospective study, generally only one **risk factor** will be of primary interest, but other risk factors to be concurrently studied also need to be properly identified. For example, in a study of maternal complications, these could be parity, nutrition status, hemoglobin level, and the nature of natal care. Decide in this case whether the study is to be restricted to women who are currently pregnant or will include all married women of reproductive age. Such specification also is important for any study.

A basic feature of a prospective study is that the incidence of outcome such as disease or relief is evaluated in those subjects who are exposed. It is often helpful to study a parallel group, also called a *control*, which is not exposed, so that a proper comparison can be made. Thus, case–control nomenclature can also be applied to prospective studies, which otherwise is restricted to retrospective studies. However, in this case, they are exposed and unexposed groups, and not diseased and nondiseased groups.

### *Comparison Group in a Prospective Study*

Proper selection of the comparison group enhances the validity of conclusions from a prospective study. Quite often, the control group comes from within the cohort, in whom some subjects are naturally exposed and some are not. For a valid comparison, the exposed and unexposed groups must be similar at baseline, particularly with regard to the factors that can influence the outcome not under study. If the objective is to study the effect of recently acquired central obesity on the electrocardiogram changes over time, factors such as age, gender, personality traits, stress conditions, and smoking need to be matched between the study group (with central obesity) and the control group (without central obesity). If complete matching is not possible, as would generally happen in practice, statistical methods are used to do the required adjustment at the time of analysis. Such an adjustment can become incomprehensible if done for a large number of factors and should be done for a few factors that are more relevant than others.



An external group can be used for comparison in some situations. An adequate number of nondiabetics may not be available in a diabetes clinic for assessing the development of coronary events. External controls can be included in such a situation; however, they should come from the same milieu and should preferably be matched for all the factors except the exposure. In a rare situation, when an appropriate external group is also not available, comparison can be done with the outcome rates in the general population. For example, the incidence of birth defects in babies born to women of age 45 years and above can be compared with that in births to women of childbearing age in the general population. The actual control group in this setup should be births to women of age less than 45 years, but a separate incidence of birth defects in them may not be easily available. The incidence in births to women of age less than 45 years may not be much different from that in all women of childbearing age since births after that age are rare. However, in many situations, the rate in the general population is not comparable with the rate in the unexposed group, and a great degree of precaution is required in using such a general group as control.

In some prospective studies, it is useful to have multiple groups for comparison. For example, the effect of profession on smoking habits can be investigated by including several categories of profession in the same study where none would be the control group in the conventional sense. Subjects with different exposure levels can also

be chosen for follow-up that would also provide multiple groups for comparison of the outcome.

It is sometimes impossible to find a group that is completely nonexposed. An example is exposure to dichlorodiphenyltrichloroethane (DDT). Even people in remote locations, such as Canada's Baffin Island, harbor traces of DDT. In such cases, the comparison effectively would be between the less exposed and the more exposed.

### Potential Biases in Prospective Studies

A large number of biases are listed under the topic **bias in medical studies**, and all those should be considered in a prospective study. Biases typically occurring in a prospective study setup are the following.

*Selection bias*: A prospective study group is rarely a random sample from the population of subjects, although this is highly desirable. Selection bias is said to have occurred when the study group has a different composition with regard to etiologic factors such as heredity, age, gender, nutrition status, and addictions compared with the composition in the target population. Studies on volunteers or on clinic subjects almost invariably suffer from such a bias. A method of selection that has a random component is considered insulation against this kind of bias. But such selection fails to take cognizance of special bias that can result from extraneous sources such as improper definition of the population. In a study on causes of psychiatric illness in old age, if the subjects are those who are single and of age 70 years or above at the time of enrollment, the bias occurs because some with severe illness may have already expired before attaining the age of 70 years. Those who remain are the ones who are robust or have minor illness.

*Bias due to loss in follow-up*: A major task in prospective studies is to accomplish successful follow-up of all the subjects. Loss occurs due to change in residence to an unknown or remote address, unrelated death, severe illness other than the one under study so that the required investigations cannot be done, loss of motivation of the patient to cooperate, fault developing in machines such as treadmill whose rectification takes time, absence of a trained technician, etc. In a clinic-based follow-up, when the patients are advised to report at periodic intervals, some may not come on the required day, and one or two follow-ups may be missed. Such loss constitutes a threat first because the size of the group shrinks and second because those lost are seldom a random subgroup. They are generally typical, for example, subjects who are seriously ill and are not hopeful of living long, or those who are mildly affected and who consider continuation in the study not worth taking the risk. If the rate of disease and the rate of severity are different in the subjects who have discontinued, it would affect the validity of the results. Where possible, a random subsample of the discontinued subjects should be investigated intensively to evaluate their characteristics versus the characteristics of those who have not discontinued. If they are really different either with respect to outcome or even with respect to the baseline information collected at the time of first contact, an adjustment may be required at the time of analysis to remove the effect of such bias. When nonrespondents could not be contacted despite best efforts, the baseline information can still be used for adjustment.

*Assessment bias and errors*: Human error can occur in assessing the condition of the patient. This can be due to either carelessness or lack of expertise of the observer. The physician may lack competence and the recording clerk may lack training or motivation. Assessment during the later part of a longitudinal study may be less accurate as fatigue sets in or may be more accurate due to learning effect.

*Bias due to change in the status*: In a prospective study on central obesity and coronary artery diseases, it is possible that some subjects of the nonobese group become obese while the follow-up is still in progress. In a study on effect of smoking, a nonsmoker at the initial stage may start smoking in the middle of the study, or a smoker may quit smoking. Exclusion of such cases is one option but is feasible only when their number is small. A long-term cohort is also affected by the environmental changes such as introduction of a new drug in the market that can influence the incidence of the disease under study.

*Validity bias*: Some prospective studies try to develop criteria to distinguish subjects with greater risk of disease from those with less risk. If sufficient distinguishing features are detected, then these criteria could indeed be developed. Such criteria may work wonderfully well on the group from which they have been derived, but they need to be externally validated on another group of similar subjects. Although statistical principles say that criteria based on representative sample should work nearly equally well on another sample from the same population, evidence of external validation is considered essential before such criteria are accepted. This validation could be done on another sample from the same population or even on a sample from a different population. The latter, of course, provides evidence that the results could be valid for other populations too.

All such biases and errors in assessment can be reduced simply by being more careful and by using precise instruments, measurements, and classification criteria that have been pretested for their validity. The identification and resolution of bias are primarily a matter of epidemiological judgment. The success of a prospective study often depends on the care taken by the investigator in recognizing and correcting these biases. Although some bias can be handled at the time of data analysis by using appropriate statistical techniques, the applicability of these techniques depends on validity considerations, particularly on the adequate number of cases with the outcome of interest in groups and subgroups. This cannot be ensured beforehand in this situation. Precautionary steps are preferable wherever feasible.

### Merits and Demerits of Prospective Studies

The advantages of prospective studies will be clearer if you are clear about other designs—retrospective and cross-sectional—of observational studies. Nevertheless, it is evident that the temporal sequence between exposure and disease can be more easily established by prospective studies than by any other format. Risk of outcome such as of disease or the chance of getting cured can be directly measured. Also, the incidence of an outcome cannot be assessed by any other method. Prospective studies are particularly well suited for assessing the effect of rare exposure such as of a specific chemical or of a new therapeutic modality because the cohort is expected to start with an adequate number of subjects. These studies allow for examination of multiple effects of a single exposure. If the outcome

of interest is death and the records are not adequate, a prospective design is the only choice.

There are several demerits too. Sometimes an outcome, such as carcinoma, may take years to appear after the exposure. It may occur only in a small percentage of subjects, which could mean a very large cohort to obtain an adequate number of subjects having the disease. Thus, prospective studies tend to be heavy on time and resources. In some situations, the natural course of the disease or the characteristics of the subjects may change during the follow-up period. As already stated, obesity, dietary pattern, and smoking can all change in a long-term follow-up. In a prospective study, subjects know that they are being observed, and this awareness may change their behavior and outcome, called the **Hawthorne effect**. If the study spans several years, it is difficult to maintain motivation and retain the trained staff. Supervision may also lose sharpness. Notwithstanding these difficulties, prospective studies are technically the most correct design because they move in the natural direction from exposure to outcome.

1. Sokal DC, Zipper J, Guzman-Serani R, Aldrich TE. Cancer risk among women sterilized with transcervical quinacrine hydrochloride pellets, 1977 to 1991. *Fertil Steril* 1995;64:325–34. http://www.ncbi.nlm.nih.gov/pubmed/7615111?dopt=Abstract
2. Pencina MJ, D'Agostino RB, Zdrojewski T. Apolipoprotein B improves risk assessment of future coronary heart disease in the Framingham Heart Study beyond LDL-C and non-HDL-C. *Eur J Prev Cardiol* 2015 Jan 29. pii: 2047487315569411. http://www.ncbi.nlm.nih.gov/pubmed/25633587
3. Tarnanas I, Schlee W, Tsolaki M, Müri R, Mosimann U, Nef T. Ecological validity of virtual reality daily living activities screening for early dementia: Longitudinal study. *JMIR Serious Games* 2013 Aug 6;1(1):e1. http://games.jmir.org/2013/1/e1/

## protective effect

The concept of protective effect arises when considering the **relative risk (RR)**. RR < 1 for any factor is an indication that this is a protective factor for the outcome being studied in the sense that the incidence of disease is lower in those with the "risk factor." RR = 1 for any factor implies that it is neither a risk factor nor a protective factor, and RR > 1 implies that it is a risk factor. These terms are being used in the usual sense as applicable to any adverse outcome. For example, eating plenty of fruits and vegetables has a protective effect on coronary events. If you calculate the RR for this factor for coronary events, you will find RR < 1. If the outcome is relief such as early discharge from the hospital, the same factor may give RR > 1 but is still protective as it increases the chance of early discharge. Thus, RRs should be cautiously interpreted. The same applies to odds ratio (OR) as well. Regression coefficients in logistic and ordinary regression may also be similarly interpreted.

Ghorbani et al. [1] have talked about protective effect of selenium on cisplatin-induced nephrotoxicity and concluded that selenium could probably prevent such injury when added to hydration therapy in cancerous patients. Akhavan et al. [2] have discussed early protective effect of hydroxychloroquine on the risk of cumulative damage in patients with systemic lupus erythematosus. There are a large number of such examples of protective effect in the literature.

1. Ghorbani A, Omidvar B, Parsi A. Protective effect of selenium on cisplatin induced nephrotoxicity: A double-blind controlled randomized clinical trial. *J Nephrol* 2013;2:129–34. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3891148/
2. Akhavan PS, Su J, Lou W, Gladman DD, Urowitz MB, Fortin PR. The early protective effect of hydroxychloroquine on the risk of cumulative damage in patients with systemic lupus erythematosus. *J Rheumatol* 2013;40:831–41. http://www.jrheum.org/content/40/6/831.short

## protocol (research) and its contents

In our context, a protocol is the document that delineates the research plan for a study. The protocol is the backbone that supports a study in all steps of its execution. Thus, sufficient thought must be given to its preparation. On many occasions, it gradually evolves as more information becomes available and is progressively examined for its adequacy. The most important aspect of study protocol is the statement of the problem, objectives, and hypotheses. Formulating a feasible scientific question is a challenge, as well as ensuring availability of resources (expertise, funds, patients, time) to complete the task. A useful first step in preparing a complete research protocol is developing a one- or two-page document that summarizes the key scientific and clinical features of the study. This document helps to structure the final protocol and can be used to communicate the scientific aspects of the study and elicit the feedback.

### *Research Problem*

It is often said that research is half done when the problem is clearly visualized. There is some truth in this assertion. Thus, do not shy away from devoting time in the beginning for identifying the problem, to thoroughly understand its various aspects, and to choose the specifics you would like to investigate.

A problem is a perceived difficulty, a feeling of discomfort about the way things are, the presence of a discrepancy between the existing situation and what it should be, a question about why a discrepancy is present, or the existence of two or more plausible answers to the same question [1]. Among the countless problems you may notice, identifying one suitable for study is not always easy. Researchability of course is a prime consideration, but rationale and feasibility are also important. Once these are established, the next important step is to determine the focus of the research. This can be done by reviewing the existing information to establish the parameters of the problem and using empirical knowledge to refine the focus. Specify exactly what new the world is likely to know through this research.

The title of research by itself is not the statement of the problem. The statement of the problem is a comprehensive statement regarding the basis for selecting the problem, the details of lacunae in existing knowledge, a reflection on its importance, and comments on its applicability and relevance. The focus should be sharp. For example, if the problem area is the role of diet in cancers, the focus may be on how consumption of meat affects the occurrence of pancreatic cancer in males residing in a particular area. For further focus, the study may be restricted to only nonsmoking males to eliminate the effect of smoking. For more depth, meat can be specified as red or white. Additionally, one can also include the amount of consumption of red and white meat and its duration. The role of other correlates that promote or inhibit the effect of meat in causing cancer can also be studied. The actual coverage of the study would depend on the availability of relevant subjects on one hand and the availability of time, resources, and expertise on the other. Such sharp focus is very helpful in specifying the objectives and hypotheses, in developing an appropriate research design, and in conducting the investigation.

### Objectives of the Study

The focus of the study is further refined by stating the objectives. These are generally divided into broad and specific. Any primary medical research can have two types of broad objectives. One is to describe features of a condition such as the clinical profile of a disease, its prevalence in various segments of the population, and the levels of medical parameters seen in different types of cases. This covers the distribution part of epidemiology of a disease or a health condition, such as what is common and what is rare, and what is the trend. It helps to assess the type of diseases prevalent in various groups and their load in a community. However, a **descriptive study** of this type does not seek explanation or causes. Evaluation of the level of $\beta_2$ microglobulin in cases of HIV/AIDS is an example of a descriptive study. A study on growth parameters of children or estimating prevalence of blindness in cataract cases is also a descriptive study.

The second type of broad objective could be to investigate the cause–effect type of the relationship between an antecedent and an outcome. Whereas cause–effect would be difficult to ascertain unless certain stringent conditions are met, an association or correlation can be easily established. Studies that aim to investigate association or cause–effect are called **analytical studies**.

A broad objective would generally encompass several dimensions of the problem. These dimensions are spelt out in specific objectives. For example, the broad objective may be to assess whether a new diagnostic modality is better than an existing one. The specific objectives in this case could be separately stated as (i) positive and negative predictivity; (ii) safety in case of an invasive procedure, or side effect of a medical treatment; (iii) feasibility under a variety of settings such as field, clinic, and hospital; (iv) acceptability by the medical community and the patients; and (v) cost-effectiveness. Another specific objective could be to evaluate its efficacy in different age–sex or disease-severity groups so that the kinds of cases where the regimen works well are identified. Specific objectives relate to the specific activities, and they identify the key indicators of interest. They are stated in measurable format.

Keep the specific objectives as few and focused as possible. Do not try to answer too many questions by a single study especially if its size is small. Too many objectives can render the study difficult to manage. Whatever objectives are set, stick to them all through the study as much as possible. Changing them midway or at the time of report writing signals that enough thinking was not done at the time of protocol development.

### Hypotheses under Investigation

A hypothesis is a precise expression of the expected results regarding the state of a phenomenon in the target population. Research is about replacing existing hypotheses with new ones that are more plausible. In a medical study, hypotheses could purport to explain the etiology of diseases; prevention strategies; screening and diagnostic modalities; distribution of occurrence in different segments of population; and the strategies to treat or to manage a disease, to prevent recurrence of disease or occurrence of adverse sequel, etc. Consider which of these types of hypotheses or any other can be investigated by the proposed study.

Hypotheses are not guesses but reflect the depth of knowledge of the topic of research. They must be stated in a manner that can be tested by collecting evidence. The hypothesis that dietary pattern affects the occurrence of cancer is not testable unless the specifics of diet and the type of cancer are specified. Antecedents and outcome variables, or other correlates, should be exactly specified in a hypothesis. Generate a separate hypothesis for each major expected relationship.

The hypotheses must correspond to the broad and specific objectives of the study. Whereas objectives define the key variables of interest, hypotheses are a guide to the strategies to analyze data. Besides objectives and hypotheses, a protocol is generally structured as described next. Most funding agencies and universities prescribe a particular format on these lines.

### Structure of the Protocol

Realize that a protocol is the focal document for any medical research. It is a comprehensive yet concise statement regarding the proposal. Protocols are generally prepared on a structured format with an introduction, containing background that exposes the gaps needing research; a review of literature, with details of various views and findings of others on the issue including those that are in conflict; a clear-worded set of objectives and the hypotheses under test; a methodology for collection of valid and reliable observations, and a statement about methods of data analysis; and the process of drawing conclusions. It tries to identify the uncertainty gaps and proposes methods to plug those gaps.

Administrative aspects such as sharing of responsibilities should also be mentioned. In any case, a protocol would contain the name of the investigator, academic qualification, institutional affiliation, and the hierarchy of advisers in the case of masters and doctoral work. The place of the study such as the department and institution should also be mentioned. The year the proposal is framed is also required. An appendix at the end includes the proforma of data collection, the consent form, etc. The main body of a protocol must address the following questions with convincing justification.

*Title*: What is actually intended to be studied and is the study sufficiently specific?

*Introduction*: How did the problem arise? In what context? What is the need of the study—what new is expected that is not known so far? Is it worth investigating? Is the study exploratory in nature, or are definitive conclusions expected? To what segment of population or to what type of cases is the problem addressed?

*Review of literature*: What is the status of the present knowledge? What are the gaps? Are there any conflicting reports? How the problem has been approached so far? With what results?

*Objectives and hypotheses*: As already described.

*Methodology*: (i) What exactly is the intervention, if any—its duration, dosage, frequency, etc.? What instructions and material are to be given to the subjects and at what time? (ii) Is there any comparison group? Why is it needed and how will it be chosen? How will it provide valid comparison? (iii) What are the possible confounders? How are these and other possible sources of bias to be handled? What is the method of allocation of subjects to different groups? If there is any blinding, how will it be implemented? Is there any matching? On what variables and why? (iv) On what characteristics would the subjects be assessed—what are the antecedents and outcomes of interest? When would these assessments be made? Who will assess them? Are these assessments necessary and sufficient to answer the proposed questions? (v) What exactly is the design of the study? Is the study descriptive or analytical? If analytical, is it observational or experimental? An observational study could be prospective, retrospective, or cross-sectional. An

experiment could be carried out on biological materials or animals, or humans, in which case it is called a trial. If experimental, is the design one-way, two-way, factorial, or what? (vi) What is the operational definition of various assessments? What methods of assessment are to be used—are they sufficiently valid and reliable? (vii) What information will be obtained by inspecting records, by interview, by laboratory and radiological investigations, and by physical examination? Is there any system of continuous monitoring in place? What mechanism is to be adopted for quality control of measurements? What is the set of instructions to be given to the assessor? (viii) What form is to be used for eliciting and recording the data? (Attach it as an appendix.) Who will record? Will it contain the necessary instructions? (ix) What is to be done in case of contingencies such as dropout of subjects or nonavailability of the kit or the regimen, or development of complications in some subjects? What safeguards are provided to protect the health of the participants? Also, when should one stop the study if a conclusion emerges before the full course of the sample? (x) What is the period of the study and the time line?

*Study sample*: What are the subjects, what is the target population, what is the source of subjects, how are they going to be selected, how many in each group, and what is the justification? What is the minimal effect you are looking for that would be considered clinically relevant, and how is this determined? What would be the statistical power or expected confidence interval, with what error? What are the inclusion and exclusion criteria? Is there any possibility of selection bias, and how is this proposed to be handled?

*Data analysis*: What estimations, comparisons, and trend assessments are to be done at the time of data analysis? Will the quality and quantity of available data be adequate for these estimations, comparisons, and trend assessments? What statistical indices are to be used to summarize the data—are these indices sufficiently valid and reliable? How is the data analysis to be done—what statistical methods would be used and are these methods really appropriate for the type of data and for providing correct answer to the questions? What level of significance or level of confidence is to be used? How are missing data, noncompliance, and nonresponse to be handled?

*Reliability and validity of results*: What is the expected reliability of the conclusions? What are the limitations of the study, if any, with regard to generalizability or applicability? What exercises are proposed to be undertaken to confirm the internal and external validity of the results?

*Administration*: What resources are required, and how are they to be arranged? How are responsibilities to be shared among investigators, supporting units (e.g., pathology, radiology, and biostatistics), hospital administration, funding agency, etc.?

In short, the protocol should be able to convince the reader that the topic is important, the data collected would be reliable and valid for that topic, and that contradictions, if any, would be satisfactorily resolved. Present it before a critical but positive audience and get their feedback. You may be creative and may be in a position to argue with conviction, but skepticism in science is regularly practiced. In fact, it is welcome. The method and results would be continuously scrutinized for possible errors. The protocol is the most important document to evaluate the scientific merit of the study proposal by the funding agencies as well as by the accepting agencies. Peer validation is a rule rather than an exception in scientific pursuits. A good research is the one that is robust to such reviews.

You can see that a protocol should consist of full details with no shortcuts, yet should be concise. It should be to the point and coherent. The reader, who may not be fully familiar with the topic, should be able to get clear answer about the why, what, and how of the proposed research. To the extent possible, it should embody the interest of the sponsor, the investigator, the patients, and the society. The protocol also is a reference source for the members of the research team whenever needed and therefore should be complete and easy to implement.

The protocol is also a big help at the time of writing of the report or a paper. The introduction and methods sections remain much the same as in the protocol, although they are in an elaborate format. The objectives as stated in the protocol help to retain the focus in the report. Much of the literature review done at the time of protocol writing also proves handy at the time of report writing.

Whereas all other aspects may be clear by themselves, special emphasis should be placed on the impartiality of the literature review. Do not be selective to include only those pieces of literature that support your hypotheses. Include those that are also inconsistent with or oppositional to your hypotheses. Justify the rationale of your research with reasons that effectively counter the opposite or indifferent view. *Research is a step in the relentless search for truth, and it must pass the litmus test put forward by conflicting or competing facts*. The protocol must provide evidence that the proposed research would stand up to this demand and would help minimize the present uncertainties regarding the phenomenon under study.

1. Fisher AA, Foreit JR. *Designing HIV/AIDS Intervention Studies: An Operations Research Handbook*. Population Council, 2002: p. 8.

## proximal and distal measures of health and disease, *see also* primordial factors



Proximal measures of health are those that directly measure the state of health in a person or a community. Common examples are various morbidity and mortality rates. Physiological and pathophysiological processes that lead to disease and infirmity are also included among proximal measures. Distal measures are those that indirectly measure health. These are the measurement of those factors that predispose the health condition and give rise to proximal measures, such as diet, smoking, and hygiene. Genetic and environmental factors such as pollution and water supply contribute to this process of reaching to proximal from distal. Assessment of social factors such as income, occupation, and education is also included among distal measures.

Proximal measures of health such as **morbidity indicators** and **mortality rates** are described separately in detail. We have also discussed **growth indicators of children** and indicators of **adolescent health**, **adult health**, and **geriatric health**. These are all proximal measures, although many of these apply to communities and not to individuals. Distal measures are also discussed separately under the topics **social health (indicators)** and **mental health (indicators of)**. Holistic definition of health includes these two aspects, among others. **Dietary indices** and **smoking index** are also presented in this volume. Among environmental indicators, the most relevant for health and medicine is **health infrastructure**. There is a full section devoted to these indicators.

## *P*-values

In the context of statistical tests of significance, the *P*-value represents the probability that our sample gives the observed effect (or more extreme) when, in reality, such an effect does not exist in the concerned population. Nonexistence of the effect is called the null hypothesis, denoted by $H_0$. Thus, the *P*-value indicates how compatible your sample values are with the $H_0$. In notations, *P*-value = $P$(sample values|$H_0$). If this probability is far too small, the null is considered implausible. Statistically, this is termed as the probability of rejecting the null when it is true. Some of you know that this is the probability of **Type I error**. The concept of *P*-value was introduced by Fisher in the 1920s, possibly to judge whether or not the evidence is significant in the sense that it requires a second look, but it has imbibed into a solid framework for evidence-based medicine. He used the lowercase *p* to denote this probability [1], which many continue to use, but we prefer the uppercase *P* to distinguish it from sample proportion *p*.

*P*-values are an integral part of hypothesis testing as a means for statistical inference from sample to population. Hundreds of papers and blogposts have been written about what some statisticians deride as "null hypothesis significance testing" [2]. They are being widely used and also widely abused. Sufficient precautions should be exercised in interpreting a *P*-value. First, the null is not necessarily zero effect. It could be equal to a certain quantity, at least this much, or at most this much. Second, this is not the probability of the null being true given the sample values, i.e., this is not $P$($H_0$ true/sample values). If you notice what we mentioned earlier, *P*-values are its statistical inverse. Thus, it is not correct to say that this is the chance that the null is true. Third, it only tells whether to reject or not reject but does not say what to accept. This is one of the severest criticisms of *P*-values. Fourth, the *P*-value is customarily compared against an arbitrary threshold such as 0.05, called the **level of significance**, and conclusions are made. Purists do feel concerned about such "arbitrary" inferences. Despite all these problems, *P*-values seem to have come to stay as the main consideration to draw data-based inferences.



### *One-Tailed and Two-Tailed* P-*Values*

*P*-values depend on whether a test is for **one- or two-tailed alternatives**. Hemoglobin (Hb) level after giving a hematinic to anemic patients is a situation where a one-tailed test is appropriate because biological knowledge and experience confirm that iron supplementation cannot reduce Hb level at least on average. Use of the two-tailed test in this case makes it unnecessarily restrictive and makes rejection of $H_0$ more difficult. However, in most medical situations, assertion of one-sided alternative is difficult and a two-sided test is needed. Most statistical packages provide two-tailed *P*-values as a default, and many workers would not worry too much about this aspect. Scientifically, a conservative test does not do much harm, although some real differences may fail to be detected when a two-tailed test is used instead of a one-tailed test. It is advised to use a one-tailed test only where a clear indication is available in its favor, but not otherwise.

### *General Method for Obtaining the* P-*Value*

Step 1. Set up a null hypothesis and decide whether the alternative is one-sided or two-sided. Also decide the level of significance $\alpha$ and thus fix the threshold of Type I error that can be tolerated for the problem in hand.

Step 2. Identify a criterion, such as Student *t* and chi-square, suitable for the setup in hand. These criteria are also called **tests of significance**. The distributional form of these criteria has been obtained and is known. The exact criterion for obtaining the *P*-value depends mostly on (i) the nature of the data (qualitative or quantitative); (ii) the form of the distribution such as Gaussian or non-Gaussian when the data are quantitative; (iii) the number of groups to be compared (two or more than two); (iv) the parameter to be compared (it can be the mean, median, correlation coefficient, etc., in case of quantitative data; it is always a proportion $\pi$ or a ratio in case of qualitative data); (v) the size of the sample (small or large); and (vi) the number of variables considered together (one, two, or more). All this is analogous to saying that the criterion for assessing the health of a person depends on the age, gender, purpose, general health or organ-focused health, etc.

Step 3. Use sample observations to calculate the value of the criterion *assuming that the null hypothesis is true.*

Step 4. Compare the calculated value with its known distribution, and assess the probability of occurrence of a value of the criterion that is *as extreme as or more extreme toward* $H_1$ than that obtained in step 3. This probability is the *P*-value. This is calculated for both the negative and positive sides when the alternative is two-sided. Since the comparison is with the distribution under $H_0$, a probability that is not very low indicates that the sample is not inconsistent with $H_0$, and thus $H_0$ cannot be rejected. In this case, sampling fluctuation cannot be ruled out as a likely explanation for the values observed in the sample. A very low probability indicates that the observed values are very unlikely to have come up from a population where the null is true—thus $H_0$ is rejected.

Statistical tables give different values of various criteria for popular threshold $\alpha$, particularly 0.05. These are called the **critical values**. In fact, what is required is the *P*-value for each value of the criterion, but that is difficult to tabulate in most cases. Most statistical software packages give exact *P*-values associated with the value of the criterion obtained.

Step 5. Reject $H_0$ if *P* is less than the predetermined level of significance. Generally, $P < 0.05$ is considered low enough to reject $H_0$. Statistical significance is said to have been achieved when $H_0$ is rejected. Such a result can be stated in a variety of ways:

- The evidence against the null hypothesis is sufficient to reject it.
- Sample values are not consistent with the null hypothesis.
- Sampling fluctuation is not a likely explanation for the hypothesized effect.
- The alternative hypothesis is accepted.
- The *P*-value is less than a predetermined threshold such as 0.05.
- The probability of wrongly rejecting $H_0$ (Type I error) is very small.
- The result has achieved statistical significance.

### P-*Values for Nonrandom Sample*

Remember that all statistical methods of inference such as confidence interval and test of significance require random sample of subjects. Patients coming to a clinic during a particular period can be considered a *random* sample from the population of patients that are currently coming to that clinic. But this limited definition of

population is sometimes forgotten, and generalized conclusions are drawn on the basis of *P*-values. This is quite frequent in medical literature and mostly accepted without question.

Two more points need to be stressed in this context. First, a sample can rarely be fully representative of the target population in a true sense. Thus, the risk of a wrong conclusion at $\alpha = 0.05$ is, in fact, slightly more in many cases than it is in case of truly random samples. This probability can also be affected by, for example, non-Gaussian distribution of the measurement under consideration. Second, *P*-values have no meaning if no extrapolation of findings is stipulated. Conclusions based on a sample *for the sample subjects* can be drawn without worrying about *P*-values.

### P-*Value Threshold: The Level of Significance*

A threshold of 0.05 in the Type I error is customary in health and medicine. Except for convention, there is no specific sanctity of this threshold. There is certainly no cause for obsession with this cutoff point. A result with $P = 0.051$ is statistically almost as significant as one with $P = 0.049$, yet the conclusion reached would be very different if $P = 0.05$ is used as the threshold. Borderline values always need additional precaution.

A value close to the threshold such as $P = 0.06$ can be interpreted both ways. If the investigator is interested in showing the presence of difference, he/she might argue that this *P* approaches significance. If the investigator is not interested, this can be easily brushed aside as not indicating significance at $\alpha = 0.05$. It is for the reader to be on guard to check that the interpretation of such borderline *P*-values is based on objective consideration and not driven by bias. The authors generally interpret $P = 0.06$ or 0.07 as encouraging though not quite good enough to conclude a difference. They can be called marginally significant. If feasible, wait for some more data to come and retest the null in this situation.

The second problem with threshold 0.05 is that it is sometimes used without flexibility in the context of its usage. In some instances, as in the case of a potentially hazardous regimen, a more stringent control of Type I error may be needed. Then $\alpha = 0.02$, 0.01, or 0.001 may be more appropriate. It is not necessary to use $\alpha = 0.01$ if a threshold less than 0.05 is required. The value $\alpha = 0.02$ can also be used. In some other instances, as in examining the presence of differences in social characteristics of the subjects, a relaxed threshold $\alpha = 0.10$ may be appropriate. For most physiological and pathological conditions, however, the conventional $\alpha = 0.05$ works fine, and that is why it has stayed as a standard for so long.

The practice now generally followed is to state exact *P*-values so that the reader can draw his/her own conclusion. A value of *P* around 0.10 can possibly be considered weak evidence against the null hypothesis, and a small *P*, say less than 0.01, as strong evidence. Any *P*-value more than 0.10 is considered as no evidence of any consequence.

### *Other Problems with* P-*Values*

Statistical *P*-values seem to be gaining acceptance as a gold standard for data-based conclusions. However, biological plausibility should not be abandoned in favor of *P*-values. Inferences based on *P*-values can also produce a biased or incorrect result just as misdiagnosis and missed diagnosis happen in clinical setting. Also see the topic **misuse of statistical tools** for examples of misuse of *P*-values.

Attempts are sometimes made to dramatize the *P*-values. James et al. [3] stated $P < 0.00000000001$ for difference in seroconversion rate against Epstein–Barr virus between patients with lupus and controls. It is pulling out a number from a computer output without being careful about its implications. Such accuracy is redundant. It really does not matter whether $P < 0.001$ or $P < 0.000001$ as far as its practical implication is concerned. Many statistical software packages rightly stop at three decimal places under default and give $P = 0.000$ when it is exceedingly small. It only means that $P < 0.0005$, and this is enough to draw a conclusion. Further decimal places are rarely required. But do not interpret $P = 0.000$ as $P = 0$ as no *P*-value can be zero.

1. Fisher RA. On the interpretation of $\chi^2$ from contingency tables, and the calculation of p. *J Royal Stat Soc* 1922;85(1):87–94. http://www.jstor.org/stable/2340521
2. Leek JT, Peng RD. Statistics: *P* values are just the tip of the iceberg. *Nature* 2015;520:612. http://www.nature.com/news/statistics-p-values-are-just-the-tip-of-the-iceberg-1.17412
3. James JA, Kaufman KM, Farris AD, Taylor-Albert E, Lehman TJA, Harley JB. An increased prevalence of Epstein–Barr virus infection in young patients suggests a possible etiology for systemic lupus erythematosus. *J Clin Invest* 1997;100:3019–26. http://www.ncbi.nlm.nih.gov/pubmed/9399948

## publication bias

There is a tendency for studies with positive results to find their way preferentially into the literature and studies with negative results not finding their place: this makes for the phenomenon named publication bias. Publication bias is a form of selection bias for studies that yield a particular type of result. This phenomenon has been convincingly demonstrated by several authors (for example, see Boulesteix et al. [1]). When published literature is solely used as the basis for drawing conclusions about a treatment, it is possible that a biased impression in favor of the treatment could result. This is possible for both informal reviews of the literature and more quantitative ones such as meta-analyses.

The fact remains that the literature is unduly loaded with positive results. Negative or indifferent results are either not sent for publication or are not published by the journals with the same frequency. Many journals are much too keen to publish reports that give a positive result regarding efficacy of a new regimen compared with the negative trials that did not find any difference. If there are 20 studies—8 with positive and 12 with negative results—may be 7 with positive results will be published, but possibly only 2 with negative results. Thus, you will get the impression that 9 studies were conducted and 7 have positive results. If a vote count is done on the basis of the published reports, positive results would hugely outscore the negative results, although the fact may be just the reverse. Thus, any conclusion based on commonality in publications can magnify this bias.

In principle, publication bias should not happen. If a trial yielding a negative finding but addressing an important therapeutic question has been conducted rigorously, and it is written up as it should be, then the resulting information will be useful to other researchers and practitioners. It deserves to be published as much as the studies with positive results that appear to show a therapeutic advance. Investigators, however, often lose enthusiasm for trials with negative results because they seem less glamorous than positive ones and may be viewed as failures by the research team and sponsors. If this happens, then it can lead to weaker reports. Song et al. [2] report a meta-analysis of different types of research studies for publication bias. The pooled odds ratio of publication of studies with positive results, compared to those without positive results, was 2.78 in studies that followed from inception, 5.00 in trials submitted to regulatory authority, 1.70 in abstracts, and 1.06 in other manuscripts. Thus, the problem seems more severe for clinical trials. Some journal editors seem to prefer to publish positive studies because in so doing

they might improve the standing of the journal. This is paradoxical, however, because if we assume that true treatment advances are uncommon, then positive reports are more likely to be in error than negative ones.

There is no one way for the readers of a report to correct for publication bias. You must recognize that selection bias may exist and retain some skepticism about a conclusion with positive results. In a few cases, readers might be aware of unpublished studies that fail to support a particular finding: these can be used to soften enthusiasm. It may be useful to ask if the paper in question would have been published if it was actually a negative trial. In other words, is the quality of the study good enough to convey externally valid information consistent with no effect? Even so, there is no way to separate true from false positive results without replicating the study several times. When conducting exhaustive quantitative overviews in systematic reviews and meta-analyses, it is important to include data from both the published and the unpublished trials to counteract such publication bias whenever unpublished reports of the trials are available. That meta-analysis tends to provide more valid and reliable results by pooling results from different publications is accepted as a good statistical technique, but the validity of the results depends on proper representation of studies with all kinds of results.

1. Boulesteix AL, Stierle V, Hapfelmeier A. Publication bias in methodological computational research. *Cancer Inform* 2015 Oct 15;14(Suppl 5):11–9. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4608556/
2. Song F, Parekh-Bhurke S, Hooper L, Loke YK, Ryder JJ, Sutton AJ, Hing CB, Harvey I. Extent of publication bias in different categories of research cohorts: A meta-analysis of empirical studies. *BMC Med Res Methodol* 2009 Nov 26;9:79. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2789098/

## purchasing power parity dollars

The most widely used measure of the level of income of a community is the per capita gross domestic product, popularly called the per capita income. Almost all countries compute this for the nation as a whole and usually also for each of their states separately. Further subdivisions may not be easily available. It is to be expected that the higher the level of income of a community, the healthier that community.

Income per se in the national currencies is seldom comparable across countries because of huge differentials in purchasing power. A Singapore dollar cannot buy as much in Singapore as the Yuan equivalent can in China. The World Bank has done considerable work in this sphere and obtained incomes of various countries in terms of purchasing power parity (PPP) in international dollars. An international dollar has the same purchasing power in the country as the US dollar has in the United States. Thus, an average per capita income of nearly 50,000 rupees per annum in 2013 in India was equivalent to Int$5350 [1]. That is, an Indian rupee in India is worth nearly 10 cents in the United States in terms of purchasing power. For Austria, the per capita income is Int$45,040. You can say that the average income of a person in Austria is nearly 8 times of that in India. Note, however, that income is mostly a family trait rather than of an individual because children are not supposed to earn. If the average size of family is four, the average family income is four times the per capita income.

PPP comes in handy for comparing cost of health care. If a cancer treatment costs Rs 500,000 in India and US$50,000 in the United States, they are nearly the same in terms of purchasing power. PPP is also used for assessing income across countries and its impact on health infrastructure or its utilization. Alkire et al. [2] have used PPP for projecting the value of lost output due to surgical conditions across 128 countries for which the data were available. Kalo et al. [3] have worked out the financial benefits of clinical trials in Hungary in terms of PPP dollars.

1. The World Bank. *GNI per Capita, PPP (Current International $)*. http://data.worldbank.org/indicator/NY.GNP.PCAP.PP.CD, last accessed February 12, 2016.
2. Alkire BC, Shrime MG, Dare AJ, Vincent JR, Meara JG. Global economic consequences of selected surgical diseases: A modelling study. *Lancet Glob Health* 2015 Apr 27;3 Suppl 2:S21–7. http://www.thelancet.com/journals/langlo/article/PIIS2214-109X(15)70088-4/abstract
3. Kaló Z, Antal J, Pénzes M, Pozsgay C, Szepezdi Z, Nagyjánosi L. Contribution of clinical trials to gross domestic product in Hungary. *Croat Med J* 2014 Oct;55(5):446–51. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4228288/

## purposive sample, *see* sampling techniques

## putative factors

"Putative" has the meaning of being part of conventional wisdom: "generally considered or reputed to be" such as putative father. This is related more to "reputation" than real. You can see that putative factors may not be the actual operational factors but are those that are widely considered to affect an outcome. These are perceived, believed, or assumed to affect an outcome, may not be necessarily supported by hard evidence, and may not be biologically related to the outcome. The term is mostly used for risk factors. On many occasions when the actual risk factors are not known, a beginning can be made with putative factors. In some situations, the interest may be restricted to putative factors to find if any or some of them are for real, have supportive evidence, and can be explained with biological mechanism.

Putative factors can best be understood by reference to the examples given by Mascheretti et al. [1] and Fechner et al. [2]. In the former, the authors found that although dyslexia runs in families, several putative risk factors, which could not be immediately identified as genetic, predicted reading disability in dyslexia patients in Italy. Putative factors they considered are younger parental age at child's birth, lower parental education, and risk of miscarriage. These factors significantly increased the odds of belonging to the dyslexia group. These findings supported the result that reading disabilities are a multifactorial disorder and may bear some importance for the prevention and/or early detection of children at heightened risk for dyslexia. In the second paper, the authors studied putative risk factors for colorectal cancer such as fecal mass, transit time, fecal pH, and secondary bile acid concentration. In so doing, they tested the hypothesis that high intake of dietary fiber has been associated with a lower risk of colorectal cancer. It was concluded that enhancing dietary fiber intake through consuming blue lupin fiber up to about 50 g/day can be recommended.

1. Mascheretti S, Marino C, Simone D, Quadrelli E, Riva V, Cellino MR, Maziade M, Brombin C, Battaglia M. Putative risk factors in developmental dyslexia: A case-control study of Italian children. *J Learn Disabil* March/April 2015;48(2):120–9. http://ldx.sagepub.com/content/48/2/120.refs
2. Fechner A, Fenska K, Jahreis G. Effects of legume kernel fibres and citrus fibre on putative risk factors for colorectal cancer: A randomized, double-blind, crossover human intervention trial. *Nutrition J* 2013;12:101. http://www.nutritionj.com/content/12/1/101

# Q

## Q–Q plot, *see* quantile-by-quantile (Q–Q) plot

## *Q* test, *see* Cochran *Q* test

## quadratic regression, *see* curvilinear regression

## qualitative measurements, *see also* quantitative measurements

The measurements that do not have a metric, such as site of injury ("measured" as head, chest, abdomen, etc.) and blood groups (measured as O, A, B, and AB), are called qualitative measurements. This is in contrast to quantitative measurements such as blood pressure, hemoglobin, creatinine, and size of kidney, which are measured in numeric metric. Qualitative measurements could be dichotomous or polytomous depending on whether the number of categories is two or more than two, and polytomous could be further divided into ordinal (such as the severity of disease being mild, moderate, serious, and critical) and nominal (with no order, such as site of cancer).

Ordinal qualitative measurements are statistically quite challenging since they actually are a manifestation of the underlying continuum. If an appropriate scale is available, we would like to measure them quantitatively. Disease severity in critical patients can be possibly measured by the APACHE score, but such scoring is not available for many diseases, and the validity of available scores is questionable in some cases. When no such scale is available, signs–symptoms and investigations are used to classify the subjects into one of the many possible ordinal groups. Even when a clear quantitative scale is available, many clinicians prefer to label their patients as mild or moderate or serious. This possibly helps them decide how to manage the patient. The most prominent examples of this are univariate diseases such as hypertension, diabetes, and anemia, which are the names of conditions indicating that a particular measurement has crossed a threshold. Problems in this are apparent since a person with a fasting blood sugar level of 119 is not diabetic and a person with a fasting blood sugar level 121 is if the threshold is 120 mg/dL. Such categories defy statistical logic, but at the same time, there is no denial that clinicians do need a threshold beyond which a treatment must be instituted.

In addition to what is just stated, note that a variable such as birth weight is also quantitative but is treated as qualitative when the interest is in knowing the percentage of newborns with weight <2500 g, 2500–3499 g, and ≥3500 g. This is an example of a quantitative variable converted to a qualitative one, where the concern is with the proportions of subjects in different categories and not mean or median. The actual scale can be metric, but the variable becomes qualitative when such categories are formed and the interest shifts to proportions. There is a rider though. For the interest in proportion to sustain, the number of such categories must be small. If, instead of three broad categories, the birth weight is divided into a large number of 100-g categories such as <2000, 2000–2099, 2100–2199, and so on, the interest would rarely be in a proportion of births in these categories and would mostly be in parameters (such as mean, median, standard deviation, percentiles, etc.) of birth weight. The categories are not qualitative then.

Statistical inferential methods for qualitative variables are different from those for quantitative variables. This is analogous to the assessment methods that are different for cardiovascular diseases from those for gastrointestinal tract diseases. These methods are based on **binomial distribution** for dichotomous categories and on multinomial distribution for polytomous categories. Dichotomous categories, where a quantitative measurement is divided into two groups, are a class apart from the nominal dichotomous categories. Examples of nominal dichotomous are male–female, case–control, and cancer of prostate or cancer of esophagus. These have no order—no single category is better or superior to the other. These are the situations where binomial distribution can be legitimately applied with full authenticity when the categories are **mutually exclusive and exhaustive**. Quantitative dichotomous can also be handled with binomial distribution, but sometimes the underlying distribution can also be used to provide additional information. For example, if you know that fasting blood glucose level follows a Gaussian (Normal) distribution, the threshold 120 mg/dL is not simply a yes/no type but has additional information such as how far this is from the mean or mode in absolute value or in SD units. The same applies to multinomial distribution for polytomous ordinal categories based on quantitative measurements.

## qualitative research, *see also* focus group discussion

This term refers to the research almost entirely based on **qualitative measurements** with an additional feature that the qualities are far too disperse and varied. This type of research does not focus on quality of instruments but generally concentrates on behavioral and mental aspects such as opinion, perception, belief, motivation, expectation, satisfaction, and reaction that cannot be easily measured in terms of quantities. The focus in qualitative research is on the why and how of the decision process instead of the what and where. Such characteristics may defy hard measurement but are important to provide the context and may help in explaining the reasons behind controversies.

The most common tool of qualitative research is **focus group discussion**, where a knowledgeable group of people is assembled and engaged in a discussion on a topic of interest. The theme of the discussion is captured by the terms and concepts most commonly used. The methodology is still developing, and it basically remains descriptive and exploratory, if not anecdotal. Many consider this unscientific because the same results cannot be obtained if the discussion is repeated; repeatability is a sheet anchor of scientific research. The extent to which biases and prejudices affect this evidence is still an area of conjecture.

Qualitative research on an audience is usually audio-recorded and transcribed. The transcriptions form the data that are then analyzed by the moderator. Group discussions can be video-recorded

and can also be observed in real time at venues specially designated for that purpose.

In many situations, medical decisions are taken on soft considerations that can be assessed only by qualitative research. Issues such as refusal to undergo a surgery that has demonstrable benefits have heavy personal overtones and can rarely be explained by hard facts. Ignoring such social issues can be perilous in some situations. The challenge is that these contexts are highly variable from person to person and can quickly change within a person also from time to time depending on the context. Experience and circumstances sometimes play an important role in determining such reactions.

Qualitative research rarely give statistically sound findings since the methods used in this kind of research, such as focus group discussion, in-depth interviews of a small, nonrandomly selected key persons, observation of activity, and reactions in a specified situation, are highly subjective. Thus, this is not a substitute for quantitative research, although it can be a useful adjunct in studies that require *soft data*. For example, in-depth interview in an intimate environment may allow the participant to talk openly about obscure issues such as about competitors with whom discussion in a group cannot be held. This can unravel facets that would otherwise rarely come to the fore.

There is no denying that we need qualitative research in health and medicine. This is a science that has profound impact on psychosocial aspects and is also deeply affected by such factors. The methodological standards and the guidelines for qualitative research in medicine and health care remain too sketchy to help one conduct such a study with conviction and critically evaluate a qualitative study. Better methodologies may develop in the future, but till such time, the results of qualitative research must be interpreted with caution. For a review of this topic, see Poses and Isen [1]. For an application, see Vazquez et al. [2] who did qualitative research to understand the phenomenon of nonadherance to dental treatment by adolescents in Brazil.

1. Poses RM, Isen AM. Qualitative research in medicine and health care: Questions and controversy. *J Gen Intern Med* 1998 Jan;13(1):32–8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1496891/
2. Vazquez FL, Cortellazzi KL, Gonçalo CD, Bulgareli JV, Guerra LM, Tagliaferro ES, Mialhe FL, Pereira AC. Qualitative study on adolescents' reasons to non-adherence to dental treatment. *Cien Saude Colet* 2015 Jul;20(7):2147–56. http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-81232015000702147&lng=en&nrm=iso&tlng=en



## quality control in medical care, *see also* control charts, lot quality assurance scheme

Quality in the statistical sense means meeting a specified standard. It is always desirable to aim at perfection, but that is quite often elusive. For example, it would be wonderful to have a hospital that discharges fully cured patients all the time without any death or disability, but that is impossible to achieve. Thus, a lower standard of, say, a 90% cure rate in casualty cases and 98% in routine admissions may have to be fixed. Limitations of knowledge and cost considerations often determine this feasibility level. Once a standard—high or not so high—is fixed, everything possible should be done to achieve it, and this is best done by controlling errors.

Errors are distinct from variation, the latter being endogenous whereas the former being exogenous. Variability remains; it is only their impact that can be minimized. On the other hand, the errors themselves can be largely avoided by being sufficiently careful, although it may not be easy to eliminate them fully. A standardized mercury manometer would still give variable readings in different individuals, but occurrence of air bubbles in the mercury column produces an error. Clinicians may genuinely differ on the dose of a drug to be given to a particular patient, but giving the drug two times when the prescription says three times a day is an error.

### *Statistical Quality Control in Medical Care*

Medical care is a big industry in many parts of the world. Providing good care is a prerequisite for this industry to maintain and increase clientele and perhaps profits, but the primary concern is with the humanitarian aspect of doing the best to prolong life and reduce suffering in an activity that deals with the life and health of people. In this context, quality of medical care assumes importance much more than in other industries. Perhaps millions of episodes of illness and deaths around the world every year can be attributed to medical errors. According to one estimate, nearly one in three patients encounter problems due to medical errors during their hospital stay [1].

A patient sometimes has to undergo an array of steps while seeking treatment: the patient describes complaints, responds to questions about the history of the problem, talks about environmental exposures, discusses extent of disabilities, and so on. The attending clinician gears questions to meet the requirement of the situation in terms of the condition of the patient, the type of complaints, and the intelligence level of the patient. The clinician also examines the patient as required and obtains data on body temperature, heart rate, blood pressure, and weight. An interim therapy is sometimes started, and laboratory and radiological investigations are ordered, if considered desirable. The laboratory and radiological units carry out these investigations and report findings. The clinician reassesses and sometimes the cycle restarts. In a hospital setup, outpatients receive their supply of medicines from a pharmacy according to the prescription and ingest the drug, whereas inpatients are administered prescribed dosages by the nursing staff. Some patients undergo surgery where a series of steps are undertaken in the preoperative ward, operation theater, and postsurgical care unit. A hospital does all this for a large number of patients day after day. It is unrealistic to expect that all steps will be correctly done in all cases all the time. Errors do occur. The question is whether these errors are far too many or too large, or are within a tolerance limit. This tolerance has a statistical nature because it is determined by the experience. Before this is discussed, a brief overview of steps that could enhance the quality of medical care may be helpful.

Edward Deming is among those who were outstanding in promoting quality control. He suggested many steps to keep a check on quality. Some of these, in the context of medical care, are as follows [2]:

- Prepare operational definitions of the services to be provided, specify the standards of service, and formulate clear guidelines about the identification of patients to be served and to be excluded by referral or otherwise.
- Refuse to accept a higher level of mistakes in diagnosis and treatment for any reason including inadequate instruments, inappropriate technology, lack of expertise, expired drugs, substandard chemicals, and so on.
- Instead of depending on ad hoc inspection, depend on statistical evidence of quality of incoming material such as pharmaceuticals, testing kits, and blood. The meaning of statistical evidence is given later in this section.
- Implement a system of self-detection of errors and do not wait for complaints to come.
- Continually update the doctors and technicians on rapidly growing developments and encourage them to acquire new skills.

- Have the objective not of finding cases of fault but of diagnosing the fault so that remedial steps can be taken.
- Drive out fear so that employees feel free to make suggestions. Assign failure to the lacuna in the system rather than to the individuals.
- Instead of setting numerical targets, set work standards in terms of quality. Thus, do not count the patients treated but assess the percentage of patients who were cured and satisfied.
- Top management must be committed to quality and should earnestly implement the preceding suggestions.

A large number of issues related to quality of medical care can be discussed. The focus here is on statistics-based issues rather than on medicine-based issues, and the following discussion is limited to some specific aspects that *illustrate* the help statistical methods can provide in improving the quality of medical care. The same kind of methodology can be used to control other kinds of errors.

If a patient with a disease, the treatment of which is known, is not cured, it would be generally correct to assume that this failure is a consequence of error on the part of the medical care chain. This can be due to carelessness or inadequate expertise of the attending personnel or to the faulty tools and equipment—in some cases to the carelessness of the patient. The present discussion is restricted to deviation such as from the prescription or from the specified procedures so that such errors can be called as medical care errors. In admitted patients, these could be in terms of administration of doses at times other than specified, wrong amount of doses, omission of a dose or giving an extra dose, administration of an unprescribed drug, and so on. Such errors can also occur with other procedures such as obtaining the history; establishing the diagnosis, measurements, and assessments; attending to the complaints of the patients; acting on warning signals; consulting a specialist; carrying out a laboratory investigation; and identifying the side effects.

Medical care errors are difficult to identify. They come to notice mostly when the outcome of a therapy is a sudden deterioration of the condition of the patient rather than gradual improvement. Listed next are some common adverse patient outcomes [3] that indicate that an error has occurred somewhere in patient management, including a clinician's error in judgment about the patient's condition.

- Admission due to adverse results of outpatient management
- Admission for complication of a problem on previous admission
- Perforation, laceration, or injury to an organ incurred during an invasive procedure
- Adverse reaction to a drug or transfusion
- Unplanned return to the operation theater
- Infection developing subsequent to admission, including infection occurring after operation
- Transfer from a general care to a special care unit
- New complications occurring after the start of therapy

This list may convince one that errors in medical care are not uncommon. It is preferable to keep separate track of each type of adverse outcome so that corrective steps can be taken accordingly. For care in the hospital, the indicators can be listed as follows:

- Medication errors per 100 patient days
- Transfusion reactions per 100 units
- Surgery site infection rate per 100 surgeries
- Ventilator-associated pneumonia per 100 ventilator days
- Incidence of bed sores per 100 patients
- Incidence of needle stick injuries per 100 patient days
- Percentage of emergency patients who had to wait for more than 30 min

A hospital's ability to provide the best care for certain types of patients can be difficult to measure because each patient's case is different and outcomes cannot always be predicted. In addition, the collection of data for health care facilities is a challenging, continuously changing science. This makes it important to compare as many quality measurement types as possible so as to form the most complete picture of a hospital's total quality of care before making a final decision.

*Monitoring Fatality*: The adverse outcomes just listed assume that the patient is alive, but some events are bound to be fatal in a hospital despite the best care. **Case-fatality** per 1000 patients admitted can be an indicator of the quality of care in a hospital. Because some hospitals are specialized to treat serious and terminal cases, interhospital comparison on the basis of gross fatality may not be valid. The same is true for different units of a hospital. However, trend within a hospital can certainly be monitored over time using this indicator. Mortality within 48 h in many cases is determined by the condition of the patient at the time of admission instead of the quality of medical care. When this is excluded, the rate is called the *net fatality rate*, a relatively more valid rate for comparison between units or between hospitals. Case-fatality rate, when calculated separately for each disease or condition, may also provide a good indication of the quality of care in different departments of a hospital and may reflect the ability to manage critical patients with different diseases. The other useful indicators of hospital mortality are general anesthesia death rate, which is based on the deaths related to general anesthesia per 1000 patients administered anesthesia. Other indicators are postoperative death rate, hospital maternal death rate (maternal deaths per 100,000 antenatals admitted), hospital early neonatal death rate, and so on. These focus on specific aspects that can be targeted for remedial steps.

It is often difficult to identify a single error that has caused a particular adverse outcome, but if the management is sufficiently careful and the staff is cooperative, analysis of each case of adverse outcome may successfully pinpoint the major cause. Thus, errors of different types can indeed be monitored. For limits of tolerance and related issues, **control charts** that can guide one on how to statistically control such errors are prepared. We have also separately discussed the method of **lot quality assurance scheme** to assess quality of lots such as laboratory kits. Also, see **six-sigma methodology**.

## *Quality Control in a Medical Laboratory*

Besides errors in analysis of the specimen, errors in a laboratory can occur due to incomplete preparation of the patient before sampling, collection of specimen in an inappropriate container, mislabeling or mishandling, delay in transport, improper storage, mistakes in data entry, and, above all, misinterpretation of the physician's order for the test. All these can add up to a formidable chance of error. Imagine the cost of these errors to the patient and medical care system. Wrong reporting can cause immense discomfort and can even cost a life. Ethics is compulsive in this activity as in all others involving health and well-being of people. If you are interacting with a laboratory where a large number of specimens are processed everyday, see if the laboratory is keeping count of such errors and has a mechanism to address the quality issues.

Laboratories differ in their methods, chemicals, skills of the staff, and so on, and thus results for aliquots of the same specimen may differ from laboratory to laboratory. Part of this variability can

be eliminated by standardization across laboratories. Differences also occur within the laboratory from time to time. If such differences are substantial, it shakes the clinician's confidence in the values reported. Quality control helps keep a check on and maintain a high level of performance. This can be done with the help of a **control chart** as discussed under that topic.

Because of widespread computerization, much of the data required for quality control are available online these days. Programs are prepared to issue warning signals when the system seems to derail. This, as well as issues related to multivariate control of quality, is discussed by MacGregor [4], although in the context of manufacturing industry. The application to medical setups is immediate.


1. Classen DC, Resar R, Griffin F, Federico F, Frankel T, Kimmel N, Whittington JC, Frankel A, Seger A, James BC. "Global trigger tool" shows that adverse events in hospitals may be ten times greater than previously measured. *Health Aff (Millwood)* 2011 Apr;30(4):581–9. http://content.healthaffairs.org/content/30/4/581.long
2. Deming WE. *Quality, Productivity and Competitive Position*. Prentice Hall, 1978:pp. 240–5.
3. Demos MP, Demos NP. Statistical quality control's role in health care management. *Qual Prog* 1989 (August):85–9. http://asq.org/qic/display-item/?item=7513
4. MacGregor JF. Using on-line process data to improve quality: Challenges for statisticians. *Int Stat Rev* 1997; 65:309–23. http://onlinelibrary.wiley.com/doi/10.1111/j.1751-5823.1997.tb00311.x/abstract


## quality of data and of measurements, *see also* fallacies (statistical)

Quality of data refers to their **reliability** and **validity**, which includes aspects such as care in obtaining and recording the data; minimization of missing values, misreporting, and biases of the observer; standardized definitions and procedures; and so on. Correct statistical decisions depend not just on choosing the right method of analysis and right interpretation but predominantly on right data so that the **garbage-in garbage-out (GIGO) syndrome** does not operate. Even the most immaculate statistical treatment of data cannot lead to correct conclusions if the data are basically wrong. Before starting an analysis, everything possible should be done to ensure that the information is correct. This mostly depends on the validity and reliability of the tools and instruments (including manpower) used for the collection of information and on the quality of data actually obtained through these instruments.

Since there are a large number of steps a patient or a subject has to pass through after entering a health care web, expertise and attention of the people doing this assessment are critical. For example, in a clinical setup, health is assessed by eliciting the history, conducting a physical examination, carrying out laboratory or radiological investigations, and, sometimes, longitudinally monitoring the progress of the subject, all of which require care and expertise. Health assessment generally requires the use of a questionnaire or a schedule for systematically eliciting and recording information; medical instruments such as a stethoscope, a blood pressure instrument, and a weighing machine; laboratory chemicals, reagents, and methods including machines such as analyzers; x-ray, ultrasound and other imaging devices, and all that go along with them; diagnostic criteria and definitions including scoring systems, if any; and prognostic indicators. Generically, all these are instruments in the statistical sense because data are obtained through them. Various procedures for patient management such as endoscopy, surgery, and treatment regimens also fall under this general term. Under certain conditions, these procedures can also be called tests instead of instruments. No medical instrument is perfect in practice because the performance of the instrument also depends, to a degree, on the human input, besides its own intrinsic properties. The quality of medical decisions depends very substantially on the dependability of these instruments in actual use. This is assessed in terms of the **validity** and **reliability** of the instruments, but quality has other dimensions as well, such as ease in use, simplicity, and being readily interpretable.

Valid and reliable instruments can also give erroneous results when not used with sufficient care. In addition, some errors creep in inadvertently because of ignorance. Sometimes, data are deliberately manipulated to support a particular viewpoint. Also remember what Tukey said: availability of set of data and aching desire for an answer do not ensure that a reasonable answer would be extracted.

### *Errors in Measurement*

Errors in measurement can arise due to several factors as listed below.

*Lack of Standardization in Definitions*: If it is not decided beforehand that an eye will be called practically blind when VA < 3/60, VA < 1/60, or any other, where VA stands for visual acuity, then different observers may use different definitions. When such inconsistent data are merged, an otherwise clear signal from the data may fail to emerge, leading to a compromised conclusion. A similar example is of the variable definition of hypertension used by different workers. In addition, special attention is required for borderline values, for example, Chambless et al. [1] use blood pressure (BP) ≥ 140/90 for hypertension, and Wei et al. [2] use BP > 140/90, where one considers BP = 140/90 hypertensive and the other normotensive. The other such example is age, which can be recorded in completed years (age last birthday) or as on the nearest birthday. These two are not necessarily the same. A difficulty might arise in classifying an individual as anemic when the hemoglobin level is low but the hematocrit is normal. Guidelines on such definitions should be very clear, and all observers involved in a particular research should follow the same definition.

*Lack of Care in Obtaining or Recording Information*: This occurs when, for example, sufficient attention is not paid to the appearance of Korotkoff sounds while measuring BP by a sphygmomanometer or to the waves appearing on a monitor for a patient in critical condition. This can also happen when responses from patients are accepted without probing and some of them may not be consistent with the response obtained on other items. If reported gravidity in a woman does not equal the sum of parity, abortions, and stillbirths, then obviously some information is wrong. A person may say that he/she does not know anything about AIDS in the early part of an interview but states sexual intercourse as the mode of transmission in the latter part of the interview. This is an example of inconsistent responses that require further probing. The observer or the interviewer has to exercise sufficient care so that such inconsistencies do not arise.

Then, there is the question of correct transfer of data in completed forms to the spreadsheet where the collection of data is not online. We have often detected errors in data entry when intimately checked after a suspicion arose. It is mostly for investigators themselves to carefully check the data and ensure that the data are correctly entered. Here is a sample of what we could detect: (i) In one instance, codes for males and females were reversed for some subjects. Such errors are difficult to detect till you find a pregnant male! (ii) In another case, the pretest values were unwittingly swapped with posttest values for some cases at the time of data entry. (iii) In yet another instance, the values for case number 27 were exactly the

same as those for case number 83, which could have occurred as a result of repeat recording of the same subject by two different workers or by wrong renumbering of forms.

To give you a few more practical examples, at the time of entering change from pre- to post-values, the minus sign was inadvertently omitted in some cases. In another instance, calculation of scores was based on wrong columns F, G, H, M, and N in the spreadsheet where the columns actually needed were F, G, H, N, and O, and the scores that turned out did not arouse suspicion. Only when a third person examined the data for some other purpose was this error detected. All these are actual instances and can happen in best of setups. Thus, proper scrutiny of data is a must for valid results. Some such errors may have crept in this book as well and we will fake it as if nothing serious has happened.

Now, some tips on how to check the data. For large data sets, double entry by two independent workers and matching may help in detecting errors that could be subsequently resolved either by going back to the forms or by reasoning. The second method is the range check. If hemoglobin level is typed as 3.2 mg/dL, birth weight as 6700 g, or age at menopause as 67 years, you know these are improbable values and need to be double checked. Whether qualitative or quantitative, frequency tabulation can help in detecting such outliers. The difficulty is in detecting age 32 years wrongly typed as 23 years—when both are equally plausible. If the stakes are high, it is a good idea to double check all the entries with the forms (assuming that information in forms is correct). If the entries are direct online, one source of errors is eliminated but the chance of detecting any error in entry steeply declines. Everything possible should be done for correct entries since these cannot be disowned later when detected.

*Inability of the Observer to Get Confidence of the Respondent*: This inability can be attributed to language or intellectual barriers if the subject and observer come from widely different backgrounds. They may then not understand each other and generate wrong data. In addition, in some cases, such as in sexually transmitted diseases, part of the information may be intentionally distorted because of the stigma or the inhibition attached to such diseases. An injury in a physical fight may be ascribed to something else to avoid legal wrangles. Some women hesitate to divulge their correct age, and some refuse physical examination, forcing one to depend on less valid information. Correct information can be obtained only when the observer enjoys full confidence of the respondent.

*Bias of the Observer*: Some agencies deliberately underreport starvation deaths in their area as a face-saving device and overreport deaths caused by calamities such as flood, cyclone, and earthquake to attract funds and sympathy. Improvement in condition of a patient for reasons other than therapy can be wrongly ascribed to the therapy. Tighe et al. [3] studied observer bias in the interpretation of dobutamine stress echocardiography and concluded that the potential for observer bias exists because of the influence of ancillary testing data such as angina pectoris and ST-segment changes. Lewis [4] found that psychiatric assessments of anxiety and depression requiring clinical judgment on the part of the interviewer are likely to suffer from observer bias. These examples illustrate some situations in which observer bias can occur.

*Variable Competence of the Observers*: Quite often, an investigation is a collaborative effort involving several observers. Not all observers have the same competence or the same skill. Assuming that each observer works to his fullest capability, faithfully following the definitions and protocol, variation can still occur in measurement and in assessment of diagnosis and prognosis because one observer may have different acumen in collating the spectrum of available evidence than the others.

Many inadvertent errors can be avoided by imparting adequate training to the observers in the standard methodology proposed to be followed for collection of data and by adhering to the protocol as outlined in the instruction sheet. Many investigations do not even prepare an instruction sheet, let alone address adherence. Intentional errors are, however, nearly impossible to handle and can remain unknown until they expose themselves. One approach is to be vigilant regarding the possibility of such errors and deal sternly with them when they come to notice. Scientific journals can play a responsible role in this respect. If these errors are noticed before reaching the publication stage, steps can be sometimes taken to correct the data. If correction is not possible, the biased data may have to be excluded altogether from analysis and conclusion.

It is sometimes believed that bad data are better than none at all. This can be true if sufficient care is exercised in ensuring that the effect on the conclusion of bias in bad data has been minimized, if not eliminated. This is rarely possible if the sources of bias are too many. Also, care can be exercised only when the sources of bias are known or can be reasonably conjectured. Even the most meticulous statistical treatment of inherently bad data cannot lead to correct conclusions.

For quality of statistical models, see the topic **robustness**.


1. Chambless LE, Shahar E, Sharrett AR, Heiss G, Wijnberg L, Paton CC, Sorlie P, Toole JF. Association of transient ischemic attack/stroke symptoms assessed by standardized questionnaire and algorithm with cerebrovascular risk factors and carotid artery wall thickness. The ARIC study, 1987–1989. *Am J Epidemiol* 1996;144:857–66. http://aje.oxfordjournals.org/content/144/9/857.full.pdf
2. Wei M, Mitchell BD, Haffner SM, Stern MP. Effects of cigarette smoking, diabetes, high cholesterol, and hypertension on all-cause mortality and cardiovascular disease mortality in Mexican Americans: The San Antonio Heart Study. *Am J Epidemiol* 1996;144:1058–65. http://aje.oxfordjournals.org/content/144/11/1058.full.pdf
3. Tighe JF Jr., Steiman DM, Vernalis MN, Taylor AJ. Observer bias in the interpretation of dobutamine stress echo cardiography. *Clin Cardiol* 1997;20:449–54. http://onlinelibrary.wiley.com/doi/10.1002/clc.4960200509/pdf
4. Lewis G. Observer bias in the assessment of anxiety and depression. *Soc Psychiatry Psychiatr Epidemiol* 1991;26:265–72. http://link.springer.com/article/10.1007%2FBF00789218#page-1




## quality of life index, *see also* physical quality of life index

This index measures quality of life of individuals either as per their own perception or on some objective basis. Thus, this section is restricted to assessment of quality of life of individuals rather than of communities. For communities, see **physical quality of life index**. The focus here is on patients, particularly those suffering from chronic conditions who are not able to lead a normal life.

Myocardial infarction (MI), breast cancer, multiple fractures, and peritoneal surgery are examples of conditions that have many survivors, but quite a few of them are not able to lead a normal life of a healthy person. The disability may be apparent, such as in walking and talking, or more subtle as in doing heavy work for long hours. Quality-of-life assessment is gaining importance as more and more people are able to live longer due to medical intervention but retain residual disability of one kind or the other. It is being increasingly assessed for the general population as well, or for patients of various other types with no disability. This is also commonly used as an outcome measure in research on the relative benefits of different treatment methods.

Quality of life is generally equated with hopes and ambitions matched by experience. It involves a person's own perception and values. Note that this is quite abstract and thus is difficult to measure. Physical, psychological, and social well-being, including functionality in daily living, are generally included in a quality-of-life assessment. In the case of chronic patients, this may contain items on sleep, appetite, sexual functions, social participation, work performance, and so on. It is often considered convenient to divide the quality-of-life questionnaire into domains such as physical health and psychological well-being and to divide a domain into facets, such as psychological well-being into negative and positive feelings, self-esteem, and memory. As the quality of life is mostly the perception of the subject, the rating sometimes may be inconsistent with the actual physical condition such as tumor stage. A patient in an advanced stage of malignancy may still report a good quality of life if he/she has learnt to live with it.

Several instruments that claim to measure quality of life in different kinds of subjects, particularly in patients with chronic ailments, are available. For cardiovascular disease, a multidimensional index based on 35 questions on different domains of quality of life was developed by Avis et al. [1]. For cancer, there is a quality-of-life questionnaire, called EORTC QLQ-C30, containing 30 items of inquiry [2]. Ferrans and Powers [3] developed a quality-of-life instrument whose different versions can be used for a variety of health disorders. Seattle Quality of Life Group has given a big list of adult and children instruments for different conditions on their website [4]. All of them translate quality into an index that can be used for statistical purposes.

For the general population, the World Health Organization has devised a quality-of-life (WHO-QOL) questionnaire with 100 items [5]. This questionnaire is considered too detailed and difficult to answer. More popular is the short form with 36 items called *SF-36*. This measures functional health and well-being. Such questionnaires can be easily downloaded from various websites such as www.sf-36.org but may have to be adapted to the local conditions. A brief questionnaire (WHO-QOL-BREF) [6] with 26 items is also available. There are other indices that measure contextual quality of life such as activities of daily living index [7] for elderly people.

Reliability and validity of such instruments have always been a concern. Reliability is statistically assessed in terms of **Cronbach alpha** for internal consistency and **test–retest** correlations for stability. All types of **validity** (face, content, concurrent, construct, etc.) are also tested in a variety of conditions before such instruments are actually used.


1. Avis NE, Smith KW, Hambleton RK, Feldman HA, Selwyn A, Jacobs A. Development of the multidimensional index of life quality: A quality of life measure for cardiovascular disease. *Med Care* 1996;34:1102–20. http://www.jstor.org/stable/3766565?seq=1#page_scan_tab_contents
2. Sprangers MA, Cull A, Groenvold M, Bjordal K, Blazeby J, Aaronson NK. The European Organization for Research and Treatment of Cancer approach to developing questionnaire modules: An update and overview: EORTC Quality of Life Study Group. *Qual Life Res* 1998;7:291–300. http://link.springer.com/article/10.1023%2FA%3A1024977728719#page-1
3. Ferrans C, Powers M. Quality of Life Index: Development and psychometric properties. *Advances in Nursing Science* 1985;8:15–24. http://www.ncbi.nlm.nih.gov/pubmed/3933411
4. Seattle Quality of Life Group. i. http://depts.washington.edu/seaqol/instruments
5. WHO. *Mental Health: The World Health Organization Quality of Life (WHOQOL).* http://www.who.int/mental_health/publications/whoqol/en/
6. Seattle Quality of Life Group. *World Health Organization Quality of Life Instruments (WHOQOL-BREF).* http://depts.washington.edu/seaqol/WHOQOL-BREF
7. Katz S, Akpom CA. Index of ADL. *Med Care* 1976;14(Suppl 55):116–8. http://www.ncbi.nlm.nih.gov/pubmed/132585


## quantal assays, *see also* bioassays

These are **bioassays** for quantal response. Quantal response are of the yes/no type, such as whether death occurred or not (within the chosen time frame), the person recovered or not, the subject developed side effects or not, the subject developed disease or not, and so on.

The response observed at different doses of a test and a standard regimen is essential to an assay. In the case of quantal assays, the response is the percentage of subjects who developed the intended outcome. For example, this could be of the type that when the dose is 5 mg, 56% responded; when the dose is 10 mg, 69% responded; when the dose is 20 mg, 88% responded; and so on. The initial objective is to find the dose–response relationship using such data. Obviously, the regimen for such assays should be such that a higher dose elicits a progressively higher or progressively lower response but not an erratic response. This would happen with insecticides and such other toxic substance, but this can also happen with drugs when the dose is given between certain limits—for example, the higher the dose between 5 and 15 mg, the higher the response, but beyond 15 mg, the dose is harmful and not tried, and less than 5 mg may be ineffective and not worth giving it a try. Anesthetic agents also have this property—the higher the dose, the more the response, although it can be lethal beyond a point.

As in all bioassays, there is a test and a standard preparation, and the ultimate objective is to obtain an estimate of the **relative potency** as done for parallel-line and slope-ratio assays for quantitative response.

Essentials of quantal assays are explained next in this section. These should make it clear what a quantal assay is and why this is called a quantal assay. We also give an outline of how to estimate relative potency with quantal assays and what are the validity conditions.

### *Setup for Quantal Assays*

As already mentioned, the basic setup in quantal assays is that different doses of a test and a standard preparation are each given to a group of subjects and the percentage who responded with different doses is noted. For example, give 2 mg to 12 subjects, 5 mg to 7 subjects, 10 mg to 20 subjects, and so on, and the number who responded could be 4, 3, and 16, respectively. In this example, the quantal response is $100 \times 4/12 = 33.3\%$ with a 2-mg dose, and similarly, 42.9% with a 5-mg dose and 75.0% with a 10-mg dose. The dose that yields response in 50% of the subjects is called effective dose 50 and denoted by $\mathbf{ED_{50}}$. If the response is death, this is called lethal dose 50 or $LD_{50}$. This is also called **median effective dose** or median lethal dose as the case may be, and this could also be a quantity of interest in addition to the relative potency.

For simplicity, consider only one preparation for the time being. Let the dose metameter be denoted by $x$ as before, which could be the dose itself or log dose or any other transformation, called metameter. The subscripts $s$ and $t$ will be introduced later for standard and test preparations. Let the $k$th dose be denoted by $x_k$ ($k = 1, 2, \ldots, K$), which means that the regimen is tried in $K$ different doses. In the example in the preceding paragraph, $K = 3$ since there are three doses, and $x_1 = 2$ mg, $x_2 = 5$ mg, and $x_3 = 10$ mg. Assume that these are arranged in ascending order so that the first group received the lowest dose, the next group received the next higher dose, the third group received the next higher dose, and so on. The first dose

is tried on $n_1$ subjects, the second on $n_2$ subjects, and so on, and the $K$th dose is tried on $n_K$ subjects. In our example, $n_1 = 12$, $n_2 = 7$, and $n_3 = 20$. Let the number of subjects who responded positively to dose $x_k$ be denoted by $r_k$, which, in our example, are $r_1 = 4$, $r_2 = 3$, and $r_3 = 16$. You can see that doses $x_k$ and number of subjects for each dose $n_k$ are fixed by the experimenter as per the design but the numbers positively responding $r_k$ depend on the sample observations and are not fixed. You can also imagine that quantal assays require many more subjects than quantitative assays because each dose is tried in a *group* of subjects of reasonable size and not just two, three, or four subjects. Thus $n_k$'s would be large.

With these notations, we move on to the usual test and standard preparation setup. Consider a hypothetical example where egg is considered a standard source of dietary protein and rice is the test preparation, both assessed for their digestibility in terms of some scoring. Suppose a digestibility score of at least 15 is considered necessary for the protein to be considered sufficiently good. This is quantal response when assessed as <15 or ≥15. The data obtained are shown in Table Q.1. In this example, there is no blank dose but some assays can have this also. It may or may not be there. For standard preparation, 5 g of egg protein was given to 20 subjects and 8 of these had a digestibility score ≥15. And so on.

For dose–response relationship, the regression required is between probabilities of positive response and the doses. This takes the form

$$\text{(A)} \quad \pi_k = \alpha + \beta x_k \, (k = 1, 2, \ldots, K),$$

where $\pi_k$ is estimated by $p_k = r_k/n_k$. In other words, $p_k$ is the sample analog of $\pi_k$. This regression is restricted to a simple linear form for simplicity of our discussion but can be more complex in some actual situations. We keep those complex situations out of our purview in this book for medical professionals.

As mentioned earlier, the dose metameter could be dose itself or its transformation such as log dose. Response is in terms of percentage (or probability), and we necessarily need a transformation to get an appropriate response metameter that could convert the probability on the left-hand side of Equation A to a continuous quantity whose distribution is known or can be conjectured and easily handled and, more importantly, to achieve linear relationship. Experience has shown that the relationship between $\pi_k$ and dose is rarely linear—it generally takes a shape of what is called a **sigmoid** (S-shaped) curve, particularly because probability has restriction to be between 0 and 1. **Probit** or **logit** transformation is commonly used to achieve linearity. Probit is preferred when the response is continuous with Gaussian distribution such as digestibility in our example and logit is preferred for genuinely binary responses. Most people use logit these days because it can be used for both setups without much compromise on the distributional properties.

**TABLE Q.1**
**Sufficiently Good Response (Digestibility Score ≥15) for Different Doses of Egg and Rice Protein**

| | Receiving Egg Protein (Standard) | | | | Receiving Rice Protein (Test) | | | |
|---|---|---|---|---|---|---|---|---|
| | Dose 1 | Dose 2 | Dose 3 | Dose 4 | Dose 1 | Dose 2 | Dose 3 | Dose 4 |
| $x_k$ | 5 g | 10 g | 15 g | 20 g | 5 g | 10 g | 15 g | 20 g |
| $n_k$ | 20 | 30 | 32 | 24 | 11 | 16 | 23 | 25 |
| $r_k$ | 8 | 12 | 15 | 17 | 3 | 7 | 15 | 20 |

### *Estimation of Relative Potency*

For two preparations, the test and the standard, you will have one logistic regression line of the type mentioned at Equation A for the test preparation and one for the standard preparation. These would take the following form:

$$\lambda_{tk} = \alpha_t + \beta_t x_{tk} \, (k = 1, 2, \ldots, K) \text{ for the test preparation}$$

$$\lambda_{sk} = \alpha_s + \beta_s x_{sk} \, (k = 1, 2, \ldots, K) \text{ for the standard preparation,}$$

where λ is the logit of the corresponding proportion. To keep notations simple, we are describing these for symmetric assays where the number of doses of test preparation is the same $K$ as for standard preparation. However, in these equations, the actual doses of test preparation can be different from the doses of standard preparation. For example, the doses of test preparation can be 2, 5, and 10 mg, and those of standard preparation can be 4, 7, and 12 mg. These correspond to the values of $x_{t1}$, $x_{t2}$, and $x_{t3}$, and of $x_{s1}$, $x_{s2}$, and $x_{s3}$, respectively, in this example. They are not necessarily the same doses but both have three doses each.

When sample values are used, these regressions can be written as

$$y_{tk} = a_t + b_t x_{tk} \, (k = 1, 2, \ldots, K) \text{ for the test preparation}$$

$$y_{sk} = a_s + b_s x_{sk} \, (k = 1, 2, \ldots, K) \text{ for the standard preparation,}$$

where $y_{tk}$ = logit of $p_{tk}$ ($p_{tk} = r_{tk}/n_{tk}$ is the proportion of positive response to the $k$th dose of test preparation) and $y_{sk}$ = logit of $p_{sk}$ ($p_{sk} = r_{sk}/n_{sk}$ is the proportion of positive response to the $k$th dose of standard preparation). The difference between λ and $y$ is that λ is the notation for the population value and $y$ is the corresponding sample value when α's and β's are replaced by their estimates $a$'s and $b$'s.

In the case of quantal assays, the convention is to restrict investigations to **parallel line assays** where the regressions are such that the lines have the same slope but different intercepts as is expected in this setup. The relative potency is estimated as (in log doses if logarithm has been used)

$$\text{Relative potency: } R = \frac{a_t - a_s}{b},$$

where $a_t$ is the estimate of $\alpha_t$ and $a_s$ is the estimate of $\alpha_s$, and $b$ is the common estimate of the regression coefficient β. The common estimate is obtained by using the data on both the preparations under a parallel-line assumption. If the dose metameter is log(dose), then this relative potency is for the log dose, and a conversion back to the original dose would be needed in this case. When the values of $a_t$ and $a_s$ are substituted (see **simple linear regression** for these values), this can also be written as

$$\text{Relative potency: } R = \left(\bar{x}_s - \bar{x}_t\right) - \frac{\left(\bar{y}_s - \bar{y}_t\right)}{b}.$$

You can use either of these expressions to estimate the relative potency. However, this is valid only when the slope of both the lines is the same and the linearity is affirmed. We illustrate this with an example.

Following are the data of a quantal assay of a test and standard insecticide showing the proportion of insecticides killed by different doses using log(dose) as the dose metameter. For simplicity in calculations, we have taken the same number of insects (1000) for every dose of the standard as well as the test preparation. In practice, they can differ—in which case, the formulas will alter. Also, in this example, only three doses of standard preparation and three doses of

test preparation are under trial, although there is no such theoretical restriction.

Plots of proportion responding ($p$) versus dose, as presented in Table Q.2, are shown in Figure Q.1a. Note two things about these: one, they are not straight lines; and two, the plots do not differ much. For the first, Figure Q.1b shows plot of logit of $p$ versus log(dose) (logit of $p$ is not shown in Table Q.2). Now, the plots are nearly a straight line. For the second, we would soon see how the results turn out in case the plots of standard and test preparations are not much different.

For estimating the relative potency, we need the estimate of the intercepts of the two regressions and an estimate of the common slope. Exact logistic calculations require a computer software that give

$$b = 3.5020 \quad \text{and} \quad R = 1.05.$$

Relative potency close to 1 indicates that there is minor difference in potency of the two preprations, if at all. This was quite evident from Figure Q.2 also since the plots for standard and test preparations hardly differed.

### *Checking the Validity of Conditions of a Quantal Assay*

As mentioned earlier, the calculation of relative potency in this case is valid only when the two regression lines have the same slope and the regressions are indeed linear. For this, the hypothesis $\beta_t = \beta_s$ (earlier called parallelism) and the hypothesis for linearity are tested using sum of squares similar to those used for ordinary regression. In the case of quantitative assays, the sum of squares is directly used as chi-square variables and tested for significance as any other chi-square with the requisite df's provided that the sample size is large. For the data in Table Q.2, a software package gives between-preparations sum of squares = 0.4585 that follows $\chi^2$ distribution with 1 df since there are two groups. The critical value of $\chi^2$ at 1 df at the 0.05 level is 3.84. Since the calculated value is less, between-preparations sum of squares is not significant, and the preparations are not different with respect to the killing of insects at different doses. This conclusion corroborates an earlier finding that relative potency of test preparation is 1.05 (nearly the same as of standard preparation).

A statistical software package gives the following for the data in Table Q.2:

Sum of squares due to regression = 6.0046 with 1 df ($P < 0.05$); that is, the slope is statistically significant.
Sum of squares for equality of slopes = 0.0139 with 1 df ($P > 0.05$); thus, equality of slopes can be assumed.
Sum of squares due to linearity = 0.0717 with $(2K - 4) = 2$ df in this example ($P > 0.05$) so that the hypothesis of linearity cannot be rejected.

Thus, this assay fulfills the condition of equality of slopes (parallelism) and linearity of regression, and the procedure followed for the calculation of relative potency is valid.

**TABLE Q.2**
**Data of a Quantal Assay on Insecticides**

| | Standard Insecticide Dose (μg) | | | Test Insecticide Dose (μg) | | |
|---|---|---|---|---|---|---|
| **Particulars** | **100** | **200** | **400** | **50** | **150** | **300** |
| Log(dose)—dose metameter ($x$) | 2.00 | 2.30 | 2.60 | 1.70 | 2.18 | 2.48 |
| Number of insects ($n$) | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 |
| Number killed (response) ($r$) | 200 | 500 | 700 | 100 | 400 | 600 |
| Proportion response ($p$) | 0.2 | 0.5 | 0.7 | 0.1 | 0.4 | 0.6 |

## quantile-by-quantile (Q–Q) plot

This is a plot of **quantiles** of one data set with the quantiles of another data set, mostly to check whether or not the two data sets follow the same distribution. Quantile is the value below which a specified fraction (or percent) of points lie. That is, for example, the 40% quantile is the value below which 40% of the values fall and 60% fall above, and the 65% quantile is the value below which 65% of the values fall and 35% fall above. If the two data sets follow the same statistical distribution, the Q–Q plot should form a diagonal line. The greater the departure from this reference line, the stronger the evidence of difference in the distribution. The two distributions could be either two groups such as males and females, or one could be sample values and the other could be theoretical population values. In the latter case, the Q–Q plot will indicate whether the sample values conform to those expected in the corresponding population.

The Q–Q plot is essentially a large sample method since these quantiles can be unstable for small samples. For this plot, instead of conventional quantiles such as deciles and quartiles, the quantiles are obtained for each value in the data set after arranging the values in ascending order. Quantiles are obtained after standardization so

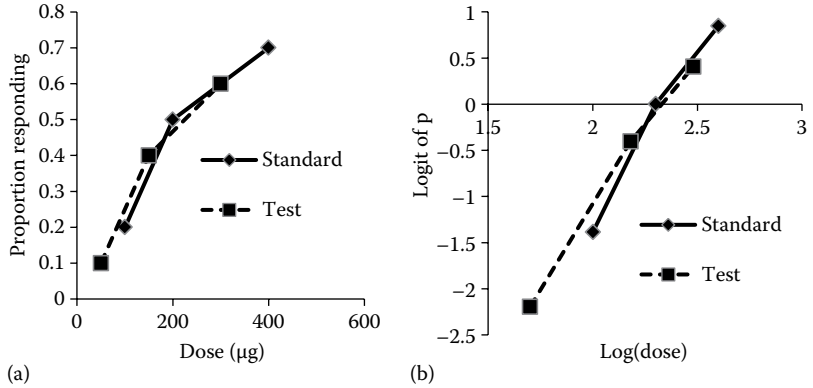



**FIGURE Q.1** Plots for the data in Table Q.2. (a) Proportion versus dose and (b) logit versus log(dose).

that the values in the two groups become comparable. This is of two types: normal Q–Q plot and detrended Q–Q plot.

### *Normal Q–Q Plot*

The feature of this plot can be best illustrated with the help of figures obtained for different types of distributions (Figure Q.2). When the sample sizes from the two distributions are the same, a Q–Q plot is essentially a plot of the sorted values in the two data sets. In case of unequal sizes, the quantiles of the larger data sets are obtained only for quantiles of the smaller data set.

It is customary to consider the plot of a Gaussian (Normal) distribution as standard for comparison. In all these figures, the horizontal axis depicts the observed values. The Q–Q plot for data being Gaussian would be a straight line as shown in Figure Q.2b except for some fluctuations due to sampling. What is important is the trend and not the individual points.

The Q–Q plot is quite sensitive to many types of differences between the two distributions under comparison. If they have different means but have the same shape, the points will still lie on a straight line but that line is displaced up or down from the reference line. For a positively skewed distribution, the plot will bend upward in the middle (Figure Q.2e), and for a negatively skewed distribution, the bow will be the reverse (Figure Q.2h). If the data are symmetric but have a flat peak and shallow tails (low kurtosis) relative to a Gaussian distribution, the plot will have its bottom end downward and its top end upward (see Figure Q.2k) and will have the reverse for relatively high kurtosis. Thus, both skewness and kurtosis can be examined by a Q–Q plot, and outliers can also be detected. Many statistical software packages have provision to provide a Q–Q plot.

Our illustration is in reference to the Gaussian distribution, but that is not a requirement. You can also compare two distributions of any shape. If these are samples, the size need not be equal as already stated. They will essentially yield the same kind of plots as just discussed, and you will be able to judge whether these distributions differ or not and in what respect. It is helpful to keep the vertical and horizontal axis on the same scale so that the diagonal is clearly demarcated at 45°.

A Q–Q plot is just an exploratory graphical device that assesses whether the two distributions have differences and, if so, locates where the differences are. This is not used for conclusive evidence. For conclusion, you would need statistical tests such as **Kolmogorov–Smirnov** and **Anderson–Darling** tests.

### *Detrended Q–Q Plot*

The detrended Q–Q plot shows the differences between the standardized values of the two distributions against the standardized values of one of the distributions. The differences remove the trend and become detrended. These are also popularly known as **worm plots**. If the distributions match, the differences will be randomly distributed along the line at difference = 0 (Figure Q.2c). Any other pattern is a clear indication of deviation from the theoretical distribution. For examples, see Figures Q.2f, i, and l. These detrended plots are even more sensitive to detect departures than the Q–Q plots.

For a useful application of these plots to child growth data, see the WHO report [2].


1. Chan P. *Interpreting the normal QQ-plot*. http://www.youtube.com/watch?v=-KXy4i8awOg
2. WHO. *WHO Child Growth Standards. Length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: Methods and development*. http://www.who.int/childgrowth/standards/technical_report/en/


## quantile regression

Quantile regression aims to estimate a specific **quantile** of a continuous response (dependent) variable in place of mean that an ordinary **regression** estimates for given $x$'s. Just change the mean of $y$ to any specific quantile of $y$ for quantile regression. The quantile can be the 95th percentile, the first quartile, the 7th decile, the 2nd tertile, or any other value. Besides genuine interest in quantile of some measurement, quantile regression can also be used when there are outliers in the data set that make mean unreliable, or when the distribution of $y$ is highly skewed. This kind of regression was introduced by Koenker and Bassett in 1978 [1]. Figure Q.3 shows regressions for 5th, 25th, 50th, 75th, and 95th percentiles versus age for children with sedentary habits. These are linear, but one can examine a quadratic form or any other form as considered appropriate.

There are situations where the interest is not in the mean but a particular quantile. For a highly skewed distribution of the dependent variable, the median may be a better indicator of the central value, and this would be the target for regression. The median is also a quantile. For growth of children, the interest may be in the factors contributing to the 90th percentile of weight of children at any particular age, and in case of terminal diseases, the interest could be in finding the factors responsible for high duration of survival reached by the top 10% of patients or in factors of critically low duration of survival. You may also like to compare the factors contributing to the 50th percentile of heart size versus those contributing to the 80th percentile of heart size. In some situations, the relationship with mean may be weak but that with quantile may be strong. All these situations would require quantile regression. These are examples of univariate regression, but you can also run regression of first, second, and third quartiles together on the same regressors by using multivariate methods. This may be the case with body mass index where both low and high values are watched. This will tell which regressors are important for which quartile. However, methods for this kind of multivariate regression are not fully developed yet.

For the sake of generalizability, let us consider the $q$th quantile. The method of quantile regression is different from comparing, say, the $q$th quantile in two or more groups such as males and females, or those receiving treatment 1, treatment 2, and treatment 3. This can be tested by an analysis of variance–like method with $q$th quantile values as dependent on the groups, keeping in mind that quantiles do not generally follow a Gaussian pattern. For small samples, **permutation tests** can be used and the confidence intervals can be obtained by other nonparametric methods such as those based on bootstrap. This will tell you whether $q$th quantiles are significantly different or not across groups but will not reveal the factors responsible for the difference. For delineating the role of factors, quantile regression can be of definite help.

The method of quantile regression is intricate but can be easily carried out with the help of appropriate software. An outline of the method of quantile regression is as follows. The least squares method followed for estimation of the regression coefficients in ordinary regression is modified to minimize the sum of squares of deviations from the concerned quantile. The solution may have to be obtained by an iterative process. Regression coefficients in quantile regression are interpreted almost the same way as in ordinary regression; for example, for a binary predictor, the intercept in quantile regression is the $q$th quantile of $y$ for the category coded as zero.


1. Koenker R, Bassett GW. Regression quantiles. *Econometrica* 1978;46:33–50. http://links.jstor.org/sici?sici=0012-9682%28197801%2946%3A1%3C33%3ARQ%3E2.0.CO%3B2-J
2. Koenker R. *Quantile Regression*. Cambridge University Press, 2005.

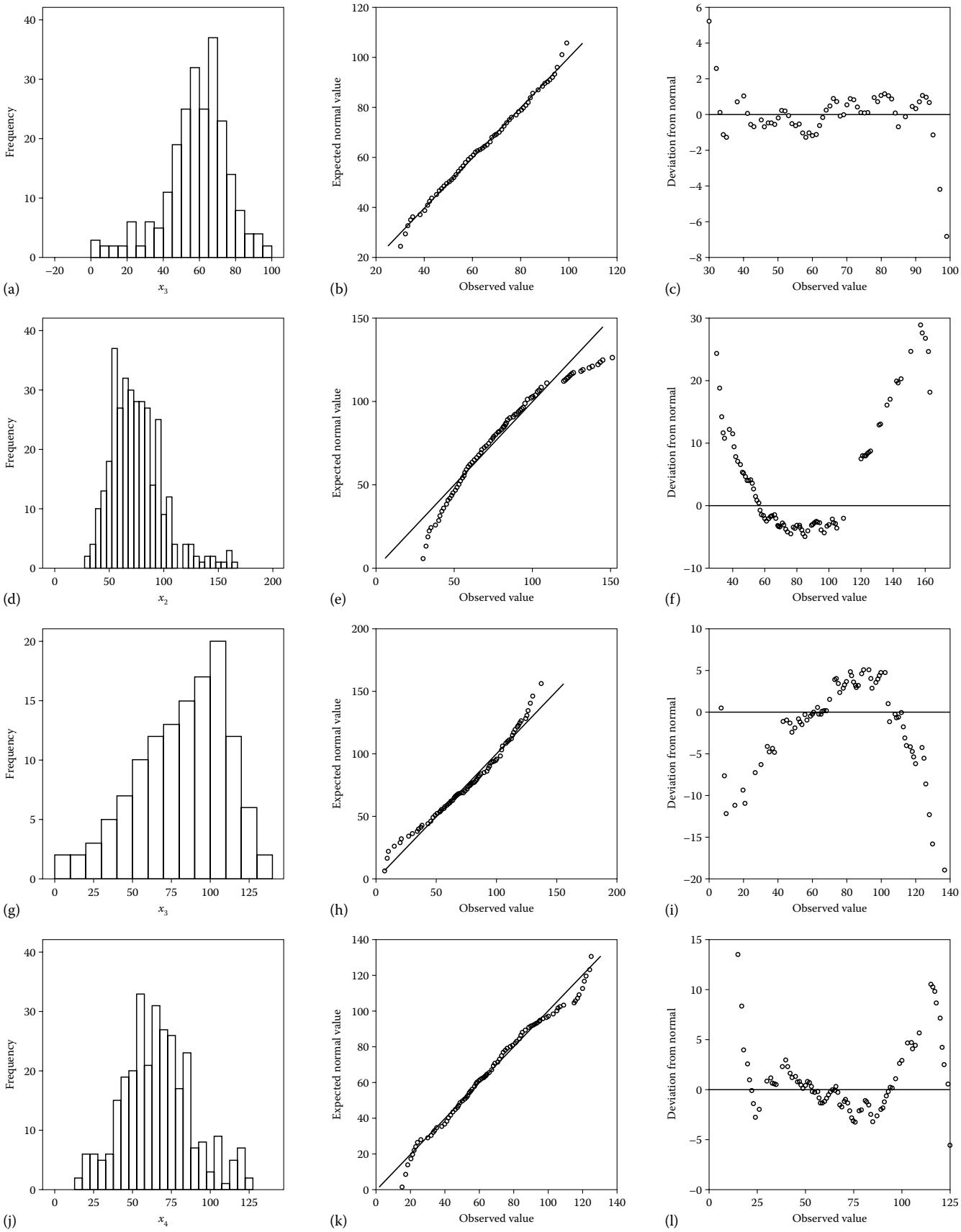



**FIGURE Q.2** Histogram, Q–Q plots, and detrended (worm) plots for different types of distributions: (a–c) nearly Gaussian distribution, (d–f) positively skewed distribution, (g–i) negatively skewed distribution, and (j–l) low kurtosis distribution.

Q

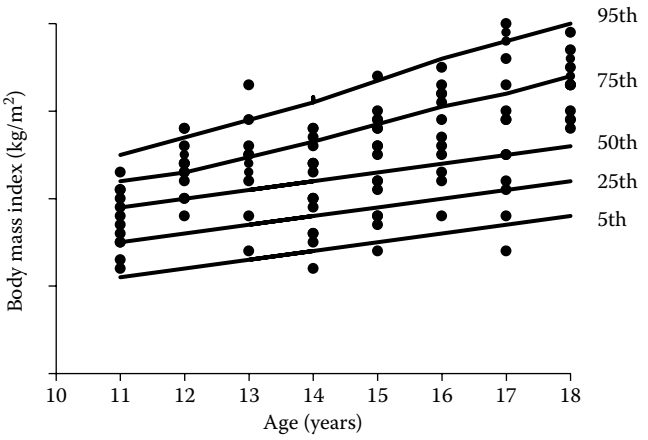



**FIGURE Q.3** Quantile regressions of BMI on age in children.

## quantiles, *see also* percentiles and percentile curves

Quantiles are the values of the variable that divide the total number of subjects into ordered groups of equal size. These are also called **fractiles**. This means that each group so formed will have the same number of subjects. In general, the values dividing subjects into $S$ equal groups may be called the $S$-tiles. The total number of $S$-tiles is $(S - 1)$ as the last $S$-tile is the maximum value itself. For example, the median is the value such that the number of subjects with a value more than the median is the same as the number of subjects less than the median—this is a quantile that divides the subjects into two equal groups. **Tertiles** divide the subjects into 3 equal groups, **quartiles** divide the subjects into 4 equal groups, **quintiles** divide the subjects into 5 equal groups, **deciles** divide the subjects into 10 equal groups, vigintiles divide the subjects into 20 equal groups, and **percentiles** divide the subjects into 100 equal groups. For example, the 9th decile is the minimum value below which at least 90% of the values lay and not more than 10% values will be greater.

Evidently quantiles are primarily meant to be used with quantitative, particularly continuous data. A value $x_p$ is the $p$th $S$-tile if $P(x \le x_p) = p/S$. For example, for a standard Gaussian (Normal) distribution, the 97.5th percentile is 1.96 since $P(z \le 1.96) = 0.975$, and the second tertile is that value of $a$ for which $P(z \le a) = 2/3$. Normal distribution gives $a = 0.43$. Note the equality sign. If the values are quantitative but discrete such as parity, months of gestation, and ranks, quantiles may not be fully defined. This can happen with continuous values also in samples, especially when the sample size is small. Quantiles in ungrouped data are obtained simply in the following manner:

$p$th $S$-tile = $(p^*n/S)$ the value in ascending order of magnitude,

where $n$ is the total number of subjects. For $n = 200$ subjects,

35th percentile = (35 × 200/100) = 70th value
7th decile = (7 × 200/10) = 140th value
3rd quintile = (3 × 200/5) = 120th value
1st quartile = (1 × 200/4) = 50th value
2nd tertile = (2 × 200/3) = 133rd value

in ascending order of magnitude.

Percentiles can be denoted by $P_1$, $P_2$, $P_3$, and so on; deciles can be denoted by $D_1$, $D_2$, $D_3$, and so on; and quartiles can be denoted by $Q_1$, $Q_2$, and $Q_3$. A feature worth noting for all quantiles is that quantile of $(x + y) \ne$ quantile of $x$ + quantile of $y$. This needs to be understood in the context of mean since mean$(x + y)$ = mean$(x)$ + mean$(y)$.

Consider the following 18 values:

23, 12, 56, 25, 34, 43, 12, 7, 49, 27, 34, 45, 28, 14, 19, 17, 16, 28.

After ordering, these are 7, 12, 12, 14, 16, 17, 19, 23, 25, 27, 28, 28, 34, 34, 43, 45, 49, 56. Thus, the 3rd quartile is the 18 × 3/4 = 13.5th value. Since the 13th value is 34 and the 14th value is also 34, the 3rd quartile is (34 + 34)/2 = 34. Similarly, the 4th quintile is the 18 × 4/5 = 14.4th value and obtained as 4/10th away from the 14th value. Since the 14th value is 34 and the 15th value is 43, 4/10th apportion of the difference of 9 between these two values is 3.6, giving the 4th quintile = 34 + 3.6 = 37.6. In most practical applications, decimals are approximated; that is, the 14th value is taken as the 14.4th value.

### *Quantiles in Grouped Data*

In case of grouped data, the calculation is based mainly on the quantile interval—the interval containing the required quantile.

$$\text{Grouped data: } p\text{th } S\text{-tile} = a_{p-1} + \frac{p^*n/\mathrm{S} - C}{f_p} * h_p,$$

where $a_{p-1}$ is the lower limit of the quantile interval. The quantile interval is the one containing the $(p^*n/S)$th observation in order of magnitude

$C$ = cumulative frequency until the quantile interval
$f_p$ = frequency in the quantile interval
$h_p$ = width of the quantile interval

Consider the duration of immobility data in Table Q.3. The 2nd tertile in this data set is approximately 2 × 38/3 = 25th value in ascending order. This is 8 days. Also, the 85th percentile = (85 × 38/100)th or the 32nd value in ascending order = 10 days. The same data are grouped in Table Q.4. The 25th value is in the interval

**TABLE Q.3**
**Duration of Immobility (Days) of 38 Patients of Acute Polymyositis**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | 5 | 9 | 7 | 36 | 4 | 6 | 7 | 5 | 8 | 3 | 6 | 5 |
| 7 | 8 | 10 | 7 | 14 | 10 | 9 | 4 | 6 | 11 | 9 | 6 | 5 |
| 8 | 8 | 6 | 7 | 5 | 5 | 12 | 3 | 5 | 9 | 10 | 7 | |

**TABLE Q.4**
**Grouping of Data in Table Q.3 on Duration of Immobility in Cases of Acute Polymyositis**

| Group $(a_{k-1}, a_k)$ | 2.5–5.5 | 5.5–8.5 | 8.5–11.5 | 11.5–14.5 | 36 | Total |
|---|---|---|---|---|---|---|
| Midpoint $(x_k)$ | 4 | 7 | 10 | 13 | 36 | |
| Frequency $(f_k)$ | 11 | 16 | 8 | 2 | 1 | **38** |
| Cumulative frequency | 11 | 27 | 35 | 37 | 38 | |

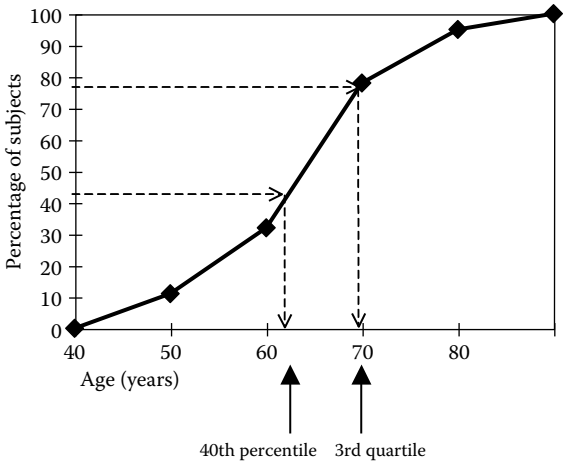



**FIGURE Q.4** Approximate calculation of some quantiles from ogive.

(5.5–8.5) days. Thus, for the 2nd tertile, $a_{p-1} = 5.5$, $C = 11$, $f_p = 16$, and $h_p = 3$. Therefore,

$$\text{2nd tertile (grouped data)} = 5.5 + \frac{25-11}{16} \times 3 = 8.1 \text{ days.}$$

Similarly,

$$\text{85th percentile (grouped data)} = 8.5 + \frac{32-27}{8} \times 3 = 10.4 \text{ days.}$$

The other method for obtaining approximate quantiles in case of grouped data is graphical. For graphical calculation of quantiles, the cumulative percentages of subjects are plotted against the upper end of the data intervals, called the **ogive**. Suppose, for an age distribution, this plot is as shown in Figure Q.4. The cumulative percentages for age 49, 59, 69, 79, and beyond 79 are 11, 32, 78, 95, and 100, respectively. To obtain a *p*th *S*-tile, draw a horizontal line at 100*p*/*S*% and read the value on the *x* axis where this horizontal line intersects the percent-based ogive. Figure Q.4 shows the 40th percentile and the 3rd quartile of age for these data. The calculations in the case of grouped data are, in any case, approximate for quantiles just as for the mean, median, and mode—thus, the graphic method may not be as bad for quantiles too.

### *Interpretation of the Quantiles*

All calculations can be done with the help of computers, but the method of computation helps in understanding quantiles and their proper interpretation. Figure Q.5 on quartiles may provide another perspective. This shows the duration of hospital stay after surgery for 30 patients. See how quartiles are determined. They divide the total number of subjects into four equal groups in terms of frequency. These frequencies may not be exactly equal in the case of the discrete data depicted in Figure Q.5 but would be equal in the case of really continuous data. Other quantiles have similar interpretations.

Quantiles are sometimes used for an objective categorization of the subjects. It may be easier to say in some situations that the bottom one-third values are low, the middle one-third are medium, and the top one-third are high. These are tertiles. In the case of data on the duration of immobility in our example, one may say that the duration of immobility up to 6 days experienced by one-third of the patients is low, that between 6 and 8 days is medium, and that above 8 days is high. A popular use of quantiles is in grading scores in educational testing where it is common to consider the scores in the top decile as excellent because only 10% of the subjects have these high scores. In medicine, scoring is sometimes used to grade the severity of disease as well. A condition with a score less than the first quartile (scored by the least affected 25% of cases) can be considered mild, one with a score between the first and second quartiles can be considered moderate, that between the second and third quartiles can be considered serious, and that beyond the third quartile can be considered very serious. Such nomenclature may or may not agree with the prognosis because the categorization is based on statistical rather than on medical considerations. The argument is that a condition with a score, for example, worse than that of 75% of cases can be safely called very serious. When the categorization is based on quantiles and not on prognosis, the interpretation is relative and not absolute. It is in this sense that percentiles are used in growth charts.

Quantiles are the *points* that divide the total number of subjects into equal groups, but it is colloquially popular to say, for example, that the scores are in the top decile as though the decile is a range. This is not technically accurate but is accepted as long as the meaning is clear.

## quantitative measurements, *see also* qualitative measurements

Any characteristic that is measured in terms of numbers is quantitative. This will have values, levels, and high–low kind of connotations that most qualitative measurements such as blood group do not have. When appropriate instrument is available, it is measured exactly numerically such as blood pressure, hemoglobin, creatinine, and size of kidney. These are the most obvious example of quantitative characteristics and said to be on the metric scale. This is further divided into interval and ratio scale—interval where the difference is more meaningful and ratio where the ratio is more meaningful.

Degree of severity of disease is also intrinsically quantitative, but in the absence of an appropriate scale, this is measured as none,

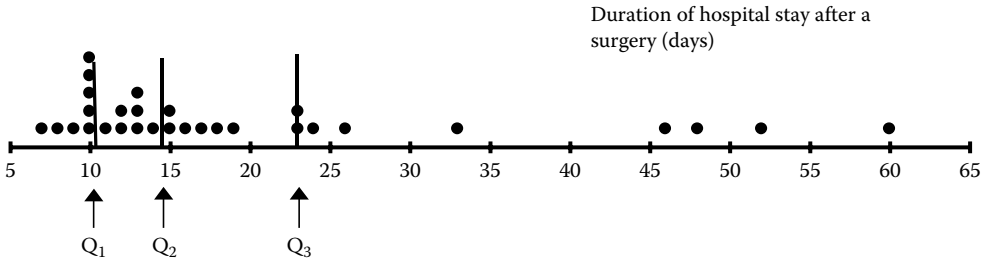



**FIGURE Q.5** Schematic of quartiles.

mild, moderate, or serious. This is called ordinal because an order is present with an underlying continuum. We may not have an adequate instrument as of now to measure them in terms of exact numbers but can hope that some are developed soon. Till such time that this does not happen, they are statistically considered ordinal and not considered quantitative. There are other characteristics that do not have continuum, and these are called nominal. Blood group mentioned earlier is a glaring example. Site of injury, site of cancer, sex, and occupation are other examples. Nominal and ordinal together form, what are called, qualitative characteristics. These are discussed in details with a different perspective under the topic **scales of measurement (statistical)**.

Characteristics that are quantitatively measured are called quantitative variables. As discussed for **variables**, these are divided into continuous and discrete. Age and blood pressure level are continuous as they can be 36.432 years and 132.57 mm/Hg, respectively, if an instrument for this kind of accuracy is available. Number of family members affected by a genetic disorder is discrete as this cannot be 2.57. Yet, both types of quantitative variables have a special place in statistics because of their amenability to mathematical manipulations. Their mean and standard deviation can be calculated, methods such as Student *t* and analysis of variance can be used, and correlation coefficient, ordinary regression, and so on can also be worked out. Durations are also quantitative, which lend themselves to survival analysis, whereas the variable that counts such as the number of persons attending a busy clinic is discrete, yet is statistically considered continuous because of the large range of values.

Even for quantitative variables, caution should be exercised when an apparently minor change in the value of a variable is highly sensitive. Body temperature may have a small range from, say, 97°F to 106°F, but a difference of 2°F or 3°F can have a profound impact on health. Also in terms of percentage, a drastic change may look miniscule in this case. This may not be so with alkaline phosphatase level—thus, biological context can never be ignored. Second, clinicians like to divide the continuum into categories that are considered relevant for patient management. For example, systolic blood pressure level may be divided as ≥140 mmHg and <140 mmHg, and body mass index (BMI) is divided into thin, normal, overweight, obese, morbid obese, and super obese categories. This categorization may look like a convenient way to decide about what to do but has undesirable statistical implications. Depending on how these categories are defined, a BMI = 29.7 kg/m$^2$ can fall into the overweight category and BMI = 30.2 can fall into the obese category. A marginal difference of 0.5 makes a big difference. View it in the light that a BMI = 25.1 and a BMI = 29.9 are both in the same category despite a relatively large difference. Moreover, one researcher may define morbid obese category to start from BMI = 40.0 and the other from BMI = 35.0. Such discrepancies are quite common in medical literature. When the values are grouped into intervals such as this, and the actual values are not available, statistical methods would consider all values concentrated at the midpoint of the interval. You can see that this is unrealistic and introduces an avoidable artifact. Because of this, statistical advice suggests that you do all the analysis on actual exact values, but for reporting in a tabular form, use categories. There are instances, though, that categorical data lead to more valid conclusions than the corresponding ungrouped data. For an example of this, see Welch et al. [1].

1. Welch HG, Schwartz LM, Woloshin S. The exaggerated relations between diet, body weight and mortality: The case for a categorical data approach. *CMAJ* 2005;172(7):891–5. http://www.cmaj.ca/content/172/7/891.long

## quartiles, *see* quantiles

## quasi-experimental design, *see also* before–after design/study

A design is called quasi-experimental when an experiment's key element—randomization—is missing although the objective remains to find the effect of one or more interventions. In a conventional experiment, the subjects are allocated to the interventions in a random manner that tends to equalize the groups under different interventions not just for known factors but also for the unknown factors. In the absence of randomization in quasi-experiments, baseline equivalence of the groups receiving various interventions is in doubt and comparability suffers. The internal validity of the experiment takes a hit and the results are rarely conclusive. Despite these limitations, a quasi-experimental design is useful in some situations.

A quasi-experimental design is used when randomization cannot be done, either due to ethical reasons or because of lack of feasibility. Suppose the study is on the effect of age on recovery after a treatment regimen, where one group is composed of subjects of age around 60 years and the other group is composed of those of age around 70 years. These groups may or may not be randomly selected. Nonrandom selection does not preclude it to be called quasi-experimental—nonrandom allocation of treatment does. Let these two groups be measured before the intervention and after the intervention. The **difference-in-differences** approach will give the estimate of the effect—in this case, the effect of age. However, this estimate will be valid only if all other factors that can affect the outcome are the same in both the groups. For example, they should have the same sex composition, the same obesity, the same nutritional level, and so on. The effect of all these factors will have to be eliminated by an appropriate analysis to reach a valid conclusion in case there is a difference.

In our example, there are two age-groups, but any **before–after design**, even for one group, is quasi-experimental because of lack of random allocation. In fact, most quasi-experimental studies are on one group with no parallel control where before treatment values serve as control, but the effect observed, purported to have been caused by the intervention, cannot be wholly ascribed to the intervention despite the same subjects being observed before and after. This is because (i) placebo effect would be confounded in this type of study; (ii) subjects at baseline may have different levels to begin with that could affect the outcome (if a person's hemoglobin level is already 13 g/dL, improvement is unlikely); (iii) there might be other concurrent changes that may be affecting the outcome; (iv) strict inclusion and exclusion criteria are needed to ensure that the outcome has not occurred before the intervention, or before the expected duration; (v) attrition during follow-up can introduce artifacts if the loss is related with intervention; (vi) if the gap between before and after measurements is substantial, the nature of measurement may change over time, and there could be regression to the mean effect in the sense that as the time passes, the subjects tend to move closer to the mean. Because of these problems, many equate quasi-experimental studies with **observational studies** instead of **analytical studies**.

Not that there are no advantages of a quasi-experimental design. One, it has wider applicability because it can be used where randomization is not practical. Second, because of nonrandomization, this design can have better generalizability—sometimes being more representative of the actual life situations. Third, if the follow-up is short, measurement of the same subject can rule out the effect of many factors. The effect of the confounding and extraneous factors can be adjusted to provide near noncontaminated results. When this

is adequately done, cause–effect conclusions can indeed be reached. Another alternative is to choose two groups that look similar at baseline, calculate the **propensity score** of each group, and use it for adjustment. Kausto et al. [1] used propensity score matching on age, sex, diagnostic category, and several other factors before intervention to study the effectiveness of legislation on sickness benefits in a quasi-experimental study.

Further details of quasi-experimental studies have been provided by Harris et al. [2]

1. Kausto J, Viikari-Juntura E, Virta LJ, Gould R, Koskinen A, Solovieva S. Effectiveness of new legislation on partial sickness benefit on work participation: A quasi-experiment in Finland. *BMJ Open* 2014 Dec 24;4(12):e006685. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4281551/
2. Harris AD, McGregor JC, Perencevich EN, Furuno JP, Zhu J, Peterson DE, Finkelstein J. The use and interpretation of quasi-experimental studies in medical informatics. *J Am Med Inform Assoc* 2006 Jan–Feb;13(1):16–23. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1380192/pdf/16.pdf

## quasi-random allocation, *see* random allocation

## questionnaire, schedule, and proforma

These are the types of forms used to record the information regarding a person, a patient, or any other unit of enquiry. The kind of information recorded in health and medical setups is one or more of the following: unique identifier; basic background characteristics such as age, sex, and education; medical history; knowledge, behavioral, attitudinal, and sometimes psychological traits; signs and symptoms; basic measurements such as blood pressure, height, and weight; investigation results; prescription and other advice; follow-up findings; and so on. These forms are generally predevised after a lot of thinking and discussion among stakeholders so that nothing of consequence is missed, and the information is available in a standardized format. Special care is taken to keep focus on the objectives of such records and to ensure that no extra information is collected or recorded. Many times, these forms are **pretested** by actually using them on a subset of cases and revised accordingly.

Many medical studies require physical examination of the subject and investigations in the laboratory or in the radiology unit. The information so received will most likely be recorded on a form containing the items on which the information is required. The recording should be accurate after following the correct and full procedure as required in the protocol, but what needs to be emphasized is the legibility of writing in case of manual forms—more so because most of the schedules for such purposes would be open ended. A computer-generated form, which allows direct entry of data without the help of intermediaries, is a big advantage in this context. However, the quality of information depends on the cooperation of the patient, the skills of the doctor, and the validity of the instruments. Ensuring all of this can be an uphill task in a given situation.

A form is a big help in retaining the focus on the objectives and in ensuring uniformity and also serves as a repository of information for later reference. In view of such important functions, ample time and thought should be invested in designing forms. The format could be a questionnaire or a schedule or a proforma for master chart. There is a considerable confusion among medical fraternity about these terms since the contents of all three formats are basically the same. Their unique features are as follows.

### *Questionnaire*

As the name implies, a questionnaire contains a series of questions that are required to be put to a subject in verbatim at the time of interview. A questionnaire could be self-administered or could be administered by an interviewer. In the case of the former, the education and the attitudes of the respondent toward the survey can substantially influence the response, and in the case of the latter, the skill of the interviewer can make a material difference. The interviewer may have to be trained on how to approach a subject; how to obtain cooperation; when to prompt, probe, pause, or interject; and so on. Some details of these aspects are described in simple language by Hepburn and Lutz [1] for health surveys.

Framing questions is a difficult exercise because the structure of the sentence and the choice of words become important. Some words may have different meanings to different people. In any case, it is evident that the questions must be unambiguously worded in simple language, and they must make sense to the respondent. The sequence must be logical, and the length of the questionnaire should be carefully decided so that all the questions can be answered in one sitting without a feeling of boredom or burden. It is always helpful for the interviewee as well as for the interviewer to divide the questionnaire into sections. Use of features such as italics, boldface, underline, and capitals can help clarify the theme of the question.

Respondents tend to be careless toward the tail end of the questionnaire if it is unduly long and not interesting, and a large number of questions may be a source of attrition for the subjects and can increase nonresponse. Thus, nonessential questions must be avoided. Ask yourself how much you lose if a particular question is not asked. If the answer is negligible, then delete the question; otherwise, do not do so. All the questions should be short, simple, nonoffending, and corroborative. They should be such that the respondent is able to answer.

A questionnaire is called *structured* when the language and the order of the questions are fixed. When the list of possible answers is also given and the respondent has to just tick one or more as applicable, this becomes *close ended*. One possible option can be "Any other (specify)" to cover the possibility of the list not being exhaustive. A question can be framed close ended only when enough is known about the type of responses so that the responses under the "other" category are minimum, say, not exceeding 10%. If they are higher than that, this reflects that the questionnaire was not prepared with sufficient thought. If the responses cannot be anticipated, do a pilot study to find out the common responses that need to be separately listed in a close-ended fashion. If substantial responses occur under the "other" category despite best efforts, invoke the "specify" component and identify which common responses can be separated. Remember that a large number of responses for "other" without identifying what can throw any research out of gear. Also, at the time of statistical analysis and conclusions, this "other" category can cause problems because it contains the assorted leftovers that defy interpretation. The list of responses in a close-ended questionnaire should be designed in such a manner that no choice is forced on the respondent. They should also be preferably mutually exclusive so that one and only one response is applicable in each case. Depending on the question, one of the responses could be "not applicable." For example, the question of results of laboratory investigation does not arise in cases where laboratory investigation was not required. This should be distinguished from cases in which the investigation was required but could not be done for some reason.

The objective of close-ended questions is to make it friendlier to the respondent, as well as for easy data processing and analysis. However, care must be exercised in providing the list of possible

answers in addition to what is already advised. The choices for "How often do you get bouts of depression?" can be never, sometimes, often, and frequent. These are subjective terms. Instead, try "less than once a year," "more than once a year but less than once a month," "more than once a month but less than once a week," and so on. In addition, for a close-ended question, the instruction on how to correct mistakes in a filled-up form without making it ambiguous should be clear.

When the response is to be recorded verbatim, the question or item is called *open ended*. This can present difficulty during interpretation and analysis of data but can provide additional information unforeseen earlier. Also, such open-ended responses can provide information regarding the quality of response and can provide results different from the results of *close-ended questions*. Suppose a question for parents is "What is the most important thing for the health of their children?" More than 60% might say nutrition when this alternative is offered on a list, but only less than 10% may provide this answer when no list is presented.

The phrasing of questions and their sequence are important since these features have potential to change the response. For side effects, for example, asking generally about any problems caused by the treatment may elicit much less response than asking separately for each possible side effect. However, in the latter case, all possible side effects should be known at the time of framing of questionnaire. Also, the last asked side effect is likely to be reported much less than the first asked. Sensitive questions such as on sexual behavior should be asked as late as possible because that can change the course of the interview. Suppose the interest is in finding whether a couple had undergone sex determination of a fetus in relation to their knowledge of legal provisions in this respect. If knowledge about legality is asked first and the respondent says "yes, he knows that it is illegal," he would be bound to say "no" to the question regarding sex determination actually practiced by him. The right sequence is to first ask about any sex determination undertaken and related questions and then ask about his knowledge regarding legal provisions in this respect.

There is also an effect of attitude of the patient toward certain ailments. An older person may perceive poor vision or infirmity as natural and may not report it at all, and a smoker might similarly ignore coughing. On the other hand, mild conditions may be exaggeratedly reported depending on the disturbance they create in the vocational pursuits of the patient. A vocalist may be worried more about the vocal cords than a fractured hand. Thus, interview responses measure the person's perception of the problem rather than the problem itself. Some conditions with a social stigma such as venereal diseases and impotence may not be reported despite their perception as important.

Avoid questions such as "Don't you use pain killers for acute pain?" They suggest an answer and are called leading questions and not recommended. Also, do not seek two pieces of information in one question. The wordings must not be ambiguous. Minimize hypothetical questions on future or regarding attitude because the response to such questions depends on the mental frame at the time of the interview, and the response may change when asked a second time. It is sensible to concentrate on questions on facts rather than on opinion. In place of asking what would he do in a particular situation, ask what he did last time in that situation. For regularity of condom use, for example, do not elicit response as never, sometimes, regular, and always, but ask how many times condom was used in the last, say, four sex encounters. Probing is allowed provided no suggestion is made regarding the reply, and the interviewer is trained.

A self-administered questionnaire must contain all the explanations and instructions for its completion. Use simple language for framing the questions. If the language of the questionnaire is different from the language of the respondent, the translation should be appropriate but that should not lose the context. Despite all these precautions, portions of a self-administered questionnaire are likely to be left blank and be prepared for this eventuality, but more accurate information can be obtained by this method on sensitive issues such as sexual behavior if the questionnaire is anonymous. Face-to-face interview may not yield the same quality of response on such sensitive issues. In postal questionnaires, chance of response can be increased by offering incentives, advance contact, personalized letter, stamped return envelope, follow-up contact, and so on.

### Schedule and Proforma

Schedule contains only the items on which the information is to be collected, and not questions. Framing questions to get that information is left to the interviewer in this case. A schedule will have an item—age. A questionnaire will have "What is your age?" The information can be obtained by observation, by interview, or by examination. Sufficient space is always provided to record the response. A bed-head ticket or a case sheet is a schedule since it contains only the items of information.

Proforma is the prototype or a sample providing the items of information on which the information is to be collected. It is neither a questionnaire nor a schedule. Perhaps this can be used to prepare columns in a register where information for each patient can be recorded in one line as for a database or master chart. The illustration in Table Q.5 would clarify the differences between a questionnaire, a schedule, and a proforma.

### Features of a Form

Whether a questionnaire, a schedule, or a proforma, the following should be kept in mind.

- It should be properly formatted with headings and subheadings, and spaces as required. Use capitals, bold, italics, and different fonts to emphasize the thrust of the information you are seeking. The layout of the form should be aesthetically pleasant.
- A printed form with a logo and an immaculate appearance instills confidence in the interviewer and the respondent alike. Photostat or cyclostyled forms do not do that.
- The unit of recording for each piece of information should be specified, for example, whether duration of symptoms is to be recorded in days or weeks or months. Age of a neonate of 4 h is recorded as 1/6 day if this information is to be used for analysis along with such information on other neonates whose age is recorded in days.
- Specify how the measurement is to be rounded off. If age is 13 years and 8 months, is it to be recorded as 14 years, which is the nearest integer, or in completed years (13) as is generally done all around the world in a medical record? Do not make categories such as 15–19, 20–24, and so on, in the data collection form. Such categories can be made later at the time of analysis or presentation.
- The questions or the items should be framed in a manner that the chance of bias in the responses is minimum.
- Questions or items that need to be skipped for being not applicable in certain cases should be specified. If there is an item on cause of death that is to be asked only if there is a death in the family, the item should have a remark that this is to be skipped in other conditions.

**TABLE Q.5**
**Illustration of Differences between a Questionnaire, a Schedule, and a Proforma**

**Questionnaire**

Unique ID_____
What is your name? _____
What is your age in completed years? _____
What is your gender? _____
What are your complaints?

(i) _____
(ii) _____
(iii) _____
(iv) _____

Does any other member of your immediate family suffer from the same complaints? Yes/No
If yes, what is your relationship with each of them?

(i) _____
(ii) _____
(iii) _____
(iv) _____

**Schedule**

Unique ID_____
Name _____
Age (completed years) _____
Gender _____
Complaints

(i) _____
(ii) _____
(iii) _____
(iv) _____

Immediate family members affected (write None if no family history)

(i) _____
(ii) _____
(iii) _____
(iv) _____

**Proforma**

**Complaints codes: 0. None, 1. Abdominal pain, 2. Vomiting, 3. Acidity, 4. —, etc.**
**Family history codes: 0. None, 1. Mother, 2. Father, 3. Brother, 4. Sister**

| Sl. No: | Unique ID | Name | Age (Years) | Sex (M/F) | Complaints Code (Enter X) 0 | 1 | 2 | 3 | 4 | Family History Code (Enter X) 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | | | |
| 2 | | | | | | | | | | | | | | |
| ⋮ | | | | | | | | | | | | | | |
| *n* | | | | | | | | | | | | | | |

- Nonstructured questionnaires are used for informal surveys. For medical research, structured questionnaires are preferred.
- If multiple responses are admissible, this should be clear to the investigator. Multiple responses imply that one question or one item can have more than one answer. Complaints could be many in one person; thus, "complaints" is a multiple-response question.
- The forms can be precoded, which will make computer entry simple. In any case, text responses such as signs and symptoms have to be coded either at the time of recording or later at the time of computer entry. However, remember that codes are not grades or scores, and they cannot be added, subtracted, and so on.
- The form should be comprehensive so that all information on one subject or one patient could be recorded together. Investigation results may have to be obtained separately but they should be transferred to the main form. Postoperative details should be recorded together with preoperative findings, and not in a separate from. Each page of the form should have the same identification number, which would be unique for each subject.
- The information should be elicited only on the items that are really required to meet the study objectives since unnecessary information tends to dilute the quality. An unduly detailed form can be tedious and boring, and wasteful of the time of the interviewer and the respondent. It is difficult to sustain the interest of a respondent for more than half an hour unless there is some encouragement. Nonetheless, obtain all the information committed in the protocol.
- As far as possible, restrict to the collection of objective information. Opinions and attitudes should be minimal and separately identified.
- Standardize the form after pretesting on a small number of subjects from the same target population for which this is going to be finally used. Pretesting helps identify ambiguities or potential biases in the responses. Investigators can be trained accordingly. It also helps assess the time and resources required to administer this kind of instrument to various types of respondents. For a large-scale study, two or more rounds of pretesting may be required.
- One aspect of some questions is the rating scale used for providing the answer. For a question such as "how fit are you after one week of a surgery?", many may rate themselves +3 if the scale is from –5 to +5 but not 8 if the scale is from 0 to 10. Also, see the topics **Likert scale** and **Guttman scale**.
- Sometimes, methods such as **Rasch analysis** are used to develop and evaluate a questionnaire or a schedule. See this topic for details.

- The questionnaire or schedule must always be accompanied by a statement of the objectives of the survey so that the respondent becomes aware. Easy-to-follow instructions for recording responses and explanatory notes where needed are always helpful. Special care may have to be taken concerning possible memory lapse if a question or item requires recall of an event. Whereas serious events such as accidents and myocardial infarctions are easy to recall even after several years, mild events such as episodes of fever or of diarrhea may be difficult to recall after a lapse of just 1 month.

1. Hepburn W, Lutz W. *Community Health Surveys—A Practical Guide for Health Workers. 5 Interviewing and Recording.* International Epidemiological Association, 1986.

## Quetlet index, *see* body mass index

## quintiles, *see* quantiles

## quota sampling

Quota sampling is one of the methods of **purposive** (nonrandom) **sampling** where a predetermined number of units is selected without adopting any specific procedure. This is generally adopted after the population is divided into strata. For example, one may decide to have 100 males and 200 females in a study on chronic kidney disease, and select them to serve a specific purpose. These quotas are generally decided on the basis of previous experience or perceptions but can also be based on statistical considerations. However, which subjects will be in the sample is left on who comes first, who are readily available, volunteers, or any such consideration. The subjects continue to be recruited till such time that the prefixed quota of subjects is met.

Quota sampling can be quicker and can be carried out without much hassle. It does not require any sampling frame that a **random sampling** does. In contrast to simple random sampling, the researcher can decide to overrepresent or underrepresent any particular group to suit the purpose. For example, it can be decided to have a minimum of 30 subjects of age 90+ years in the sample, or to have not more than 40 subjects of age 40–49 years in the sample.

Statistically, the demerits hugely outscore the merits of this kind of sampling. Because of nonrandom selection, none of the statistical methods of inference such as confidence interval and test of hypothesis would be applicable and the generalizability would suffer.

There are many examples in the medical literature where this kind of sampling has been used. Niedźwiedzka et al. [1] used quota sampling for assessing the validity of self-reported height and weight in elderly people of Poland. Note how random sampling in this kind of study was not a strict requirement. Hongthong et al. [2] also used quota sampling for finding factors that influence the quality of life of elderly people in a rural area of Thailand. Because of purposive sampling, their findings are suspect and cannot be generalized even to the rural area of Thailand studied by them.

1. Niedźwiedzka E, Długosz A, Wądołowska L. Validity of self-reported height and weight in elderly Poles. *Nutr Res Pract* 2015 Jun;9(3):319–27. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4460065/
2. Hongthong D, Somrongthong R, Ward P. Factors influencing the quality of life (Qol) among Thai older people in a rural area of Thailand. *Iran J Public Health* 2015 Apr;44(4):479–85. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4441960/

# R

## radar graph

A radar graph provides a suitable presentation when the performance of one group on several variables is to be compared with another group. As many axes are drawn as the number of variables, and a polygon is drawn according to the values as shown in Figure R.1. You can see that this looks like a radar. The larger the number of variables, the bigger the radar should be so that all axes can be clearly seen. This depiction seems most appropriate for four to eight variables.

The first figure compares average age, percent females, percent hypertensives, percent obese, and average hemoglobin level in cases and controls at baseline in a randomized controlled trial (RCT). The graph reveals that percent hypertensives are very different in the two groups, but all other values are not much different. The second figure compares male and female diabetic patients for six measurements. Notice the marked difference in percent on diet control and with positive history. If one radar graph portrays general population averages for five or six variables and another graph the same variables for a specific group such as those suffering from cancer, such a graph can highlight the variables, if any, on which the cancer patients exceed the general population and on which variables they have less value.

Radar graph seems like a good tool for multivariate comparisons, although only two groups can be cleanly compared. For three or more groups, this becomes too cluttered. As with almost all other graphs, a radar graph too can mislead depending upon the scale chosen for depiction. In Figure R.1b, the axis for average age at detection can be stretched, and then the difference between the two groups will magnify. Any axis can be stretched or compressed as we wish. The figure is most effective when all variables have the same scale such as percentage or **standardized** score because then all axes can also have the same scale.

Kalonia et al. [1] have effectively used a series of radar graphs to visualize subdivided particle concentration and size data for IgG1 mAb solution, and Oowada et al. [2] have used radar graphs to depict multiple free-radical scavenging capacity in serum of chronic kidney disease patients.


1. Kalonia C, Kumru OS, Kim JH, Middaugh CR, Volkin DB. Radar chart array analysis to visualize effects of formulation variables on IgG1 particle formation as measured by multiple analytical techniques. *J Pharm Sci* 2013 Dec;102(12):4256–67. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3856888/
2. Oowada S, Endo N, Kameya H, Shimmei M, Kotake Y. Multiple free-radical scavenging capacity in serum. *J Clin Biochem Nutr* 2012 Sep;51(2):117–21. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3432821/


## radix, *see* expectation of life and life table

## random allocation, *see also* block, cluster, and stratified randomization and minimization

This is the method that randomly allocates eligible subjects to different groups such as treatments and control in a trial. The objective is to achieve baseline equivalence of the groups by providing same chance to each subject to be assigned to any of the groups under study and to provide a foundation for the application of a statistical test of significance. A big advantage of random allocation is that the groups tend to have equal distribution not just for the known factors but also for the unknown factors. Thus, it covers the **epistemic uncertainties** as well. Random allocation distributes the effects of concomitant variables (covariates), both observed or unobserved, in a statistically acceptable way. However, baseline equivalence with random allocation is more likely to be achieved with a large sample than with small samples since small samples can have bias even when randomly allocated. If the number of available subjects is



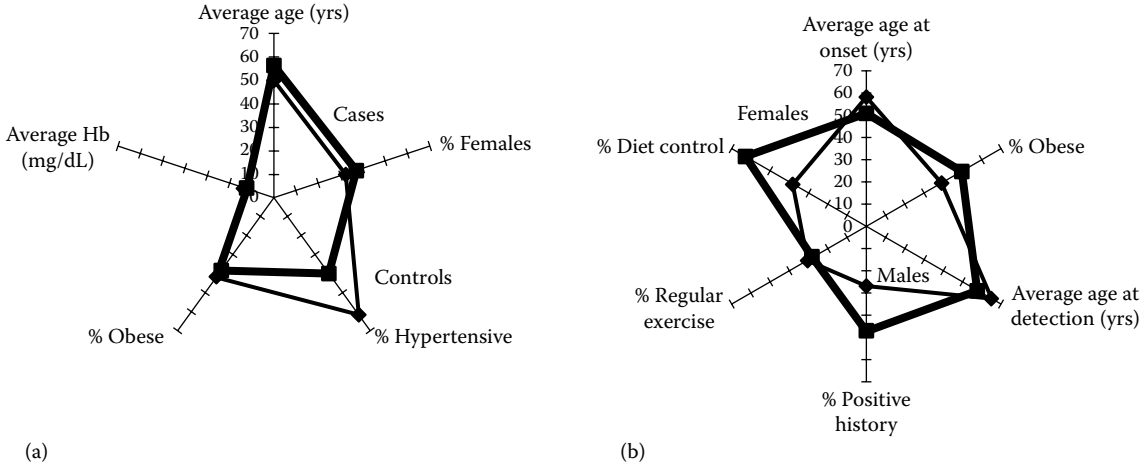



**FIGURE R.1** Radar graphs: (a) compares some characteristics of cases (thick lines) with controls (thin lines) in a randomized controlled trial (RCT); (b) compares some characteristics of male cases (thin lines) of diabetes with female cases (thick lines).

small or the resources do not permit a large trial, consider the strategy of matching instead of randomization, as this at least equalizes the groups for known important covariates.

Random allocation is advocated only when the treatments are harmless—for example, it is not feasible to allocate subjects to smoking and nonsmoking groups as an intervention. It is necessary for random allocation to meet ethical considerations, and it cannot be used if a patient's condition requires a specific treatment.

A large number of methods are available for random allocation as discussed next, but not many of these methods are adequate. The general practice is to conduct a trial on consecutive patients reporting in a clinic after excluding those who are not eligible or do not provide consent. If the patients indeed come in random order and the consent does not introduce any bias, the even-numbered patients can be assigned to one group and the odd-numbered patients to the other group. For three or more groups, the first patient can be randomly allocated to any group by, say, a draw of lots and the remainder allocated in a systematic way. For example, if there are four groups and the first patient is randomly allocated to group 3, then the allocation for the incoming patients will be group 4, 1, 2, 3, 4, 1, etc. This systematic scheme will fail in achieving unbiased allocation if the subjects follow an unknown or known design in the sequence of their arrival. Also, it is difficult to enforce **blinding** with this kind of allocation.

In situations where there are only two groups such as treatment and control, the participants can receive one of these on the basis of the outcome of a chance event, such as tossing a coin behind the curtain. This has the potential to provide an impartial procedure for the allocation of treatment to individuals, free from personal biases. Another method for two groups could be to draw a single digit random number and allocate participants corresponding to odd digits to group 1 and even digits to group 2. If the first two digits are odd, these participants go to group 1; if the third digit is even, this participant goes to group 2; and so on. This process can give an unequal number of participants in the two groups; however, we generally want the groups to be of equal size. This is a worry to researchers who are concerned that the randomization is not as it should be.

It is sometimes considered convenient to include patients who report to a clinic on alternate days or during a specified time of the day. This may introduce bias because some patients may choose a time and day according to the availability of a particular physician, and the management of cases by this particular physician may have prognostic implications. Allocation of subjects on the basis of even or odd date of birth may apparently look random but can be misused. All such methods are called **quasi-random allocation**. As already noted, the biggest problem in such allocation is that it cannot be kept blind, and the chances of bias in assessment remain. The investigator knows which treatment the next patient will receive, and that can influence the decision to enter him/her into the trial. Random allocation can be kept open so that the concerned subjects know that they are randomized to a particular group, but the strategy of **concealment of allocation** is the gold standard in clinical trials.

A more acceptable method is to draw or generate a **random number** between 1 and $K$ (both inclusive), where $K$ is the number of groups. When an eligible subject arrives, assign him/her to the group bearing this number. If the random number drawn is 3, then the subject is allocated to group 3; if the random number is 1, then the subject goes to group 1; and so on. In place of generating a random number between 1 and $K$, it may be convenient to generate a two-digit random number. Divide it by $K$ and add 1 to the remainder to get the group. (This assumes that the number of groups is $K \leq 9$.) If $K = 3$ and the generated random number is 52, then assign the subject to group 2 (remainder of 52/3 is 1, and addition of 1 gives 2). In fact, a small computer program can be developed that randomly allocates subjects to the specified number of groups. This program can be used over and over again to provide fresh allocation whenever a new trial begins. This is called *simple randomization*. However, there are problems with this also as discussed next.

A difficulty in simple randomization is that one particular group may have its full quota of subjects much before the other groups. This is called imbalance since the subjects appearing late will not have the chance to be allocated to the already completed group. In that case, the process of allocation may be continued as follows: The subjects for whom the allocated group as per the existing randomization scheme turns out to be the already completed group are excluded from the trial. Suppose there are three groups, and each is planned to have 30 subjects. It is possible in this scheme that the full 30 are allocated to group 2 when group 1 has only 22 and group 3 has only 26 after allocation of 78 subjects. In this case, ignore the 79th subject if he/she is randomly allocated to group 2. This subject will not be included in any group. If the random number for the 80th subject is 3, then the subject is assigned to group 3. This process continues until each group has its full quota of 30 subjects. This procedure may mean wastage of some eligible subjects, but it ensures fair allocation. It is not considered fair to assign the last few subjects to only one or two groups and deny them a chance to be theoretically in the other groups.

By avoiding purely physical methods, such as the one just mentioned, randomization can be achieved with the help of random number tables, as follows: List the first $K$ participants and assign them numbers from 1 to $K$ if there are $K$ groups. Randomly permute these $K$ numbers to get the group to which these would go. If you have $K = 3$ groups in your trial, the random permutations could be (2,1,3), (3,2,1), (2,3,1), (1,3,2), etc. A total of $n$ such permutations, each a triplet of the type just mentioned, will determine allocation of $3n$ subjects in 3 groups of $n$ subjects each. The easiest way is to use the website *randomization.com*, where you can specify the number of groups and the number of subjects, and get the random allocation.

Note that random allocation is not haphazard allocation. It follows a well thought-out scheme. The most preferred method of randomization is **block randomization**, which is discussed separately under the topic **block, cluster, and stratified randomization and minimization**, where other methods of randomization are also discussed.

## random effects, *see* fixed and random effects

## random effects ANOVA

**Analysis of variance (ANOVA)** of data arising from a setup where all factors have random effects is called random effects ANOVA. This is different from the usual ANOVA where all factors are fixed, particularly in the way the $F$-ratio is calculated. Review the topic **fixed and random effects** if you want to understand these terms. In brief, if the observed levels of a factor represent many other levels not actually observed, it is a random effect factor. Suppose the interest is in finding whether the duration of hospitalization after a particular surgery is the same or not across hospitals. The interest in this case is in finding the hospital effect: whether or not it varies from hospital to hospital. If you select 4 hospitals randomly out of, say, 80 hospitals in a state, these 4 are the levels of the factor "hospital." The interest in this case is not restricted to these 4 hospitals but also in the variability across all hospitals. The analysis of data from such a study would require random effects ANOVA.

The ANOVA for random effects is not the same as for fixed effects, although luckily the computation of the mean sums of squares is the same. In the case of fixed effects, the mean sums of squares due to factors are divided by the mean square error (MSE) to get the *F*-ratio, but this is not the case for a random effects model. Details are as follows.

### *One-Way ANOVA with Random Effects*

In the case of fixed effects, the model for a one-way ANOVA is $y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$ ($j = 1, 2, \ldots, J$; $i = 1, 2, \ldots, n$), where $J$ is the number of levels of factor A, $\mu$ is the overall mean, $\alpha_j$ is the effect of the $j$th level of factor A, and $\varepsilon_{ij}$ is the error term, generally required to have **Gaussian** (normal) **distribution** with mean 0 and variance $\sigma^2$, written as $\varepsilon_{ij} \sim N(0, \sigma^2)$, whereas all other terms in the model are fixed. The $j$th level effect $\alpha_j$ is generally defined such that $\Sigma_j\alpha_j = 0$. When factor A is random, the model changes to $y_{ij} = \mu + a_j + \varepsilon_{ij}$, where $a_j$ continues to be the effect of the $j$th level of factor A, but is now random. This is required to be $N\left(0,\sigma_A^2\right)$. The mean continues to be 0 just as $\Sigma_j\alpha_j = 0$ in fixed effects model, but there is a variance attached to the effects now because they are random. In our example, if the hospitals are purposively selected, the effect of these hospitals can still be studied but the conclusion will be valid for these four hospitals and not for the other hospitals. You would not be able to say anything about the hospitals not in your sample. No general statement regarding whether duration of hospitalization across hospitals in the state is different or not can be made. For this kind of conclusion, the hospitals have to be selected randomly and the analysis done by random effects ANOVA. The difference would be clearer if you look at the null hypothesis in the two setups:

$$\text{Null hypothesis in case of fixed effects: } \alpha_1 = \alpha_2 = \ldots = \alpha_J;$$

$$\text{Null hypothesis in case of random effects: } \sigma_A^2 = 0.$$

The alternative hypothesis in the fixed effects case is that at least one $\alpha_j$ is different from any other, whereas in the random effects case, the alternative hypothesis is $\sigma_A^2 > 0$.

The calculation of mean sum of squares is the same in random effects model as in fixed effects model. The calculation of *F*-ratio also remains the same with the same dfs for one-way ANOVA (but not for higher-way ANOVA): only the interpretation changes. We now explain two-way ANOVA with random effects to highlight the difference.

### *Two-Way ANOVA with Random Effects*

For two fixed factors, the usual ANOVA model is $y_{ij} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk}$ with the conditions that $\Sigma_j\alpha_j = 0$, $\Sigma_k\beta_k = 0$, $\Sigma_j(\alpha\beta)_{jk} = 0$, and $\Sigma_k(\alpha\beta)_{jk} = 0$; where we assume that there are same $n$ number of subjects for each combination of factor levels ($i = 1, 2, \ldots, n$), $\alpha_j$ is the effect of the $j$th level of factor A, $\beta_k$ is the effect of the $k$th level of factor B, $(\alpha\beta)_{jk}$ is the interaction, and $\varepsilon_{ijk}$ is the error term with $N(0, \sigma^2)$ distribution. When both factors are random, this model becomes $y_{ij} = \mu + a_j + b_k + (ab)_{jk} + \varepsilon_{ijk}$ where now $a_j \sim N\left(0,\sigma_A^2\right)$, $b_k \sim N\left(0,\sigma_B^2\right)$, and $(ab)_{jk} \sim N\left(0,\sigma_B^2\right)$. In our example of duration of hospitalization in different hospitals after a particular surgery, two-way random effect would be applicable if you want to consider treatment modality also as the second factor. Let there be a large number of treatment modalities practiced in all hospitals. Of all the levels of this factor, include $K$ of them by random selection with the objective of finding whether these treatment modalities differentially affect the outcome. Based on statistical consideration of what is called expected mean squares in this setup, the *F*-ratios are calculated as follows when both factors have random effects:

$$\text{Random effects: } F_A = \frac{\text{MS}_A}{\text{MS}_{AB}},\ F_B = \frac{\text{MS}_B}{\text{MS}_{AB}},\ F_{AB} = \frac{\text{MS}_{AB}}{\text{MSE}},$$

where MS is for mean square. The degrees of freedom for these are as per the respective numerators and denominators in these ratios. The calculation of mean squares remains the same as in fixed effects. For fixed effects, the denominator is MSE in all these *F*-ratios, but not so for random effects. The null hypotheses under tests are $\sigma_A^2 = 0$, $\sigma_B^2 = 0$, and $\sigma_{AB}^2 = 0$, respectively, in random effects. If any of these is rejected, the conclusion is that the effects of different levels of the concerned factor have significant variation. In our example, if $K = 3$ treatment modalities are included randomly, $n = 10$ patients in each cross-tabulation of levels of factor A and factor B, and $F_A = 17.43$ at df = (3, 108), then $P < 0.05$, and thus the mean duration of hospitalization is different across hospitals. This conclusion is for all hospitals in the state and not just for the four hospitals in the sample.

An interesting situation arises when one factor is random and the other is fixed. This can happen in our example when factor A continues to be different hospitals but factor B is sex of the patients, which is fixed anyway. It would be interesting to find whether different hospitals treat males differently than females with respect to duration of hospitalization, called interaction. The data where some factors are fixed and others random are analyzed by **mixed models** ANOVA as per the details mentioned for that topic.

## random errors

These are inherent, unpredictable errors that arise when repeated measures of the same characteristics are taken in identical conditions and exclude bias or man-made error. Random errors occur due to imprecision of the process—in methods, measurements, etc., such as in staining, magnification, rotation, counting, and timing in a laboratory setup; in the environment such as room temperature, humidity, or minor variation in chemicals, reagents; or in the varying care adopted by different observers. All measurements are prone to random error, but they are particularly prominent in health and medicine.

Random errors can also arise due to (i) shifting goal posts so that similar performance is rated differently, (ii) measurements taken in a hurry or too much slackness, (iii) tendency to assign some middling value to most response, (iv) tinkering values perceived as unduly large effects, (v) trying to correlate response to different items when none exist, (vi) tendency to compare one person with others that can modify the reporting or measurement, and (vii) fatigue, boredom, lack of concentration, carelessness, etc. All these errors may occur unwittingly in the subconscious mind without the investigator really knowing.

**Replication** is an effective method of reducing the effect of these errors on the conclusions since these tend to average out in replications. The effect of random errors can be reduced through ordinary statistical methods such as using the mean. If you measure a pharmaceutical tablet eight times with an accurate balance and get slightly different values, the mean weight will be very close to the actual weight of the tablets if these differences are really random. If not, suspect that the errors are systematic.

It may be a surprise to many that random errors are not so random after all. They follow a pattern, called **Gaussian**. This means that most random errors will have a value around zero with a plus or minus sign, and a few will be large. This is also the basic assumption in most statistical methods. For example, in **multiple linear**

**regression**, the errors are $\varepsilon = (y - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \ldots - \beta_K x_K)$, and they are required to follow a Gaussian distribution with mean zero and variance $\sigma^2$ for validity of the procedures such as confidence interval and test of hypothesis. In case of samples, these errors are called **residuals**. The least squares method finds those estimates of regression coefficients β's that minimize the sum of squares of residuals. Goodness of fit of any statistical model is assessed by the magnitude and nature of these residuals. If many of these residuals are large, the model is not considered a good fit even if the mean is zero. If the mean is not zero, then it is called bias and it will require adjustment in the model.

In the context of experiments, random error is called *experimental error*. This can be controlled by appropriate choice of the experimental design. For example, experimental error could be caused by interobserver variability. Blocking by observers would control this error (see **randomized block design**). If it is due to age, sex, severity of disease, etc., these are controlled by considering them as factors in the statistical analysis so that their effect can be isolated and removed.

Opposite to random errors, there are systematic errors, which are also called **bias**. Systematic errors are reproducible inaccuracies that are consistently in the same direction. They can persist throughout. If you measure vials as part of the process of filling them but the measurements are 0.10 g higher because the scale was not properly tared, it would be a systematic error. If there is an air bubble in the mercury column of a sphygmomanometer, the blood pressure reading will be consistently high than the actual. Systematic errors are difficult to detect; worse, they cannot be analyzed statistically. Spotting systematic error and correcting requires a lot of diligence, and it should be done before starting the formal measurement for the study subjects.

## randomization, *see* random allocation

## randomized block design

A design is called a randomized block when the treatments are randomly allocated to the subjects after dividing them into some homogeneous groups, called blocks. The objective is to reduce the experimental error. Consider an experiment on the effect of three doses on duration of recovery that depends on the severity condition of the patients, among other factors. Divide the condition into mild, moderate, serious, and critical. Note that this division into homogeneous blocks is not random; instead, within each severity, the patients are allocated to receive one of three doses randomly. You may have 12 cases with mild disease, 21 cases with moderate disease, 15 cases with serious disease, and 6 cases with critical disease. We are taking these sizes in multiples of three since there are three doses. It is decided randomly which of the 15 cases with serious disease will get dose 1, dose 2, and dose 3. Thus, the random assignment is restricted to the doses within blocks as illustrated in Table R.1.

**TABLE R.1**
**RBD with Disease Severity as the Block**

| Dose | Mild | Moderate | Severe | Critical |
|---|---|---|---|---|
| Dose 1 | 2, 3, 6, 10 | 4, 8, 9, 12, 13, 20, 21 | 1, 7, 9, 10, 12 | 1, 5 |
| Dose 2 | 4, 5, 9, 12 | 1, 3, 7, 10, 14, 15, 18 | 3, 4, 8, 11, 14 | 2, 4 |
| Dose 3 | 1, 7, 8, 11 | 2, 5, 6, 11, 16, 17, 19 | 2, 5, 6, 13, 15 | 3, 6 |
| **Total number of subjects** | **12** | **21** | **15** | **6** |

*Note:* Table entries are subject numbers.

Sometimes a one-way experiment is repeated two or three times in different groups of subjects to increase the level of confidence in the findings. This is also blocking, where the blocking factor is time. However, one has to make sure that conditions such as reagents, calibration, and environment do not change in such repetitions.

Blocking characteristics can be those associated with subjects, such as age, sex, nutritional status, and severity of disease, or those associated with experimental setting, such as hospitals, observers, time, method of assessment, etc. The former controls intersubject variability, and the latter helps in achieving standardized conditions. In medical setups, blocking by age and sex is common but is useful only if this reduces experimental error—thus, the responses of subjects within each block must be substantially homogeneous. When this is so, the results will be more reliable. If any one particular block is to be excluded because of huge nonresponse or any other spoil, the results for other blocks would still be valid. However, if there are a few missing values in one or more blocks, this causes imbalance due to unequal $n$ and may also cause bias if missing values are related to response. Results from a randomized block design (RBD) can be difficult to interpret if there is an interaction between the blocking variable and the treatment levels. Yet, in some cases, this interaction can provide useful information on how a treatment works in different blocks.

In most applications of RBD, the two factors under study have fixed effects. In our example, the interest could be in just those three doses that are included in the experiment and in those four disease severity levels as included. But that is not necessary, and any or both factors can be random. When both are fixed, analysis of data is done by the ordinary **two-way ANOVA**. However, in this case, if you have just one subject in each combination of factors, you would not be able to study interaction in any ordinary manner. A special method called **Tukey test for additivity** is used for this purpose. If any one of the two factors is random, **mixed effects** ANOVA would be used, whereas if both factors are random, **random effects ANOVA** would be used. In an RBD, usually the interest is in the treatment effect—blocks are just for reducing the experimental error.

Distinguish RBD from a factorial experiment. In the case of the latter, there are two treatments and they can have $J$ and $K$ levels, respectively. This is like using two drugs: drug A in doses 5, 10, and 15 mg and drug B in doses 10, 50, and 100 mg. If it is a factorial experiment, there will be a group receiving 5 mg of drug A and 10 mg of drug B, another group receiving 5 mg of drug A and 50 mg of drug B, etc. The allocation of "treatments" is in combinations rather than one at a time.

When one or more treatment–block combination is missing, this is called *incomplete block design*. For example, you may not have any patient with critical condition receiving treatment 3. Incomplete blocks require a special method of analysis. For details, see Ruxton and Colegrave [1].

When time is used as a blocking variable on the same subjects, beware that the measurements at different points in time in the same subjects would be related. This violates the basic requirement of independence and invalidates the ANOVA unless special methods such as **repeated measures** are involved.

1. Ruxton G, Colegrave N. *Experimental Design for Life Sciences*, Third Edition. Oxford University Press, 2010.

## randomized consent design, *see* Zelen design

## randomized controlled trials (RCTs)

A clinical trial with random allocation of the subjects to the treatment and the control arms is called a randomized controlled trial (RCT). In the case of therapeutic modalities, this is the phase III trial (see **phases of (clinical) trials**) once the early trials (phase I and phase II) have established the overall safety of the regimen, its basic clinical pharmacology, its therapeutic properties, and its most important side effects. This kind of trial sits at the top level of the evidence pyramid (see **evidence (levels of)**) with the most credible and the least biased results. Employing an RCT is also a prerequisite to meet the regulatory standards of licensing for efficacy as well as safety of a new regimen.

For credibility, there must be a valid control group and allocation of subjects should be fully **randomized** in this phase to the test arm and the control arm; also **blinding** is done where feasible. Through such strategies, a phase III trial is expected to provide compelling evidence of the **efficacy** of the regimen or its lack thereof. When benefits are explored, proper assessment of harm is equally crucial—thus, safety is also assessed. In fact, these days, when efficacy of many regimens is exceeding 80%, safety is an overriding consideration in many trials.

An RCT in drug development is a large-scale trial with generally 300–1000 subjects recruited for each arm. The follow-up must be sufficiently long for efficacy and side effects to emerge, and to rule out that any relief is transient. For side effects, a larger sample may be needed then to evaluate efficacy because critical side effects can be rare. An RCT is expected to provide a full picture of the clinical performance of the treatment under test. Specifically, this can include

- Exact identification of the diagnostic group that responds reasonably well, and comparison of the beneficial and adverse effects with those of the existing treatment, if any.
- An increase in patient exposure in terms of both the number of patients and the length of follow-up over phase I and phase II trials so that less common and late side effects can also be identified.
- More evidence on possibility of adverse interaction with other treatment regimens with which the new treatment is likely to be prescribed.
- The ideal dose regimen for different types of patients with regard to age, body weight, severity of disease, etc.
- Further pharmacological studies.
- Acceptability to the treatment regimen of communities with different medical cultures. This last objective can be achieved by conducting a multicentric trial.

Among thousands of studies, just to give you an example, Gianotti et al. [1] performed one in Italy in which a total of 305 patients of cancer of the gastrointestinal tract with preoperative weight loss <10% were randomized to receive either preoperative artificial nutrition supplementation, postoperative jejunal infusion (perioperative group), or no artificial nutrition (conventional group). There are three groups in this RCT including one control (the conventional group). The outcome variables were postoperative infections and length of hospital stay. **Intention-to-treat analysis** and differences between the groups showed that preoperative supplementation was as effective as perioperative administration, and both strategies are superior to the conventional approach. This illustrates how an RCT provides an opportunity to reach to a definitive and useful conclusion. However, among other considerations, the quality of evidence provided by an RCT depends on the proper choice of the cases and controls.

### *Selection of Participants for RCT*

When an inordinately large number of subjects are available that pass the **inclusion and exclusion criteria**, they must be randomly selected for inclusion in the trial. This allows generalizability. When the patients arrive randomly as they ordinarily do and there is no filter except the eligibility, one method of random selection could be systematic (e.g., every fifth), and another method is to include consecutive eligible patients arriving in a clinic or hospital within a specified period. Otherwise, random numbers can be used for selection. *Note that this random selection is different from random allocation.*

**Ethical** considerations such as **informed consent** can preselect a biased group. Some patients or some clinicians may have strong preference for a particular therapy, and they can refuse randomization. Some eligible patients may refuse to participate when they are told that they could be randomized for a **placebo** or a new untested therapy. Some may refuse because it is a trial and not treatment per se. Considerable efforts may be needed to keep such refusals at a minimum.

In addition, the groups should be such that there is an a priori uncertainty about the efficacy of the test therapy in them. This is called patient equipoise and helps to ensure that the patients are homogeneous. Patient equipoise implies guarding against unwitting tendency to include subjects who are likely to benefit from the drug under trial without declaring that the trial is restricted to such specific groups. Sometimes health-conscious people agree to enter a trial while, in fact, participants should be fair representatives of the class of patients that are finally targeted to benefit from the regimen in case the trial is successful. The participants should be genuinely uncertain about the outcome of the trial so that the results are unbiased. See the topic **equipoise** for other such requirements.

Be extra cautious in trials on severe cases. Beside medical surveillance that these cases require, note that some such cases sometimes hardly have scope to deteriorate further in case the drug is ineffective—they can only improve. If the patients have very high blood pressure (BP), they may remain so, whereas patients with relatively low BP can easily show a rise. When comparing cases of high BP with cases of low BP for efficacy of a drug, the drug may be unnecessarily considered to have caused the rise in the low BP group. In some situations, it could be just the reverse.

**Bias** can still occur in subtle or unknown ways in an RCT despite random allocation and blinding. A major source of bias is negligence to follow-up. If the follow-up requires recalling or revisiting the patients, some may not turn up or may refuse to cooperate, some may be untraceable, and some could die from unrelated causes. Even if the outcome assessment is within the hospital stay, some can leave against medical advice. Another factor that could affect a clinical trial is the need to change the treatment modality midway if any patient develops a serious illness. In addition, there could be patients who did not follow the full regimen. This is called partial **compliance**. Take preemptive steps to minimize such losses and plan to adjust the results if needed. Also make sure that the experience gained on previous subjects does not alter the method of assessment of subsequent cases. The fundamental assumption of independence of the subjects must not be violated.

### *Control Group in a Clinical Trial*

As in other experiments, clinical trials can have one or more treatment regimens, but a parallel or concurrent **control group** is almost invariably required except in **crossover trials** and **before–after**

**design**. As the name implies, controls are necessary in an RCT. Real controls are those that are similar but follow the natural course of disease without any intervention. In practice, however, the reference control group is either treated with an existing regimen or administered a placebo so that (i) any **Hawthorne effect** is neutralized and (ii) any psychological effect of a regimen is accounted. See the topic **placebo** for details of the conditions where placebo could be justified. There is a debate whether surgical trials need a group with sham surgery as placebo. Perhaps evidence is not enough that sham surgery has the same psychological benefits as a placebo in a drug trial. Nevertheless, a sham surgery group can be adopted for a setup where it is not too expensive and is harmless to the participants.

For some medical maneuvers such as renal dialysis and fitting of an artificial limb, for which a placebo group is nearly impossible, the existing regimen is given to the control group—now called active control. But the control group is now widely considered as a scientific necessity. Such a group too should have similar baseline and must be subjected to the same maneuvers as the test group, except for the regimen itself. The control group should undergo the same medicinal rituals, such as dietary regulations, as the treatment group. If a nonparallel group is a control, then appearance, schedule of administration, discomfort, etc., may cause differential compliance. Such a trial cannot be **double blind**. Finding a strategy that minimizes bias in such cases can be a challenging task. Whatever bias creeps in will have to be tackled at the time of analysis of data, and this too may not be easy to handle.

An apparently simple approach is the comparison of the current test group with a group previously treated with the required control regimen including placebo. This is called historical control. This may be derived from previous clinical trials or records such as registries and databases. The advantage is that no concurrent control is required—thus, the requirement of cases reduces by half and the cost also reduces accordingly. Ethical issues regarding recruiting and exposing subjects to the control regimen are also avoided. Historical controls may be appropriate for a disease that has a relatively stable natural history, and understanding of prognostic aspects has not changed. Multiple controls in this setup may help increase the level of confidence if the results replicate in each group of controls.

Despite demonstrable equivalence, the results are rarely accepted as definitive when based on historical controls. The flaw is that some factors may have changed over time such as diagnostic techniques, and evaluation procedures may have improved over time. Known changes can be accounted for in the interpretation of results, but there might be some obscure changes that could affect the results. Lack of randomization also compromises the credibility of results in this setup. Historical controls may not have been monitored with the same keenness as one would with concurrent controls, causing bias in the results.

Notwithstanding the strong argument made above for some kind of controls, there might be situations where they are not needed. If a treatment is being tried for a rapidly fatal disease such as tuberculous meningitis, where is the need for a control group? Saving of some cases is enough evidence of efficacy. Drugs with dramatic effects such as penicillin do not need a control. Utility of Pap smear was established without recourse to a controlled trial. However, such instances are rare.

### *Some Subtleties of Statistical Analysis of RCT*

Perhaps it needs to be emphasized that *the difference between treated and control must be statistically significant for it to be medically relevant*. If it is not statistically significant, nobody can be confident that the difference is actually present in the corresponding population. Thus, the first step is to assess statistical significance. If it is not significant and the statistical power is adequate, there is no need to worry about its medical relevance in most cases because it is likely to be there in the sample by chance. If significant, further statistical testing is used to judge if it reaches a medically relevant threshold. This threshold comes from medical acumen, and value judgment is required. In a rare case, statistical nonsignificance can be ignored. If in an RCT, 3 deaths occur in 100 subjects in the placebo group and none out of 100 in the treatment group, the difference is not statistically significant, but would you not try this treatment for your family if everything else has failed? Additional factors such as environment, family condition, and availability of health infrastructure are also considered while taking a final decision regarding the management of a patient.

When considering analysis, **intention-to-treat (ITT)** is an ingenious strategy to partially circumvent the limitation imposed by a specific kind of distortion in the data. This strategy is particularly advocated for RCTs. See that topic for details.

Some professionals feel that an RCT is the only mode to provide scientific evidence of a cause–effect relationship. A moment's reflection will convince you that this is not so. Most accept that cigarette smoking *causes* lung cancer, but no trial has ever been conducted. The controlled experiments do provide direct evidence, but evidence from observational studies can be equally compelling when confounders are really under control and the results replicate in a variety of settings. In any case, an RCT is almost never undertaken for a potentially harmful substance or regimen.

Sample size, $n$, is just about the first thing that comes to mind when planning a trial. It is among the most important considerations that determine the utility of a trial. For a discussion on this aspect, see the topic sample size for study formats, where a subsection about clinical trials is given.

For those who want to know more, Ref. [2] is a seminal text on RCTs. Also see the topic **adaptive designs for clinical trials**.

1. Gianotti L, Braga M, Nespoli L, Radaelli G, Beneduce A, Di Carlo V. A randomized controlled trial of preoperative oral supplementation with a specialized diet in patients with gastrointestinal cancer. *Gastroeneterology* 2002;122:1763–70. http://www.ncbi.nlm.nih.gov/pubmed/12055582
2. Piantadosi S. *Clinical Trials: A Methodologic Perspective*. Wiley, 2013.

## randomized response technique

This is a procedure to collect usable information on sensitive items that the respondents are generally hesitant to answer correctly. Such items could be regarding extramarital sex, abortions, consumption of illegal drugs, going to quacks, drunk driving, etc. Many persons are hesitant to reveal their correct income, and for some women even their exact age. They may elect not to reply at all to such questions or reply with incorrect answers. The resulting bias is difficult to assess. It can be removed by allowing the respondent to answer privately through a randomizing response technique (RRT). This is done through a device in his/her control as per the details given next. The method was proposed by Warner [1] in 1965.

Take an example of a survey about abortion. Let the objective be to estimate the proportion of women who have ever had an abortion. This is a sensitive issue and a correct answer may be evasive. Under the RRT, two inverse questions are framed: (a) "Have you ever had an abortion?" and (b) "Have you never had an abortion?" Answer to both is either yes or no. Thus, the answer does not reveal which question has been answered. The RRT is implemented as follows. Ask the respondent to roll a die away from the interviewer and tell her to correctly answer (a) if the die face is 1–4 and answer (b) if the die

is 5 or 6. Assure the respondent that no one will know which question she is answering. Only the respondent knows which question is answered, and the investigator does not know the outcome of the die. But the investigator knows that the probability of answering (a) is $\theta = 2/3$ and that of answering (b) is 1/3. Since the answer to both (a) and (b) is either yes or no, the respondent is assured that her answer would remain confidential and so might answer correctly.

Let $\pi$ be the true proportion in the population that we want to estimate. This means that the probability of a "yes" response to question (a) is $\pi$ and the probability of a "yes" response to question (b) is $(1 - \pi)$. These are mutually exclusive answers. Thus, the probability of "yes" response to either (a) or (b) by addition rule is

$$P(\text{yes}) = \theta\pi + (1 - \theta)(1 - \pi).$$

This gives $\pi = \dfrac{P(\text{yes}) - (1-\theta)}{2\theta - 1}$ and its estimate $p = \dfrac{x/n - (1-\theta)}{2\theta - 1}$, where $x$ is the number of "yes" responses by $n$ women. In our rolling a die method, we know that $\theta = 2/3$. Thus, for this method, $p = \dfrac{x/n - 1/3}{1/3}$. Therefore, the proportion in the population with abortion can be estimated. In place of rolling a die, any other randomized mechanism that protects the privacy can be used, but this method requires that $\theta$ should neither be 0 nor 1 nor ½. Thus, tossing a coin is not an option. When $\theta = 1/2$ as in tossing a coin, another method called innocuous question method [2] is used. Blair et al. [3] have developed a multiple regression technique for analyzing randomized responses and have also proposed power analysis.

Despite the availability of this method for a long time, there is a dearth of substantive applications. Perhaps the reason is the use of an outside method of randomized device that can confuse the respondents. Respondents possibly will not feel confident in answering sensitive questions despite this device.

Chen et al. [4] used randomized response method to estimate commercial sex proportion among men having sex with men and condom use in China. Wolter and Preisendörfer [5] compare the RRT to direct questioning with a survey that contains sensitive questions. Their data came from 552 face-to-face interviews with subjects who had been convicted in a metropolitan area in Germany for minor criminal offences. They found that the RRT does not lead to higher prevalence estimates of sensitive behavior than does direct questioning. Secondly, effects of individual and situational determinants of misreporting differ between the two question modes. The effect of need for social approval is stronger in RRT than in the direct question mode.


1. Warner SL. Randomized response: A survey technique for eliminating answer bias. *J Amer Stat Assoc* 1965;60(309):63–9. http://www.ncbi.nlm.nih.gov/pubmed/12261830
2. Lensvelt–Mulders GJL, Hox JJ, van der Heijden GM. How to improve the efficiency of randomized response designs. *Qual Quant* 2005;39:253–6. http://joophox.net/publist/qq05.pdf
3. Blair G, Imai K, Zhou Y–Y. Design and analysis of the randomized response technique. *J Amer Stat Assoc* 2015;110(511):1304–19. http://imai.princeton.edu/research/files/randresp.pdf
4. Chen X, DU Q, Jin Z, Xu T, Shi J, Gao G. The randomized response technique application in the survey of homosexual commercial sex among men in Beijing. *Iran J Public Health* 2014;43(4):416–22. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4433722/
5. Wolter F, Preisendörfer P. Asking sensitive questions: An evaluation of the randomized response technique versus direct questioning using individual validation data. *Sociol Methods Res* 2013;42:321–53. http://smr.sagepub.com/content/42/3/321.full.pdf


## randomness (statistical tests for)

When dealing with data, you might want to determine whether they are indeed random. This kind of problem generally arises while generating random numbers such as for simulations. First, consider just what is meant by **random**. If the data are single numbers, a datum is random when it is drawn from an equally probable set of possible values. Statistically, it will have a **uniform distribution**. Also, if you take a sequence of random numbers, each number drawn must be statistically independent of the others. That is to say that drawing one value does not make that value less likely or less likely to occur again. So with an unloaded die, if you roll a six, that does not mean that the chance of rolling another six changes.

Among many tests available to check randomness of data, runs test is the most popular.

### *Runs Test for Randomness*

In the case of tossing a coin, randomness implies that there are no big sequences of consecutive heads or tails. If a coin shows up 8 consecutive times, would you not doubt the fairness of the coin or of the toss? A consecutive similar outcome is called a run. In the following sequence of 15 tosses, there are 8 runs—4 of heads and 4 of tails:

| T T | H H H | T | H H | T T | H H H | T | H |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |

In case of numerical values, a series of increasing values or a series of decreasing values would be a run. These can also be divided as below median and above median, ignoring the values exactly equal to median (hoping that these are not many) as – and +, or 0 and 1. Too few runs indicate nonrandomness. But there is a problem with too many runs also. The sequence with many runs such as 0 1 0 0 1 1 0 1 is not necessarily random—it could be cyclical. Thus, the number of runs should not be too many either. Runs test for randomness is based on the basic premise that the runs should be neither too few nor too many. For large samples,

$$\text{Runs test for randomness: } t = \frac{R - \bar{R}}{s_R} \text{ with } (n_1 + n_2 - 1) \text{ df,}$$

where

$n_1$ = number of positive runs and $n_2$ = number of negative runs,
$R$ = number of runs in your data ($R = n_1 + n_2$),

$$\bar{R} = \frac{2n_1n_2}{n_1 + n_2} + 1,$$

$$s_R^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}.$$

Reject the null hypothesis of randomness if $|t|$ exceeds the critical value at $\alpha$-level of significance. This could be easily done with the help of software. In the literature, you may find $z$ instead of $t$, although it legitimately is $t$ because of the estimated standard error (SE) of $R$ in the denominator. For large $n$, it really does not make much of a difference. For small samples, there are separate tables [1] to determine critical values that depend on values $n_1$ and $n_2$.

A **Cusum chart**, as described separately in this volume, can also be used as a visual test for randomness. If the plot follows a discernible pattern such as most values on one side or continually increasing or decreasing, the randomness is in doubt. Another test

for randomness uses frequency within a block. This exploits the premise that the number of 0's and 1's expected in different blocks of values should be the same if the data are random. For details, see Kenny and Mosurvski [2]. Despite these tests, it is hard to find that a sequence of numbers is really random. All possible sequences are equally likely, and it is possible to get six zeroes in a row. Some perfectly random sequences may appear nonrandom and may fail the randomness test, but the suspicion rises if many sequences fail the test.

1. Mendenhall W, James R. *Statistics for Management and Economics*, Fourth Edition. Duxbury Press, 1982.
2. Kenny C, Mosurvski K. *Random Number Generators: An Evaluation and Comparison of Random.org and Some Commonly Used Generators*. Trinity College Dublin Project Report, April 2005.

## random numbers

What do we mean by random numbers? These are those numbers that do not follow any predictable pattern. There are a number of methods of building up a series of random numbers for use in selection of cases or otherwise. Apart from all physical methods, the method of building up a sequence from a random number table is conceptual.

With two participants, the tried and tested method, beautiful in its simplicity, is to toss a coin. Other methods are dice, cards, and drawing lots. A different and possibly more scientific method is to use a random number table of the type shown in Table R.2, where three-digit random numbers are arranged in 10 columns and 30 rows. This table can be used to generate one, two, three, or higher digit random numbers. Each digit is random and there are 900 digits in this table. You can start at any random point and pick up one, two, or more consecutive numbers to form a random number of the size you want.

It would be more accurate to call these as pseudo-random numbers as they are generated by a mathematical process. These tables are to be found in statistics books, or they can be produced by computers and some types of calculators.

Although random allocation of subjects to the regimen in an experiment or a trial is directly done by computer programs, this can also be done by the random number table of the type shown in Table R.2. If there a total of 60 subjects for allocation to two groups of 30 each, select the first 30 distinct random numbers beginning at any random point in the table that are ≤60. If any number repeats, ignore this number and also exclude 000. These 30 will be in group I and the remaining 30 in group II. You can move across or downward in the random number table for reading the numbers; that does not matter because the numbers are not just random but also randomly arranged. But decide in advance how you plan to go from one number to the next since no change is permissible after seeing the numbers.



## random sampling, *see* sampling techniques

## random sampling methods

A sample is called random when inclusion or exclusion of a particular eligible subject in the sample depends on chance and cannot be predicted in advance. However, as will be illustrated shortly, the chances are not necessarily equal for all subjects for inclusion in the sample. For this reason, it is sometimes prudent to call it a **probability sampling**.

Random selection is just a strategy to get a representative sample. If representativeness can be achieved by any other method, the same can be adopted. The larger the sample relative to the size of the population, the more the chance of it being representative—random or not. A large number of methods are available for choosing a random (or probability) sample. However, only some are commonly used. For details, refer to the following topics: **simple random sampling**, **stratified random sampling**, **multistage random sampling**, **cluster random sampling**, and **systematic random sampling**. Note that all of these methods include the term "random." This is because choosing a sample at random is an easy way of getting a truly representative sample. Not choosing at random can result in the sample being biased, although this kind of sample may still be applicable in some situations. For an overview of nonrandom methods, see the topic **sampling techniques**. Some of these such as **haphazard sampling**, **snowball sampling**, and **volunteers** are also discussed as a separate topic.

Typically the simple random sampling (SRS) allows the same chance of selection to each unit in the population. Other sampling procedure can have unequal chance despite being random. For example, in stratified random sampling, if you are randomly selecting 20 out of 100 mild cases and 20 out of 40 serious cases, the chance of selection of each mild case is 1/5 and the chance of each serious case is 1/2. These chances are not equal. This can be so with multistage and cluster random sampling also. For this reason, proportionate sampling is sometimes advocated that assigns the same chance to all individuals, but that will have a relatively small sample of cases that are rare in the population—defeating one of the purposes of stratified sampling of adequate size from each stratum.

A visual comparison of some methods of random sampling is shown in Figure R.2. This may help to distinguish among various methods. The method of choice is always SRS unless it is difficult to adopt. The estimate provided by SRS is generally the most precise. Some problems with SRS are the following:

- Nonavailability of **sampling frame**. This is the list of items of subjects to be sampled and is needed to adopt SRS.
- SRS is likely to select widely dispersed subjects, making the approach difficult. In case of choosing people residing all over a state by SRS, they may be too far off from one another.
- It is possible in SRS that specific groups of interest are not adequately represented. For example, you may find after random selection that the females are underrepresented in your sample.
- If you are using a random number table for selection, obtaining so many distinct random numbers may be difficult as the numbers tend to repeat after a while.

The first two problems can be handled by either cluster or multistage sampling. The answer to the third is stratified sampling, and the remedy for the fourth is systematic sampling because it requires only one random number. The systematic method also does not require the full frame. In any case, with the wide availability of computers, generating distinct random numbers is not a problem. Although small-scale studies may use one of the preceding methods, medium- and large-scale studies would generally be based on a mix of two or more methods.

**TABLE R.2**
**Random Number Table Used in Random Number Generation**

| Row # | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 939 | 720 | 840 | 376 | 644 | 997 | 729 | 271 | 249 | 224 |
| 2 | 987 | 736 | 804 | 840 | 389 | 290 | 704 | 784 | 115 | 613 |
| 3 | 527 | 613 | 677 | 893 | 900 | 473 | 198 | 448 | 673 | 798 |
| 4 | 142 | 979 | 621 | 927 | 805 | 920 | 993 | 806 | 467 | 438 |
| 5 | 763 | 871 | 634 | 286 | 280 | 535 | 310 | 727 | 711 | 978 |
| 6 | 302 | 047 | 470 | 350 | 551 | 412 | 899 | 364 | 325 | 597 |
| 7 | 881 | 205 | 178 | 361 | 133 | 151 | 045 | 218 | 963 | 549 |
| 8 | 572 | 977 | 367 | 011 | 614 | 266 | 605 | 767 | 105 | 296 |
| 9 | 653 | 330 | 748 | 906 | 656 | 989 | 554 | 614 | 369 | 417 |
| 10 | 708 | 165 | 950 | 994 | 733 | 688 | 320 | 839 | 825 | 149 |
| 11 | 213 | 006 | 336 | 356 | 414 | 878 | 990 | 608 | 652 | 849 |
| 12 | 321 | 435 | 091 | 069 | 588 | 425 | 670 | 070 | 177 | 134 |
| 13 | 367 | 707 | 172 | 814 | 490 | 158 | 856 | 202 | 078 | 089 |
| 14 | 823 | 437 | 303 | 187 | 698 | 503 | 479 | 434 | 535 | 018 |
| 15 | 581 | 490 | 346 | 215 | 530 | 901 | 135 | 519 | 411 | 689 |
| 16 | 427 | 885 | 343 | 678 | 577 | 763 | 542 | 579 | 598 | 043 |
| 17 | 305 | 202 | 479 | 187 | 006 | 826 | 951 | 091 | 075 | 857 |
| 18 | 161 | 107 | 533 | 548 | 342 | 011 | 841 | 103 | 645 | 382 |
| 19 | 854 | 543 | 696 | 987 | 119 | 131 | 502 | 323 | 062 | 555 |
| 20 | 582 | 220 | 831 | 889 | 106 | 101 | 917 | 910 | 929 | 340 |
| 21 | 428 | 905 | 595 | 105 | 917 | 960 | 156 | 730 | 126 | 086 |
| 22 | 602 | 718 | 449 | 096 | 995 | 106 | 631 | 222 | 198 | 856 |
| 23 | 241 | 696 | 018 | 818 | 995 | 539 | 640 | 857 | 317 | 261 |
| 24 | 758 | 636 | 411 | 373 | 370 | 867 | 578 | 682 | 357 | 486 |
| 25 | 755 | 741 | 465 | 871 | 434 | 091 | 747 | 724 | 084 | 336 |
| 26 | 423 | 189 | 866 | 011 | 814 | 893 | 871 | 688 | 790 | 656 |
| 27 | 419 | 519 | 836 | 596 | 570 | 023 | 087 | 385 | 379 | 665 |
| 28 | 712 | 159 | 418 | 025 | 208 | 068 | 824 | 877 | 069 | 689 |
| 29 | 490 | 611 | 925 | 282 | 289 | 800 | 582 | 306 | 443 | 009 |
| 30 | 469 | 581 | 449 | 922 | 938 | 729 | 328 | 722 | 782 | 623 |

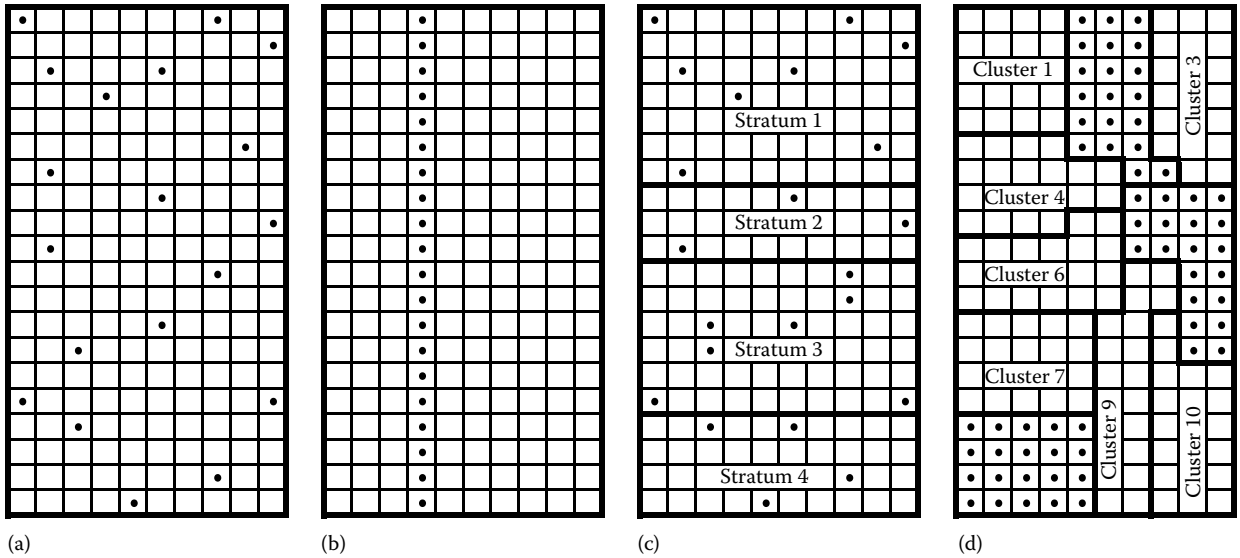



**FIGURE R.2** Visual illustration of some random sampling methods: (a) simple; (b) systematic; (c) stratified; (d) cluster (squares with dots are in the sample).

### *Other Methods of Random Sampling*

The most prominent among other random methods of sampling are the **probability proportional to size (PPS) sample**, **area sample**, **inverse sample**, and **sequential sample**. Refer to these topics.

Thompson [1] is a useful reference for further details of random sampling methods.

1. Thompson SK. *Sampling*, Third Edition. Wiley, 2012.

## random variables, *see* variables

## range, *see* variation

## rank correlation

This is the correlation between **ranks** of quantitative values in two series rather than the values themselves. In some cases, two variables $x$ and $y$ clearly increase or decrease together, but the relationship is not necessarily linear. This is called a **monotonic** relationship (Figure R.3). The dependence of height on age of children is an example of such a relationship. This is not linear but definitely exists. The relation of visual acuity with age after 60 years may be monotonically decreasing with a sharp drop after the age of 80 years. The usual method to measure the strength of *linear* relationship for such data is the Pearson **correlation coefficient**. This is adequate in most situations but not for the situations just listed. The linearizing effect of the product–moment correlation coefficient in such cases amounts to an oversimplification of the relationship. Also it can provide misleading results when outliers are present. Its test of statistical significance is valid only when at least one of $x$ and $y$ has Gaussian distribution. When these conditions are not met, rank correlation is used.

The best known measure of rank correlation is the **Spearman correlation coefficient**. For this correlation, the values of $x$ and $y$ are separately ranked from 1 to $n$ in increasing order of magnitude, and the ordinary product–moment correlation coefficient is computed between ranks of $x$ and ranks of $y$. This simplifies to the following:

$$\text{Spearman rank correlation coefficient: } r_S = 1 - \frac{6\Sigma d^2}{n(n^2-1)},$$

where $d$ = (rank of $y$ – rank of $x$) and where the summation is over all pairs of observations. This coefficient always lies between –1 and +1. When computed for all subjects in the population, it is denoted by ρ. Calculation of rank correlation is illustrated in Table R.3 for the height and weight of eight children. This is $r_S = 1 - \frac{6 \times 0.50}{8 \times 63} = 0.994$. This example illustrates the following:

1. Equal ranks are assigned to the tied observations. In this example, the second child and the fifth child have the same weight. They would have received ranks 4 and 5 but now received a midrank of 4.5 each because of the tie.
2. The ordinary product–moment correlation coefficient between weight and height in this example is 0.967. The rank correlation is slightly higher. This would have been exactly +1 if the weight of the second child were 13 kg and the height the same 92 cm. The rank correlation can sometimes overrate the strength of the relationship because it partially disregards the actual magnitude of $x$ and $y$. This is clear from the explanation given next.
3. The value of rank correlation would not change if, for example, the height of the third child were 114 cm instead of 127 cm. This is because the rank is not changed by such an alteration in this case. Rank correlation, thus, is not fully sensitive to the exact values of the variables.

One big advantage of rank correlation is that it attenuates the effect of outliers. You may be aware that only one outlier can substantially distort the value of the product–moment correlation coefficient. The value of rank correlation is not so much affected because a high value is converted just to the next rank. This is a nonparametric procedure, and the test of its significance does not require a Gaussian pattern of either $x$ or $y$. This correlation is thus preferable when the distribution pattern of both $x$ and $y$ is far from Gaussian.

Rank correlation, being a nonparametric procedure, requires a **permutation test** for checking its statistical significance for small $n$. For large $n$, the same Student $t$-test with df = $(n - 2)$ as used for the product–moment correlation coefficient can be used, namely, $t = r_S\sqrt{\frac{n-2}{1-r_S^2}}$. For the confidence interval, **Fisher $z$ transformation** can be used for large $n$ just as for the product–moment correlation coefficient.

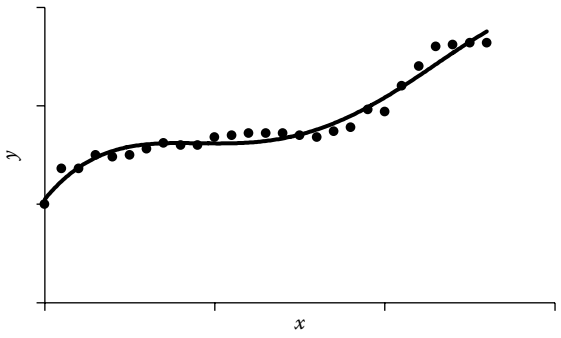



**FIGURE R.3** Monotonic but not linear relationship.

**TABLE R.3**
**Calculation for Rank Correlation Coefficient**

| | Child 1 | Child 2 | Child 3 | Child 4 | Child 5 | Child 6 | Child 7 | Child 8 | |
|---|---|---|---|---|---|---|---|---|---|
| Weight (kg)—$y$ | 10 | 14 | 26 | 6 | 14 | 18 | 32 | 4 | |
| Height (cm)—$x$ | 65 | 92 | 127 | 63 | 95 | 96 | 128 | 55 | |
| Rank ($y$) | 3 | 4.5 | 7 | 2 | 4.5 | 6 | 8 | 1 | |
| Rank ($x$) | 3 | 4 | 7 | 2 | 5 | 6 | 8 | 1 | |
| Difference ($d$) | 0 | +0.5 | 0 | 0 | –0.5 | 0 | 0 | 0 | $\Sigma d^2 = 0.50$ |

**TABLE R.4**
**Minimum Value of $r_s$ for Different $n$ (≤ 10) That Is Statistically Significant for Different α Levels (Two-Tailed)**

| | Minimum Value of $r_s$ | | | |
|---|---|---|---|---|
| **Sample Size ($n$)** | **α = 0.10** | **α = 0.05** | **α = 0.02** | **α = 0.01** |
| 1, 2, or 3 | None | None | None | None |
| 4 | 1.000 | None | None | None |
| 5 | 0.900 | 1.000 | 1.000 | None |
| 6 | 0.771 | 0.886 | 0.943 | 1.000 |
| 7 | 0.714 | 0.786 | 0.892 | 0.929 |
| 8 | 0.643 | 0.738 | 0.810 | 0.857 |
| 9 | 0.600 | 0.683 | 0.783 | 0.817 |
| 10 | 0.564 | 0.648 | 0.733 | 0.781 |

*Source:* Snedecor GW, Cochran WG. *Statistical Methods*, Eighth Edition. Iowa State University Press, 1980; p. 478. With permission.
*Note:* For $n \geq 11$, use Student $t$-test.

For samples of 10 or fewer pairs, the minimum values of $r_S$ for different significance levels are given in Table R.4. For $n \geq 11$, the Gaussian pattern holds reasonably well, and the Student $t$-test can be used with df = $(n - 2)$ as just stated.

1. Snedecor GW, Cochran WG. *Statistical Methods*, Eighth Edition. Iowa State University Press, 1980; p. 478.

## ranking and selection

Ranking of groups is achieved by putting groups in order from minimum to maximum with respect to mean or proportion, marking those that are not significantly different. Selection is identifying the best or the worst or the median. The statistical method for ranking and selection is generally traced to a paper by Bechhofer [1] in 1954.

Ranking and selection problem usually originates from a question such as, "Which diet, out of $K$ available options, is the best to prevent cancer?" and "Which is the next best?" In statistical terms, the objective is to find the order of the groups such that $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_K$ or $\pi_1 \geq \pi_2 \geq \ldots \geq \pi_K$. Whereas some of these can be strict inequalities, others may be equalities only. When such ordering is done, the group with the largest or the smallest mean or the smallest proportion is automatically identified. The constraint imposing the challenge is that we should be able to do this with probability of Type I error less than the level of significance α.

Ranking and selection can also be understood as identifying the best set of alternatives from the given options based on their expected performance. The methods of **order statistics** are used to address the problem. These methods are based on the distribution of any order statistics of interest. For example, $\bar{x}_{[k]}$ is the $k$th-order statistics if $(k - 1)$ means are less than this and the remaining $(K - k)$ are more than or equal to this. Median and percentiles are the popular example of order statistics. Statistical distribution of such statistics is not straightforward and may require some intricate mathematics.

Complexities in ranking can be illustrated as follows. Suppose that the largest sample mean is $x_{[K]}$. Because of sampling fluctuation, this could have come from the population whose mean is just the middling type. What is the chance that this indeed has come from the population with largest mean $\mu_{[K]}$? If values in all groups follow a Gaussian pattern with the same variance $\sigma^2$, this could be shown to be [2]

$$P(\bar{x}_{[K]} \text{ from the population with mean } \mu_{[K]}) = P\left(z_k < z_K + \frac{\mu_{[K]} - \mu_{[k]}}{\sigma/\sqrt{n}}, k = 1, 2, \ldots, K-1\right),$$

where $n$ is the sample size from each population and $z_k = \frac{\bar{x}_k - \mu_k}{\sigma/\sqrt{n}}$. It is with the help of such probabilities that ranking and selection is done.

The situation is illustrated in Figure R.4, which shows the sample means on the lower side and the population means on the upper side. In practice, we will have only the lower half as the upper half on population will be unknown. It is possible that the sample mean of group 2 is much lower than its population mean and becomes substantially larger in another sample.

When considered in an oversimplified manner, ranking can be done on the basis of **multiple comparisons**. If there are five groups A, B, C, D, and E, multiple comparisons methods would determine which groups are significantly different from others, and can provide the results of the following type:

C B E A D

This is the presentation when they are ordered by the values of sample mean, and the mean in group B is found not statistically significant from group E, and group E not significantly different from group A and group D. However, group B is significantly different from group A, and group C with the minimum mean is significantly different from all other groups. In this manner, we are able to rank these groups with respect to their mean. When the results of multiple comparisons are as shown, it is possible to say that group C has minimum mean, and there is no clear ranking between group B and group E: they are in the indifference zone. Similar conclusions can be drawn for ranks of other groups as well. For example, in this case, no clear maximum could be established by the amount of available data.

However, the multiple comparisons just mentioned are oversimplification because the question in selection is identifying the maximum irrespective of statistical significance [3]. Multiple comparisons unnecessarily increase the required sample size. Thus, alternative and more robust procedures are used that consider the minimum specified difference between the "best" and the second best. These are classified as subset selection procedures and

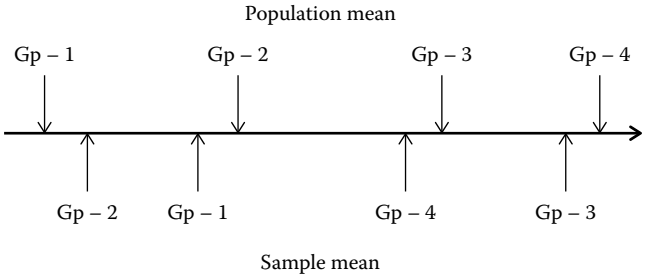



**FIGURE R.4** Order of sample means and population means in four groups.

indifference-zone selection procedures. For details, see Gupta and Panchapakesan [4].

Cui et al. [5] provide a good example of microarray experiments where the interest was in a set of genes with the largest difference in gene expression under different methods—not just in significant genes. Thus, the method of ranking and selection was required for this problem.


1. Bechhofer RE. A single sample multiple decision procedure for ranking means of normal populations with known variances, *Ann Math Stat* 1954;25(1):16–39. http://projecteuclid.org/download/pdf_1/euclid.aoms/1177728845
2. Dudewicz EJ. *Introduction to Statistics and Probability.* Holt, Rinehart and Winston, 1976: p. 345.
3. Rasch D, Kubinger K, Yanagida T. *Statistics in Psychology Using R and SPSS.* Wiley, 2011.
4. Gupta SS, Panchapakesan S. *Multiple Decision Procedures: Theory and Methodology of Selection and Ranking Populations.* Society for Industrial and Applied Mathematics, 2002.
5. Cui X, Zhao H, Wilson J. Optimized ranking and selection methods for feature selection with application in microarray experiments. *J Biopharm Stat* 2010;20(2):223–39. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2909494/


## ranks

These are the relative positions of the observation in a sample with respect to their value. The most obvious example can be found in sports: first, second, and third. Similarly, the positions in the examination results are another example. As shortly spelled, such rankings are not rare in medical setup—thus, the idea behind them should be clear. Suppose there is a database that includes some extreme values that would bias the arithmetic mean of these values. The way forward is to rank the data and then use the median to get an idea of the "middle" data point. In addition, for medical measurements that defy exact quantitative measurement or are too expensive or inconvenient, ranks are easy portal for valid inference. It is easier to assess that body burns are extensive than to say that these are 88%. All ordinal measurements are ranks, and these are quite common in medical setups.

Statistical application of ranks is extensive for **nonparametric methods**. Quantitative values are assigned rank after ordering them from minimum to maximum. Consider the following values of fasting blood glucose level (mg/dL) of 12 subjects attending a diabetes clinic:

117, 124, 148, 208, 169, 209, 124, 110, 115, 124, 169, 151.

This would have ranks as in Table R.5 after arranging from the maximum to the minimum.

*Ties* are values that are quantitatively the same and given the same ranks. There are three persons with a value of 124. If they are distinct, they would get ranks 4, 5, and 6. But now because of the tie, they get a rank of 5.0 each. Rank 5.0 is the average of their supposed ranks of 4, 5, and 6. Next rank would be 7. Similarly, two persons with a value of 169 each get the rank of 9.5, which is the average of their supposed rank of 9 and 10. While using nonparametric methods based on ranks, special adjustment is made for such ties.

Realize that the maximum value in our example is 209. Had it been 221, the rank would still be the same. Similarly, the value 148 will continue to get rank 7 even if it is 150, and the value 117 will continue to get rank 3 even if it is 123. Thus, the ranks do lose some information, and this is what makes ranks-based nonparametric methods less efficient in many applications.

## rapid assessment method, *see* cluster sampling

## Rasch analysis

Georg Rasch was a Danish statistician known mostly for the development of a class of measurement models now known as Rasch models. These models are used for analyzing the ability of a test, such as a questionnaire, to measure what it intends to measure. The objective of the Rasch analysis (RA) is that the total score of any test measures the target ability of the subjects and nothing else. That is, the questionnaire or any such test should not be affected by the person's age, sex, or any such characteristics not of interest. For example, the questionnaire should be such that two physicians assessing severity of disease in a patient come up with the same answer.

Georg Rasch

The purpose of RA is to increase the chance of a more able person correctly answering any item compared with a less able person, and to increase the chance of getting a more difficult item correct by a more able person compared with a less able person. In this process, some items of questionnaire may be recalibrated. The procedure has limitations, however, as it depends on who and how many respondents attempt different items. Also, it generally assumes that all items have equal correlation with the latent trait intended to be assessed. RA also helps in reducing redundancy in the items of the questionnaire to yield a more valid and simple questionnaire without sacrificing the requisite information.

The principle that RA follows is that the questionnaire should fit the Rasch model instead of the usual in which the model

**TABLE R.5**
**Ranks of 12 Values of Fasting Blood Glucose Level**

| Value | 110 | 115 | 117 | 124 | 124 | 124 | 148 | 151 | 169 | 169 | 208 | 209 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rank | 1 | 2 | 3 | 5.0 | 5.0 | 5.0 | 7 | 8 | 9.5 | 9.5 | 11 | 12 |

should fit the data. Rasch is a predetermined model. If the fit is not good, the questionnaire is changed. The reason is that Rasch modeling is seen as the most rational way for measuring ability or trait in a linear and continuous way. It follows the principle that a more able person has a higher chance of correctly responding to more difficult items. Thus, the probability of the correct response solely depends on the person's ability and the difficulty level of an item.

The fit statistics obtained from RA indicate whether to delete, rescore, or reword an item. Answers to items may be ordinal, such as extremely dissatisfied to extremely satisfied, but RA transforms this to exact quantitative measures that can be used for more exact statistical analysis. When properly done on an adequate sample, RA provides a valid measure that can be extrapolated to its population.

General statistical software packages may not have the provision to do RA, and special software packages for this purpose may have to be accessed. This analysis flags the items that do not fit the model. Perhaps no set of items can exactly fit the Rasch model, and the items that do not fit are dropped or modified. Calculations such as mean square fit are used to find out whether the items fit the model or not. This measure is independent of the sample size, whereas the statistical tests of significance are intimately dependent. Those items with mean square fit between, say, 0.7 and 1.3 are considered adequate. A lower value indicates that the item is undesirably correlated with extraneous characteristics, whereas a higher value indicates that the item is too unreliable.

da Rocha et al. [1] give further information on RA. They aimed to present the main characteristics of RA in the context of patient-reported outcomes in psychiatry as follows, using as an example depressive symptoms, measured by using the Beck Depression Inventory (BDI). In fitting data to the Rasch model, the authors aimed to demonstrate the structural validity of the BDI scale. They also illustrate how RA can inform on how best to use scales: the meaning of the numbers therein; the latent traits that such numbers represent; and how best statistical operations could be used to analyze them. They illustrate how fitting data to the Rasch model has become the measurement gold standard for patient-reported outcomes in general and, more specifically, how it facilitates a quality improvement of outcome instruments in the field of psychiatry.

For further details on RA, see Bond and Fox [2].

1. da Rocha NS, Chachamovich E, de Almeida Fleck MP, Tennant A. An introduction to Rasch analysis for psychiatric practice and research. *J Psychiatric Res* 2013;47:141–8. http://www.ncbi.nlm.nih.gov/pubmed/?term=An+introduction+to+Rasch+analysis+for+psychiatric+practice+and+research
2. Bond TG, Fox CM. *Applying the Rasch Model: Fundamental Measurement in Human Sciences*, Second Edition. Lawrence Erlbaum, 2007.

## rate

Rate is a measure of the frequency of occurrence of a phenomenon. It counts the number of events occurring in a specific period per unit of time. Time is an essential element of this concept—per day, per month, per year, etc. *Since this frequency can change over time, rate is time specific.* If 8 deaths occurred per 1000 population in a particular year, this becomes the death rate for that year. If, on average, 2 new cases of cirrhosis appeared per 1000 population in the year 2012 in an area, we call it the incidence rate per 1000 for that year. When the variation from time to time is not large, the reference to the point in time can be omitted. If the death rate from accidents remains fairly constant at about 14 per million population per year after year in an area, a general statement can be made without mentioning any particular year. But new discoveries and advances in technology can alter any rate even when the ground situation remains the same. Thus, exercise caution in comparing rates over time.

Essential components of a rate are (i) a numerator, (ii) a denominator, (iii) a specification of duration, and (iv) a multiplier such as percent or per thousand. The multiplier helps to convert an awkward-looking fraction to a convenient and nice number. The denominator is generally the number of subjects in the group *at risk* for the events counted in the numerator. Measures such as birth rate per thousand are exceptions because the children born are not part of the entire population, and yet it is called a rate. Since the essential feature of a rate is the frequency of occurrence over time, it is legitimate to call it a birth *rate*.

The main purpose of computing a rate is to be able to validly compare two or more groups, times, places, etc. The denominator is population, patients, subjects, etc., which vary periodically and geographically. Rate makes the denominator uniform and thus allows valid comparison. However, a rate is a valid tool only for stable conditions over the period for which it is calculated. If the population varies within that period, such as the beginning of the year and the end of the year, it is conventional to use the average population, say at the middle of the year. Herein lies the catch. If a city has a population of 10,000 at the beginning of the year and 6000 perish due to a severe earthquake in the month of February, the midyear population would be nearly 4000 (some would die due to routine causes and some births may take place). In that case, the death rate, when based on the midyear population, for that year would be nearly 6000 × 1000/4000 or 1500 per 1000 population, or 1.5 per person: ridiculous! This rate would be entirely different if earthquake deaths occur in the latter half of the year. This dramatic example illustrates that the midyear or average population can be fallacious if deaths, or for that matter any event for which rate is required, do not occur regularly throughout the period at nearly a uniform rate. The concept of rate is not applicable when such unusual spikes or dips occur. Also, when the duration of observation is not uniform for all the subjects, rate per **person-time** is calculated with the limitations mentioned in that topic.

Rates can be averaged so long as the denominator is the same. If the incidence rate of a disease in a vulnerable section of population is 20 per 1000 per year and in robust section is 10 per 1000 per year, the rate for both together will not be 15 unless both groups are equal. If 70% of the population is robust and 30% vulnerable, the average incidence rate is 13 per 1000 population, which is the **harmonic mean**. Similarly, if the incidence rate is 5 per 1000 per year for 2 years and 8 per 1000 per year for 5 years in the same total population, the average for 7 years is not (5 + 8)/2 = 6.5 but is 7.14 per 1000 per year. This can be checked by simple arithmetic by calculating the total number of new cases (10 + 40 = 50 cases per 1000 in 7 years).

When rate is a dependent (outcome) variable in a regression, the ordinary least squares method will not be generally applicable. Instead, use **Poisson regression**.

## rate of homogeneity, *see* design effect and the rate of homogeneity

## ratio

Generally speaking, a ratio is one quantity relative to another. It can be expressed as *a:b* or as *a/b*. In a broad sense, all rates are also

ratios because there is a numerator and a denominator. In practice, the usage of the term *ratio* is restricted to a situation in which the numerator is not a part of the denominator, nor is the denominator a part of the numerator. Both are separate and distinct entities. The ratio of white blood cells to red blood cells is 1:600. The sex ratio in a general population is computed as the number of females per 1000 males, and the dependency ratio as the number of dependents (say, of age –15 and 65+ years) in a population to the working (15–64 years) population. Doctor–population, bed–population, waist–hip, acid output basal to maximal, and albumin–globulin ratios are other examples. All these ratios have biological meaning—they do convey the state of health or disease in some sense.

"Ratio" is also encountered in the term "ratio scale." A scale is an instrument on which the characteristics are measured. Those characteristics that can be exactly measured in terms of a quantity are said to be on a metric scale. The metric scale can be further subdivided into interval and ratio scales. In an interval scale, there is no absolute zero: for example, body temperature. On the other hand, in a ratio scale, a zero point can be meaningfully designated. It is correct to say that the duration of survival of 6 years is twice as much as 3 years, and parity 3 is thrice as much as parity 1. On the other hand, body temperature rise from 100°F to 105°F cannot be called a mere 5% rise.

Among statistical ratios, the most prominent is the *F*-ratio, commonly used in the analysis of variance and regression. Factually speaking, even Student *t* is a ratio of mean and its standard error (SE) adjusted for the degrees of freedom. Odds ratio is frequently used to measure risk in retrospective studies. Relative risk used in prospective studies is also a ratio.

Ratios can rarely be averaged; therefore, be careful if you come across any average ratio or if you want to use average ratio.

## ratio estimator

This topic introduces the procedure for using auxiliary information about the population to estimate a given unknown population parameter of interest with better precision. This auxiliary information may be for a known variable to which the unknown variable of interest is related. The auxiliary information may be straightforward to measure, while measurement of the actual variable of interest might be expensive. Ratio estimator uses the ratio of auxiliary information to the sample values to provide a better estimate of the parameter of interest.

Suppose the variable of interest is *y*, which is strongly (linearly) correlated with another variable *x* through origin. The latter implies that if *x* is zero, *y* also is zero. Suppose *x* is the auxiliary variable whose even population mean $\mu_x$ is known. You observe both *x* and *y* for a random sample of *n* subjects. The usual estimate of mean of interest is $\bar{y}$. But the ratio estimate of mean of *y*:

$$\hat{\mu}_y = \frac{\bar{y}}{\bar{x}} * \mu_x \text{ if } x \text{ and } y \text{ are positively correlated, and}$$

$$\hat{\mu}_y = \frac{\bar{y}}{\mu_x} * \bar{x} \text{ if } x \text{ and } y \text{ are negatively correlated,}$$

would have far more reliability compared with $\bar{y}$ alone. Means can be replaced by totals. The higher the correlation between *x* and *y*, the more the reliability. But this comes at a cost. This estimate requires that (i) *x* is also measured along with *y* for the sample subjects, (ii) the population mean $\mu_x$ of *x* is known, and (iii) the relationship between *x* and *y* is linear through origin. Ratio estimate is used only when the first two pieces of information can be easily obtained without much effort, or when they are already available, and when the third condition holds.

The auxiliary variable (*x*) could be the baseline values, and the variable of interest (*y*) may be the values at the end point after follow-up. Baseline values may be available for all patients admitted to a clinic such as their age. This could be useful if the outcome of interest is intimately related with age. For example, lung functions after the age of 40 years are intimately related to age. Suppose you know from records that the average duration of workers exposed to pollutants in a factory is 12 years and their average total lung capacity at recruitment is 4.6 L. In a sample of 30 such workers, the average duration of exposure was found to be 10 years and their total lung capacity 4.1 L. This sample shows a decline of 0.5 L. How much is the decline on average in all the workers? Here, the duration of exposure is *x* and $\mu_x = 12$ years. In the sample, $\bar{x} = 10$ years, and the average decline in lung capacity in the sample $\bar{y} = 0.5$ L. If the decline is linearly dependent on the duration of exposure, $\bar{\mu}_y = \frac{0.5}{10} \times 12 = 0.6$ L.

This innocuous-looking procedure can mislead if the relationship between *y* and *x* is not linear and is not through origin. In our example, the relationship will not be linear if the decline in total lung capacity tends to be increasingly steep with duration of exposure. That is unlikely in this example, but keep this in mind in applications. Note that this is regression through origin since when the duration of exposure is zero, the decline in lung capacity is also zero.

Jeelani et al. [1] listed a large number of modified ratio estimators for various conditions. For example, Subramani [2] proposes a modified ratio estimator for the estimation of the population mean of the study variable when the population median of the auxiliary variable is known. The bias and mean squared error of the proposed estimator are derived and are compared with those of existing modified ratio estimators for certain known populations. Further, the author has also derived the conditions for which the proposed estimator performs better than the existing modified ratio estimators. From the numerical study, it was observed that the proposed modified ratio estimator performs better than the existing modified ratio estimators for certain known populations.


1. Jeelani MI, Mir SA, Nazir N, Jeelani F. Modified ratio estimators using linear combination of co-efficient of skewness and median of auxiliary variable under rank set sampling and simple random sampling. *Indian J Science Technol* 2014;7(5):723–8. http://www.indjst.org/index.php/indjst/article/viewFile/50159/40871
2. Subramani J. A new modified ratio estimator for estimation of population mean when median of the auxiliary variable is known. *Pak J Stat Operation Res* 2013;9:137–45. http://www.pjsor.com/index.php/pjsor/article/viewFile/486/301


## ratio scale, *see* scales of measurement (statistical)

## RCT, *see* randomized controlled trials (RCTs)

## recall lapse

A recall lapse is the genuine failure to remember when reporting past events or characteristics. This can introduce bias in a medical setup in a variety of ways.

In a **case–control study**, cases may easily recall past events since they are sick or affected, whereas controls may fail to do so. The opportunity of accurately observing the events as they unfold, as in a prospective study, is not available in a case–control study. Dependence is entirely on the reporting done by the subject or on the records. Recall lapse can occur, and the records may be incomplete,

adversely affecting the results. This is commonly seen in smoking studies where retrospective recall method is used to gather data on daily cigarette consumption [1].

Despite widespread and fairly well-known phenomenon, nothing much seems to have been studied regarding the impact of recall lapse on study findings and how to adjust it. It stands to common sense that the events far back in time would be forgotten more frequently than recent occurrences; that is, recall lapse is directly proportional to the recall gap. It also stands to reason that the events making serious impact on health will be remembered, and the mild events would lapse. Adjustment can be made in situations where the correct values can be obtained by some intensive efforts, but not otherwise.

Wang et al. [1] analyzed smoking history by recall and ecological momentary assessment that records each cigarette as it is smoked. The analysis suggests that recall lapse contributes substantially to the distribution of self-reported cigarette counts. They also found that recall lapse was more common in females than males. Thus, recall lapse can also depend on the host characteristics such as age and sex in addition to the physical features such as recall time and severity of health consequences of the event. Note that recall lapse is different from intentional misreporting. Some details of recall lapse in the context of demographic surveys have been described by Som [2].

1. Wang H, Shiffman S, Griffith SD, Heitjan DF. Truth and memory: Linking instantaneous and retrospective self-reported cigarette consumption. *Ann Appl Stat* 2012;6(4):1689–706. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3889075/
2. Som RK. *Recall Lapse in Demographic Enquiries*. Asia Publishing House, 1973.

## receiver operating characteristic (ROC) curve, *see also* *C*-statistic

This is the plot of **sensitivity** of a quantitative test versus its (1 – **specificity**) (true positive rate vs. false positive rate) for different values of the test as illustrated with an example later in this section. Receiver operating characteristic (ROC) curve is used to assess the overall validity of the test and to find the best threshold that has maximum sensitivity and specificity. It can also be used to compare the validity of two or more quantitative tests.

Medical tests such as white blood cell count for appendicitis are continuous as a range of values are obtained. Tests such as presence or absence of abdominal pain are discrete—mostly dichotomous. The ROC curve deals with medical tests that give continuous measurements and cannot be used for qualitative characteristics.

Thresholds that minimize the sum of false positives and false negatives can also be obtained by a procedure called **discriminant analysis**. This procedure is particularly useful when the number of possible categories is three or more, say cirrhosis, hepatitis, and malignancy of liver. For discriminating between two competing diagnoses, two approaches are available as discussed next. The first approach based on the sensitivity–specificity ROC curve is more popular but less valid, and the second based on **predictivities** is a new and more valid approach.

### *Sensitivity–Specificity Based ROC Curve*

Is T4 better or TSH better in discriminating between hyperthyroid and euthyroid cases? Comparison of the performance of two quantitative tests can be obtained by the area under the ROC curve. See the topic ***C*-statistic** that measures this area. Historically, the name ROC comes from signal detection developed during World War II. For details of ROC curve, see Sackett et al. [1]. A brief is given as follows.

Consider the example of many full-term births in a hospital requiring induction of labor that succeeds in most cases but fails in a few. In case the induction fails, a Cesarean is done for delivery. This involves pain, time, and money and also requires mental preparedness. It would be nice for the woman and the family as well as the attending obstetrician if they are able to anticipate a Cesarean delivery on the basis of the patient characteristics. The traditional method for this is to compute the Bishop score based on dilatation, effacement, consistency, and position. Other parameters that influence the success of induction of labor are maternal age, parity, BMI, and amniotic fluid index. Suppose a study was carried out in $n = 166$ cases with prelabor rupture of membrane to find if the duration since rupture can help in predicting the Cesarean delivery. Although the study was prospective, sensitivity and specificity were calculated for different durations of rupture. There were 36 distinct durations. The values obtained for each minimum duration are shown in Table R.6.

**TABLE R.6**
**Sensitivity and (1 – Specificity) for Cesarean Delivery at Different Duration of Rupture of Membrane**

| Duration (h) Greater Than or Equal to | Sensitivity | 1 – Specificity |
|---|---|---|
| 0.00 | 1.000 | 1.000 |
| 0.63 | 1.000 | 0.976 |
| 0.88 | 1.000 | 0.969 |
| 1.25 | 1.000 | 0.890 |
| 1.75 | 1.000 | 0.866 |
| 2.13 | 1.000 | 0.819 |
| 2.38 | 1.000 | 0.811 |
| 2.75 | 1.000 | 0.780 |
| 3.25 | 1.000 | 0.717 |
| 3,75 | 1.000 | 0.709 |
| 4.50 | 1.000 | 0.646 |
| 5.13 | 0.971 | 0.583 |
| 5.38 | 0.971 | 0.575 |
| 5.75 | 0.971 | 0.551 |
| 6.25 | 0.914 | 0.378 |
| 6.75 | 0.914 | 0.346 |
| 7.13 | 0.857 | 0.291 |
| 7.38 | 0.857 | 0.283 |
| 7.75 | 0.857 | 0.276 |
| 8.25 | 0.800 | 0.189 |
| 8.75 | 0.800 | 0.181 |
| 9.25 | 0.743 | 0.110 |
| 9.75 | 0.743 | 0.102 |
| 10.25 | 0.543 | 0.039 |
| 10.75 | 0.543 | 0.031 |
| 11.50 | 0.457 | 0.024 |
| 12.50 | 0.400 | 0.008 |
| 13.50 | 0.343 | 0.000 |
| 14.50 | 0.286 | 0.000 |
| 15.50 | 0.257 | 0.000 |
| 16.50 | 0.200 | 0.000 |
| 17.50 | 0.171 | 0.000 |
| 18.50 | 0.143 | 0.000 |
| 19.50 | 0.114 | 0.000 |
| 20.25 | 0.057 | 0.000 |
| 21.50 | 0.000 | 0.000 |

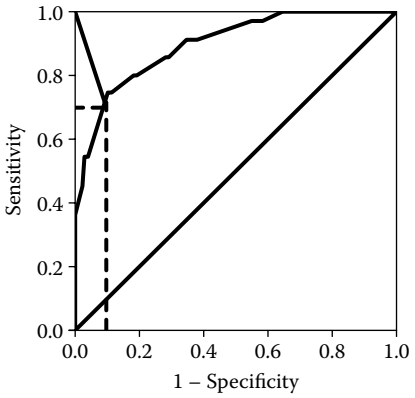



**FIGURE R.5** ROC curve of duration since rupture of membrane for Cesarean delivery.

The ROC curve obtained by plot at different cutoffs is shown in Figure R.5. The diagonal is the line of equality. A statistical software package found that the area under the curve (AUC) is $C = 0.898$ with its standard error (SE) = 0.029, and 95% CI from 0.841 to 0.956. With $C = 0.898$ out of a possible 1.0, it seems from this ROC that duration since rupture of membrane itself is a good indicator to anticipate Cesarean delivery. The best cutoff that maximizes (sensitivity + specificity) is 9.75 h. At this duration, the sensitivity is 0.74 and the specificity is 0.90 (1 – specificity = 0.10).

Although not shown in this example, the Bishop score was not found to be as good an indicator of an impending Cesarean as was duration since rupture in this example. Otherwise also, the Bishop score is subjective and nonreproducible because of higher interobserver and intraobserver variability. Our example illustrates how ROC can be effectively used for medical decisions. The example is illustrative only and should not be construed to mean that duration since rupture can be solely used to anticipate Cesarean. For this, studies in different locales are needed.

The ROC curve is useful in

- Evaluating the discriminatory ability of a test to correctly pick up known diseased and nondiseased subjects
- Comparing efficacy of two or more medical tests for assessing the same disease
- Comparing two or more observers measuring the same test (interobserver variability)
- Finding the optimal cutoff point of a quantitative test to least misclassify the diseased and nondiseased subjects.

The first of these—discriminatory ability—is assessed by sensitivity–specificity at different cutoffs as already explained, and the overall performance is measured by the area under the ROC curve. Efficacy of two or more medical tests is also compared by the AUC—better test will have a higher area. The same is used to compare two or more observers. The details of the last, namely, the optimal cutoff point, are given next.

### *Methods to Find the "Optimal" Threshold Point*

The ROC curve can help to identify a threshold that gives the highest sum of sensitivity and specificity in situations where these two are available for a large number of values. In our example, the values are on a continuous scale, but ROC can also be obtained when they are in spaced categories. The number of such categories must be at least five for the ROC curve to be adequate. In our sample, there are 36 categories.

Three criteria are used to find the optimal threshold point from the ROC curve. These are known as points on the curve closest to (0, 1), the **Youden index**, and the minimize cost criterion, respectively. The first two methods give equal weight to sensitivity and specificity and impose no ethical cost and no prevalence constraints. The third criterion considers cost, mainly the financial cost, for correct and false diagnosis, cost of discomfort to person caused by treatment, and cost of further investigation when needed. This method is rarely used in medical literature because it is difficult to estimate the respective costs and prevalence is often difficult to assess. The other two are described next.

If $s_n$ and $s_p$ denote sensitivity and specificity, respectively, the **Euclidean distance** between the point (0, 1) and any point on the ROC curve is $d = \sqrt{[(1 - S_N)^2 + (1 - S_P)^2][(1 - S_N)^2 + (1 - S_P)^2]}$. To obtain the optimal cutoff point to discriminate the disease with nondisease subject, calculate this distance for each observed cutoff point and locate the point where the distance is minimum. Most of the ROC analysis software packages calculate the sensitivity and specificity at all the observed cutoff points, allowing you to do this exercise.

The second, the Youden index [2], maximizes the vertical distance from line of equality to the point $(x, y)$. The main aim of the Youden index is to maximize the difference between true positive rate ($s_n$) and false positive rate ($1 - s_p$), and a little algebra yields $J = \max(s_n + s_p)$. This is discussed as a separate topic.

These procedures for finding the optimal threshold are applicable when both sensitivity and specificity are equally important. If they are not, expert judgment may be required to find an appropriate cutoff depending upon whether sensitivity or specificity is more important.

For details of the area under the ROC curve and how this is used for inference, see the topic ***C*-statistic**. A word of caution is in order. Nothing is impossible, but it is extremely unlikely in health and medicine that you get area under ROC = 1. But the question has tremendous significance. Sometimes the sample values are such that they give you an area nearly equal to 1. This is more common with small samples where it is most likely a case of overfitting. If you get an area nearly equal to 1 in large samples, be assured that the evidence is firmly for an excellent test.

### *Predictivity-Based ROC Curve*

The threshold based on the conventional ROC would be valid across the population since sensitivity–specificity is not affected by prevalence of the disease. However, the diagnostic efficiency is obtained by the predictivities and not by the sensitivity–specificity, and predictivities are affected by the prevalence of the disease. Thus, the ROC curve between positive predictivity and (1 – negative predictivity) may be more useful in a local setup as it takes prevalence into account and uses the right kind of indicators.

When predictivities are not known, the relationship between predictivities and sensitivity–specificity and prevalence can be used to estimate these (see **predictivities**). These can be used to find the criterion that is best to confirm the diagnosis (i.e., maximally increase the positive predictivity) and to exclude the disease (i.e., maximally increase the negative predictivity). Consider the following example of creatine phosphokinase (CPK) in myocardial infarction (MI) to understand how this is done.

Figure R.6 is drawn for thresholds 350, 250, and 150 U/L of CPK level for detecting MI assuming that (sensitivity, specificity)

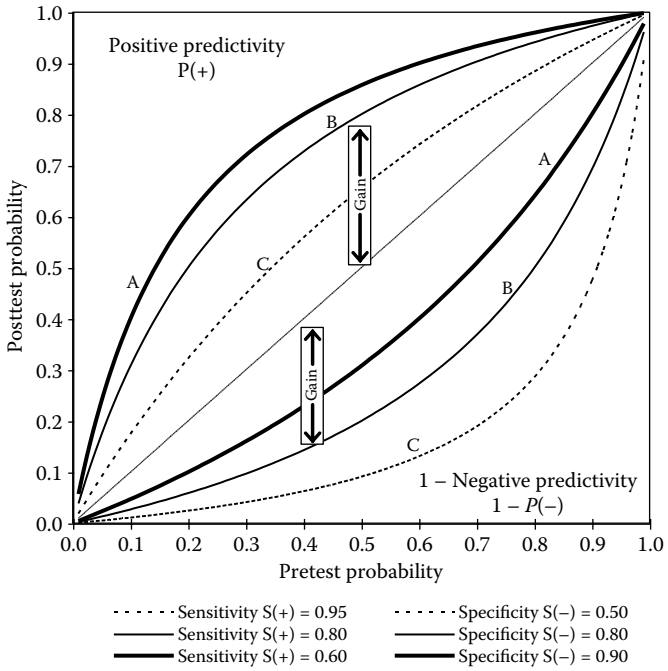



**FIGURE R.6** Illustration of the relationship between pretest and posttest probability of presence of disease (positive predictivity) and absence of disease (negative predictivity).

for these thresholds, respectively, are (0.60, 0.90), (0.80, 0.80), and (0.95, 0.50). Shown are positive predictivity on the upper left side of the diagonal and (1 – negative predictivity) on the lower right side of the diagonal for different prevalences (pretest probability). Thus, this is a predictivity counterpart of the ROC curve. The curves are marked as A, B, and C for the three pairs of sensitivity–specificity levels, respectively, corresponding to the three CPK levels under consideration. When sensitivity and specificity are equal, the curves are symmetrical as illustrated by curve B in this figure. Depending on the pretest probability for the patient in hand, which could be either the known prevalence of infarction in these types of cases or the **personal probability** on the basis of history and signs and symptoms, and the CPK level present, you can immediately obtain the posttest probability (or predictivity) of the presence of infarction with such curves.

If the chance of infarction in a patient with specific signs and symptoms is 60% (pretest probability 0.60) and the CPK level is found to be 150 U/L, then curve C applies and the posttest probability of MI can be read as 0.70. This is a gain of merely 10% over pretest probability. If the CPK level is 350 U/L, then curve A applies and the posttest probability is 90%—a handsome gain of 30% over the pretest probability for the presence of the disease. These gains can be utilized to find a threshold CPK level that is best in the sense of highest gain over a particular pretest probability. Now, note the following.

All the discussion of the sensitivity–specificity and predictivities assumes that these can be exactly obtained. In practice, these will be based on the study of a sample and are subject to sampling error. Similar variation is also expected in the prevalence rate, and this too will be generally based on a sample. Thus, caution should always be exercised in interpretation of the predictivities. Even when a pretest probability is based on the clinician's belief concerning the presence of disease after taking the history and examination into consideration, it would most likely vary from clinician to clinician. Thus, the values of sensitivity–specificity and predictivities serve only as guidelines and do not have much utility in an absolute sense. The ultimate decision, as always, rests with the attending clinician and with whether to give credence to such indicators.

1. Sackett DL, Haynes RB, Gyatt GH, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine*, Second Edition. Little Brown, 1991.
2. Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biom J* 2008;50:419–30. http://www.ncbi.nlm.nih.gov/pubmed/18435502

## record linkage, *see also* medical records

Where you have two or more records on any given patient for the same or disparate disease, record linkage is a method of assembling the information contained therein such that the same individual is counted only once and the record is comprehensive. This also helps in removing duplicates and for confirming one source with another. Knies et al. [1] provide a good example of record linkage for community-level surveys.

When unique identifiers are present, the linkage is immediate and causes no further challenge. This is called exact matching. Records of the patients for laboratory investigations, radiological examinations, clinics, and wards are linked with this identifier. If this comes up automatically through computer databases, as in most countries around the world, the linkage is easy. In a setup where unique identifiers are entered manually, the probability of wrong punch is not ruled out.

The process of record linkage becomes challenging in setups where there are no such unique identifiers. This can happen when trying to link birth and death registration records with hospital records or with census records, or while linking one hospital record with health center records. In such cases, vital characteristics of the persons are matched. For example, the name in one record may be Mark JH and in another record as Jay H Mark, and there might be two or more persons with the same name even when the parents' names are matched. Information such as dates of hospitalization and diagnosis are checked with the corresponding information in the other record, besides age and sex. Such matching may have a chance of error and is called statistical matching. In this case, the first step is to frame the rules regarding the characteristics to be exactly matched so as to minimize the chance of mislinkage and the characteristics for which flexibility is allowed. The linkage is done in the second step.

1. Knies G, Sala E, Burton J. *Consenting to Health Record Linkage: Evidence from the British Household Panel Study.* Institute for Social & Economic Research. https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2011-27.pdf

## records (medical), *see* medical records

## reestimation of sample size

This is encountered in **adaptive designs** of clinical trials. Suppose after due consideration of available knowledge, you plan a trial on 1000 subjects in each arm with 1 year follow-up. After 2 months into the trial, you find from **interim appraisal** that your anticipations at the planning stage were incorrect and the trial will give you confirmatory results about efficacy (or lack of it) in just 6 months

(or that you need to extend it to 16 months to get the adequate number with the desired end point); or only 700 subjects would suffice (or 1500 would be needed); or that doses you are trying are too high (or too low) and you need to have a middling dose; or that a concomitant treatment is needed; or that a particular subgroup such as males of age 70+ years needs to be excluded. Sometimes even the baseline information on the enrolled subjects may indicate that modifications in the design are needed. Such questions specially occur in a **group sequential design**.

Among various adaptations just listed, the statistically most relevant is the reestimation of the sample size on the basis of the actual **effect size** found at interim stages. Reestimation requires intricate statistical inputs as this is done to preserve the planned **level of significance** and the statistical **power** for detecting a specified effect. Such adaptation does not cause much of ethical problems; rather it seems to enhance ethics by keeping provision to stop the trial early in case convincing evidence of efficacy or of futility appears. Although the general perception is that sample size reestimation can increase or not increase but can never decrease (except in the case of stopping for futility), there is no statistical justification for regulators to consider this. If a trial is inappropriately curtailed and consequently fails to get a significant result, the loss is of the trialist and not of the regulatory authorities. Thus, the authorities may be unconcerned about reduction in sample size—their concern is that the trial remains fairly powered and the **Type I error** remains under control. Reduced sample size is more of a concern for ethics committees who may worry about wastage of efforts and unnecessary exposure to the patients if the reduced size does not produce convincing results.

McClure et al. [1] give an example of how all this works in practice. The authors point out that sample size estimation at the time of planning is often based on educated guess of the parameter values, and that these may in fact prove to be overestimates or underestimates. For example, after the study is started, the published data may indicate that the recurrent stroke rates might be lower than initially planned for the study. Failure to account for this could result in an underpowered study. The researchers may perform a sample size reestimation and evaluate different scenarios based on the reestimated overall event rate, including increasing the sample size and increasing the follow-up time, and determine their impact on both the Type I error and the power to detect the initially planned treatment difference. They illustrated that by increasing the sample size from 2500 to 3000 and by following the patients for 1 year after the end of recruitment, they would be able to maintain the planned Type I error rate and also increase the power of the study.

As another example, consider a trial planned on 330 subjects in each group, where this sample size is based on the efficacy of 70% and the power 80% to detect an effect (difference in efficacy) of 10%. After conducting the trial on 200 subjects in each group, suppose you find that the efficacy is just 40%. If that were so, the power with a sample size 330 would be only 73% for detecting a difference of 10% in efficacy. If the power 80% is to be retained, the sample size must be increased. The answer to "how much increase" comes from sample size reestimation. Important features of this procedure are given next for quantitative outcome with a **Gaussian** (normal) **distribution**.

Let the number of subjects already assessed at interim appraisal stage be $n_1$ in each group. Obtain the sample means ($\bar{y}_1$, $\bar{y}_2$) and the sample standard deviations (SDs) ($s_1$, $s_2$) for two groups in the trial. Since they would be generally required to have a common SD, a pooled estimate of $\sigma^2$ is $s_p^2 = \left(s_1^2 + s_2^2\right)/2$. The difference that this sample would be able to detect is $t_{(2n_1-2)}\sqrt{2s_p^2/n_1}$. If the difference you planned to detect is more than this, the sample size $n_1$ is already adequate, and there is no need to go any further. If not, additional subjects are included such that the final sample size $n$ per groups is at least as much as $t^2_{(2n_1-2)}\left(2s_p^2/d^2\right)$, where $d$ is the minimum medically relevant difference that you propose to detect [1]. This will ensure the power originally planned and the probability of Type I error not exceeding prefixed $\alpha$. This procedure requires that the trial is unblinded. If blinding is important, this exercise is done by a third group such as the **Data Safety and Monitoring Board** (DSMB). Gould and Shih [2] have developed a procedure that obviates the need to break the codes.

Bowden and Mander [3] have provided an example of a trial comparing medical treatment of knee osteoarthritis with surgery to relieve the pain. There were doubts about the effect size likely to be seen in the RCT context. Thus, a trial with interim analysis was planned. They followed a different and more complex procedure of sample size reestimation, but the interim analysis after 50 patients in each arm was revised to include 140 per arm—an addition of 90 patients per arm to meet the power requirement and to keep the Type I error within the specified threshold.

We have presented an overly simplified version of the complex topic. For further details of sample size reestimation, see Shih [4].


1. McClure LA, Szychowski JM, Benavente O, Coffey CS. Sample size re-estimation in an ongoing NIH-sponsored clinical trial: The secondary prevention of small subcortical stroke experience. *Contemp Clin Trials* 2012;33:1088–93. http://www.ncbi.nlm.nih.gov/pubmed/22750086
2. Gould AL, Shih WJ. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Comm Stat* 1992;21(10);2833–53. http://www.tandfonline.com/doi/abs/10.1080/03610929208830947#.VeARM_mqpBc
3. Bowden J, Mander A. A review and re-interpretation of a group-sequential approach to sample size re-estimation in two-stage trials. *Pharm Stat* 2014 May–Jun;13(3):163–72. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4288989/
4. Shih WJ. Sample size re estimation in clinical trials. Chapter 18 in: *Biopharmaceutical Sequential Statistical Applications* (Ed. Peace KE). CRC Press, 1992.


## reference category

A reference category is one that is used as the baseline for comparison with other categories. In a case–control study, the control group is the reference against which the performance of the case group is studied. It is generally the normative category. The concept of reference category is commonly used in **logistic regression**. When logistic coefficients are available, the corresponding **odds ratio (OR)** relative to a reference can be immediately computed. It is this property of the logistic regression that has made this method so popular. The category coded as $x = 0$ serves as the reference, and $e^b$ is the ratio of odds for category $x = 1$ relative to the reference category where $b$ is the logistic coefficient. For example, if $b = 1.3$ for a positive family history in case of breast cancer, OR = $e^{1.3} = 3.67$. This means that the odds of breast cancer in those with a positive family history are 3.67 times the odds in those without a positive family history.

For real nominal categories, suitable **contrasts** of interest or **indicator variables** define reference category. These contrasts are generally defined in terms of the difference of one group from one or more of the others. If the objective is to examine the relationship of 5-year survival with site of malignancy, you can have one particular site, say lung, as the reference category and compare this with oral, prostate, esophagus, etc. This can be done either by forming contrasts such as (lung—oral), (lung—prostate), etc., or by coding. In the case of coding, the computer package may have to be

told that these are categorical; else use the indicator variables. If you consider it more appropriate, you can have "prostate cancer" as the reference category. In this case, all other sites will be compared with this cancer. This requires coding accordingly with the help of indicator variables. The advantage with these kinds of coding is that each category will have its own logistic coefficient, and the coefficient for one category can differ from that of another category. Thus, differential effects of the categories on the response, if present relative to the reference category, will emerge. Since other categories are compared with the reference, the conclusions are more reliable when the number of subjects in the reference category is reasonably larger relative to the other categories.

In some situations, such as categories of subjects with no disease, disease 1, disease 2, etc., it is obvious that the category with no disease will be the reference. But there are situations where the reference category is not obvious as in the cancer example in the preceding paragraph. While comparing males with females, any sex can serve as the reference. Similarly, for studying risk of disease in a different blood group, any blood group can serve as the reference for comparison. In the case of marital status categorized as never married, currently married, divorced/separated, or widowed, you may have to make a judicious choice.

Statistically it does not matter what the reference category is. The results are going to be the same. It is just that the proper choice of reference category makes it easier to interpret the results. When this is important, do not leave that to the "default" category of the software package. Some packages use the highest category (the one that is stated last in your data or with the highest code in case of ordinal categories) as the default reference, whereas others use the lowest. When you are leaving it to the software package, just be more careful in interpreting the results.

## reference values, *see* normal range

## registration of births and deaths, *see* birth and death registration

## registration of trials

The conventional medium for publication of clinical trial results is a peer-reviewed journal. Such publications have many merits, but they also suffer from restricted access, fixed format, and limited space. Thus, many trials, particularly those with negative findings, do not find much favor with the editors, and such trials remain obscure. Also, trials that are partially reported because of limited space can accentuate bias. It is widely recognized that researchers have an ethical obligation to make the full results of human research public, whether or not the results of the study are positive.

Several innovations have evolved to meet this need. The World Health Organization (WHO) has led an international effort to promote registration of clinical trials at the time of initiation. International Committee of Medical Journal Editors (ICMJE) has issued its own registration requirements [1]. These requirements include submission of the complete protocol that would make it obligatory not just to publish but also not to miss out any inconvenient findings. This may also enhance the quality of the trial, because the protocol has to follow a standard format, and any deviation from the protocol would have to be explained. Selective reporting and suppression of unfavorable findings can be easily detected. This also requires public sharing of the data (except confidential and proprietary information) soon after publication of the results. The data can be posted on the individual websites for public viewing.

Besides WHO and ICMJE, different countries have agencies for registering a trial. In the United States, the National Institutes of Health has International Standard Randomized Controlled Trial Number (ISRCTN) with nearly 200,000 registered trials as of 2015, half of them outside the United States. The National Health Service of the United Kingdom has Clinical Research Collaboration (UKCRC). They provide support for high-quality, efficient, effective, and sustainable clinical trials research in the United Kingdom. Indian Council of Medical Research also has a registry, and there is a Brazilian Clinical Trials Registry.

ICMJE registry is for the publication of trial results in regular journals, but Wager [2] has argued that peer review has a poor record in detecting incorrect or fabricated data: the interpretation could be subjective, and claims can be unfounded. Wager made a passionate plea for open access as advocated by the Public Library of Science (PLoS). If the results are freely accessible on publicly funded websites, the cost and delays can be reduced. Journals can publish reviews and critiques of those results, and can also provide interpretation for different audiences such as researchers, clinicians, and patients. The journals may also relax its norms and not preclude articles based on web-posted results if a researcher chooses to do so.

BioMed Central publishes a large number of open-access journals. Among these is the *Journal of Negative Results in BioMedicine*. This is in realization of the need of professionals to know about negative trials as much as for positive trials.

1. De Angelis C, Drazen JM, Frizelle FA et al. Clinical trial registration: A statement from the International Committee of Medical Journal Editors. *N Engl J Med* 2004;351:1250–1. http://www.ncbi.nlm.nih.gov/pubmed/15364170
2. Wager E. Publishing clinical trial results: The future beckons. *PLoS Clin Trials* 2006. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1626095/

## regression coefficients, *see also* regression models (basics of)

Regression coefficients are the constants through which the relationship between a quantitative variable $y$ and a set of regressors ($x_1$, $x_2$,…, $x_K$) is expressed. These are best understood in a linear regression where the equation is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \varepsilon$. In this equation, ($\beta_1$, $\beta_2$,…, $\beta_K$) are the regression coefficients. This can be extended to include $\beta_0$ also, which is an intercept. The sample estimates of these regression coefficients are denoted by $b_0$, $b_1$, $b_2$,…, $b_K$, respectively, and are generally obtained by the **least squares method**. This method finds those values of $b$'s that minimize the sum of square of deviations of the observed values from the regression equation. Mathematical derivation of these estimates requires matrices and their inversion, requiring the level of mathematics we are avoiding in this book. However, this is easy to explain when there is only one regressor. For this, see the topic **simple linear regression**. Also see the topic **logistic coefficients**, which also are regression coefficients albeit in logistic regression.

### *Interpretation of Regression Coefficients*

Coefficients in the case of regression for quantitative variables have a very simple interpretation. In the regression equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_K x_K,$$

if $x_2$, for example, increases by unity, then $\hat{y}$ increases by $b_2$ when the other $x$'s remain the same. Thus, $b_k$ for any $k$ is the independent contribution of one unit of $x_k$ when the other $x$'s remain constant. (This in a way also explains why $x$'s should not be strongly correlated among themselves, called **multicollinearity**. If they are, it is difficult to imagine that only one varies without any change in the others.) Thus, these coefficients are "adjusted" for other $x$'s in the model. When a simple regression is obtained between $y$ and any one $x_k$ excluding the other $x$'s, the regression coefficient $b_k$ obtained is called unadjusted. The adjusted $b_k$ can be more than the unadjusted $b_k$, less than the unadjusted, or may not change, depending on how other $x$'s affect $y$.

Each regression coefficient measures the *average* contribution of the corresponding variable. The actual contribution in a subject can vary. This could be partly due to deviation from average and partly due to the influence of factors not included in the model. The attempt is to manage variations and consequent uncertainties by using statistical methods so that a valid conclusion can still be drawn. A **confidence interval (CI) for the regression coefficient** can be used for this purpose.

If the data are log-transformed, the interpretation of regression coefficients can dramatically change. See the topic **logarithmic scale/transformation** for details.

### *Statistical Significance of the Regression Coefficients*

When all specified regressors are forced into the model, then some of these may not be statistically significant and can be dropped without significantly affecting the efficiency of the model. In other words, some regression coefficients could be nearly zero. The test of $H_0$ that a particular regression coefficient is zero or not is done by using the **Student *t*-test** under some conditions such as Gaussianity, independence, and homoscedasticity. This is obtained as $t = b/\text{SE}(b)$, where SE($b$) is the standard error of $b$ and same as mentioned for its CI. Most software packages directly give $P$-values for each regression coefficient. Statistical significance is a reasonable assurance that the regression coefficient is not zero and that the corresponding regressor is indeed making some contribution in determining the value of the dependent variable.

A computer output for regression would generally also provide the $P$-value for the intercept $b_0$. But this needs to be interpreted with added caution. This $P$-value corresponds to $H_0$: Intercept = 0. This is plausible only when $y$ is expected to be zero for $x = 0$. This is not plausible, for example, when regression is between age and weight. Birth weight is the weight at age = 0 but is not expected to be zero. A zero intercept will not be of any interest except in a few special cases such as regression through origin.

### *Standardized Regression Coefficient*

A regression coefficient is called standardized when the corresponding regressor is **standardized**. This implies that in place of regressor $x_k$ in the regression equation, $x_k^* = \dfrac{x_k - \bar{x}_k}{s_k}$ is used. This standardized variable has mean 0 and variance 1 in the sample. The values of $\bar{x}_k$ and $s_k$ can be easily computed from the sample values; else the software can be asked to use standardized values in the regression. If $x_k$ is body mass index (BMI) and you have a total of $n$ subjects, $\bar{x}_k$ is the mean BMI of these $n$ subjects and $s_k$ is the standard deviation (SD). When the regressors are standardized, the dependent variable should also be standardized before running the regression. This will make the *regression through origin* with $b_0 = 0$.

Standardization removes the unit of measurement. If birth weight is measured in kilograms and if the ordinary regression coefficient, now called the metric regression coefficient, is 1.7, this would become 0.0017 when birth weight is measured in grams. There is no need to make this transformation when standardized values are used. But the interpretation of standardized regression coefficients is in terms of SD units. The second advantage of standardized regression coefficients is that they are comparable across the regressors. If one regressor is BMI and the other is calorie intake per day, it would be legitimate to conclude about which one is contributing more by comparing their standardized coefficients. This now is possible because the standardized coefficients are independent of units of measurement of the variables. Whereas this comparison would be statistically correct, the actual assessment of which regressors are more important than others would depend on their biological context. More important ones not only would have statistical significance but also would be those that are easy or less expensive to elicit and contextually relevant. The other problem with standardization is that this uses sample mean and sample SD, which themselves could be subject to large sampling fluctuation, particularly when the sample size is small. Also, mean and SD may not be valid in case the distribution is highly skewed. Quite often in a regression setup, the values of the regressors are deliberately chosen. Thus, the spacing could be quite arbitrary. This would affect the SD and correspondingly the standardized regression coefficient. The comparison of two standardized regression coefficients for assessing which one is contributing more could be jeopardized in case of arbitrary spacing of $x$'s.

Standardized regression coefficients can be obtained directly without standardizing the regressors and the dependent. It can be shown that

$$\text{standardized regression coefficient: } b_k^* = b_k * \frac{s_k}{s_y},$$

where $b_k$ is the ordinary (metric) regression coefficient, $s_k$ is the SD of $x_k$, and $s_y$ is the SD of the dependent $y$. When there is only one regressor (simple linear regression), the standardized regression coefficient is the same as the (Pearsonian) correlation coefficient.

Besides their usefulness in interpreting the importance of the regressors in explaining or predicting an outcome, regression coefficients are also sometimes utilized in developing scoring. A common and acceptable method of assigning scores to individual characteristics is based on the regression coefficients that are estimated using a multiple regression method when the outcome is quantitative. The method of assigning scores proportional to the regression coefficients is valid only when the values of the predictive factors are standardized to mean 0 and variance 1 before the regression is run. The regression coefficients, without this standardization, are not comparable and are severely affected by the unit of measurement. The method of assigning scores for qualitative dependent can be similarly based on the logistic coefficients as done for developing APACHE score.

## regression diagnostics, *see* regression fit (adequacy of), regression requirements (validation of)

## regression fit (adequacy of), *see also* regression models (basics of)

When performing regression analysis, there are a number of procedures and tools that can be used to test that the regression is as it should be: together, these are termed **regression diagnostics**. These diagnostics aim at two levels. First, check that all the requirements for fitting a regression model are adequately satisfied, and second, that the model obtained is indeed adequate or not. This section

discusses the second point. For the first, see the topic **regression requirements (validation of)**.

The most common criterion for judging adequacy of a regression fit is the proportion of total variance explained by the regression. This is named as squares of the **multiple correlation coefficient** ($R^2$) for linear regression and **coefficient of determination** ($\eta^2$) for other regressions. See these topics for details. The higher the variance a regression is able to explain, the better the regression becomes. Generally a model with $R^2$ less than 0.60 is considered not adequate, $R^2$ in the range (0.60–0.69) is considered reasonable to pursue, $R^2$ in the range (0.70–0.79) is considered as desirable, $R^2$ in the range (0.80–0.89) is considered good, and $R^2 \geq 0.90$ is excellent. Rarely ever $R^2$ will become as high as 0.90 in medical applications because of high variability among individuals. Statistical significance of $R^2$ is tested by the ***F*-test**.

The second criterion for adequacy of a regression model is its biological relevance. It should include all the characteristics as regressors that are considered important contributors to the outcome as per the current knowledge, although it may include others also that are statistically significant—biologically relevant or not in the hope that the relevance may emerge in course of time. More important is that the regression process must start with all relevant variables and nothing important is left out. For details, see the topic **regressors (choice of)**.

Not all observations involved in fitting a regression model contribute equally to the result. *Leverage* is the name given to the influence of each observation on the fitted values. While checking the validity of a model, the aim is to examine any or all observations with a high leverage. To check the effect of any such observation, refit the model without that observation and calculate what is called the *Cook distance*, which is a measure of the change that would occur in the estimated parameters in the model of interest if a particular observation were omitted. The larger the Cook distance, the more important the observation.

Further diagnostics, similar to Cook distance, are named *DFBETA and DFFIT*. DFBETA shows how much each coefficient changes in standard deviation units when an observation is excluded. The influence of any observation on the estimated mean response is measured by DFFIT. This is the number of standard deviations the *estimated mean value* moves if each observation, in turn, is excluded. Note that these are individual values and not for variables. Generally, apparent outliers are subjected to this analysis, and this can be done with the help of an appropriate software package. For variables, statistical significance as obtained by the Student *t*-test is used. For further details of DFBETA and DFFIT, see Montgomery and Peck [1].

The most popular criterion for assessing the adequacy of a regression model is the plot of the residuals of the regression. These deserve a topic of their own: refer to the topic **residuals**. If the fit is good, the residuals would be random with small variance and no outliers. Cook and Weisberg [2] have described such graphic methods for assessing the adequacy of regression models.

1. Montgomery DC, Peck EA. *Introduction to Linear Regression Analysis*, Fifth Edition. Wiley, 2012.
2. Cook RD, Weisberg S. Graphic for assessing the adequacy of regression models. *J Amer Stat Assoc* 1997;92(438):490–9. http://www.jstor.org/stable/2965698

## regression fitting (general method of), *see also* regression models (basics of)

Before the method of fitting a regression, consider the basic requirements. To obtain a regression equation (this is also called fitting a regression), the data required are the values of dependent $y$ corresponding to different values of the regressor $x$'s. For one set of values of $x$'s, one or many values of $y$ can be observed. For example, in a regression of body surface area (BSA) on age in children, you may have three children of the same age: 7 years. This is the value of $x$. Their BSAs are likely to be different. These are the values of $y$. In this case, for one $x = 7$, there are three values of $y$, one corresponding to each child. But $y$ should be available for several distinct values (at least four for the regression to be meaningful) of $x$. The values of the $x$ variables can be *deliberately* chosen to serve the purpose since they are considered fixed in a regression model.

The total number of individuals, as usual, is denoted by $n$. The number of regressors, $K$, should preferably be considerably less than the number of subjects $n$ (i.e., $K \ll n$), and the same $K$ regressors must be available for each subject in any analysis. The exact method of fitting varies according to whether $y$ is qualitative or quantitative, but the basic steps are common. For qualitative $y$, the dependent actually is the proportion of subjects with the specific quality. This gives rise to **logistic regression** and is discussed as a separate topic in this book. For quantitative $y$, generally ordinary **least squares** regression is obtained. Common steps for fitting for both ordinary and logistic regressions are as follows.

Step 1. Identify the dependent variable of interest. It should ideally be an outcome or response of the regressors. In the case of risk of lung cancer and smoking, smoking does not depend on the risk, but risk depends on smoking. Thus, risk is $y$ and smoking is $x$. In some cases, the direction of dependence may not be clear. Between cholesterol level and obesity, there could be a debate regarding which is a precursor of the other. Abnormal histology of the liver can precede cholecystitis or can follow cholecystitis. Both are admissible. In such cases, either could be investigated for its dependence on the other. Regression in these cases is just an expression of relationship and not of dependence.

Step 2. Recognize the nature of the dependent variable—that it is whether quantitative or qualitative. If quantitative, check whether it is likely to have a **Gaussian** (normal) **distribution** or a **Poisson** or any other distribution. If it is qualitative, check whether it is dichotomous or polytomous, and if polytomous whether it is ordinal or nominal. For details of these terms, see the topic **scales of measurement (statistical)**. These considerations will determine that ordinary quantitative regression is required or logistic, Poisson, or any other, and whether or not any transformation of the dependent is needed. Generally, most continuous dependents such as creatinine level and glucose level will follow a Gaussian distribution as such or after necessary transformation, small counts such as the number of family members with positive history follow a Poisson distribution, and dichotomous variables a binomial distribution.

Step 3. Identify the set of regressor variables. The number of regressors must be substantially less than the proposed sample size. The regressor set ideally should consist of all those that are known or suspected to influence the response $y$. It may not always be feasible to study all possible variables, and some selection may be necessary. Choose those that influence more than the others and those that are directly related in place of those that cause indirect influence. These should preferably be unrelated to one another, at least not intimately related. If they

are intimately related, **multicollinearity** is said to exist and can adversely affect the reliability of the regression. Regression equation will make more sense when only one of the highly correlated variables is retained. See the topic **regressors (choice of)** for further details.

Step 4. Decide the general form of relationship that you want to investigate, i.e., decide whether you want to restrict to linear regression, curvilinear regression, or a nonlinear form of regression. See the topic **regressions (types of)** for details of various types of regressions. The best guide for selecting the right type is the **scatter diagram**, particularly where there is only one regressor, and both $y$ and $x$ are quantitative. If there are many regressors, then many plots would be required: one corresponding to each regressor. Also decide if any **interaction** term is to be included. The basic function of regression analysis is to provide the best estimate of the regression coefficients once the form of regression, the dependent, and the regressors are specified. However, several forms can be explored, and the best can be objectively identified. Remember though that all regressions are compromises, and they never depict the data exactly.

Steps 1, 2, 3, and 4 together grossly specify what is called the **regression model**.

Step 5. Estimate the **regression coefficients**. This is done in such a manner that the fitted line or the curve passes closest to the points in the scatter plot. The mathematical method generally employed to obtain such regression coefficients is called the **maximum likelihood method**. This method finds those values of the regression coefficients that make our observed sample most likely to happen. The other method commonly used in regression is called the **least squares method**. For this reason, the whole method of regression with quantitative dependent is sometimes referred to as ordinary least squares. This finds a regression that is closest (least sum of the squared distances) to the points in the scatter diagram. Both these methods tend to produce an equation that *regresses* toward the mean of $y$ across different values of $x$ (see the topic **regression to the mean**). These two methods in many practical situations lead to the same estimates particularly when $y$ follows a Gaussian distribution for each set of $x$'s. Many software packages are available to perform the calculations required to estimate the regression coefficients. A feature of estimation of the regression coefficients is that some residuals are positive and others negative such that their sum is always zero. See the topic **residuals**. When the estimates of the regression coefficients are obtained, the regression model can be fully specified in view of steps 1 to 4.

Step 6. Test the goodness of fit of the obtained regression model. The method for this test depends on whether $y$ is quantitative or qualitative. See the topic **regression fit (adequacy of)**.

Step 7. The overall regression model may or may not be a good fit, but it is also important that individual regressors are tested separately for their significance. The method for testing this is also different for qualitative $y$ than for quantitative $y$. The regressors found statistically significant are considered to contribute to the relationship and are retained in the model. The "not significant" ones can be dropped without substantially affecting the utility of the model. Such deletions help to enhance the simplicity of the model, called parsimony.

Step 8. Calculate **residuals** and check for validity of requirements of the regression model. Residuals are the differences between the fitted value of $y$ and its observed value. The validity requirements are stronger for quantitative $y$ than for qualitative y. One of them is **homoscedasticity**, i.e., uniform variance of $y$ for different values of the regressors. A transformation such as the logarithm, square, and square root may be sometimes needed to meet this requirement. The other way to overcome this is to modify the method of estimation from least squares to *weighted least squares*. Also for testing or for computing confidence interval (CI), the residuals must follow nearly a Gaussian pattern in the case of the usual quantitative data. If the residuals are skewed, other distributions such as gamma, Poisson, or negative binomial are tried. These methods are intricate and not discussed in this book. A scatter plot of residuals versus each $x$ is often helpful in assessing the validity of requirements and in suggesting improvements to the model.

Sometimes a relation exists, but the regression fails to detect it. The success depends mostly on the adequate specification of the model in steps 1–4 and the sample size. Thus, failure to find a relationship through regression does not necessarily mean that no relationship exists.

One of the important functions of regression is to provide an estimate or to predict the value of $y$ for a given set of $x$'s. There is a fine distinction between an estimated value and a predicted value. The former term is generally used for the *mean* of $y$ for a group of individuals with a specific value of $(x_1, x_2,\ldots, x_K)$, and the latter is used for the value of $y$ for a single individual. The former has less variance than the latter, although both have the same value. Because the estimated and predicted values are the same, these terms are interchangeably used in this text.

Besides several topics in this book related to regression, a large number of books are available just on regression. Consult any of those in case you want more details. One among them is by Harrell [1].

1. Harrell F. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, Second Edition. Springer, 2015.

## regression models (basics of), *see also* regression fitting (general method of)

Regression models are the ones that depict relationship between two or more variables. Relationships are inherent in medicine and health, and it is common to investigate them in quest of new understanding. Eosinophil count normally decreases from birth to adulthood, and the risk of lung cancer increases with the amount and duration of smoking. Total leukocyte count (TLC) and hemoglobin (Hb) levels decrease up to the age of 6–8 years and show some increase thereafter in healthy subjects. Height velocity is at its peak during adolescence. Blood pressure (BP) levels have a positive relationship with the waist–hip ratio (WHR), and the incidence of dental caries in children depends negatively on the socioeconomic grade of their family. Such relationships are everyday occurrences in the practice of medicine and health, so much so that they are sometimes taken for granted. There are two primary purposes of studying such relationships: (i) to be able to predict, with some degree of precision, the value of one with the help of the others; and (ii) to help understand

the underlying mechanisms in such a relationship, particularly by studying the effect of alteration in the value of one variable when other conditions do not change. The latter can provide useful clues about what determines the outcome and how much influence each variable has.

It is known that height and weight are correlated with age in children, but visual acuity is independent of age in this group. Hb level, serum ferritin, and mean corpuscular volume are related to one another, but each is unrelated to WHR. Risk of breast cancer may be related to dietary constituents but unrelated to Hb level. But the relationship varies from person to person, and presence or absence of relationships is not absolute. Some relationships show up sometimes but vanish at other times. Variation and uncertainties play a prominent role in the presence or absence of such relationships. Some individuals or groups defy the nature and strength of relationships from the one seen in other subjects. Thus, care is needed in their use and interpretation. As always, the attempt is to identify the trend and separate it out from fluctuations. Relationship is just another term for such a trend.

What do we statistically mean by relationship? Consider the example of cholesterol level affected by obesity and hypertension. Cholesterol level is a continuous variable in this case. The degree of obesity can be measured on a continuous scale by body mass index (BMI), but it can also be divided into four groups, namely, thin, normal, overweight, and obese. Then the quantification of BMI is lost, although the gradient remains. That cholesterol is affected (or not affected) by obesity is a perfect conclusion in this case, but relation in statistical sense is the *nature or form of relationship.* Thus, this is an expression of quantitative change in one variable per unit change in the other. In the cholesterol level/obesity example, a relationship would be able to indicate how much increase in cholesterol level is expected when BMI increases from, say, 25 to 26, or from overweight to obese category. Such a form of relationship is called **regression** when the outcome is stochastic, i.e., varies from person to person. An equation such as BMI = weight/height$^2$ is not a regression equation. BMI does depend on weight and height, but it is not stochastic; it is not subject to any variation for given weight and height. Similarly, the equation for aortic valve area (AVA) in square centimeters, AVA = π*(LVOT/2)$^2$*SVTI/MVTI, where LVOT is the left ventricular outflow tract diameter (cm), SVTI is the subvascular velocity time integral (cm), and MVTI is the maximum velocity time integral across the valve (cm), is mathematical and not statistical since the dependent is not subject to any variation. The term regression is applicable only when the dependent variable is stochastic, i.e., it depends on chance.

In general, a regression model is expressed as

$$\hat{y} = f(x_1, x_2, \ldots, x_K),$$

where

$\hat{y}$ is the predicted value of the outcome or response of interest, called dependent

$f$ is some function that specifies the actual form

$x_1$, $x_2$,..., $x_K$ are the variables that may predict the variable $y$, called independents.

See the topic **dependent and independent variables** for various names and forms of these variables. Regression, in fact, is a representation of the actual relationship in the "population," but a reasonable sample helps us to get an estimate. The difference between the regression-predicted value of $y$ and the observed value of $y$ is called **residual** and is denoted by $e$, i.e., $e = (y - \hat{y})$. For the methods of this section, each value of $y$ must be independent of other values of $y$. For example, they cannot be the same measurement belonging to different sites of the same body. $K \geq 2$ variables on the right side of the regression equation make it a *multivariable* setup, although not a *multivariate* setup. As explained next, this allows us to study the "net" effect of each and joint effect of two or more independent variables on the dependent. While such a multivariable setup seems like the only feasible method to assess joint effect, net effect of each individual variable can be studied by stratified analysis also. If BMI and sex are suspected to affect cholesterol level, you can stratify by sex and study the effect of BMI separately for males and females. If BMI is in 4 categories, you need a total of 8 categories of subjects for a stratified study. If hypertension is also added as a factor and is in 3 categories (e.g., normotension, borderline, and clear hypertension), you need data in a total of 24 categories. This multiplies fast as more factors are added. On the top of this, stratified analysis requires adequate sample size in each such category. Also stratified analysis is not feasible if any of these factors is on continuous scale unless it is categorized. Multivariable setup obviates these problems.

As explained later, such a relationship in health and medicine is never exact, and the right-hand side of the regression equation provides only an estimated or predicted value of $y$. For this reason, this is denoted by $\hat{y}$. In the example just cited, cholesterol level can be denoted by $y$ and obesity and hypertension status by $x_1$ and $x_2$, respectively. In general, there can be many $x$ variables. The interest actually may be in change in $y$ for change in one particular $x_k$, but other $x$'s that may influence $y$ are also kept in the regression so that the adjusted effect or independent effect of $x_k$ can be studied. Other $x$'s are called **concomitant** or **confounding** variables. For example, in a study on change in triglyceride levels with age in children, the possible concomitant variables are BMI, fat intake, insulin level, physical activity, etc. In some situations, the interest may be in the independent effect of each of these $x$'s instead of any particular $x_k$.

Conceptually, regression requires observation of several values of $y$ for each value of $x$. For example, for regression of diastolic BP ($y$) on age ($x$), you should have, say, 10 individuals of age 20 years, 15 of age 21 years, 6 of age 25 years, 12 of age 28 years, etc. These should be random samples from the population of subjects of different ages. Essentially, *mean* diastolic BP for each age is used to obtain the regression.

Regression seeks a trend in *means* of $y$ at different values of $x$. This is illustrated for linear regression in Figure R.7, where the distribution of $y$ for each $x$ is Gaussian. There are only four values of $x$ in this figure, but generally there are many. Regression is the best fit to the means. This may not pass through all the means even when population means are considered and can definitely happen in the case of sample means. For regression of diastolic BP on age, supposing the average diastolic BP in a sample of males of age 30 years is 79.5 mmHg, in males of age 40 years is 82.7, in males of age 50 years is 85.2, etc., then the *trend* is DiasBP = 71.0 + 0.3(age).

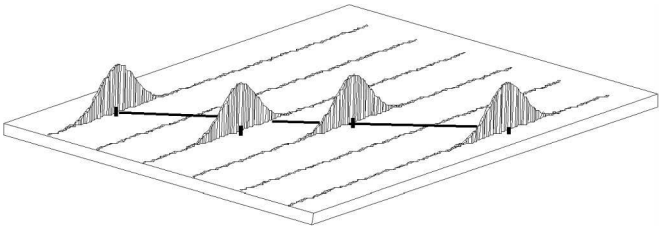


**FIGURE R.7** Conceptual framework of distribution of $y$ for each $x$ in a linear regression setup.

The level expected from the trend could differ from the observed means. The second component of residuals arises from the difference of the trend from the observed means.

There are several types of regression: linear, curvilinear, or nonlinear, depending upon the shape of its graph as explained under the topic **regression (types of)**. It could be quantitative (ordinary least squares), **logistic**, **Cox**, **Poisson**, or any other type depending upon the nature of the dependent variable. If the number of independent variable is just one, it is called simple, and in case of more than one independent variable, it is called multiple. See the topics **simple linear regression** and **multiple linear regression**, both of which are restricted to linear models. When the dependent is more than one for simultaneous consideration, it is called **multivariate regression**.

Irrespective of the type of regression, all regressions are only an expression of the nature of a statistical relationship as revealed by a set of data. They merely indicate how the dependent variable generally behaves when the independent variables change in value. This relationship is not necessarily causal and could be entirely incidental.

## regression requirements (validation of), *see also* logistic regression (requirements for)

So long as there is a measurement ($y$), called dependent, that is suspected to be determined or associated with another set of measurements ($x$'s), called independents or the regressors, and the interest is in studying their relationship, regression is always the model that can be used. There is no further underlying requirement. The question of requirement arises when you want to know whether this is an adequate model or not, practically useful or not, serves the purpose or not, etc. The whole exercise of fitting a regression can be a waste of effort and can even mislead unless its underlying requirements are satisfied. These requirements are different for different types of data and different types of regression, but all these may ultimately come down to the sample size, choice of the regression model, and the choice of the regressors besides Gaussian pattern, independence, and homoscedasticity of the dependent variable. All these together come under the generic of **regression diagnostics**, although this also contains methods to check that the regression is adequately fit or not.

Assuming that the sample of subjects is representative of the target population and is preferably randomly chosen, the first requirement is a large-enough sample size so that the results are valid and reliable. A rule of thumb floating around is that there must be at least 10 observations per regressor; if there are 5 regressors, $n$ must be at least 50. This rule seems more valid for **logistic regression** with an additional constraint that it is the limiting size—applicable to the rarest combination of the regressors. For quantitative regression, where the regressors are also quantitative, this requirement is not as strict as it is for logistic. Our experience suggests that as the number of regressors increases, the sample size requirement progressively declines. For 10 regressors, instead of a minimum of $n = 100$, perhaps a sample size of 80 would provide fairly reliable estimates of the regression coefficients. However, if the interest is in statistical power to detect a certain correlation, the sample size requirement can be higher. As always for sample size, the larger is better.

The choice of regression model depends mostly on the type of the dependent variable. A separate section on the topic **regression (types of)** provides details of what type of regression can be used in which situation. For example, there are conditions where curvilinear or nonlinear is more adequate than linear, where multiple is more adequate than simple, and where logistic, Cox, Poisson, or other regressions are more adequate than ordinary least squares.

Choice of **regressors** is also discussed separately. This has focus on the issues such as **multicollinearity** and the number of regressors, particularly in view of size of sample on one hand and parsimony of the model on the other.

### *Gaussian Pattern, Independence, and Homoscedasticity*

Whereas the method of **least squares** for fitting a quantitative regression is nonparametric and does not require any specific pattern of $y$ values, it works well only when there are no outliers. Real requirements come when trying to obtain confidence intervals (CIs) and for testing statistical significance. These procedures ordinarily require that the residuals follow a Gaussian distribution. They also require that the values of $y$ are independent of one another (except when **generalized estimating equations** are used), and the distribution of $y$ for various sets of values of $x$'s has the same variance, called **homoscedasticity**.

Independence of residuals can be violated in a case such as a time series where the value of $y$ depends serially on its value on the previous occasion. Such **autocorrelation** can occur because the time effect by itself may not be able to take care of other factors that may be simultaneously changing, such as population size and technological development. The effect of these changes spills into the residuals and causes autocorrelation. In this case, the residual plot may track snakily or follow some other pattern. The **Durbin–Watson test** is used to detect any such correlation.

**Homoscedasticity** can be checked by a scatter plot of residuals for different values of $x$. A random pattern is an indication of uniformity of variance. If the spread of residuals is found to follow a discernible pattern, a **transformation** such as the logarithm, inverse, square, or square root may help. Calculation for regression coefficients and for their significance will have to be done all over again after such a transformation.

Gaussian pattern is a requirement for finding the CIs and the test of hypotheses on the regression coefficients. This requirement is actually for the values of $y$ for each set of the values of $x$'s but is best examined for residuals because then there is no need to consider values at different values of $x$'s and all residuals can be combined. Residuals must exhibit a random pattern. As discussed under the topic **residuals**, any special or identifiable pattern in residuals can either be in the variability or in the values themselves or in both. Also there might be identifiable outliers that need to be closely examined, and excluded if found to be erroneous values. Groups of apparent outliers indicate that an important covariate is missing from the regression model. No observation should be dropped unless there is good nonstatistical evidence that the observations should be excluded. Note that if you delete valid data with large residuals, this will artificially improve the model fit, leading to a false impression of the precision of the parameter estimates.

Let us reemphasize that Gaussian pattern, independence, and homoscedasticity are requirements for a valid test of statistical significance but not for estimating the regression coefficients. These estimates can still be obtained by the method of least squares and interpreted in the usual way. But the coefficients so obtained lack credibility if the requirements are violated. For example, they may not be unbiased or may have inflated variance that makes them less reliable. Also, the practical utility of the model is greatly diminished if one is not able to test its statistical significance, and this can happen if the requirements are violated.

An example where all the common requirements of a regression model are violated arises when studying coronary artery

calcification (CAC). The amount of CAC helps in predicting future risk of coronary artery disease (CAD) morbidity and mortality. The distribution of CAC would contain a high relative frequency of zeroes and would be highly skewed. Even logarithmic transformation [ln(CAC + 1)] does not help. (To get an interpretable logarithm, 1 is added, since log of zero is $-\infty$.) Also, the variance of the quantity of CAC increases with age. If it is measured at multiple locations in the same subject, the independence is also lost. This will continue to be so for residuals too. In a situation like this, a complex model, namely, the generalized estimating equations, may be appropriate.

With some experience, you can reasonably judge the validity of these requirements by examining the process of generation of the data and by looking at the patterns in the data. Transformation, if needed, can be incorporated into the model itself. For further details, see Montgomery and Peck [1].

1. Montgomery DC, Peck EA. *Introduction to Linear Regression Analysis*, Fifth Edition. Wiley, 2012.

## regression splines, *see* spline regression

## regression to the mean

This phenomenon was first noted by Galton [1] in 1886 when he observed that *each peculiarity in man is shared by his kinsmen, but on the average to a less degree*. With regard to height, for example, tall parents will tend to produce tall offspring, but these, on average, will be relatively shorter than their parents. Similarly, short-statured parents will have children who are relatively taller. The term is now in general usage for the phenomenon that a variable that has an extreme value on its first measurement will tend to be closer to the center of the distribution in a subsequent measurement. For example, in a screening program for hypertension, persons with high blood pressure (BP) are asked to return for a second BP measurement. On the average, the second measure taken will be lesser than the first. This also explains why the term "regression" is used, which literally means "going back."

Regression to the mean can result in reduced BP in those who have a high level even without treatment. This may seem like people with higher BP benefitting more than those with lower BP. Thus, the difference between pre- and post-treatment values cannot be taken on its face value. This highlights a limitation of **before–after studies**. In place of difference, if the data are analyzed by using baseline values as a covariate, the effect of regression to the mean would ameliorate to a great extent.

Wilcox et al. [2] have reported regression to the mean phenomenon for birth weight, and Kario et al. [3] for BP. This can happen in any situation where repeat measurements are taken, but is not necessary to happen. The hypothesis behind this phenomenon is that nature tries to settle down to its normal course when upheavals occur. For details, see Morton and Torgerson [4].

It should be clear that regression to the mean is more pronounced with extreme values as they subsequently try to catch up with the mean. If a regimen produces excellent effect in one patient, be prepared to accept that it may not be so in another patient due to the regression to the mean effect. Given a chance, the same patient may also not replicate the effect. If the performance is consistently high, you are confident that the regimen is good. An initial high value and the subsequent moderation can be due to regression to the mean effect and not due to the treatment. This is a ubiquitous phenomenon and should always be considered as a possible cause of the observed change.

A statistical explanation is not far to seek. Bland and Altman [5,6] have explained it well. The slope in simple linear regression of $y$ on $x$ is computed as $\rho\frac{\sigma_y}{\sigma_x}$, where $\rho$ is the correlation coefficient between $y$ and $x$, and $\sigma_x$, $\sigma_y$ are the respective standard deviations (SDs). Thus, change of one SD in $x$ affects change of $\rho$*SD in $y$ on average. Since $\rho \le 1$, $y$ is closer to its mean than $x$ is. This would not be so when $\rho = 1$: an almost impossible value in a medical setup. They have given examples of comparison of two methods of measurement and of publication bias where regression to the mean can spoil the relationship. Also, see the article by Barnett et al. [7].

1. Galton F. Regression towards mediocrity in hereditary stature. *J Anthropol Inst* 1886;15:246–63. http://www.jstor.org/stable/2841583?seq = 1#page_scan_tab_contents
2. Wilcox MA, Chang AM, Johnson IR. The effects of parity on birthweight using successive pregnancies. *Acta Obstet Gynecol Scand* 1996 May;75(5):459–3. http://www.ncbi.nlm.nih.gov/pubmed/8677771
3. Kario K, Schwartz JE, Pickering TG. Changes of nocturnal blood pressure dipping status in hypertensives by nighttime dosing of alpha-adrenergic blocker, doxazosin: Results from the HALT study. *Hypertension* 2000 Mar;35(3):787–94. http://hyper.ahajournals.org/content/35/3/787.long
4. Morton V, Torgerson DJ. Effect of regression to the mean on decision making in health care. *BMJ* 2003 May 17;326(7398):1083–4. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1125994/
5. Bland JM, Altman DG. Statistics Notes: Regression to the mean. *BMJ* 1994; 308:1499. http://www.bmj.com/content/308/6942/1499
6. Bland JM, Altman DG. Statistics Notes: Some examples of regression to the mean. *BMJ* 1994;309:780 http://www.ncbi.nlm.nih.gov/pubmed/7950567
7. Barnett AG, Van Der Pols JC, Dobson AJ. Correction to: Regression to the mean: What it is and how to deal with it. *Int J Epidemiol* 2015 Oct;44(5):1748. http://ije.oxfordjournals.org/content/44/5/1748.long

## regression trees, *see* classification and regression trees

## regressions (types of)

Among various types of ordinary least squares regressions are linear, curvilinear, nonlinear, and asymptotic, as well as simple and multiple. Among several other types are logistic, Poisson, and Cox regressions. There are several others as listed at the end of this section. This section provides an overview—specific types are discussed separately under the relevant topic.

### *Linear, Curvilinear, and Nonlinear Regressions*

Although most regressions used in practice are linear, the utility of curvilinear and nonlinear regressions in health and medicine has been increasingly realized. An understanding of the distinction between these types of regressions is important.

For $K$ regressors, the regression

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_Kx_K.$$

is called linear since it is linear in the coefficients $b_1$, $b_2$, …, $b_K$. Geometrically, this equation defines a hyperplane in the $(K + 1)$-dimensional space and reduces to a line in two dimensions (length and breadth) when $K = 1$. When there is only one independent variable $x_1$, the regression equation is $\hat{y} = b_0 + b_1x_1$, which depicts a line. It is more convenient to write it as $\hat{y} = a + bx$. This

is the equation for **simple linear regression**. When the number of regressors is more than one, this is called **multiple linear regression**. See these topics for details.

The constants $b_1$, $b_2$,…, $b_K$ in the regression equation are called **regression coefficients**, and $b_0$ is called the **intercept**. These, in fact, are estimates of actual regression coefficients in the target population. The corresponding regression parameters in the population are denoted by $\beta_0$, $\beta_1$, $\beta_2$, …, $\beta_K$, respectively.

Curvilinear regressions are those where the regression coefficients are linear but the regressors are not. They may be square, log, reciprocal, or other types. Graphically, they show curves instead of line or plane. For further details, see the topic **curvilinear regression**. A good example of a curvilinear relationship is that between diastolic BP and age in female adults in a general urban population in India [3]. (The regression for males was linear.) The regression obtained for females is

$$\hat{y} = 58.18 + 1.668t - 0.0453t^2 + 0.000439t^3,$$

where $\hat{y}$ denotes estimated diastolic BP in mmHg and $t$ is age in years up to 60 years. The shape is as shown in Figure R.8.

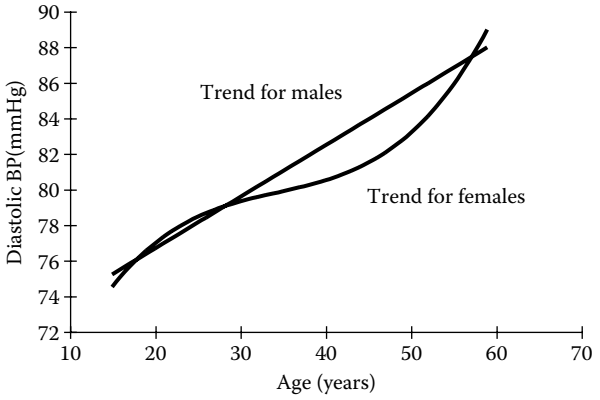



**FIGURE R.8** Diastolic BP in relation to age in adult males and females.

This is a curve but can be considered linear in $x_1 = t$, $x_2 = t^2$, and $x_3 = t^3$. Each power of $t$ can be regarded as a separate regressor. The regression thus is intrinsically linear. For this reason, such a regression is called curvilinear. *All polynomials are curvilinear.* The above equation is a polynomial of degree 3. The exponential growth curve $y = ae^{bx}$, seen in multiplication of organisms when not interrupted as in a laboratory (Figure R.9a), is apparently nonlinear but is intrinsically linear. In this case, $\ln(y) = \ln(a) + bx$, which takes the form $z = b_0 + bx$ where $z = \ln(y)$ and $b_0 = \ln(a)$. This is then a straight line but is between $\ln(y)$ and $x$. There are several other forms of the equation that are intrinsically linear. An example of another type of curvilinear relationship is that between increase in the Hb level in anemic women (say, less than 8 g/dL) and duration of iron supplementation (Figure R.9b). The initial increase is high but soon levels off. This can be represented by $y = a - be^{-cx}$, where $x$ is the duration of supplementation and $y$ is the Hb level. This is called **asymptotic regression** and is linear between $y$ and $z = e^{-cx}$.

Certain relationships are genuinely nonlinear in parameters and cannot be reduced to a linear form. A popular example is the relationship between height and weight in children represented by a growth chart. This possibly is so complex that it cannot be expressed by an equation and can be depicted only graphically. The other example is the equation

$$y = a/(1 + be^{ct}),$$

which represents a population growth model where $t$ is the time or the year (Figure R.9c). There are three parameters, $a$, $b$, and $c$, in this model. No transformation of variables $y$ and $t$ can reduce this equation to a linear form. Thus, this is a genuine nonlinear relation. The other nonlinear form is cyclic such as between estrogen levels in menstrual cycle (Figure R.9d). See the topic **nonlinear regression** for more details.

Despite all-round advancements, statistical software packages have still not given the intelligence to choose an appropriate form of regression. The user must specify the broad shape. The software finds only the best coefficients that would make the specified shape closest to the scatter. The shape of regression is decided by the following considerations.

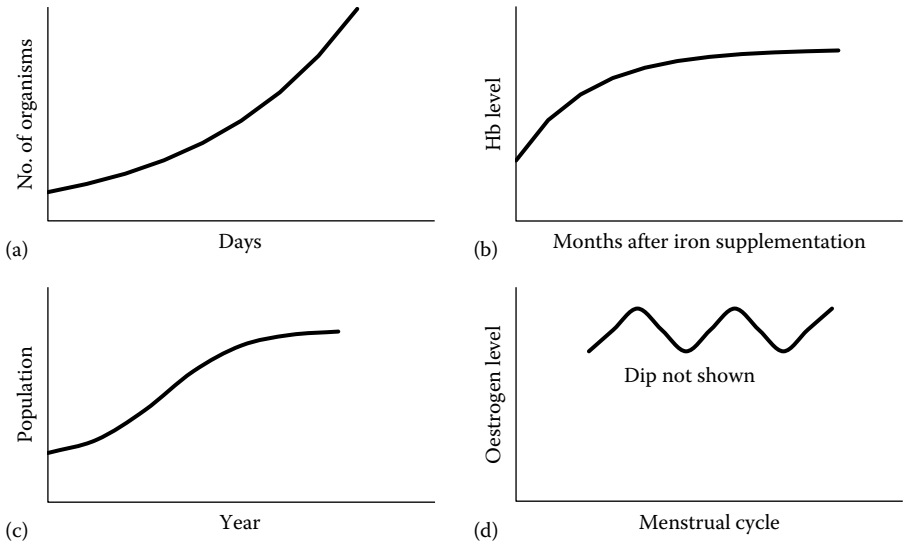



**FIGURE R.9** Some examples of curvilinear and nonlinear relationship: (a) multiplication of organisms over time (exponential curve); (b) rise in hemoglobin level after iron supplementation (asymptotic curve); (c) population growth curve; and (d) estrogen level in menstrual cycle (cyclic curve).

- Start from the simple linear regression with the most prominent factor as the only regressor. Simple linear is easy to understand and easy to explain. Check if (i) the value of the **multiple correlation coefficient** $R^2$ is reasonably high, (ii) the regression is close to the scatter, (iii) the residuals are random with small variance, and (iv) the relationship is biologically explainable. If satisfied on all these three counts, you are done. These three remain cardinal indicators for judging the adequacy of any regression.
- If simple linear regression is inadequate, examine the scatter closely for any specific pattern. It could be parabolic, cyclical, exponential, or any other. Such patterns can be fitted by using terms such as square, cube, reciprocal, and logarithm of the regressor variable. For cyclical factors such as estrogen level in women of reproductive age, try the sine function. If needed, approach a statistician to identify the form of regression for investigation.
- If the outcome measurement looks appreciably affected by two or more factors, consider multiple linear regression. Include as regressors only those factors that can really affect the outcome and are relatively independent of each other (no **multicollinearity**). If there are many candidates, use the **stepwise method** to select a few significant ones. Do not forget to consider interaction—for this, use the product (multiplication) of the variables that you think can interact.
- If multiple linear regression also does not give satisfactory results in terms of the four cardinal indicators listed earlier, try multiple nonlinear regression. But this can become too complex to explain, and the parsimony is lost. Weigh the cost–benefit of this exercise. Sometimes it is worth spending time and resources on investigating a complex relationship. Perhaps many relationships are complex, and many researchers probably waste resources on exploring simple relationships where in fact the relationship is complex.
- If that also fails, conclude that either the choice of regressors was inappropriate or not enough information is available about the factors responsible for that outcome. In this situation, the first step is to do basic research and find factors that could be rightfully conjectured as explanatory of the outcome.

Note also that choosing a form of regression implies that you know enough regarding which form is appropriate. Quite often the mechanics of how predictors affect the outcome are not known. In this situation, various forms are tried in the hope that one of these would fit the data, and then a biological explanation is sought for that kind of relationship. If the likely form of relationship as indicated by biological processes is known or can be conjectured, try to obtain a regression of this form. Either way, epistemic uncertainties can be prominent because both are based on the existing knowledge—and the existing knowledge can be inadequate and does not end the uncertainties. No matter what form of regression is chosen, part of variation will always remain unexplained. Good researchers tighten the control on this unexplained part so that medical decisions are more valid and reliable.

### *Logistic, Poisson, and Cox Regression*

The terms linear, curvilinear, and nonlinear regression generally apply to the setup where the dependent is quantitative, mostly continuous. What if the dependent is qualitative? In a medical setup, clinicians are often interested to know whether or not a set of regressors can predict presence of disease, or that a critical patient is going to survive or has benefitted from a treatment, etc. For such dichotomous outcomes, the regression of choice is **logistic**. This can be extended to polytomous (nominal or ordinal) outcomes. For details, see the topic **logistic models/regression (basics of)**.

There could be another type of dependent that counts events of interest—the number of patients waiting for kidney transplant, the size of sibship, the number of episodes of diarrhea last month in a child, etc. These counts are quantitative but are rarely large in magnitude. For such dependents, the appropriate model is **Poisson regression**. This can also be used for rates per unit of time such as deaths per year by accidents in a city and incidence of a disease in a segment of a population. For fitting a Poisson regression, the method of **generalized linear regression** can be used with a log-link. See the topic **link functions**.

Another type of regression popularly used in medical setups is the **Cox regression**. This is used when the hazard ratio of occurrence of an event is to be modeled to depend on a host of regressors. Hazard is a concept intimately related with time—thus, this method is used where durations are studied.

The regression types we have mentioned are for univariate dependent. In some situations, the interest is in multivariate dependent such as simultaneous consideration of various lung functions. In this situation, a **multivariate regression** is obtained. This is just multivariate analog of the corresponding univariate regression but requires heavy calculations and careful interpretation of the computer output. See that topic for details.

### *Other Types of Regressions*

One can think of other types of regressions depending upon the peculiar characteristics of the dependent variable. Examples are **ridge regression**, **quantile regression**, and **spline regression**, although these are not common in medical setup. In addition, we also have **harmonic regression**, **hierarchical regression**, and **meta-regression**, all of which are separately discussed.

## regressors (choice of), *see also* stepwise methods, best subset method of choosing predictors

Regressors are the variables that are included in a regression model to study their effect on the outcome. If the interest is in finding the effect of age and sex on the systolic level of blood pressure (BP), age and sex are the regressors. In general, one can think of a large number of factors that can affect any outcome of interest, but all of them can be rarely included in a regression. Adequacy of the regression and its practical utility mostly depend on proper choice of the regressors.

Among several considerations, choice of regressors depends on the purpose of running the regression. One purpose could be to examine how the regressors affect the outcome, called explanatory regression. In this regression, the explanation you get from one set of variables may be different from what you get from another set of variables. Such differing explanations may not be biologically plausible. In this setup, the choice of variables is crucial and requires deep thinking so that the effect of regressors has medical meaning. The second purpose could be prediction wherein any set of variables that predicts close to the observed values is good. This set of variables may or may not have much biological meaning: this does not matter for prediction. Two different sets of regressors can be equally good in the prediction of the outcome, and any can be chosen for this purpose.

The next step is to decide how many regressors should be included and which ones. Generally all those that are suspected to affect the outcome are included, but the size of the sample can place a severe restriction on how many regressors can be included. In ordinary least squares setup with quantitative dependent, the number of variables should be substantially less than the sample size $n$, but in logistic regression, a ratio of 1:10 is advocated between the limiting sample size and the number of regressors. *Limiting sample size* is the least number of subjects in cross-classification of the regressor values. Our experience suggests that this is too conservative, and progressively less $n$ is required as the number of regressors increases.

Theoretically, it is possible to specify a large number of regressors and ask the computer program to identify those that contribute significantly in statistical sense. **Stepwise methods** are used for this purpose. A more prudent method is to examine the biological plausibility of each and select those that can really affect the outcome. This is like not worrying about the size or shape of the stones for producing fire by rubbing them but worrying about their dryness. The key is relevance for the generalization to be valid. Remember in this context that separation of relevant from irrelevant factors is the beginning of knowledge. Include interaction terms where appropriate by considering the product of the regressors as a new variable. This product would be as good a regressor as any other. The difficulty with this approach is that the present knowledge about regression may be inadequate—many others could be important but not fully known yet. Nonavailability of data on some relevant regressors may be another limitation.

The other statistical method for selecting the right variable in a regression model is to check the statistical significance of each individual regressor by running a simple regression. Those not significant can be excluded without much loss. However, there are examples when a variable by itself is ineffective but becomes important when another variable with synergy is present. Another method is to consider the square of the **multiple correlation coefficient** ($R^2$). You can compare one model with $K_1$ regressors with another model with $K_2$ regressors by using **adjusted $R^2$**. This can also help in choosing the right regressors.

The methods just described tend to automatically exclude variables that have high **multicollinearity** with other regressors. Including too many regressors can cause multicollinearity that considerably reduces the efficiency of the regressions. If systolic BP is among the regressors for, say, serum glucose level, then inclusion of diastolic BP among the regressors may not serve much purpose. This is because systolic and diastolic levels are highly correlated. If body fat can be calculated from skinfold at triceps and thigh, there is no use in keeping all three as regressors; keeping body fat as a regressor is enough. An association or correlation exceeding 0.8 on a scale of 0–1 is considered enough to exclude the variable, particularly when the form of relationship to be examined is linear.



## rejection region, *see* acceptance and rejection regions

## relative efficiency of estimators

A combination of sample observations is considered an *estimator* as long as you use a notation and do not plug in the values. When the values are substituted, it becomes an estimate. Mean $\bar{x}$ is an estimator of population mean μ, but the average hemoglobin level 12 g/dL is an estimate. The concept of efficiency is for the estimator and not for the estimate. An estimator is considered statistically more efficient compared with another estimator of the same parameter if it has a smaller standard error (SE), because it provides more precise information about the **parameter**. Relative efficiency is the ratio of the sampling variances (square of SE) of the two estimators of the same parameter. Generally, it is obtained in comparison with the most efficient estimator. The most efficient estimator is the one that has minimum variance among all possible estimators. For example, statisticians found long ago that sample mean is the most efficient estimator of the central value in a Gaussian distribution. In comparison, sample median has variance $\pi/2$ times as large as the sample mean. Thus, the relative efficiency of the sample median is $2/\pi$ or 64% for Gaussian distribution. For another distribution, such as double exponential (two-sided exponential), the median is twice as efficient as the mean.

It is customary to talk about the **asymptotic relative efficiency** (ARE) of an estimator instead of just efficiency. This is the relative efficiency when $n$ is infinity—practically meaning a large sample and can be obtained for any statistical procedure. The ARE is an indicator of how good a procedure is relative to another when the sample size is sufficiently large. This is generally obtained for **tests of hypothesis** rather than for the estimators. You may be aware that statistical tests tend to provide "significant" result as the sample size increases even if the effect size is small. Even a minor effect can become statistically significant when the sample size is sufficiently large. Thus, in the case of tests of hypothesis, relative efficiency is the ratio of sample sizes required to reject the null by the two methods under comparison. For example, it is established that the ARE of the Wilcoxon–Mann–Whitney signed rank test is 95.5% compared with the one-sample Student $t$-test when the distribution is Gaussian; $n = 955$ will give you the result with the same reliability by the Student $t$-test as a signed rank test will give with $n = 1000$ from a Gaussian distribution. This, however, does not tell you the full story. The signed rank test is mostly used for small samples, and for these, the Student $t$-test will perform much better if the underlying distribution is Gaussian. Many other nonparametric tests do not have such high ARE and thus are advocated only when the sample size is small and the underlying distribution of values is far from Gaussian.

## relative potency, *see also* bioassays

Potency is a measure of the targeted change due to a regimen—the higher the potency, the lower the requirement of dose of the regimen, and relative potency is the change brought about relative to a standard regimen. This is more relevant when the regimens are two preparations that are identical but differ in their concentration. The cardinal method for estimating the relative potency is **bioassay**, which helps to characterize and compare the potency of a variety of regimens.

When **dose–response** curves for the test and the standard are parallel as required for **parallel–line assays**, relative potency is generally estimated by the ratio of $ED_{50}$ of a well-characterized standard and $ED_{50}$ of the test preparation. That is, when parallelism holds,

$$\text{relative potency: } \rho = \frac{ED_{50} \text{ of standard}}{ED_{50} \text{ of test}}.$$

Xu et al. [1] used this method to conclude that the potency of hyperbaric ropivacaine is 0.67 relative to that of isobaric ropivacaine, in a subarachnoid block for knee arthroscopy—thus, a higher dose of hyperbaric is required for the same anesthetic effect. In place of $ED_{50}$, $ED_{20}$ or any other defined magnitude of response can be used where parallelism holds. If not parallel, the ratio $ED_{20}$ can give very different results. Parallelism must be validated before relative potency is calculated in this manner. When parallelism is violated, the method based

on multiple-point estimates over the range of responses from $ED_{20}$ and $ED_{80}$ can be used as proposed by Villeneuve et al. [2].

$ED_{50}$ can be replaced by the mean dose in most cases, maybe after log or any other suitable transformation. If this is accepted, an alternative in terms of dose–response relation can be explained as follows. In case of parallel–line arrays, the dose–response equations are

$$y_s = \alpha_s + \beta x_s \text{ for the standard preparation}$$

$$y_s = \alpha_t + \beta x_t \text{ for the test preparation}$$

In this case, the intercepts differ but the slopes are the same, and

$$\text{relative potency: } \rho = \frac{\alpha_t - \alpha_s}{\beta}, \text{ which is estimated as}$$

$$R = (\bar{x}_s - \bar{x}_t) - \frac{(\bar{y}_s - \bar{y}_t)}{b}.$$

For quantal (yes/no type) of responses, logit transformation is used, and this changes to $y_t = \text{logit}(p_t)$ and $y_s = \text{logit}(p_s)$, $p_t$ and $p_s$ being the proportion response (Figure R.10a).

Now consider **slope–ratio assays** where the response lines have the same intercept but different slopes. In this case, the basic equations are

$$y_s = \alpha + \beta_s x_s \text{ and } y_t = \alpha + \beta_t x_t.$$

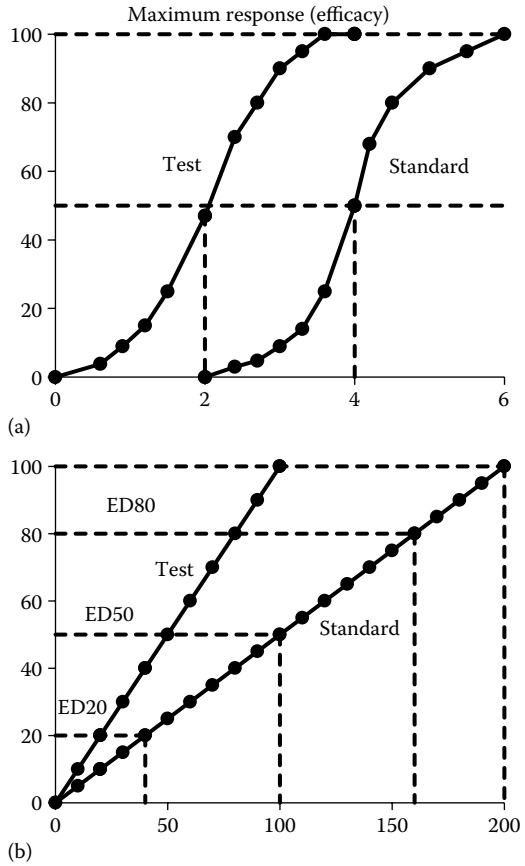



**FIGURE R.10** Relative potency in (a) parallel–line and (b) slope–ratio assays. (Redrawn from Villeneuve DL et al., *Env Toxicol Chem* 2000:19(11):2835–43. http://www.usask.ca/toxicology/jgiesy/pdf/publications/JA-253.pdf.)

In this case, the relative potency $\rho = \beta_t/\beta_s$, and the sample estimate is $R = b_t/b_s$, where $b_t$ and $b_s$ are the estimated respective regression coefficients on the basis of the actual data (Figure R.10b).

Consider an example of digestibility of egg and rice protein where digestibility score measures how much protein is really absorbed in the body. Both are given in doses of rice and egg with 0, 5, 10, 15, and 20 g of protein to three persons each, with all other diets standardized. This means that 15 persons receive egg protein and another 15 receive rice protein. Let the estimated regression equations be $y_t = 5 + 1.0x_t$ for rice (test) and $y_s = 5 + 0.5x_s$ for egg (standard) protein. This gives relative potency of rice protein = 1.0/0.5 = 2.0 relative to egg protein for digestibility. Thus, rice protein is twice as "potent" with regard to digestibility score as egg protein in this example.

Confidence interval for relative potency, being a ratio, can be obtained by the **Fieller theorem**. The methods we have discussed are valid mostly for Gaussian distributions. If the response distribution is far from Gaussian and transformation does not help, nonparametric approach similar to the one proposed by Sen [3] can be adopted.


1. Xu T, Wang J, Wang G, Yang QG. Relative potency ratio between hyperbaric and isobaric solutions of ropivacaine in subarachnoid block for knee arthroscopy. *Int J Clin Exp Med* 2015 Jun 15;8(6):9603–6. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4537999/
2. Villeneuve DL, Blankenship AK, Giesy JP. Derivation and application of relative potency estimates based on in vitro bioassay results. *Env Toxicol Chem* 2000:19(11):2835–43. http://www.usask.ca/toxicology/jgiesy/pdf/publications/JA-253.pdf
3. Sen PK. On the estimation of relative potency in dilution (direct) by distribution–free methods. *Biometrics* 1963;19:532–52. http://www.jstor.org/stable/2527532


## relative risk (RR)

Relative risk (RR) is the ratio of the risk of developing an outcome such as disease ($D$) in those with an antecedent factor ($A$) compared with those without this factor. This is also called the **risk ratio**. Risk can be measured by incidence at any defined point in time. In terms of probabilities,

$$\text{RR} = \frac{P(D+/A+)}{P(D+/A-)},$$

where $D$ is for disease and $A$ for antecedent. The + sign is for presence and the – sign for absence. Antecedents of interest can be the exposure to a risk factor.

The concept of relative risk obviously requires a **prospective study** that gives rise to the data as in Table R.7, where $\pi$'s are the group probabilities in the population. The dot in the subscript indicates the total for that particular subscript. In terms of these notations, $\text{RR} = \pi_{11}/\pi_{12}$ subject to the condition that $\pi_{11} + \pi_{21} = 1$ and $\pi_{12} + \pi_{22} = 1$. It measures the degree of association of outcome with the antecedent factor. If the incidence of lung cancer among heavy smokers for 10 years is 3% and among nonsmokers 0.5%, then RR = 6, i.e., heavy smokers have six times as much risk of developing lung cancer as nonsmokers.

The numbers $a$, $b$, $c$, and $d$ in Table R.7 are for the number of subjects observed in the sample. The value of RR in the target population can be estimated from a sample of subjects as follows:

$$\text{Estimate of RR}: \widehat{RR} = \frac{a/(a+c)}{b/(b+d)} = \frac{p_1}{p_2},$$

**TABLE R.7**
**Structure of a Study for Relative Risk (RR): Independent Samples**

| | Antecedent | | |
|---|---|---|---|
| **Outcome** | **Present** | **Absent** | **Total** |
| Present | $a(\pi_{11})$ | $b(\pi_{12})$ | $\boldsymbol{O_1.(\pi_1.)}$ |
| Absent | $c(\pi_{21})$ | $d(\pi_{22})$ | $\boldsymbol{O_2.(\pi_2.)}$ |
| **Total** | $\boldsymbol{n_1(1)}$ | $\boldsymbol{n_2(1)}$ | |

None of the cell frequencies should be very small, say less than five, so that the estimates are not unstable. If any cell frequency is zero or very small, say less than 1, then a modified estimate of RR is

$$\widehat{RR}_{\text{mod}} = \frac{(a+0.5)/(a+c)}{(b+0.5)/(b+d)}.$$

This modification is just a ploy to be able to compute RR but does not make it stable. These and several other expressions in this topic are estimates, but we ignore the hat (^) sign for simplicity. Even in the literature, an indication that these, in fact, are estimates is rare.

Consider 120 women of parity ≤ 2 and 120 women of higher parity who are subsequently assessed for anemia. The study design is **prospective** instead of cross-sectional. Parity in this case is the antecedent and anemia is the outcome. If the incidences of anemia in the two groups are 23% and 40%, then $a = 28$ and $b = 48$ since the sample size is 120 each. These give an estimate of

$$\text{RR} = \frac{28/120}{48/120} = 0.58.$$

Thus, the risk of anemia in the lower parity women is nearly half of that in the higher parity women.

The following comments are helpful in clarifying certain issues regarding RR:

- RR = 1 implies independence, i.e., the risk in exposed subjects is the same as in nonexposed subjects. It might seem paradoxical that correlation = 1 means perfect relationship, but RR = 1 means no relationship; however, that is how RR is calculated. RR > 1 means a higher risk in the exposed subjects, and RR < 1 means a lower risk. RR < 1 can be interpreted as a protective effect in place of risk as in our parity–anemia example. RR cannot be negative but can be 100 or more. RR alone is rarely enough to draw a valid conclusion about a cause–effect type of relationship unless other possible explanations are ruled out.
- An RR does not tell what the actual risks were. For example, the same RR can be obtained as 27/9, 4.5/1.5, 0.045/0.015, etc. Actual risks are entirely different in these cases, but RR = 3 in all of them. Risk may decrease over time without an alteration in the relative risk. This will happen when the risk in the comparison group also declines in the same proportion as in the test group.
- Context can be important for the interpretation of an RR. If you find that diabetes doubles the risk of stroke, you would be careful without worrying about baseline. If the risk of vehicular death in regular conditions is 1/10,000 and in bad weather conditions 2/10,000, you may still take it lightly despite double risk.
- Assessment of an RR is considered important to discuss the consequences with the patient and in preparing a management strategy. Remember, however, that one great limitation of an RR is that the relative risk can be a high value if the risk in the unexposed group is low, giving a fallacious impression of a very high risk. If the risk in the unexposed group is only 2% and in the exposed group 70%, the RR is 35. On the other hand, if the risk in the unexposed group is high, say 60%, the RR cannot exceed 100/60 = 1.67. It cannot exceed 2 if risk in unexposed is 0.5 or more. Thus, the RR heavily depends on the risk in the unexposed group. In view of this limitation, interpretation of the RR requires caution.
- For the RR to be a valid measure, it is necessary that the two groups under comparison are identical except for the presence of risk factor in one and its absence in the other. Thus, the design of the study that ensures this equivalence is important.
- A value of the RR between 1 and 2 may look small but can have tremendous public health importance if exposure is widespread. For example, in some countries, more than one half of the adult population use tobacco (smoke + smokeless). An RR of 1.20 of premature death in these adults is just 20% increase in chance of early death, but this can translate into thousands of additional early deaths due to the use of tobacco in a population.
- **The confidence interval for an RR** can be easily obtained for large samples as explained separately.

### *Test of Hypothesis on RR*

The null hypothesis is mostly RR = 1 that says that the risk is the same in the two groups under comparison. This null can be tested for large samples by the usual chi-square test. We illustrate this with the help of the data in Table R.8 on maternal age and wheezing lower tract illness (LRI) during the first year of life of 500 boys [1].

For the data in Table R.8, $\text{RR} = \dfrac{48/165}{65/335} = 1.68$. This means that higher maternal age increases the risk by 68%. For testing its statistical significance, the expected frequencies under $H_0$ can be calculated as usual, and chi-square can be calculated as follows:

$$\chi^2 = \frac{(48-37.29)^2}{37.29} + \frac{(117-127.71)^2}{127.71} + \frac{(65-75.71)^2}{75.71} + \frac{(270-259.29)^2}{259.29} = 5.93.$$

A statistical software package gives $P = 0.015$. Since this is less than the usual 0.05, the null hypothesis is rejected, and it is concluded that RR ≠ 1.

Note that chi-square is a two-tailed test and so the conclusion too is two-sided. If there is a priori reason to ensure that RR would be more than 1, then $H_1$: RR > 1. For such one-sided alternative,

**TABLE R.8**
**Maternal Age and LRI in Infant Boys**

| | LRI | | |
|---|---|---|---|
| **Maternal Age (Years)** | **Yes** | **No** | **Total** |
| <26 | 48 | 117 | **165** |
| ≥26 | 65 | 270 | **335** |

**TABLE R.9**
**Matched Pairs with Dichotomous Antecedent and Dichotomous Outcome: Prospective Study**

| | Partner 1: Antecedent Not Present (Not Exposed or Control) | | |
|---|---|---|---|
| **Partner-2: Antecedent Present (Exposed or Experiment)** | **Positive Outcome (Disease+)** | **Negative Outcome (Disease–)** | **Total** |
| Positive outcome (disease+) | a | b | ***a + b*** |
| Negative outcome (disease–) | c | d | ***c + d*** |
| **Total** | ***a + c*** | ***b + d*** | ***n = a + b + c + d*** |

it may be prudent to use the formula $z = \frac{\ln RR}{SE(\ln RR)}$ and refer it to the Gaussian distribution to get the one-sided *P*-value. For a large sample, $SE(\ln RR) = \sqrt{\frac{1}{a} - \frac{1}{n} + \frac{1}{b} - \frac{1}{n}}$. For these data, this is $\sqrt{\frac{1}{48} - \frac{1}{165} + \frac{1}{65} - \frac{1}{335}} = 0.1648$, and $z = \frac{0.4055}{0.1648} = 2.46$. From Gaussian distribution, $P(z \geq 2.46) = 0.0069$. This *P*-value is not only less than 0.05 but also less than 0.01. The relative risk can be said to be *highly* significantly more than RR = 1. There is only an exceedingly small chance that RR = 1 will give frequencies as extreme as observed in this sample.

### *RR in Matched Pairs*

A general situation for matched pairs with regard to antecedent and outcome in a prospective study is shown in Table R.9.

In this case, the RR is estimated as follows:

$$\text{Relative risk (matched pairs): } RR_M = \frac{a+b}{a+c},$$

where the notations are the same as in Table R.9. The meaning of *a*, *b*, *c*, *d* is different here than in Table R.8 (in independent samples). The numerator of this equation is the number of subjects developing the disease among those exposed, and the denominator is the number of subjects developing the disease among those nonexposed. The exact expression for the SE of $RR_M$ is complex, but it is known that the odds ratio for matched pairs, $OR_M$, approaches $RR_M$ for rare outcomes. Thus, the same CI and test criterion can be used for $RR_M$ as for $OR_M$ in most practical situations. For $OR_M$, see the topic odds ratio.

1. Martinez FD, Wright AL, Holberg CJ, Morgan WJ, Taussig LM. Maternal age as a risk factor for wheezing lower respiratory illnesses in the first year of life. *Am J Epidemiol* 1992;136:1258–68. http://www.ncbi.nlm.nih.gov/pubmed/1476148

## reliability, *see also* repeatability and reproducibility

Reliability is the property of giving the same result on repeated measurement in identical conditions. For true reliability, the values must vary considerably from subject to subject, although the repeat values for each subject should be nearly the same as before. This property generates confidence as you know that the result would not be any different if the measurement is taken again. Besides measurements, in a medical setup, the question of reliability arises in a large number of situations such as instrumental reliability, reliability of conclusions, and statistical reliability of estimates. A brief of each of these reliabilities is given in this section, but first realize that the variation in a repeated measure can be due to one or more of the following reasons:

- Chance or unsystematic events.
- Systematic inconsistency.
- Actual change in the underlying event being measured.
- Measurement reliability is not just the property of an instrument. It also depends on the conditions (busy day at the clinic, stress, new application) in which measurement is done. Factors due to subjects (e.g., patient fatigue, mood) and factors due to ability of the observer/rater/interviewer to elicit correct response from the respondent also affect the repeated measurements.
- Instrument reliability, i.e., the research instrument or measurement approach (e.g., poorly worded questions, quirk in mechanical device).
- Data processing reliability, i.e., manner in which data are handled (e.g., miscoding).
- Lack of clarity of instructions, i.e., amenable to varying interpretation by different observers.
- Besides fatigue, variations can also arise due to boredom, memory lapse, carelessness, lack of motivation, lack of concentration, etc.

Colloquially, reliability is sometimes equated with success. If a system fails often, we say that it is unreliable. If the response is 30% consistently across groups of patients, it is consistent but is still perceived unreliable as only a few respond. Sometimes, it is related to a time period such as considering a regimen reliable if it is able to get response within, say, a week, and unreliable if it is not able to do so. Our concern in this section is with statistical reliability, which is consistency in varying conditions.

### *Reliability of Measurements*

Measurement is a process by which the dimension of an object or event is either quantified or classified. Numeric assignment is done for dimensions that signify magnitude or the quantity, whereas the classification is done for qualitative attributes. Reliability of a measurement is the degree to which consistent values are obtained in repeated applications in identical conditions. Conceptually, this is distinct from other reliabilities such as instrument reliability since measurement reliability is the inherent property of a measurement and refers to the consistency when the instrument is fully reliable.

For example, weight of a person is a reliable measurement compared with a pain score that can quickly change depending upon the instability exhibited by the person.

Statistical reliability of measurements was initially defined as the ratio of true variance to the observed variance and is the quantification of the consistency of values across replications. True variance may never be known so that the reliability of measurements is assessed in terms of agreement of values taken at two or more occasions. The extent of agreement is measured by **intraclass correlation** if the measurements are quantitative and by **Cohen kappa** if they are qualitative. An intraclass correlation exceeding 0.90 can be considered excellent, that between 0.80 and 0.89 good, and that between 0.70 and 0.79 tolerable. Values less than 0.70 are suspect for reliability, although they might be good in other contexts. For Cohen kappa, the thresholds are milder: <0.3 is considered unreliable, 0.3–0.5 tolerable, and higher values progressively good. These measures are valid if the errors in measurements are independent of their quantitative level. If higher errors are anticipated for larger values, the errors may have to be converted to, for example, percentages.

Sometimes reliability of measurements is confused with **inter-rater reliability**. This occurs because errors in measurement are considered to have arisen due to differential skills of the observers used for repeated measurements, and not because of inherent variations in the measurement.

### *Reliability of Instruments/Devices*

Reliability of an instrument or device is the stability of the response with repeated use of the instrument in identical conditions. A clinician uses various tools during the course of practice such as signs–symptoms syndrome, physical measurements, laboratory and radiological investigations, and intervention in the form of medical treatment or surgery. Besides his or her own skills in optimally using what is available, the efficiency of a clinician also depends on the **validity** and reliability of the tools he or she uses. Unlike validity, the concept of reliability is not related to any gold standard. For example, we know that a clinical thermometer is a reliable tool for measuring body temperature since it behaves the same way under a variety of conditions. Reliability of an instrument is its ability to be consistent across varying conditions in the sense of providing the same results when used by different observers or in different setups. The performance of a reliable tool is not affected much by the external environment in which the tool is used. This kind of reliability across repeated measurements can be studied by agreement assessment such as by the **Bland–Altman method** or by computing the **intraclass correlation coefficient**.

An instrument such as a questionnaire can have multiple items. Reliability may suffer if the wordings are not precise and instructions are not explicit so that there is room for subjective interpretation either by the assessor or by the assesse, or by both. How does one assess the reliability of such an instrument? It has two components: (i) internal consistency and (ii) stability across repeated use. The first concerns the consistency of responses across various items in the same test. A popular measure of internal consistency of a multi-item instrument is **Cronbach alpha**, which assesses congruence of items with one another: whether or not they amiably hang out together. All the items should be tapping different aspects of the same attribute and not different attributes. In effect, the items of a scale should be moderately correlated with each other. Also, each item should correlate well with the total score. This is what Cronbach alpha broadly assesses. Another method is **split-half consistency**, which measures agreement between two random halves of the instrument. The next is repeatability that essentially involves administering the test twice to the same subjects—thus also known as **test–retest reliability**. It can be calculated for an instrument that does not provide any learning to the respondent, and the second-time responses are not affected by the first-time responses. The scores or the measurements obtained on the two occasions are checked for agreement. This can be checked itemwise in case of multi-item instrument (such as a questionnaire), but generally the total score is compared.

The choice of items included in a multi-item tool makes or mars the reliability. You may have to examine various criteria for determining which items to keep and which to discard. Eliminate any items that are ambiguous or do not make sense to the respondent. Generally speaking, an item should have a correlation of more than 0.20 with the total score; if it is lower, that item should be discarded. Also, the reading level or the understanding required to respond should not be too high. For this, it is advisable to **pretest** the tool on a group of people comparable to those who will actually be the ultimate targets, checking that they do, in fact, understand each item.

Another type of instrument commonly used in medical assessment is a scoring system. This is similar in nature as a questionnaire and can provide varying results in repeated use when the wordings, the sequence, and the method of eliciting response are not standardized. The reliability of a scoring system is also assessed the same way as of a multi-item instrument discussed in the preceding paragraphs.

Consider an example on nutritional rickets. Severity of this disease can be assessed by the degree of metaphyseal fraying and cupping, and the proportion of the growth plate affected, based on radiographs of wrists and knees. Thacher et al. [1] evaluated the utility and reliability of a 10-point scoring system that progresses in half-point increments from 0 (normal) to 10 points (most severe). They found that the interobserver correlation of the score was 0.84 or greater for all observer pairs they used, and the intraobserver correlation was 0.89 or greater for each observer. Thus, there is a fair amount of consistency. The authors conclude that this score should be useful to objectively assess the severity of rickets. This illustrates how reliability measurement can be useful.

### *Reliability of Estimates*

A major statistical exercise in many medical studies is to estimate the effect size of an intervention and to estimate incidence, prevalence, etc., of a condition. The question of estimation arises because the entire population is almost never studied and a sample is used for inference. All samples are notorious for providing varying estimates from sample to sample. Thus, it is necessary to assess the reliability of the estimates.

Reliability of estimates is evaluated by their **standard error (SE)**. See this topic for details if you are not familiar with it. These SEs almost invariably depend on two things: (i) the intrinsic variation from subject to subject and (ii) the sample size. For example, $SE(\bar{x}) = \sigma^2/n$, which indicates that the higher the sample size, the greater the reliability. This sounds intuitive also: an estimate based on a sample size of 100 would not vary as much from an estimate based on another sample of this sample size as an estimate based on a sample of size 10 would.

In the case of multiple linear regression, the reliability of the regression can be measured by the **multiple correlation** coefficient ($R$). Most researchers present results of such regression in terms of $R^2$. Generally speaking, the higher the $R^2$, the greater the reliability. But high $R^2$ for a small sample size is not really helpful because standard errors are still high and the reliability of the estimates is low. A high $R^2$ can also occur in the case of **multicollinearity**, where many individual regression coefficients are not statistically significant.

In the context of senile dementia, van Belle et al. [2] argue that the concept of the reliability can also be applied to the change over time. The reliability of the estimates of change is shown to depend primarily upon the *length* of time of observation, not the *number* of observations made. The authors introduce the concept of signal-to-noise ratio to compare reliabilities in change scores.

### *Reliability of Conclusions*

Reliability of statistical conclusions mostly depends on the sample size, given that the data are correct and analyzed properly. Sometimes, as in a case–control study, a large number of controls when easily available and more cooperative are included, which helps to increase the reliability of the results. A study by Kornitzer et al. [3] on the effect of serum selenium level on cancer mortality is an example in which this strategy was adopted, and three controls were included for each case. Consistent results in repeat studies, where feasible, also increase the confidence level.

Nonresponse in clinical trials and many other studies requiring follow-up has an adverse impact on the results: the ultimate sample size available to draw conclusions reduces, which affects the reliability of the results. This deficiency can be remedied by increasing the sample size corresponding to the anticipated nonresponse. However, this does not remedy the bias in case the nonresponse is selective by those who are critically sick, or have more robust health, etc.

### *Reliability versus Agreement*

The reliability is consistency, whereas the agreement is similarity of values. Apparently, there seems hardly any difference. However, Kottner and Streiner [4] make reference to a study by Costa-Santos et al. [5] to provide a valuable example about the difficulties in comparing and interpreting reliability and agreement coefficients arising from the same measurement situation. According to them, much of the confusion around reliability and agreement estimation is caused by conceptual ambiguities. In case of agreement, the absolute degree of measurement error is of interest. If all values are high but agree, the agreement would be high. On the other hand, reliability would be low if all values are high (and consequently have high variance) since reliability also considers interindividual variability that agreement does not. A clear distinction between the conceptual meanings of agreement and reliability is necessary to select appropriate statistical approaches and adequate interpretation.

Streiner and Norman [6] have discussed some of these issues in more detail.


1. Thacher TD, Fischer PR, Pettifor JM, Lawson JO, Manaster BJ, Reading JC. Radiographic scoring method for the assessment of the severity of nutritional rickets. *J Trop Pediatr.* 2000; 46:132–9. http://www.medicalbiostatistics.com/scoring.pdf
2. van Belle G, Uhlmann RF, Hughes JP, Larson EB. Reliability of estimates of changes in mental status test performance in senile dementia of the Alzheimer type. *J Clin Epidemiol* 1990;43:589–5. http://www.jclinepi.com/article/0895-4356(90)90163-J/abstract
3. Kornitzer M, Valente F, de Bacquer D, Neve J, de Backer G. Serum selenium and cancer mortality: A nested case-control study with age- and sex-stratified sample of Belgian adult population. *Eur J Clin Nutr* 2004; 58:98–104. http://www.ncbi.nlm.nih.gov/pubmed/14679373
4. Kottner J, Streiner DL. The difference between reliability and agreement. *J Clin Epidemiol* 2011;64:701–2. http://www.jclinepi.com/article/S0895-4356(10)00433-6/pdf
5. Costa-Santos C, Bernardes J, Ayres-de-Campos D, Costa A, Costa C. The limits of agreement and the intraclass correlation coefficient may be inconsistent in the interpretation of agreement. *J Clin Epidemiol*. 2011; 64:264–9. http://www.ncbi.nlm.nih.gov/pubmed/20189765
6. Streiner DL, Norman GR. *Health Measurement Scales-A Practical Guide to their Development and Use,* Second Edition. Oxford Medical Publications, 1999.


## repeatability and reproducibility, *see also* reliability

According to Peng [1], repeatability and reproducibility are the two foundational characteristics of a successful scientific research enterprise. In common parlance, repeatability and reproducibility are interchangeably used, but these have slight technical difference in statistical context. Repeatability of a study is the property that it will produce a result consistent with the original when another study is conducted in the same conditions. This is also called **replicability**: the keyword being "consistency." The result of the second study may not be exactly the same but would be consistent. This slight leeway is allowed since the subjects could be different, and there might be other subtle unrecognizable differences that could cause minor variations despite essential similarity. Reproducibility is the ability to get the same result when the same set of data is analyzed again with the same or another method. This can be assessed only when the raw data are accessible. If the data analysis is poor, the results will not reproduce, which will indicate lack of expertise or poor judgment of what method should be used and how. Thus, reproducibility is more statistical in nature than repeatability. This also points to some false positive and false negative results arising due to inadequate analysis, although false results can also be due to several other reasons such as poor quality of data, poor design, and inadequate definitions.

Whereas repeatability is regularly assessed by conducting similar studies in varying setups, assessment of reproducibility has not been easy. A major bottleneck in this is the nonavailability of raw data for reanalysis. Many journals now encourage researchers to store the data at publicly available websites, and some provide facility to warehouse the data at their own website; however, there is a considerable reluctance by the researchers to share the data that can be unscrupulously used for other analyses and conclusions. According to Peng [1], the answer is to build trust, and this would come from the credibility of the researchers as established by evidence of their skills and expertise.


1. Peng R. The reproducibility crisis in science: A statistical counterattack. *Significance* 2015;12(3):30–5. http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2015.00827.x/abstract




## repeated measures ANOVA

**Repeated measures** are those that are taken at different fixed points in time for each subject. Repeated measurements could be heavily correlated—thus, they require special methods. Repeated measures analysis of variance (ANOVA) is used to analyze data arising from repeated measures where the outcome is continuous and is studied in different groups of subjects. The objective in this analysis is to find whether or not the average values differ among groups.

Medical studies are commonly conducted for investigating changes over time in one or more outcome variables. Firstly, consider the simplest case of two measurements per subject, that is, there is only one group of subjects and they are measured on two occasions only. For example, in a clinical trial, this could be before and after treatment. The analysis of such data is straightforward—you would use a paired ***t*-test** (or the nonparametric equivalent if the requirements of the paired *t*-test are not met). However, in the case of more than one group, there are other options, discussed below,

for analysis of repeated measures data when there is just one repetition [1].

- Ignore the baseline measurements and simply compare the final measurements with the usual tests used for non-repeated data, i.e., **one-way ANOVA**. To take account of other covariates affecting the outcome, general linear regression could be used. Obviously, this takes no account of the baseline measurements and is a feasible option (although not necessarily recommended) if there is no reason to believe that the groups differed at baseline, such as in a study where subjects are randomly assigned to groups.
- Calculate the change for each subject at each time point and compare the average change across groups using the same methods as just described. Unfortunately, this method tends to overcompensate for any baseline differences.
- Use the method of **analysis of covariance (ANCOVA)**. This compares the final values across groups while also taking account of the baseline values. Essentially, it is done by carrying out a multiple regression, putting in the final values as the dependent variable and the baseline values, as well as the variable used to define the groups, as the independent variables. If applicable, other covariates can also be taken into account by adding them into such multiple regression.

Now consider several repetitions. The temptation may be to compare subjects at each time point separately, perhaps with a series of unpaired *t*-tests. We will come to the setup with several groups in a short while, but first consider only one group of $n$ subjects, each measured at the same $T$ time points, such as measuring hemoglobin (Hb) level at monthly intervals in $n = 60$ pregnant women. If the measurements are done from the third month to the ninth month, then there are $T = 7$ measurements per woman. In this special case, one can think of analyzing the data by two-way ANOVA considering each subject as one factor with 60 levels and time points as the second factor with 7 levels, provided that the validity conditions hold. The null hypothesis of interest in this case would be equality of means at seven time points. This will have one observation per cell, and the interaction such as the fast rise in Hb in one woman and the slow rise in the other cannot be studied—only the average at seven time points can be compared. This analysis will also ignore the trends over time, if any, since ANOVA considers all levels (in this case, time points) nominal and not metric. This analysis will also ignore that some women have a low level to begin with whereas others have a high level. Thus, this analysis has severe limitations.

Suppose these women are divided by their parity as 0, 1, 2, and 3+. This makes it a real repeated measures ANOVA setup. If there are 30 women of each parity, there will be a total of 120 women and each will be measured seven times at monthly intervals. The objective now is to compare the mean Hb level across parity groups and not across time points. Two-way ANOVA is adequate for this kind of repeated measures design subject to the condition that the difference $(x_t - x_{t'})$ has the same population variance for every pair of occasions: the subscript $t$ is for the time point. This is achieved when all $x_t$'s have the same variance, and the covariances between all pairs of $x_t$ and $x_{t'}$ are equal (see **sphericity**). Small differences do not matter if $n$ is large, but the differences could be very large in some situations. Measurements close in time may be more highly correlated than those widely separated in time. Heart rates (HRs) 1 and 5 min after anesthesia could be highly correlated, and HRs after 1 and 30 min poorly correlated. Thus, the covariances will not be uniform, and the condition for using repeated measures ANOVA will be violated. In this case, use the **Huynh–Feldt correction** if the **Mauchly test** for sphericity is significant.

Another alternative is to consider repeated observations on each subject in multivariate setup and do **multivariate analysis of variance (MANOVA)**. This works well for balanced design with large $n$ and does not require sphericity of repeated values, but the software may produce a complex-looking output. This output generally mentions several tests, but **Wilks lambda** is preferred in most cases as this could be exact under regular conditions such as multivariate Gaussian distribution. Although MANOVA is a preferred method for equal $n$, if you find MANOVA output difficult to decipher, do univariate repeated measures ANOVA as stated earlier. Both these analyses test equal means response at different time points and across groups, but not time trends. You can test gross differences in time trends by including the time*groups interaction term, but that really does not test any specific time trend such as linear, quadratic, or any other polynomial form. Special commands can be given in most standard software packages for testing the presence of such a trend. If this interaction is significant, conclude that the average time trend in at least one group is different from the trend in other groups.

Complexities seem to never end with repeated measures. What happens if the design is unbalanced (unequal $n$'s in different groups) and covariance matrices across groups are also unequal? For this situation, the **Welch test** is recommended in place of $F$. This is similar to the Welch test for the two-sample situation. But the $n$ in the smallest group should preferably be at least $3*(T - 1)$ for testing main effects and at least $5*(T - 1)$ for testing interaction, where $T$ is the number of time points for repeated measures. Also consider the following.

- In case the groups are randomly selected from a large population, use **mixed effects** ANOVA.
- **Multilevel modeling** is another possible approach for analyzing repeated measures data. This requires considerable statistical expertise and a good statistical package.
- All this is for a metric, preferably, continuous outcome. If the outcome is binary, the data can be analyzed by repeated measures **logistic regression** using logit **link** in **generalized linear models**.

## Sphericity and Huynh–Feldt Correction

Equality of covariances together with equality of variances within a matrix is known as *compound symmetry*. A slightly relaxed condition is that the differences between all pairs of times are independent (covariance = 0) and have same variances. This is called **sphericity** and is the actual required condition for repeated measures ANOVA. This is tested by the **Mauchly test**.

The Mauchly test for sphericity works well for Gaussian data and reasonably well for minor departures from Gaussianity, but not when departures are gross and the dataset is small. When sphericity is violated, too many hypotheses are falsely rejected. When in doubt, the safe bet is to use **Huynh–Feldt correction** to df's. This correction can be used directly without worrying about Mauchly test results, and is explained separately. Many statistical packages provide results with Huynh–Feldt correction and also the corresponding $P$-value. The value of $F$ remains the same, but the correction factor reduces both the df's of $F$ by a factor called epsilon. You may find other types of epsilon also in your software output, but the Huynh–Feldt epsilon is widely accepted. In a rare case, this correction factor can exceed one, in which case no correction to df is done.

Besides sphericity, which is for variances–covariances among values at different points of time within each group, you also need

to worry about homogeneity of covariance matrices across groups when groups are two or more. For this, the **Box *M* test** is used. This also requires Gaussian distribution of the data.

For further details of repeated measures ANOVA, see Verma [1]. In case the data have many zeroes, special methods such as the one discussed by Berk and Lachenbruch [2] can be used.


1. Verma JP. *Repeated Measures Design for Empirical Researchers*. Wiley, 2015.
2. Berk KN, Lachenbruch PA. Repeated measures with zeroes. http://support.sas.com/rnd/app/stat/papers/repeatedmeasures.pdf


## repeated measures studies

In many medical situations, as in the case of administering an anesthetic agent, it is necessary to monitor a subject by repeatedly observing vital signs such as heart rate and blood pressure at specified intervals. The basic feature of repeated measures is longitudinal follow-up, although it can be short as in the case of trials on anesthetic agents or can last for years as for quality of life after surgical interventions. The purpose is to measure the change with respect to the previous values, to assess the trend, and sometimes to identify the time point when the changes are significant. This is valid only if there is no natural or man-made change in values over time, that is, when any change can be assigned purely to the intervention. Guard against subjects showing improvement because of self-regulating mechanism in the body without intervention because that may contaminate the effect of intervention. Beware of fatigue setting in so that readings at later instances are taken with the same earnestness as in the earlier phase. Also, there should not be any **Hawthorne effect**. Analysis of such repeated measures requires special methods because these are not independent. Subsequent values depend on previous values, and this violates the validity of most statistical methods such as **analysis of variance (ANOVA)**. Special methods such as **repeated measures ANOVA** are used for this setup. The follow-up in repeated measures design is done at predefined intervals for each subject. If the time points are different for different subjects, it is called a **longitudinal study** instead of a repeated measures study.

As an example, consider a carcinogenic marker measured just before administration of a carcinogen in mice and after 1, 2, 5, and 10 days. This can also be done separately for the control group that receives the placebo. The changes are noted in the experimental and the control groups. In case of anesthetic agents, repeated measures help to study the short-term outcome separately from that emerging in the long term and the trend can be studied. **Bioequivalence** studies necessarily use a repeated measures design, because only then can the **pharmacokinetics** and the course of disease can be studied.

A repeated measures study may have two or more groups such as different dose groups or subjects with different forms of a disease. Each subject in each group can be measured at several time points to study the trend of quantitative outcome such as various enzyme levels. The objective in this case is to find if the average time trend in one group is the same as in the other groups. This is the same as assessing absence of the interaction group*time. For studying such interaction, $n$ in each group must be sufficiently large so that it has power to detect this kind of interaction.

Repeated measures have the potential to provide unnecessary satisfaction because of apparently having a large volume of data. If blood pressure of 10 subjects is measured hourly 12 times during day time for diurnal variation for a week, you will have a total of 840 data points for systolic and 840 for diastolic level. This can mislead you to believe that a large sample of values is available. Actually, there are only $n = 10$ subjects. Reliability of conclusions will not be high in this case despite a large number of data points because of small sample size.

## replicability, *see* repeatability and reproducibility

## replication

Replication is repeating the study. It is through replication that considerable confidence is added when the same results are obtained. Replication by another worker in a different setting helps to confirm that the results were not dependent on local conditions that were unsuspectingly assisting or hindering the outcome. Such replication also provides evidence for the robustness of the results. Sometimes the investigator himself wants to replicate the experiment to be on firm footing, particularly when the response varies widely from subject to subject. The variability across subjects remains unaltered, but replication reduces variability in experimental results; when the results between replicates are similar, their reliability naturally increases.

Replication also fulfills an important statistical requirement. It helps quantify the **random error** in the experimental results. It is also an effective instrument in reducing the effect of this error. Random errors occur due to imprecision of the process—in methods, measurements, etc., such as in staining, magnification, rotation, counting, and timing; and in the environment such as room temperature, humidity, or minor variation in chemicals, reagents. They also occur due to the varying care adopted by different observers. Killeen [1] explores the concepts in more detail. Macdonald [2] published a rejoinder to Killeen, discussing replication probabilities. For more details, see these references.


1. Killeen PR. Replicability, confidence and priors. *Psychol Sci* 2005; 16:1009–12. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1440522/
2. Macdonald RR. Why replication probabilities depend on prior probability distributions: A rejoinder to Killeen. *Psychol Sci* 2005;16:1007–8. http://www.ncbi.nlm.nih.gov/pubmed/16313668


## reproducibility, *see* repeatability and reproducibility

## reproduction rates (net and gross), *see* fertility indicators

## reproductive number (of a disease)

The basic reproductive number of a disease is the average number of people to whom an affected person is able to spread this disease during its life course. The concept applies to communicable diseases, particularly infectious diseases. If an infected person is able to infect at least one person on average during the entire period of infectivity, then the infection will sustain or increase. In countries where hepatitis B infection is on the rise, this rate is more than one. The 2014 Ebola virus disease had this number between 1.71 and 2.02 in different countries of West Africa [1] causing the spread in epidemic proportions. If the reproductive number is less than one for any disease, expect that the infection will die down or stabilize at a low level. This underscores the need to control transmissibility of an infectious disease in order to reduce it to incidence. Dietz [2] gives a survey of various methods of estimation of the reproductive number. One particular method is as follows.

The definition implies that the reproductive number depends on the duration of infectivity, the infectiousness of the organism, and the number of susceptible persons that come in contact with the index case. This is calculated as follows:

$$\text{reproductive number: } R_0 = C^*P^*D,$$

where

$C$ = the average number of contacts an index case has with the susceptibles during the infective period.
$P$ = probability of transmission per contact per unit of time: this is the transmissibility.
$D$ = duration that the index case remains infectious.

Transmissibility depends on the infectiousness of the disease, susceptibility of the contacts, and the *herd immunity*. For example, it is surmised that diphtheria transmission throttles if effective vaccine coverage is 85% or higher. This is the herd immunity as this limits the scope of spread of infection. The gross reproductive number of this disease is estimated as 6–7, that is, an infective case of diphtheria can infect 6–7 persons if enough susceptibles come into contact. The reproductive number of a disease varies from population to population mostly due to the variation in the number of contacts and transmissibility. Diphtheria is not able to spread in most parts of the world because children are immunized and there are not enough susceptibles, which makes the basic reproductive number of this disease quite low.

The gross reproductive number assumes that all contacts are susceptible. In practice, some contacts may be immune either because of previous infection or because of immunization. If one third of contacts are immune, the effective reproductive number of diphtheria would be nearly 4.7 in place of 7. It can be used to estimate the heard immunity required to stop the spread of infection, shown as follows:

$$\text{required herd immunity} = 1 - \frac{1}{R_0},$$

where $R_0$ is the reproductive number. For diphtheria, for example, if $R_0 = 7$, required herd immunity is $(1 - 1/7) \times 100 = 86\%$. If the reproductive number for any disease is just 2, the herd immunity required is only 50%. Immunization plans can be drawn accordingly. This calculation assumes that all the vaccinations are fully effective and the coverage is fairly spread out so that it does not leave out pockets of susceptibles.

Holtgrave [3] provided an interesting projection for human immunodeficiency virus (HIV) infections in the United States that shows that the transmission rate of the disease in the country has substantially decreased. So, is elimination of HIV infection feasible in the foreseeable future? The author demonstrated that if the HIV transmission rate were reduced by 50%, then the reproductive number of HIV infection would drop below unity. Experts have asserted that the HIV transmission rate can be halved by 2020, if not earlier, provided that sufficient investment is made toward achieving this goal. This should lead to eventual elimination of HIV infection in the United States.

1. WHO Ebola Response Team. Ebola virus disease in West Africa—The first 9 months of the epidemic and forward projections. *N Engl J Med* 2014;371:1481–95. http://www.nejm.org/doi/full/10.1056/NEJMoa1411100
2. Dietz K. The estimation of the basic reproductive number for infective diseases. *Stat Methods Med Res* 1993;2:23–41. http://smm.sagepub.com/content/2/1/23.full.pdf
3. Holtgrave D. Is the elimination of HIV infection within reach in the United States? Lessons from an epidemiologic transmission model. *Public Health Rep* 2010;125:372–6. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2848261/

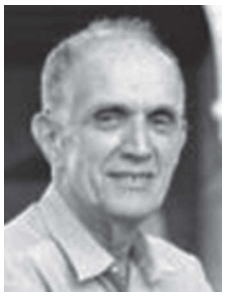
## resampling

Under this method, subsamples from an existing sample are drawn that help to draw statistical conclusions and also help to check if the results replicate on such subsamples. Thus, this method assesses the reproducibility or **reliability** of the results besides helping to tackle difficult setups where lack of standard conditions hampers valid conclusions. Two popular methods of resampling are **bootstrapping** and **jackknifing**. The following is a brief and introductory account; see Good [1] for details.

### *Bootstrapping*

You probably know that bootstrap is the loop at the back of a boot that is used to pull the boot on. In science, the term bootstrap is used for an algorithm with initial instructions such that subsequent steps are automatically implemented. In statistics, bootstrap is generating thousands of samples from the available sample. If you have a sample of size $n$, another sample of size $n$ can be drawn with replacement. This means that the sampled value is noted and replaced back as a candidate to be resampled. Theoretically, the same value can repeatedly appear in a bootstrap sample.

The actual procedure of bootstrap is something like this. If you have a sample of size $n = 18$, select one at random from this sample. Replace this value back so that you again have 18 values. Select another from these 18, which is the second value in the generated pseudo-sample. Since the sampling is with replacement, the same value can be selected again. Do this 18 times and you have a *pseudo-sample* of size 18. In a rare case, the same value can occur 18 times in one pseudo-sample. The process can be repeated to build thousands of pseudo-samples from one sample. These will be called bootstrap samples and can be drawn easily with computer-generated random numbers through an automated procedure. These samples build a proxy universe for valid inferences. However, the inferences will be valid only for the population represented by the initial sample. For this reason, it is advised that the initial sample is a genuine random sample from the target population. The method was first proposed by Bradley Efron in 1979 [2] as a variation of jackknife resampling.


Bradley Efron

Each of the bootstrap samples can give the mean, median, quartile, standard deviation, or whatever **statistic** is required. You will have thousands of these values, one corresponding to each pseudo-sample, and can generate a pseudo-distribution of sample median, sample quartile, etc. This distribution can be used, for example, to find a range such that 2.5% medians are less than the lower limit of this range and 2.5% are more than the upper limit of this range that

will provide 95% *bootstrap* confidence interval for median. This can be done for any parameter of interest of any population. The method is particularly useful where the conventional methods are not applicable due to unknown distributions.

Note how the available data are used to tell more about the data through resampling. No extra help is needed—thus the name bootstrap. But also note that bootstrapping of a bad sample (not representative) cannot yield valid results. In this case, bootstrap may give you a false sense of security.

To understand the actual usage of this method in statistics, consider the following. It is well known that the statistic ±1.96SE as a 95% confidence interval (CI) comes from the **Gaussian distribution**. This can be used for the CI for mean from a non-Gaussian distribution also when *n* is large because then the **central limit theorem** comes to the rescue, but cannot be ordinarily used for estimating nonadditive parameters of interest such as median (e.g., $\mathbf{ED_{50}}$) and **quartile**. The method of bootstrap was originally devised to find the CI for such parameters. The proxy universe generated by bootstrap samples is used for finding the CI as stated in a previous paragraph by basically generating 1000 bootstrap samples. This method is **nonparametric** and can be used to estimate the parameter of any distribution of the data.

The bootstrap procedure is also used for assessing **robustness** of results. This involves taking several samples from the same group that was actually studied and examining if the results replicate. This is different from what is just described. For robustness, the same size is generally preferred so that the varying *n* does not cause problems, but you can have a bootstrap subsample of, say, 150 each time without replacement when you actually have a sample of size 200. Although such resampling can be done many times, generally three or four subsamples of this type would be enough to assess robustness.

### *Jackknife Resampling*

Originally, the jackknife is sampling $(n - 1)$ values without replacement from the available *n* sample values that would construct *n* samples of size $(n - 1)$ after deleting one at a time. The first jackknife subsample would comprise all the values except the first, the second subsample will have all the values except the second, etc. In other words, one value is left out each time in turn, and *n* subsamples are created from just one sample. If a large sample is available, two or more values can be dropped each time. For dropping two at a time, the procedure would be as follows. If you have a sample of $n = 212$ subjects, drop numbers 1 and 2 and recalculate the result on the basis of numbers 3–212. Then drop numbers 2 and 3 and recalculate the result based on numbers 1 and 4–212, and so on. Generally, all possible subsamples are considered, and these subsamples can be used for inference regarding the parameters of interest. For example, the mean or average of these *n* subsamples would have better reliability than the mean of just one original sample.

The term jackknife signifies a ready simple tool that can be folded and opened, and can be used in a variety of situations. The technique was first proposed by Quenouille [3] in 1956 as a tool to estimate bias, although he did not use the term jackknife.

The method is primarily used to detect outliers that have major effect on the results. It is evident that jackknifing is very effective in detecting outliers. If, for example, the mean of the fifth subsample that has dropped the fifth value is very different from the other means, you know that this fifth value is an outlier. The method can be used for assessing robustness also. For robustness, there is no need to study all possible samples—possibly three or four subsamples are enough. If the results replicate in repeated subsamples, there is evidence that they are robust. The method works well for summaries that use all the values, such as mean and variance, but can fail for median, which does not use tail values. For further details of the method, see Shao and Tu [4].

The only difference between bootstrap and jackknife is that, in the latter, one or more values are serially dropped at a time and the results recalculated. Thus, in jackknife resampling, the subsamples are not necessarily random. This method can be used even when the sample is relatively small such as 10 or 15.

The following example illustrates another use of jackknife resampling for deriving various validity measures of a new method of classification of breast cancer cases. Breast Imaging Reporting and Data System (BI-RADS) is a conventional method of mammographic interpretation. Buchbinder et al. [5] used a computer-aided classification (CAC) that automatically extracted lesion-characterizing quantitative features from digitized mammograms. This classification was evolved on the basis of 646 pathologically proved cases (323 malignant). The jackknife method (100 or more subsamples by a computer) was used to calculate that the sensitivity of CAC is 94%, specificity is 78%, positive predictivity is 81%, and the area under the ROC curve is 0.90.

Resampling methods can be used for a variety of purposes and obviate the need to depend on Gaussian distribution. Many enthusiasts believe that there would be a paradigm shift, and statistical methods would be primarily based on resampling instead of the conventional methods. These methods show promise today as much as they did 25 years ago but have not been able to make much headway so far.


1. Good P. *A Practitioner's Guide to Resampling for Data Analysis, Data Mining, and Modeling*. Chapman & Hall/CRC Press, 2011.
2. Efron B. Bootstrap methods: Another look at the jackknife. *Ann Statist* 1979;7:1–26. http://projecteuclid.org/download/pdf_1/euclid.aos/1176344552
3. Quenouille MH. Notes on bias in estimation. *Biometrika* 1956;43(3–4):353–60. http://www.jstor.org/discover/10.2307/2332914
4. Shao J, Tu D. *The Jackknife and Bootstrap*. Springer, 1995.
5. Buchbinder SS, Leichter IS, Lederman RB, Novak B, Bamberger PN, Sklair-Levy M, Yarmish G, Fields SI. Computer-aided classification of BI-RADS category 3 breast lesions. *Radiology* 2004;230:820–3. http://pubs.rsna.org/doi/full/10.1148/radiol.2303030089, last accessed September 19, 2014.


## research

Research is either discovery of new facts, enunciation of new principles, or fresh interpretation of the known facts or principles. It is an attempt to reveal to the world something that either was never thought of or was in the domain of conjectures—at best being looked at with suspicion. It is a systematic investigation to develop or contribute to generalizable knowledge. Research is a step in relentless search for the truth—it is an organized and systematic approach to finding answers to questions or to refute the existing knowledge. The basic function of research is to answer the why and how of a phenomenon, but searching answer to what, when, how much, etc., is also part of research endeavors. All these questions have relevance to any discipline, but medicine seems to have special appetite for such enquiries. The goal of medical research is to improve health, and the purpose is to learn how systems in the human body work, why we get sick, and how to get back to health and stay fit. It is the logical process to better understand the etiology, pathophysiology, diagnosis, therapy, and prognosis of health aberrations. Research is the very foundation of improved medical care. It can also provide evidence for policies and decisions on health development.

Although research brings dividends in terms of satisfaction, eminence, and funding, does it really benefit the medical science?

According to Chalmers et al. [1], of more than 25,000 reports published in six leading basic sciences journals between 1974 and 1983, just 100 included confident claims of new discoveries, mere 5 resulted in licenses by 2003, and only 1 intervention is being widely used. Thus, it looks as though the focus is lost somewhere along the process. Poor methodology aggravates the research wastage, and this could well be a reason for such dismal performance.

Perhaps we need to focus on patient outcome in terms of not just physical improvement but also meeting their mental and social expectations, and the perception of the doctors and other health care providers. This may require indulgence in soft data as opposed to the present emphasis on hard data, in realization that inaccurate measurement of the real outcome of interest can be more beneficial than accurate measurement of unimportant, perhaps superficial, outcomes.

Whereas outcomes or the results of research endeavors are important by themselves, the process of research determines the credibility. The planning should be based on critical appraisal of the available evidence in the context of the methodology adopted to create that evidence. The deficiencies should be objectively evaluated and remedial actions taken. While interpreting the results, it is important to distinguish the quality of evidence between an aggregation of small-scale studies and a single large-scale study with immaculate methodology.

1. Chalmers I, Bracken MB, Djulbegovic B et al. How to increase value and reduce waste when research priorities are set. *Lancet*, 2014 Jan 11;383(9912):156–65. http://thelancet.com/journals/lancet/article/PIIS0140-6736(13)62229-1/fulltext

## research synthesis, *see also* systematic reviews, meta-analysis

Synthesis is the process of combining and reconciling varied and sometimes conflicting evidence. Although statistical *analysis* is acknowledged as an essential step in empirical research, the importance of *synthesis* is sometimes overlooked. Research synthesis is a global term that includes **systematic reviews** and **meta-analyses**.

The findings of an investigation do not often match with those of another similar investigation. Diabetes, smoking habits, and blood pressure levels were found to be significant factors of mortality in Italy in one study, but not in other studies in the same country [1]. Prevalence of hypertension in India was found to range widely from 0.36% to 30.92% in a general population of adults [2]. Sometimes the results are at odds with one another. These anomalies occur for a variety of reasons such as genuine population differences; sampling fluctuation; disparate definitions, methodology, and instruments; and differences in the statistical methods used. A major scientific activity is to synthesize these varying results and arrive at a consensus. The discussion part of most articles published in medical journals tries to do such a synthesis, but this can be biased toward the finding of the study and prone to subjective judgments. The objective of most review articles is basically to present a holistic view after reconciling the varying results in different studies. In addition, techniques such as meta-analysis seek to combine evidence from different studies. These synthesis methods are primarily statistical in nature and are important components of medical research endeavors.

Varying results by different workers put a question mark on what conclusion one should draw. The effect of lack of uniformity is a waste of intellectual efforts. Synthesis helps to retrieve at least part of the disparate and conflicting results by analyzing in specifics that seem to have caused such results. Another related term is integration that takes us a step further in science by accumulation and refinement of information.

Research synthesis is like putting together stones of different sizes and shapes for constructing a fairly smooth wall. This is difficult but doable by filling up gaps with knowledge that sticks like cement. According to Barnett-Page and Thomas [3], this may require (i) cross-disciplinary approach to build a solid framework of evidence; (ii) different synthesis methods depending on the type of data; (iii) innovative thinking for filling up the gap and, more importantly, for identifying the gaps; (iv) critical thinking to be able to decide what to include and what to exclude; and (v) skills in putting together disparate pieces of information in a manner that commands trust and is considered credible. Thus, synthesis requires rigorous skills of distilling and combining the ideas.

For further details of research synthesis, see Cooper et al. [4].

1. Menotti A, Seccareccia F. Cardiovascular risk factors predicting all causes of death in an occupational population sample. *Int J Epidemiol* 1988;17:773–8. http://www.ncbi.nlm.nih.gov/pubmed/3225084
2. Gupta R. Meta-analysis of prevalence of hypertension in India. *Indian Heart J* 1997;49:43–8. http://europepmc.org/abstract/med/9130424
3. Barnett-Page E, Thomas J. *Methods for Synthesis of Qualitative Research: A Critical Review.* ESRC National Centre for Research Methods—NCRM Working: Paper Series Number (01/09). http://eprints.ncrm.ac.uk/
4. Cooper H, Hedges LV, Valentine J (Eds.). *The Handbook of Research Synthesis and Meta–Analysis*, Second Edition. Sage, 2009.

## residuals, *see also* regression models (basics of)

All **models** are simplified version of the actual process—thus, there would always be some difference between the values obtained from the model and the actual observations. This difference is called error in the context of the population and residuals in the context of the sample. The models are built in a manner that these residuals are minimal without sacrificing the simplicity. The **least squares method** is a case in point that minimizes the sum of squares of residuals in regression models.

In the case of **multiple linear regression**, the model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \varepsilon,$$

where the $x$'s are the regressors and $y$ is the dependent. The term $\varepsilon$ is the part that remains unexplained. This notation is for the population and is called error. The corresponding sample analogue for the model is $y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_K x_K + e$, where the $b$'s are the estimates of β's and now $e$ is called the residual.

Consider the simple regression model given by BSA = 1321 + 0.3433(Wt in g) for children where BSA is the body surface area. If this equation holds, it tells what BSA to expect in children of different weights. With this model, the estimated BSA of a child of 20 kg is 8187 $cm^2$. The actual values would most likely differ from these predicted values. If the actual BSA of a child of 20 kg is found to be 8093 $cm^2$, then the difference from the expected is –94 $cm^2$. If another child of 20 kg has BSA 8214 $cm^2$, then the difference is +27 $cm^2$. The residual for the first child in this example is –94 $cm^2$, and for the second child it is +27 $cm^2$. In general, for the subjects giving rise to the regression, some residuals will always be positive and some negative in such a manner that their sum, and hence mean, is always zero.

The magnitude of residuals is crucial in judging the adequacy of a model. If the model is a good fit to the data, the residuals will be

mostly small. Because their mean is zero, small residuals amount to a small standard error (SE). Conversely, a small SE of residuals is an indication that the model obtained is a good fit to the data. If not, examine if addition of more relevant regressors can reduce the SE. Examples of unusual residual plots are given in Figure R.11c and d, where the residuals have a U-shape in one of the figures indicating that the model is inadequate and incorporation of the $x^2$ term might help. In the second figure, the residuals are almost equally divided on either side of zero, but the variability increases as $x$ increases. This is called *fanning* and indicates lack of **homoscedasticity**. A log or square root transformation of $y$ may help in achieving uniformity of variance over $x$ in this case. Thus, the residual plot can provide important clues for improving the model.

Ordinary quantitative regression methods require that the residuals follow a Gaussian pattern. This is especially important for obtaining confidence intervals on various parameters and for test of hypothesis on them. If the residuals are indeed random, most will be around zero and progressively less and less with larger values in absolute sense. Thus, they will follow a Gaussian pattern. If not, the usual practice is to try transformation depending upon the shape of the plots as explained in the preceding paragraph.

Consider a study on relationship between cholesterol level and body mass index in morbidly obese people (BMI ≥ 40 kg/m$^2$). In such subjects, the cholesterol level does not increase with BMI as found for non-obese people; rather it somewhat decreases, and the slope in **simple linear regression** is negative. Figure R.11c shows the plot of residual versus $x$ when such a regression analysis of cholesterol level ($y$) on BMI ($x$) was run, and Figure R.11d shows the plot versus $y$ values. The first plot shows that the residuals are almost equally and randomly (no discernible pattern) divided across the zero line, and they are concentrated more around the zero line with a few points far away, confirming that the residual distribution is not far from Gaussian and no residual seems to be an outlier. The scatter of the plot does not really differ much with increasing $x$, confirming homoscedasticity. Large values of residuals indicate that the BMI itself is not enough to explain the cholesterol level in these subjects—other regressors are needed for the regression to be an adequate representation of the data. In case of multiple regression, similar plots can be done versus each regressor, and conclusions as just illustrated can be drawn.

The residuals versus $y$ values give a neat straight line (Figure R.11d) with increasing residuals as $y$ increases. This shows that the model overpredicts for low values of cholesterol level and underpredicts for high values, and indicates a room for improvement either by incorporating new regressors or by including powers of $x$. Transformation of $y$ can also help. There is no outlier in our sample, but this kind of plot is helpful in detecting outliers as well when present.

It is customary to examine the plot of residuals versus the predicted $y$. In case of adequate fit, this plot is also expected to randomly hover around the zero line just as it does versus $x$ in Figure R.11a. In case there is a discernible pattern, there is a scope for improving the model.

Perhaps it is more informative if the residuals are Studentized, which is obtained by dividing each by the SE since the mean is already zero. Under the Gaussian conditions, 95% of these residuals should lie between –2 and +2.

## residual sum of squares, *see* error sum of squares

## response rate

Statistically, this is the proportion of subjects who really responded to a question or a therapy. This generally does not have a time component but is still called a rate. The following example illustrates that the response rate could be anomalous in some situations.

Consider a clinical trial on a new therapeutic regimen that is expected to improve the response rate in patients with

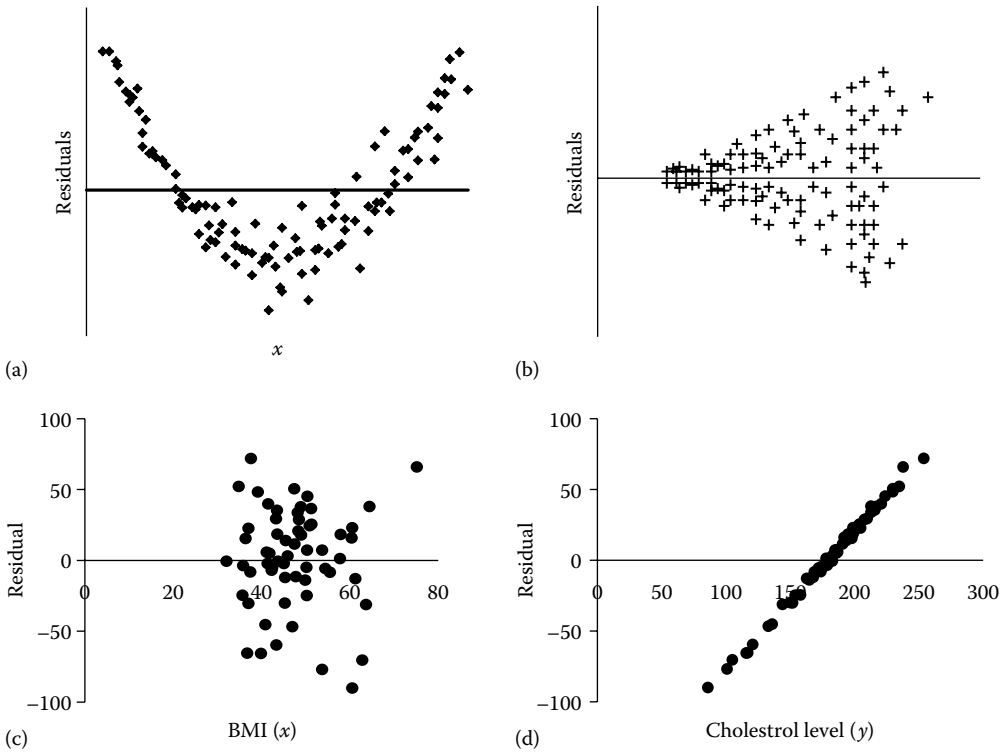



**FIGURE R.11** Residual plots: (a) U-shaped; (b) fanning: (c) versus $x$; and (d) versus $y$.

hyperthyroidism within a stipulated period. A total of 40 cases with varying grades of disease were put on trial, and an equal number receiving the existing therapy were used as controls. The allocation of subjects into treatment and control groups was randomized. The data in Table R.10 were obtained when the response of subjects was subsequently divided according to the severity of their condition.

The response rate in the treatment group was 80.0% and in the control group 62.5%. This difference is statistically significant ($P < 0.05$). When the subjects were divided according to the severity of the disease, it turned out that the therapy has a differential effect in mild than in severe cases. Randomization did not match the control group with the treatment group for severity of the disease in the subjects. When mild and severe cases from the control group were compared with those of the treatment group, respectively, the new treatment was found to be of no significant value.

This example illustrates that response rates as arrived at in a study or as reported in a publication cannot be taken granted on face value. Beside the case mix as in our example, there could be other problems. Penwarden [1] gives an example of online surveys where this problem is particularly acute. Such surveys are common but are notorious for low response rate, and the response can be severely biased. Even a rate of 25% in online surveys is amazing. Telephone surveys generally yield just about 10% response. Compare this with a rate of 90% that people can easily achieve with captive population who are asked to respond to a questionnaire as part of their duty such as nurses in a hospital. The response rate also depends on how useful the survey is perceived by the respondents. If they think that this is going to help them, they will respond in large proportion. If a survey is based on three or four questions, perhaps almost everybody will respond. Larger questionnaires tend to distract the respondents and may yield unreliable information, particularly toward the end of the questionnaire. Even in face-to-face interviews, fatigue sets in with the interviewer as well if it is long, and the survey loses quality.

The response rate also depends on how the questions are framed. A similar question when put in an interesting manner may yield a response, even a correct response, compared with a badly worded question. Rapport established by the interviewer with the respondent could be a key parameter determining the response rate as well as the quality of responses. It helps to have an interviewer from the same milieu as the respondent, speaking the same dialect, perhaps of the same intelligence.

Now let us talk a bit about responses to medical regimens. This has a different meaning since this response implies that the patient has shown perceptible improvement. It is in this sense that the term response is used in Table R.10. This kind of response rate is generally believed to vary either because of difference in the potency of the intervention or because of severity of disease. But there might be other factors as well that tend to be ignored such as nursing care, family support, inner motivation to fight the disease, etc. Thus, one should apply sufficient caution in interpreting the response rate of an intervention.

1. Penwarden R. *Response Rate Statistics for Online Surveys—What Numbers Should You Be Aiming for. Fluid Surveys.* http://fluidsurveys.com/university/response-rate-statistics-online-surveys-aiming/

**TABLE R.10**
**Response Rate in Mild and Severe Hyperthyroid Cases**

| Group | Number of Subjects | Number Responded | Response Rate (%) |
|---|---|---|---|
| I. Treatment group | 40 | 32 | 80.0 |
| Mild | 30 | 26 | 86.7 |
| Severe | 10 | 6 | 60.0 |
| II. Control group | 40 | 25 | 62.5 |
| Mild | 8 | 7 | 87.5 |
| Severe | 32 | 18 | 56.2 |

## response surface

Consider regression of the birth weight of full-term healthy babies on the weight of the father and mother. Suppose the regression equation obtained on the basis of a random sample is

$$\text{BW} = 2.65 + 0.008\ (\text{MW}) + 0.004\ (\text{FW});$$
$$55 \le \text{MW} < 80;\ 60 \le \text{FW} < 90,$$

where BW is birth weight, MW is mother's weight, and FW is father's weight. All weights are in kilograms. Since there are two regressors in this equation, graphically a plane is obtained in place of a line. This is the response surface of this equation as shown in Figure M.8 in the topic **multiple linear regression**. This is just about the most simple response surface. The response in this equation is considered to be linearly affected by the weight of the father and mother. In practice, there will be many factors that will affect the response; it will not be linear and could be quite complex.

Response surface methodology is a step further. It does not restrict to depicting to response surface but extends to designs that optimize the response for different levels of the factors. This requires investigations into how the response varies as the values of the factors are changed, and also requires picking up a combination of factors that optimizes the response. The factor values in this case must be manipulative, that is, they are not something like age and sex, which are not in our control, but are something like types of treatment, dose levels, route of administration, and intake schedule. All these are within our control, and we can find which combinations of these factors really minimize, say, duration of hospitalization. The response surface in this case is some function of the levels of factors, say $f(x_1, x_2, \ldots, x_K)$, and may look like a regression model. The difference is that in case of regression, best-fitting response surface is obtained for given values of $(x_1, x_2, \ldots, x_K)$, whereas here the regressor values are altered to optimize the response.

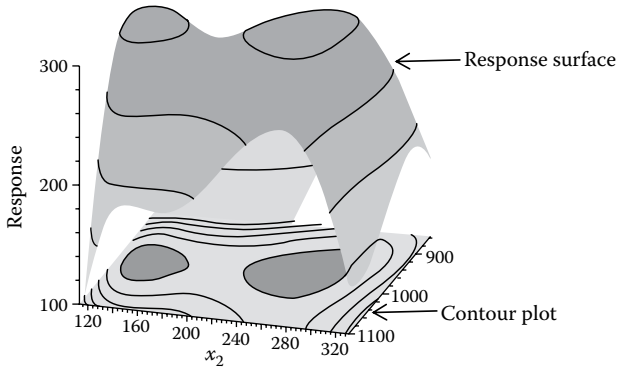



**FIGURE R.12** Three-dimensional response surface and the corresponding contour plot. (From Alvarez LF. Response surface methodology, in: *Approximation Model Building for Design Optimization Using the Response Surface Methodology and Genetic Programming* (PhD Thesis). University of Bradford, U.K., 2000. http://www.brad.ac.uk/staff/vtoropov/burgeon/thesis_luis/chapter3.pdf.)

Response surface methodology involves obtaining contour curves between levels of two factors, say between $x_k$ and $x_{k*}$, keeping all other factors fixed. The kind of contour you can obtain and the corresponding response surface are shown in Figure R.12 for variables $x_1$ and $x_2$.

Ref. [2] is a useful resource for understanding the response surface methodology.


1. Alvarez LF. Response surface methodology, in: *Approximation Model Building for Design Optimization Using the Response Surface Methodology and Genetic Programming* (PhD Thesis). University of Bradford, U.K., 2000. http://www.brad.ac.uk/staff/vtoropov/burgeon/thesis_luis/chapter3.pdf
2. Myers RH, Montgomery DC, Anderson-Cook CM. *Response Surface Methodology*: *Process and Product Optimization Using Designed Experiments*. Third Edition. Wiley, 2009.


## retrospective studies

Retrospective studies are those that investigate antecedent–outcome relationship by going from known outcome to the unknown antecedents. A format for examining gestational age in relation to maternal anemia is that births with different gestation periods are chosen and anemia status of the mothers during the antenatal period is retrieved from records. The first assessment in this format is the outcome, and the antecedents are subsequently assessed for each type of known outcome. Note the temporal difference between this and a **prospective study** (Figure R.13).

Investigation in the case of retrospective studies is in the reverse direction—from outcome to the antecedent. This may seem unnatural but is generally considered more efficient. Note the quickness with which a study in this format can be carried out. There is no need to wait for the outcome to develop or not develop. The outcome is already known, and the information on antecedent is obtained either from records or inquiry. The cases, and in most studies controls, are assembled, and information regarding their past exposure to risk factors is collected. The cases can arise or can be recruited in the future such as cases of breast cancer coming to a clinic, yet the study is technically retrospective so long as it investigates antecedents for a known outcome.

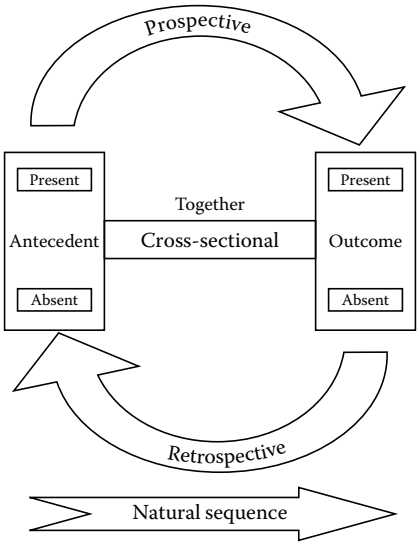



**FIGURE R.13** Schematic representation of prospective, retrospective, and cross-sectional studies.

Only subjects for whom the outcome is already known can be studied with a retrospective design. The outcome is the factor in this format, and the antecedent, which is elicited in retrospective studies, is the response. In prospective studies, on the other hand, antecedent is the factor, and outcome is the response since that is what is elicited.

Many studies do not proceed from the outcome to the antecedent, yet are termed as retrospective in the medical literature. This usage indicates the time frame and not the etiological sequence. For example, Hilska et al. [1] in 2001 reported an analysis of 150 patients with primary proximal colon cancer in Finland who were operated upon during 1981–1990. But the outcome measure was a 5-year survival rate. The study still is from the antecedent (colon cancer) to the outcome (survival). It is a prospective study going by the terminology we use, and could be called a retrospective cohort, but not technically a retrospective study.

A retrospective study can also be conducted in the current time frame in special situations. Women with preeclampsia can be assessed for their present nutrition level and parity as risk factors in the hope that the level of such risk factors was the same before the occurrence of preeclampsia and contributed to the condition. The direction of investigation in this study is also from the outcome to the antecedent—and thus is technically retrospective.

Sampling of subjects in a retrospective study is based on the outcome; thus, the first step is to identify and define the outcome of interest. This could be negative such as disease or death, or positive such as relief or a specified minimum reduction in cholesterol level. The criteria for diagnosis of a disease and its severity must be fully specified. For example, in cancer cases, the stage of disease and, of course, the affected site must be specified. Depending on the amount of information available, it is sometimes useful to keep a separate track of the definite from suspected cases. Obtaining information on possible risk factors that could have given rise to this outcome is the next step. The most common approach for eliciting such history is interviewing the respondent. Because of the **recall lapse** and the intentional misreporting, it is better to depend on records if they are relatively complete.

Most retrospective studies are in the **case–control** format where one group is with disease and the other is without disease. The frequencies of antecedent in these two groups are compared, and the analysis is mostly by **logistic regression** where the **odds ratio** (OR) of each antecedent is obtained. However, some retrospective studies are done without controls. Obtaining history of cancer cervix patients regarding infections, diet, and use of oral contraceptives is retrospective but not case–control if noncancer subjects are not part of the study.

Retrospective studies have many inverse properties to prospective studies. A retrospective study can be accomplished with a relatively small number of subjects in less time and resources. It is efficient for rare outcomes because it can begin with a sufficient number of cases. It can simultaneously evaluate many causal hypotheses and is efficient also in the evaluation of interaction between different risk factors. A case–control study also allows easy control of confounders. All such advantages accrue mostly because a large number of affected cases are generally available in this format.

On the downside, recall lapse is common in a case–control study that can bias the results. Differentials such as the ability of cases to recall events easily than controls can cause additional bias. It can also be biased because only those who already have had the required outcome can be included. Many severe cases may have already died and cannot be a part of this type of study. A case–control format will not be able to establish the sequence of events.

**TABLE R.11**
**General Performance Comparison of Prospective, Retrospective, and Cross-Sectional Designs**

| Criteria | Prospective | Retrospective | Cross-Sectional |
|---|---|---|---|
| Cost and time | High | Low | Low |
| Number of subjects required | Large | Small | Large |
| Suitability for rare exposures | Good | Poor | Poor |
| Suitability for rare outcomes | Poor | Good | Poor |
| Spectrum of etiologic factors that can be investigated | Small | Large | Large |
| Spectrum of outcome factors that can be investigated | Large | Small | Large |
| Recall lapse and other biases | Not likely | Very likely | Not likely |
| Completeness of information | High | Low | Full, but only cross-sectional |
| Dropouts | More | Less | None |
| Changes in the characteristics of the subjects over time | More likely | Less likely | None |
| Assessment of temporal relationship | Good | Difficult | Not possible |
| Suitability for assessment of sensitivity and specificity | No | Yes | Yes, if the sample is representative |
| Suitability for assessment of predictivities | Yes | No | Yes, if the sample is representative |
| Evaluation and control of confounders | Poor | Good | Fair |
| Assessment of risk | Direct by relative risk | Indirect by odds ratio | Approximate by prevalence rate ratio |
| Assessment of cause–effect relationship | Good | Fair | Poor |

Table R.11 is a useful summary comparing the performance of prospective, retrospective, and cross-sectional designs with regard to their performance.

### *Sampling in Retrospective Studies*

Since a retrospective study is based on cases with disease to begin with, the sample size must be adequate that can represent the entire spectrum of subjects and can provide reliable results suitable for generalization. Cases included should ideally be representative of all persons with the specified outcome, but many case–control studies are carried out on a nonrandom sample. Random selection is especially important for **descriptive studies** but probably not so important for **analytical studies** including retrospective studies. Experience suggests that the relationship between antecedent and outcome can be adequately assessed despite a nonrandom sample in many situations as long as the bias is under check, the basic requirement for which is the baseline equivalence of the cases and controls.

At the same time, these studies also involve estimating a parameter such as OR, finding a confidence interval (CI), and testing a hypothesis. These statistical procedures do require a random sample of the subjects. Indeed, a random sample should be taken whenever feasible. In addition, many of these statistical methods require a large sample.

### *Nested Case–Control Studies*

An extension of retrospective studies is **nested case–control studies**. In this setup, a cohort of subjects is followed up, and cases (and controls) are chosen from this cohort. See the example below for a nested case–control study that uses stratified sampling to assess the effect of serum selenium levels on cancer mortality.

The serum selenium level is widely suspected to affect cancer mortality, although the results across studies are not coherent. Belgium has a system to follow up each patient till death. From this "cohort," a stratified (by sex: male and female) random sample of 201 cancer deaths of age 25–74 years out of a total of 343 during a 10-year period was studied by Kornitzer et al. [2] for their selenium level as well as some other factors. Three controls were also selected from the cohort for each case, and these were matched for age and gender. Thus, a total of 603 controls were also studied. The serum selenium level was found to be a significant predictor of cancer mortality in males but not in females.

Also note that the investigations proceeded from outcome (cancer death) to an antecedent (serum selenium level), and thus the study is retrospective in nature. Since controls were also investigated, it is a case–control study. It is nested because follow-up of each person is routinely done in Belgium, and cases and controls were chosen from this follow-up. Controls were easily available, and choosing three controls per case helped to increase the reliability of results without the corresponding increase in cost.

On the flip side of this study is the sample of 201 cancer deaths out of 343 and the claim that it is a random sample. Such 60% sample is not a norm: one can legitimately wonder why nearly all 343 could not be included in the study. Had all these been investigated, it would still be a sample in the sense that they occurred in a 10-year period that would naturally exclude past and future deaths.

1. Hilska M, Gronroos J, Collan Y, Laato M. Surgically treated adenocarcinomas of the right side of the colon during a ten-year period: A retrospective study. *Ann Chir Gynaecol* 2001;90 (Suppl 215):45–9. http://www.ncbi.nlm.nih.gov/pubmed/12016748
2. Kornitzer M, Valente F, de Bacquer D, Neve J, de Backer G. Serum selenium and cancer mortality: A nested case-control study with age- and sex-stratified sample of Belgian adult population. *Eur J Clin Nutr* 2004; 58:98–104. http://www.ncbi.nlm.nih.gov/pubmed/14679373

## reverse causation

We all grew up thinking that cause precedes the outcome. But it is possible that the time frame of cause and effect might be reversed of what is generally thought of. In some cases, the effect, although not causation, seems to precede the cause, although that might be temporary. Also called retrocausation or backward causation, reverse causation refers to a situation where apparently the outcome precedes the exposure against the conventional relationship from antecedent to outcome. This may look strange but is not an altogether

unlikely situation. For example, the studies might indicate that obese people are more depressed as though obesity contributes to depression. Actually, depression might change dietary preferences and can cause obesity. This is the reverse causation. A study by Marquis et al. [1] reported that increased breastfeeding did not cause stunting of children in a shanty town in Lima, but stunted children generally were increasingly breastfed.

Flegal et al. [2] discussed the phenomenon of reverse causation in the context of effect of illness-related weight loss on the relation between body weight and mortality. A low weight can cause illness rather than the usual paradigm of illness causing low weight. This is an example where it may be difficult to establish what caused what, and reverse causation might be operative. For a good discussion on causality, see Pearl [3].

No clear approaches are yet available for studying reverse causation. Flanders and Augestad [4] suggest that the analysis should be balanced with **sensitivity analysis** and **simulations**. Lawlor et al. [5] used the Cox proportional hazards model for studying reverse causality between weight and mortality.


1. Marquis GS, Habicht J, Lanata CF, Black RE, Rasmussen KM. Association of breastfeeding and stunting in Peruvian toddlers: An example of reverse causality. *Int J Epidemiol* 1997;26:349–56. http://ije.oxfordjournals.org/content/26/2/349.full.pdf+html
2. Flegal KM, Graubard BI, Williamson DF, Cooper RS. Reverse causation and illness-related weight loss in observational studies of body weight and mortality. *Am J Epidemiol* 2011;173(1):1–9. http://aje.oxfordjournals.org/content/early/2010/11/07/aje.kwq341.full
3. Pearl J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.
4. Flanders WD, Augestad LB. Adjusting for reverse causality in the relationship between obesity and mortality. *Int J Obes (Lond)* 2008 Aug;32 Suppl 3:S42–6. http://www.ncbi.nlm.nih.gov/pubmed/18695652
5. Lawlor DA, Hart CL, Hole DJ, Davey Smith G. Reverse causality and confounding and the associations of overweight and obesity with mortality. *Obesity (Silver Spring)* 2006 Dec;14(12):2294–304. http://www.ncbi.nlm.nih.gov/pubmed/17189558


## reviews, *see* systematic reviews

## ridge regression, *see also* regression models (basics of)

This is a method of fitting a regression to minimize the effect of **multicollinearity** on the estimates when it exists among the regressors.

Any regression method, when widely used, is looked at very critically and the inadequacies noted. Thus, several modifications have been suggested for the regression to meet specific requirements. Among those used sometimes in health and medicine is the ridge regression. Consider predicting body fat with the help of skinfold thickness at the triceps, mid-arm, and thigh. These three are highly correlated, and the regression coefficients are likely to have an increased variance because of such multicollinearity among the regressors. The regression model loses its sheen and its utility is compromised because of such instability of the estimates. This can also happen when the sample size is too small for the number of regressors. Ridge regression is one of the modifications that can remedy this problem to a great extent. Instead of using the method of least squares, a method called ridge is used that produces slightly biased estimates of the regression parameters but substantially reduces the standard errors (SEs). Some statistical software packages provide an option to run this kind of regression. A brief is as follows, but see Montgomery and Peck [1] for more details.

As a safety against blown-up estimates (that can happen due to multicollinearity or due to a large number of regressors relative to the sample size $n$), ridge regression puts a constraint that $\Sigma b_k^2 < C$, where $b_k$'s are the estimated regression coefficients and $C$ is some predefined constant, known as the ridge parameter. This estimation is done after the regressors are standardized, and there is no need of any intercept. This constraint implies that the least squares method that we generally use for estimating a regression coefficient is penalized if the sum of squares of regression coefficients becomes too big. The net result is that the estimates of the regression coefficient reduce in magnitude, but the cost is that they are no longer unbiased. That is, the usual property of the estimates of regression coefficients reaching the corresponding population regression coefficients in the long run does not hold. This is the price paid to get more reliable estimates in such situations.

de Vlaming and Groenen [2] discussed the use of ridge regression for prediction in quantitative genetics using single-nucleotides polymorphism data, which have a high number of variables and high multicollinearity. They also discussed how **LASSO**-type methods can be helpful in running a ridge regression.


1. Montgomery DC, Peck EA. *Introduction to Linear Regression Analysis*, Fifth Edition. Wiley, 2012.
2. de Vlaming R, Groenen PJ. The current and future use of ridge regression for prediction in quantitative genetics. *Biomed Res Int* 2015;2015:143712. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4529984/


## risk difference, *see* attributable risk

## risk factors

These are those factors that directly or indirectly affect the outcome of interest. In the context of diseases, a risk factor can be defined as a "measurable characteristic that is associated with increased disease frequency and that is a significant independent predictor of an increased risk of presenting with the disease" [1]. In common parlance, the term *risk* connotes an adverse outcome, but this term in statistics is used even for favorable factors. Thus, the effect of risk factors on the outcome can be negative or positive. The Framingham Heart Study is generally attributed to have used this term for the first time [2], although the underlying concept has been widely applied earlier.

Risk factors are distinguished from a cause. The cause of typhoid is *Salmonella typhi*, and the cause of thyroidism is iodine deficiency. Cause is easily identifiable for most infective diseases, but that is not so for most chronic diseases. In chronic diseases, "causes" are usually multifactorial; thus, the concept of risk factor was born, which works in probability senses—some people are affected and some not—that also when other conducive factors are present. When risk factors are under control, the disease incidence is lower. For example, the risk factors of cardiovascular disease are hypertension, obesity, sedentary habits, and imprudent diet. They increase the risk but do not cause cardiovascular disease.

The concept of risk factors is applicable only to the **analytical studies** and not to the **descriptive studies**. These factors should be identified before data are collected by using previous results, clinical insight, and thorough understanding of the disease process. One easy method of identifying relevant risk factors for a study is to draw up a list of all factors that could possibly influence the outcome of interest based on your own knowledge, wisdom of seniors, or a review of the literature, and choose the more relevant ones that would be studied as risk factors or antecedents for their role in the outcome in the context of the study.

As the knowledge expands, the list of risk factors for various health outcomes increases, and they are now increasingly identified with more emphasis on their exact quantitative contribution. This quantification helps in their **ranking and selection**, thereby helping to devise more rational strategies for control of diseases. The most common statistical methods used to study the risk factors are **logistic regression** for qualitative outcome (yes/no or (mild/moderate/severe/critical), ordinary **regression** for quantitative outcome, and **Cox regression** for hazards. When properly analyzed and correctly interpreted, these regressions will quantify the contribution of each risk factor to the outcome, both in the presence of other risk factors and in their absence. Remember that some risk factors produce synergistic effect when other favorable conditions are present—thus, the interpretation requires care. For a discussion on mediator, independent, overlapping, and proxy factors, see the article by Kraemer et al. [3].

An important issue often forgotten while interpreting the effect of risk factors is that our knowledge is limited, and that all studies are done as per the existing knowledge. There may be important risk factors about which we do not know yet. Thus, the implication of findings remains truncated. Terms such as *net effect* of a risk factor, for example, in case of multivariable logistic regression, are inappropriate since the regression is based only on the known risk factors, and the effect of unknown factors continues to haunt the results. In addition, many risk factors still are in the hypothetical instead of the real domain as they are just suspected and not proven.

Besides logistic regression, among other statistical methods available to identify important risk factors is **sensitivity analysis**. For example, risk of coronary disease can be modeled to depend upon the presence or absence of diabetes, hypertension, and dyslipidemia. This model might be able to correctly predict 10-year risk in 62% of the cases. However, addition of smoking and obesity may increase this to 70%. This addition of 8% is substantial. Thus, predictivity of coronary disease is *sensitive* to the choice of risk factors. The first model is based on three risk factors and the second on five risk factors. Had the contribution of smoking and obesity been only 2% or 3%, the conclusion would be that prediction of coronary disease is insensitive to smoking and obesity when diabetes, hypertension, and dyslipidemia are known. This kind of argument can be used to establish robustness of the results on risk factors.

Risk factor finding can be tricky sometimes as illustrated by the following example provided by Yusuf et al. [4]. More than 80% of the global burden of cardiovascular disease occurs in low-income and middle-income countries, but knowledge of the importance of risk factors is largely derived from developed countries. Therefore, the effect of such factors on risk of coronary heart disease in most regions of the world is unknown. To delineate these, the authors established a standardized case–control study of acute myocardial infarction in 52 countries, representing every inhabited continent. They managed to enroll 15,152 cases and 14,820 controls. They found that smoking, history of hypertension or diabetes, waist/hip ratio, abnormal lipids, smoking, hypertension, diabetes, abdominal obesity, psychosocial factors, consumption of fruits, vegetables, and alcohol, and regular physical activity accounted for most of the risk of myocardial infarction worldwide in both sexes and at all ages in all regions. Their research suggested that approaches to prevention can be based on similar principles worldwide and have the potential to prevent most premature cases of myocardial infarction anywhere in the world.

1. O'Donnell CJ, Elosua R Cardiovascular risk factors. Insights from Framingham Heart Study. *Rev Esp Cardiol* 2008;61(3):299–310. http://public-files.prbb.org/publicacions/737f77a0-08a4-012b-a773-000c293b26d5.pdf
2. Stampfer MJ, Ridker PM, Dzau VJ. Risk factor criteria. *Circulation* 2004 Jun 29;109(25 Suppl 1):IV3–5. http://circ.ahajournals.org/content/109/25_suppl_1/IV-3.long
3. Kraemer HC, Stice E, Kazdin A, Offord D, Kupfer D. How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *Am J Psychiatry* 2001 Jun;158(6):848–56. http://www.ncbi.nlm.nih.gov/pubmed/11384888
4. Yusuf S, Hawken S, Ounpuu S, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): Case-control study. *Lancet* 2004;364:937–52. http://www.ncbi.nlm.nih.gov/pubmed/15364185

## risk ratio, *see* relative risk

## robustness

Robustness in modeling implies consistency, sustainability, and stability of results in less than ideal conditions. For statistical procedures, it connotes certain resilience to deviations from their underlying requirements. Robustness also refers to insensitivity of the overall conclusion to various limitations of data, limited knowledge, and varying analytical approaches. By showing endurance of results to withstand the pressure of minor variations in their validity underlying conditions, robustness establishes wider applicability of the results.

The results that risk of coronary disease is higher in persons with higher blood pressure and risk of lung cancer is higher in persons smoking heavily are robust. The presence or absence of other factors does not affect these results much, and this is seen in all population segments. On the other hand, the conclusion that urinary sodium excretion is dependent on salt intake is not robust since it is so easily affected by metabolism, creatinine excretion, and body mass index. The result that the risk of lung cancer increases by 3% by smoking 100 additional cigarettes-years is also not robust. This chance can easily be 2% or 5% depending on the person's nutrition, exposure to kitchen smoke, etc.

Models, by definition, are relatively simple statements of complex processes. They are man-made manifestations of nature. By their very nature, models are imperfect representations of the real world and, at best, are only approximations of the actual situation. Statistical models are no different as they too often work only in very specific situations, and slight variation in the conditions can produce weird results. In any case, they are supposed to work in the long term and may fail in individual cases. Thus, it is important that robustness of statistical models is established before they are put to use.

Good researchers exhibit robustness of their results to rule out challenges from critics. The evaluation could be done for the set of underlying conditions as a whole but is generally done for its components such as robustness to choice of variables, to instrumentation, and to the methodology. Individual conditions are altered, in turn, to examine if the broad conclusion still remains the same. If you are using a published result for your practice, examine if the result is sufficiently robust to the altered conditions under which you intend to apply it. This also is a way to minimize the impact of uncertainties. **Sensitivity analysis** is one of the many methods to assess robustness, where it is assessed by varying the conditions such as excluding or including some risk factors, altering their validity conditions, changing the basic structure, etc. The other method is **resampling** where subsamples are analyzed to examine if the results replicate.

Many statistical studies of robustness examine the behavior of the estimates with and without outliers in the data. They are sometimes

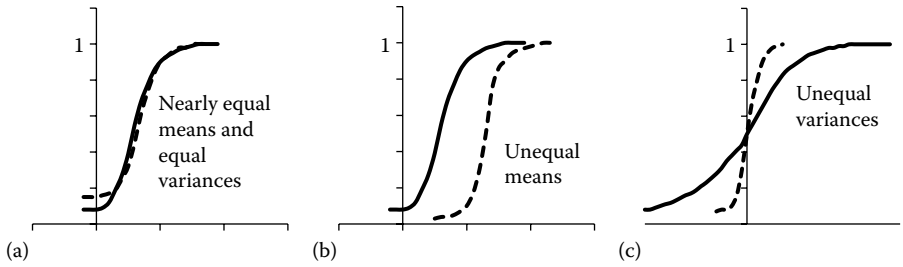



**FIGURE R.14** Plot of cumulative distribution functions in two groups when (a) the distributions are nearly equal, (b) the distributions have unequal means, and (c) the distributions have unequal variances.

extended to examine the effect of serial dependence, skew distribution, varying variance, etc. For example, the Student *t*-test is considered robust to departure from its validity conditions. In a slight variation, the term robust model is used in medicine also for experimental unit such as whether mouse is a robust model to study advanced diabetic kidney disease, and whether cardiomyocytes is a robust preclinical model for assessing atrial selective pharmacology. Our concern in this section was not with medical but with statistical robustness, as may be evident from our discussion.

## ROC curve, *see* receiver operating characteristic (ROC) curve

## $R^2$, *see* multiple correlation

## runs test for equality of distributions

Runs test is used for two purposes and the test takes a different form for these two cases. First is runs test for randomness as discussed under the topic **randomness (statistical tests for)**, and the second is for testing equality of two distributions as discussed next. This also is very similar to the test for randomness.

Run is a sequence of similar values such as increasing or decreasing order, or similar response in case of qualitative data. Consider a sample of size $n_1$ from group 1 ($x$'s) and of size $n_2$ from group 2 ($y$'s). If $n_1 = 4$ and $n_2 = 5$, the following is one of the possibilities when arranged in increasing order:

$$\underline{x\,x}\;\underline{y\,y\,y}\;\underline{x}\;\underline{y\,y}\;\underline{x}$$

This configuration has five runs based on the group to which they belong as indicated by the underlines. If the distribution of values in these groups does not differ much (Figure R.14a), the values in these groups, when arranged in, say, increasing order, would thoroughly mix and the number of runs would steeply rise. As shown for groups with unequal means in Figure R.14b, and with unequal variances in Figure R.14c, you can imagine that the number of runs would be much smaller in these situations. Thus, if the number of runs is smaller than expected, the null hypothesis of equality of the two distribution can be rejected. This is a nonparametric method and can be used in a variety of situations.

This test uses the following argument. Let the number of obtained runs in our sample be denoted by $R$. The

$$P(R \text{ runs}) = \frac{\text{number of ways of getting } R \text{ runs}}{\text{total number of ways of arranging } x\text{s and } y\text{s}}.$$

The denominator is always $(n_1 + n_2)!/(n_1! + n_2!)$. If $n_1 = 4$ and $n_2 = 5$, this is 11!/(4!*5!) = 13,860. This multiplies fast. The numerator depends on whether $R$ is even or odd, and requires knowledge of permutations and combinations. The test is easy for large samples since, in this case,

$$\text{mean of the runs, } \bar{R} = \frac{2n_2n_2}{n_1+n_2} + 1, \text{ and the SE,}$$

$$s_R = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}.$$

For a large sample, Student $t = \dfrac{R - \bar{R}}{s_R}$ with df = $(n_1 + n_2 - 1)$. Reject the null of equality of two distributions if the calculated value of $t$ is less than the critical value at the predetermined $\alpha$ level of significance. Since Student $t$ also approaches Gaussian $z$ for large samples, this is sometimes stated in terms of ***z*-test**.

# S

## sample

A sample is that fraction of the **population** that is actually studied. In almost all situations, the statistical population of interest is big, and we rarely have resources to study all the subjects. Even if the resources are unlimited, we can only study existing cases—the future cases and many of the past cases cannot be studied. And the irony is that the conclusions from the empirical studies are drawn, and we think they would be applicable to future cases. Thus, the sample remains an integral part of medical research framework.

Two questions primarily arise while studying a sample: (i) whether the sample is an adequate representation of the population it is supposed to represent and (ii) whether it is large enough to ensure adequate reliability of the conclusions. A large number of random and nonrandom methods are available to address the former question as discussed under the topic **sampling techniques**, and the **sample size** is a big separate topic in itself to address the latter. We are not repeating those arguments here.

## sample design, *see* sampling techniques (overall)

## sample size determination (general principles)

Sample size, $n$, is just about the first thing that comes to mind when planning an investigation. Indeed, it is among the most important considerations that determine the utility of a study. A good researcher would always provide sample size calculation for his/her study and would give full justification of the values used as inputs to these calculations. This exercise is done in advance (before conducting the study) and stated in the study **protocol** not just for record but also to plan the resources, for setting the time frame, etc., accordingly, and all this is stated again at the time of reporting of results. Charles et al. [1], in a statistical review of 215 published reports of randomized controlled trials, found that 5% did not report calculation for sample size, and 43% did not include full details.

Statistically, the larger the sample, the better the **reliability** of the results. This comes from the fact that the **standard error (SE)** of estimates has $n$ or its function in the denominator, and the inverse of the SE measures the reliability of the estimate. A smaller SE results in a narrow width of the confidence interval (CI)—this increased precision provides more confidence in the results and better statistical power to detect any specified effect. Yet, the sample size should not be too large as this could mean a waste of resources. If a study of only 200 subjects can give a reliable answer, why spend resources to study 250? An unduly large sample is unethical in experiments because it means that some subjects are unnecessarily exposed to an intervention whose utility is in doubt. A small sample, on the other hand, may not give evidence one way or the other, and the study may fail to achieve its objectives. Thus, this also is a waste of resources.

Administratively, a large sample is hard to execute as it tends to become a burden on the investigator in most situations (Figure S.1), and small samples are not useless altogether as they are valid for a pilot, exploratory, or feasibility study, including phase I trials.

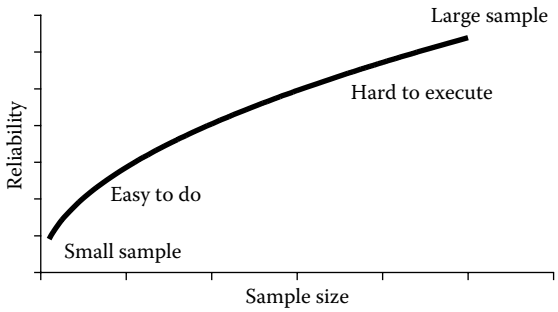



**FIGURE S.1** Reliability of a study and difficulty in its execution with increasing sample size.

When honestly done, small sample studies do allow us to move forward from knowing less. A small sample can be adequate if the anticipated effect size is large. In most other situations, small samples fail to provide a definite conclusion.

A statistician is sometimes expected to be able to suggest an adequate sample size immediately on being told of the title and the objectives of a study. This is like expecting a physician to prescribe a treatment regimen for pain in the abdomen without going into the details. A physician would need some more information such as the duration and intensity of pain, exact location, palpability or tenderness, any accompanying complaint, sometimes an x-ray image, etc. Similarly, the question about sample size can be answered only after some deliberation on the variability of the observations, precision required, confidence desired, etc. These general principles are explained in this section. Most biostatistics books do not emphasize the common structure behind sample size calculations, although many give details of sample size determination in specific setups. The general procedure for determining the sample size is different for estimation as compared to that for testing of hypothesis as follows.

### *Sample Size Required in Estimation Setup*

Estimation of a parameter such as mean and proportion is generally the primary objective of **descriptive studies**. A descriptive study could be a case series, but that is rarely based on a random sample. This leaves **sample survey** as the only relevant format of descriptive studies for calculation of sample size in estimation setup. But estimation is done in **analytical studies** as well, where parameters such as difference in efficacies and odds ratio are estimated. The following are the details of the concerns that need to be addressed to arrive at a sample size for estimating a population parameter. These details focus on mean and proportion, but the same sort of arguments applies to the other parameters. These may look like commonsense considerations in determining the sample size for estimation.

1. The first consideration for the calculation of sample size is the primary *variable of interest*. In antecedent–outcome setup, the sample size will be based on the outcome

variables. If blood pressure, body mass index, blood glucose, and cholesterol level are measured as outcomes in a study on noncommunicable diseases, the sample size can be based on any of them, and different variables will give different answers. You may like to calculate for all of them and choose the maximum as the required sample size. Generally, though, sample size calculation is done for a specific target variable of primary interest. It is helpful to specify this in any case so that the focus of the study is clear.

2. The second point of consideration is the nature of the primary variable: whether it is discrete or continuous, what its statistical distribution is, and whether the distribution can be approximated by Gaussian (normal) for large samples. Almost all formulas for determining sample size are valid under **Gaussian conditions**.
3. The next is the **parameter** of interest—whether it is mean, proportion, difference, odds ratio, correlation coefficient, or what. Sample size formula will depend on the choice of such parameter.
4. What is the **variability** between subjects in the population? Cholesterol levels tend to differ widely even among healthy subjects. Thus, a small sample of, say, seven subjects would not be adequate to reveal the full spectrum of cholesterol levels present in the target population. That is, another sample of seven subjects can yield an entirely different picture. Body temperature, on the other hand, is relatively stable in healthy subjects, and a small sample can be adequate. If the material to be sampled is as homogeneous as blood in the body, even one drop may be enough to reveal nearly the full picture in a person. The sample size requirement increases as the variance increases. In the case of sample proportion $p$, the variance depends on the value of the population proportion $\pi$. Thus, the question becomes, what proportion of subjects are expected to possess the characteristic under investigation? In relative sense, a large sample is required if $\pi$ or $(1 - \pi)$ is small.
5. What is the minimum degree of **precision** required? A pharmaceutical company may be interested to know very precisely the percentage of those with tuberculosis who are becoming drug resistant. This can help the company to assess the exact market for a new antitubercular drug. On the other hand, in surveys on opinions, attitudes, and behaviors, an error in the estimate to the extent of as much as 10% or sometimes even higher may be tolerable. The sample size should evidently be larger when greater precision is required.
6. What *least confidence* in the estimate would be tolerable? No empirical conclusion is 100% dependable, but the investigator may wish to be sufficiently confident that the conclusion will be replicated in repeated samples. There is always a chance of it being false. The chance of reaching a wrong conclusion can be kept low, generally less than 5%, by including a reasonably large number of subjects in the sample. In some cases, such a large sample may not be feasible because of constraints of time and resources, and a larger error may have to be tolerated. If a chance of as much as 10% or 20% of being wrong can indeed be tolerated, as in the case of some opinion surveys, then a small sample may be adequate.
7. Are there *any subgroups of interest?* If conclusions are to be drawn for various subgroups, as in stratified sampling in some cases, then each subgroup needs to be adequately represented. The sample size quickly multiplies and can become an enormous number if cross-classifications of a number of factors are under consideration. For only two factors, viz., age at menarche as <12 and ≥12 years, and age at first live birth as <25 and ≥25 years, the four cross-classifications are (i) menarche at <12 years and first live birth at <25 years, (ii) menarche at <12 years and first live birth at ≥25 years, (iii) menarche at ≥12 years and first live birth at <25 years, and (iv) menarche at ≥12 years and first live birth at ≥25 years. If there are four factors and all are dichotomized, the number of cross-classifications is $2^4 = 16$. Results would be reliable if each of these classifications has an adequate number of subjects.
8. How much, if any, **nonresponse** is expected? Although nonresponse raises questions about the validity of the survey because of bias it introduces, it also reduces the effective number of subjects that can be utilized to draw conclusions. Thus, a larger sample is planned if it is feared that many subjects may not be available, may not cooperate, or may have to be dropped because of development of, say, undesirable side effects. If the subjects are healthy and an invasive procedure is involved, even as small as a pinprick, the attrition could be large.
9. What **sampling technique** is to be used? For example, cluster sampling may require double the size or even larger relative to simple random sampling (SRS) because of the clustering effect. Different procedures have different limitations. Procedures such as stratified and two-stage can require smaller samples to achieve the same precision if the units can be rationally divided into homogeneous groups. This may be difficult to achieve when sufficient information is not available or when the subjects are widely scattered. Then the sample size could increase and will depend on the **design effect**.
10. Although ignored in most practical situations, the sample size also depends on the sampling distribution of the summary measure under study. This section assumes Gaussian (normal) distribution, which is likely at least for mean and proportion when the sample size is large.
11. Sample size also depends on the number of variables to be *simultaneously* considered. Because of complexity, this book discusses only one variable at one time setup for calculation of sample size.
12. Sample size also depends on the design of your study. Designs such as matching and repeated measures require different approach from that in independent samples.

In summary, a bigger sample is required for estimation if

i. Interindividual variability is higher or expected prevalence is low.
ii. More confidence or higher precision is required in the result.
iii. There are subgroups for whom results are to be applied.
iv. Higher nonresponse is expected.
v. Sampling method is other than simple random (in most situations).

Let the population parameter under estimation be denoted by $\tau$ and its sample estimate by $t$. Let $\delta$ be the difference between the two, i.e., $\delta = |\tau - t|$. Suppose the investigator requires that this difference does not exceed a specified limit $L$ in at least $100(1 - \alpha)\%$

of repeated samples. The quantity $L$ is called precision and is the half-width of the CI. The quantity $(1 - \alpha)$ is the confidence level. If a Gaussian form of distribution can be assumed for the sample estimate $t$, which, in fact, could be so in most cases when the sample size is large, it can be shown that

$$L = z_{1-\alpha/2}*\text{SE}(t),$$

where the coefficient $z_{1-\alpha/2}$ is taken from the standard Gaussian distribution such that the probability between $-z_{1-\alpha/2}$ and $+z_{1-\alpha/2}$ is $(1 - \alpha)$. The exact value of $z_{1-\alpha/2}$ for $\alpha = 0.05$ is 1.96; for $\alpha = 0.10$, $z_{1-\alpha/2} = 1.645$; and for $\alpha = 0.01$, $z_{1-\alpha/2} = 2.58$. Thus, for confidence level 90%, $L = 1.645*\text{SE}(t)$; for 95%, $L = 1.96*\text{SE}(t)$; and for 99%, $L = 2.58*\text{SE}(t)$.

This equation is basic for the calculation of sample size in an estimation setup. Let us call this the general equation. In this equation, SE($t$) would almost invariably have $n$ in the denominator, which could then be worked out when other values are known. But a difficulty is that SE($t$) would also contain an unknown parameter such as $\sigma$, which is to be replaced by its estimate. Where would you get this estimate before the survey is conducted? This is obtained either from a previous study or from a pilot study. SE($t$) also depends on the sampling method proposed to be followed. The sample size so obtained may have to be inflated to adjust for expected nonresponse. If estimates for various subgroups are required, then this calculation is done separately for each subgroup.

If there are many parameters under estimation, two approaches are available. The first is to calculate the sample size for the most important parameter if that can be identified. The second is to calculate the size for all the parameters and use the one that is the largest. The latter would give better-than-required precision of some estimates, but each estimate will have *at least* the specified precision. The following example may clarify the procedure to be followed in some cases.

A sample survey is planned to estimate the average level of fasting blood glucose in surviving and apparently healthy females of age 60 years and above of a particular ethnic group. It is desired that the estimate should be within 1.2 mg/dL of the population mean with probability 0.95. A previous study revealed an estimate of standard deviation (SD) = 6.3 mg/dL. How big a sample is required if an SRS is to be followed?

In this case, $L = 1.2$. For SRS, $\text{SE}(\bar{x}) = \sigma/\sqrt{n}$. If we replace $\sigma$ by its estimate $s = 6.3$, then our general equation gives $1.2 = 1.96 \times 6.3/\sqrt{n}$, or $n = 105.9$. This is always approximated upward, in this case to $n = 106$. (A purist would legitimately argue that the Student $t$ value should be used in place of $z_{1-\alpha/2}$ since $\sigma$ is being replaced by $s$, but it does not matter much since we are restricting this discussion to large samples.)

If greater precision (lower $L$) is required, say $L = 0.4$, then $n = (1.96 \times 6.3/0.4)^2 = 953$. Note how rapidly the sample size increases when the required precision is enhanced. If the confidence level $(1 - \alpha)$ is raised to 0.99, then, for $L = 1.2$, $1.2 = 2.58 \times 6.3/\sqrt{n}$ and $n = 184$. If the subject-to-subject variability in the population is smaller, say $s = 2.4$ only, then for $L = 1.2$ and $\alpha = 0.05$, $n$ is 16.

A bigger sample is required whenever (i) greater precision is needed, (ii) more confidence is necessitated, or (iii) the variability in the population is large. The quantity that affects it most is the precision you desire of the estimate. For double the precision (i.e., half $L$), the sample size required becomes four times.

If separate estimates for two age groups such as 60–69 and 70+ years are required, and if SD and other values are the same for each age group in our example, the same sample size is required for subjects of 60–69 years and for those of 70+ years. The total sample size doubles.

For proportions and $(1 - \alpha) = 0.95$, the general equation becomes $L = 1.96*\text{SE}(p)$. If the objective is to estimate the proportion of apparently healthy subjects who can be suspected as diabetic on screening using the cutoff 120 mg/dL of fasting blood glucose level and if it is desired that the estimate should not differ by more than one-fifth of the proportion of $\pi$ in the population, then $L = \pi/5$. This is the relative precision that can be expressed as 20% of $\pi$ or $0.20\pi$. Since $\text{SE}(p) = \sqrt{\pi(1-\pi)/n}$, we need a value of $\pi$ for the sample size calculation. If it is anticipated to be between 0.02 and 0.03, both limits can be used to calculate the sample size. For $\alpha = 0.05$, our general equation gives $0.02/5 = 1.96\sqrt{0.02 \times 0.98/n}$ for $\pi = 0.02$, and $0.03/5 = 1.96\sqrt{0.03 \times 0.97/n}$ for $\pi = 0.03$. The former gives $n = 4706$ and the latter $n = 3106$. The bigger of the two is statistically safer. Thus, a sample of size nearly 4700 is required.

Sometimes it is desirable to do reverse calculations. If resources permit a specific $n$, the general equation can be used to find the precision corresponding to that $n$. For example, if not more than 30 subjects can be studied due to resource limitations in our example on fasting blood glucose level, and if the values of the SE and the level of confidence remain the same, then $L = 1.96 \times 6.3/\sqrt{30} = 2.3$ mg/dL. Thus, the estimate of mean fasting blood glucose level can be more than 2.3 mg/dL away from its true value in the population 5% of the time. If this error is acceptable, go ahead with a sample of size 30. If this is too high, reconsider the idea of carrying out the study on only 30 subjects, as a larger sample may be required.

The level of confidence can also be calculated corresponding to a given size of sample and fixed $L$. For $n = 30$ and $L = 1.2$, we get $1.2 = z_{1-\alpha/2} \times 6.3/\sqrt{30}$, or, $z_{1-\alpha/2} = 1.04$. From the Gaussian distribution, get $\alpha/2 = 0.1492$. Thus, $\alpha = 0.30$ and the confidence level is only 70%. When $n$ is 30, there is only 70% chance that the estimated mean would be less than 1.2 mg/dL away from the actual mean in the population. This chance may be too low to proceed with a study on such a small sample.

As mentioned earlier, the sample size formulas follow a similar general pattern but differ in detail in different situations. We are devoting separate sections in this book on the sample size determination in different setups.

### *Sample Size for Testing a Hypothesis with Specified Power*

The primary aim of analytical studies is to investigate antecedent–outcome relationship. Although other setups are possible, the most common is to find whether two groups are different or not with respect to an antecedent factor in the case of retrospective studies, and with respect to an outcome in the case of prospective studies. Statistically, this leads to testing of hypothesis setup.

The method for determining sample size in the testing-of-hypothesis setup depends on the actual criterion used for testing. Various criteria are discussed under different topics in this volume. At this stage, it may be stated that the sample size calculation requires the following information in most situations encountered in practice. These are mostly the same as in the estimation setup, but there are some changes.

1–4. Same as in the estimation setup.

5a. How much minimum *difference* ($\delta$) between the actual value of the parameter and its value under the null hypothesis is **medically important**? This is the test of hypothesis counterpart of "degree of precision" in the estimation setup. A medically important difference is specified under the alternative hypothesis $H_1$. A small difference is difficult to detect, and a large sample would be required for a

small difference to be statistically significant. If the minimum difference to be detected were large, a small sample would be adequate. This difference is determined on clinical considerations and not on statistical considerations. Do not confuse it with actual effect size, which could be more or could be less than the minimum medically important effect.

5b. What is the statistical **power** required? This is the probability of detecting a specified difference when present and calling it statistically significant. The notation for this is $(1 - \beta)$. Power depends on the magnitude of the difference specified in (5a) above. If the power is to be 99% for the specified difference, obviously a bigger sample is required than for power of 80%. Power and $n$ have a direct relationship for any fixed level of significance.

6a. What **level of significance** is required? In the testing-of-hypothesis setup, the complement of confidence level is the level of significance $\alpha$. On the analogy stated for CI, a large sample is required if $\alpha$ is to be kept small. If a large $\alpha$ can be tolerated, a relatively small sample would be enough.

6b. Is the test a **one-tailed test or a two-tailed test**? A one-tailed test with $\alpha = 0.05$ is equivalent to a two-tailed test with $\alpha = 0.10$ in most situations. A high $\alpha$-level in a two-tailed setup is a relatively small $\alpha$-level in a one-tailed setup. This is true for ***P*-values** also. Thus, one-tailed testing requires a smaller sample size compared to what a two-tailed setup requires. If $n = 300$ gives $P = 0.03$ in a two-tailed test, the same $n$ would generally give $P = 0.015$ in a one-tailed test. For one-tailed $P = 0.03$, the required $n$ would be smaller than for a two-tailed $P = 0.03$.

7–12. Same as in the estimation setup.

In the case of large $n$, when Gaussian conditions prevail, these considerations lead to an equation of the following type in most situations for a two-sided test for detecting a difference of at least $\delta$ between the values under the null and alternative hypotheses (if present) with at least $100(1 - \beta)\%$ power:

$$z_{1-\alpha/2} * \mathrm{SE}_0 + z_{1-\beta} * SE_1 < \delta,$$

where $\mathrm{SE}_0$ and $\mathrm{SE}_1$ are the standard errors of the estimate of this parameter under the null and alternative hypotheses, respectively. If these two SEs are the same (this would be so when the variances in the two groups are the same and the sample sizes also are the same), the relation can be approximated as $(z_{1-\alpha/2} + z_{1-\beta}) * \mathrm{SE} < \delta$, or in variance terms, as $(z_{1-\alpha/2} + z_{1-\beta})^2 * \mathrm{Var(estimate)} < \delta^2$. The sample size $n$ would occur in the expression of the SE. This can be solved to obtain $n$ when everything else is specified. This takes a different form in different setups; see the topic **sample sizes for statistical analyses**, for example. A detailed and unified account of sample size calculations is given by Hanley and Moodie [2].

### *Some Comments*

It is ironic that sample size calculations require the value of the SD of the observations even before the sample is studied. As advised, this may be anticipated from a value reported in a previous study or from the estimation arrived at from a **pilot study**. There would be variations that are best resolved by discussion with a biostatistician. Sometimes it may involve guesswork, and the sample size so calculated will be approximate. To be safe, it is better to inflate this $n$ slightly to compensate for possible error in the estimate of the SD. When SD from previous studies is not available and the range is available instead, a conservative guess for SD for calculation of sample size is range/4, particularly if the underlying distribution is nearly Gaussian. For proportion $\pi$, a conservative estimate is 0.5 that will give a higher sample size than any other value of $\pi$.

The other, possibly more valid, option is to calculate sample size for a range of values of parameters whose values are uncertain, and choose the one that looks feasible without compromising the scientific requirement of the study. This is easy these days as repeated sample size calculations can be done with many readily available online calculators.

For sampling methods other than SRS, the SEs of those sampling methods should be used for these calculations. This, for example, in a two-stage sample, can raise questions on balancing the number of primary units for first-stage sampling and the number of subjects to be selected at the second stage. The calculations can become complex, and the information required for computing the size can also become difficult to manage. Besides the sampling method, remember that whatever sample size you come up with after calculations would need upward revision for uncertainty in the values used in the calculations, nonresponse, number of subgroups, design of the study, etc. If the nonresponse is anticipated to be completely at random in $100*p\%$ of subjects, the adjusted sample size is $n/(1 - p)$. For example, if the calculation reveals that $n = 200$ and the expected nonresponse is 15%, the required $n$ is $20/(1 - 0.15) = 236$. After 15% nonresponse, the final sample would be 200.

The formulas and procedures generally available for determining the sample size are valid for large samples where Gaussian approximation is valid. With advances in computer technology, it is now possible to use exact methods for small samples. However, these methods are still sketchy and available for a restricted class of situations. For example, Rahme and Joseph [3] worked out exact methods for determining the sample size for binary outcomes, Royston [4] published tables of sample sizes for pair-matched case–control studies, and Hilton and Mehta [5] devised an algorithm for ordered categorical data for small samples.

An additional consideration not accounted for in the usual formulas is the number of concomitant variables or covariates that are proposed to be studied together. The higher this number, the bigger the requirement of the sample size. The sample size quickly multiplies and can become an enormous number if cross-classifications of a number of factors are under consideration as already illustrated. Results would be reliable only if each of these classifications has an adequate number of subjects. No clear guidelines are available about the sample size that would be adequate to take care of, say, five risk factors as opposed to two risk factors, but a rule of thumb is that each cross-classification should have a sample of at least 30 subjects. This is applicable when the disease prevalence is large, say ≥20%. A size of at least 30 implies for four cross-classified binary factors that there must be at least a total of $16 \times 30 = 480$ subjects, evenly distributed to the 16 categories. All these comments are summarized in Table S.1.

The sample size many times depends on the resources and time available for a project. For a master's thesis, only 1 year is generally available with practically no funding. Thus, the sample size would be small regardless of what the formulas say. Our advice in such situations is to choose a topic for which an adequate number of subjects are available within 1 year, and restrict to the investigations that are feasible. Perhaps the master's thesis should not be done on rare diseases. Else, be clear that it would be a pilot study in nature whose results would not have much reliability but can provide important clues to plan a major study.

**TABLE S.1**
**Requirement of Larger Sample**

| Requirement | Sample Size |
|---|---|
| Smaller Type I error | Larger |
| Smaller Type II error (higher statistical power) | Larger |
| Smaller difference to be detected | Larger |
| Higher inter-individual variability or smaller proportion | Larger |
| Higher anticipated nonresponse | Larger |
| Higher number of subgroups | Larger |
| Higher number of variables to be considered together | Larger |

Even for big studies, resources and time constraints sometimes dictate the sample size. Vickers [6] gives this interesting example: A colleague points out that the drug under trial is safe and inexpensive, and could be advocated if it is able to reduce average pain score by even half a point. A recent paper showed an SD = 2 for the change, for which the formula gives $n$ = 774. Fund limitation may not allow one to do such a big trial. What if SD = 1.5? The sample size reduces to 380, but it is still high. Lower the bar and aim at detecting pain score reduction by 0.75 instead of half a point. Now $n$ = 170, which is doable. If this is chosen, the size of the study is dictating the research objective, whereas in reality research objective should dictate the sample size. In practice, this can happen and is accepted.

In situations where a large sample is not feasible, a small-scale study is better than no study at all with two precautions. First, a study based on a smaller-than-desired sample is pilot in nature that may not give conclusive results. Second, a separate class of statistical procedures such as nonparametric and permutation tests are sometimes used to analyze data based on small samples. These methods require extensive calculations—much more than the usual methods—and separate software packages are used for this purpose. Generally, only very large difference would turn out to be statistically significant in case of small samples, except for experiments on animals because of highly standardized conditions in the laboratories.

### *Nomograms and Tables of Sample Size*

Many medical researchers find sample size formulas difficult to adopt. Most statistical software packages incorporate tools to calculate sample size based on one's requirement. Two other alternatives are available, although they may not be readily available. One is **nomogram**—a graph containing lines and curves, and needs only a ruler to read the sample size that would meet the specifications. Altman [7] gave a nomogram for the size of SRS for comparing two groups, and Kumar and Indrayan [8] developed one such nomogram for reading the sample size required in cluster sampling (see the topic **nomogram**). Neter and Wasserman [9] gave nomograms for sample sizes for analysis of variance situations where three or more groups are compared for means. If you are weary of the formulas and the relevant software is not available, try to locate a nomogram for your situation. Nomogram is especially helpful to find sample sizes for several scenarios and to choose the scenario that is considered most suitable for the problem in hand. It can be easily used and reused, and can be carried in your pocket when going to field areas if necessary.

The second alternative is a table of sample sizes. Some authors have worked out sample sizes for various situations and tabulated them such as what Lwanga and Lemeshow [10] did for commonly occurring simple situations. The sample size required for specified values can be directly read from such tables. An interpolation may be required when the desired exact specification is not present in the table.

### *Rules of Thumb*

Sometimes approximate calculations are done for simplicity. For example, for power of 80%, $1 - \beta = 0.80$ and $z_{1-\beta} = 0.84$, and for $\alpha = 0.05$, and two-sided hypothesis, $z_{1-\alpha/2} = 1.96$. Thus, the formula for detecting a mean difference of $\delta$ in the two groups, under the usual regularity conditions, becomes

$$n = \frac{(1.96 + 0.84)^2 * \sigma^2}{\delta^2} \approx 8 \frac{\sigma^2}{\delta^2}.$$

For power = 90%, this becomes $\approx 11\sigma^2/\delta^2$. These are easy to remember and can be used for 80% or 90% power and two-sided level of significance 0.05 without being too far off. These values of power and the level of significance seem to occur commonly in medical research studies. Just take the ratio of variance to the minimum difference to be detected and multiply by 8 to get the sample size per group [11].

In addition, there are rules of thumb. They lack scientific basis and many scientists dislike them. When no baseline information for computation of sample size is available and the constraints do not permit pilot study either, the following rules of thumb can be used.

A large-sized medical trial should include nearly 300 subjects in *each* group, a midsized trial nearly 100 in each group, and a small-sized trial at least 30 in each group. The last can be used for master's theses where time and resources are limited. A bigger study is multicentric with these numbers in each center. Same norms can be used for a retrospective or case–control study. However, in the case of a prospective study, the number to be followed up should be such that at least 30 persons are finally available with the outcome of interest in *each group*. This applies to field trials also. In this case, an extremely large group may be needed to yield an outcome such as hepatitis B infection in at least 30 subjects after administration of a protective vaccine. To calculate exact numbers, use the exact formulas. A study of these formulas would indicate that a laboratory experiment on animals could be done on a smaller number because of fairly standard and controlled conditions in a laboratory that minimize variations.

For a descriptive study that seeks to find normal levels in healthy subjects, the rule of thumb is to include at least 200 subjects in each group for which norms are required, although in this case the exact number can also be calculated using an appropriate formula. For pathological levels in patients, the group size could be smaller. Depending upon the targeted reliability of the results, exact sample size requirement can be calculated. However, all these may have to be modified because of feasibility considerations when resources and time are limited. Such limitations obviously compromise the reliability.

1. Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P. Reporting of sample size calculation in randomised controlled trials: Review. *BMJ* 2009 May 12;338:b1732. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2680945/
2. Hanley JA, Moodie EEM. Sample size, precision and power calculations: A unified approach. *J Biomet Biostat* 2011;2:124. http://www.omicsonline.org/2155-6180/2155-6180-2-124.pdf

3. Rahme E, Joseph L. Exact sample size determination for binomial experiments. *J Stat Plan Inference* 1998;66:83–93. http://www.medicine.mcgill.ca/epidemiology/joseph/publications/methodological/binexact.pdf
4. Royston P. Exact conditional and unconditional sample size for pair-matched studies with binary outcome: A practical guide. *Stat Med* 1993 Apr 15;12(7):699–712. http://www.ncbi.nlm.nih.gov/pubmed/8511446
5. Hilton JF, Mehta CR. Power and sample size calculations for exact conditional tests with ordered categorical data. *Biometrics* 1993;49(2):609–16. http://www.jstor.org/stable/2532573?seq=1#page_scan_tab_contents
6. Vickers AJ. *Let's Dance! The Sample Size Samba*. Medscape Multispecialty. http://www.medscape.com/viewarticle/584026
7. Altman DG. *Practical Statistics for Medical Research*. Chapman and Hall, 1991: p. 456.
8. Kumar R, Indrayan A. A nomogram for single-stage cluster sample surveys in a community for estimation of a prevalence rate. *Int J Epidemiol* 2002; 31:463–7. http://ije.oxfordjournals.org/content/31/2/463.full
9. Neter J, Wasserman W. *Applied Linear Statistical Models: Regression, Analysis of Variance and Experimental Designs*. Richard D. Irwin, Inc., 1974: pp. 827–8.
10. Lwanga SK, Lemeshow S. *Sample Size Determination in Health Studies: A Practical Manual*. Geneva: World Health Organization, 1991. http://apps.who.int/iris/handle/10665/40062
11. Lehr R. Sixteen *S*-squared over *D*-squared: A relation for crude sample size estimates. *Stat Med* 1992;11(8):1099–1102. http://onlinelibrary.wiley.com/doi/10.1002/sim.4780110811/abstract


## sample size for odds ratio and relative risk, *see also* sample sizes for study formats, sample sizes for statistical analysis methods, sample sizes for simple situations (mean, proportion, and differences)

Odds ratio (OR) in case–control studies and relative risk (RR) in prospective studies are among the most common parameters studied in medical context. In view of their special importance, we devote this full section to the sample size determination for estimating and testing of hypotheses of OR and RR.

You may want to review the topic **sample size determination (general principles)** for an overview of the basics of sample size determination. The procedure and the requirements for estimation (**confidence intervals**) are different from those for **testing of hypotheses**. If the objective is estimation, then it is necessary to specify the confidence level ($1 - \alpha$) and the precision desired on either side of RR or OR. This is generally specified as a proportion of RR or OR. The anticipated RR or OR or the anticipated probability of disease (or of exposure) in the respective groups should also be specified. If the objective is to test a hypothesis, then the specification of the relative precision is not required, and the **significance level** is required in place of the confidence level. Specification of the statistical power ($1 - \beta$) is needed for a testing-of-hypothesis setup, and for this, the difference to be detected must also be specified.

While $\alpha$ and $\beta$ are mostly determined externally, it is important to select suitable values for effect size for these calculations. For this, review the literature fully and locate a study in a similar setting so that the estimates are plausible for your study. The choice of the minimal effect size to be detected should be reasonable and suitable such that it is practically meaningful, and conducting the study is feasible. The factors time, cost, and recruitment of subjects should all be considered together with the power and minimal difference. Also consider the statistical procedure to be used—for example, a different procedure may be needed if a logistic regression is planned rather than a simple $2 \times 2$ table analysis. In addition, consider the ethical and financial cost also for finalizing the sample size.

The formulas are given in Table S.2 for large-scale studies as they are based on **Gaussian conditions**. They require the anticipated probabilities in the two groups under comparison. If only one of these probabilities can be reasonably anticipated, the anticipated value of RR (or of OR) would be needed. In that case, the other probability can be calculated by using the definition of RR or of OR. The notation now is $\pi_1$ and $\pi_0$ for disease probability in the exposed and unexposed groups, respectively, and $\pi_1'$ and $\pi_0'$ for exposure probabilities in the case and control groups, respectively. In terms of these probabilities, RR = $\pi_1/\pi_0$ and OR = $\dfrac{\pi_1'/(1-\pi_1')}{\pi_0'/(1-\pi_0')}$.

An astute reader would note that formula (b) in Table S.2 for test of hypothesis on RR uses $\overline{\pi} = (\pi_1 + \pi_0)/2$, but such an average is not used in formula (d) for OR. Instead, only $\pi_0'$, which is the exposure rate among the controls, is used. Quite often, this rate is reliably known in the case–control setup, and there is no need to use the average. If you are not fully confident about $\pi_0'$, the first part in formula (d) should be $\sqrt{2\pi(1-\pi)}$ just as in formula (b).

The formulas in Table S.2 are restricted to the following: (i) two independent samples (not matched pairs), (ii) a large $n$ so that the Gaussian pattern is applicable, and (iii) for OR, a small proportion of the subjects have disease. Relative precision $\varepsilon$ is on either side of RR or OR, and the alternative hypothesis is RR $\neq 1$ or OR $\neq 1$. Both these inferences are two-sided. For the one-sided alternative, replace $z_{1-\alpha/2}$ by $z_{1-\alpha}$. The formulas for matched pairs are mentioned later in this section. The following examples illustrate the calculations.

Consider a clinical trial in which a new regimen is compared with an existing regimen for their role in control of nausea while treating depression. The objective is to estimate the RR of nausea in depression cases. The anticipated values are as follows:

Proportion with nausea in the patients on new regimen $\pi_1 = 0.40$
Proportion with nausea in the patients on existing regimen $\pi_0 = 0.50$
Relative precision desired, $\varepsilon = 0.10$ (i.e., estimated RR should be within 10% of its actual value)
Confidence level 95%, i.e., $\alpha = 0.05$

Thus, $z_{1-\alpha/2} = 1.96$, and, from formula (a) in Table S.2,

$$n = \frac{(1.96)^2}{[\ln(1-0.10)]^2}\left[\frac{1-0.40}{0.40} + \frac{1-0.50}{0.50}\right] = 865$$

with upward approximation.

This is the number of subjects required for each of the regimens. If the relative precision is relaxed to $\varepsilon = 0.20$, then $n = 193$. The sample size requirement declines steeply when the precision requirement is relaxed.

As another example, consider the plan of a case–control study of people with beta-thalassemia major to evaluate their thyroid status. This is a rare disease. Thyroid status is divided into only normal and abnormal categories but under very stringent criteria. The objective is to assess whether there is any association between thyroid status and beta-thalassemia. Thus, this is a testing-of-hypothesis setup. Suppose the anticipated values are as follows:

Proportion with abnormal thyroid among cases, $\pi_1' = 0.25$
Proportion with abnormal thyroid among controls, $\pi_0' = 0.10$
[Thus, anticipated OR is (0.25/0.75)/(0.10/0.90) = 3.]

S

**TABLE S.2**
**Sample Size Required for Two-Sided Inference on Relative Risk (RR) and Odds Ratio (OR)**

| Problem | Formula for Computing $n$ (Large $n$) per Group | Explanation of the Notations |
|---|---|---|
| | **Relative Risk** | |
| (a) Estimating RR | $\frac{z^2_{1-\alpha/2}}{[\ln(1-\varepsilon)]^2}\left[\frac{1-\pi_1}{\pi_1}+\frac{1-\pi_0}{\pi_0}\right]$ | $\pi_1, \pi_0$ = anticipated probability of disease among the exposed and nonexposed subjects, respectively<br>$\varepsilon$ = relative precision in terms of proportion of RR |
| (b) Hypothesis testing for RR = 1<br>Equal exposed and unexposed groups | Or, $\frac{1}{(\pi_0-\pi_1)^2}\left[z_{1-\alpha/2}\sqrt{2\bar{\pi}(1-\bar{\pi})}+z_{1-\beta}\sqrt{\pi_1(1-\pi_1)+\pi_0(1-\pi_0)}\right]^2$<br>$\frac{\left[z_{1-\alpha/2}\sqrt{2\bar{\pi}(1-\bar{\pi})}+z_{1-\beta}\sqrt{\pi_0(1+r)-\pi_0^2(1+r^2)}\right]^2}{[\pi_0(1-r)]^2}$ | $\bar{\pi}=(\pi_1+\pi_0)/2$<br>$\pi_1, \pi_0$ as above<br>Medically relevant least RR is $r=\pi_1/\pi_0$ |
| | **Odds Ratio (Rare Disease)** | |
| (c) Estimating OR | $\frac{z^2_{1-\alpha/2}}{[\ln(1-\varepsilon)]^2}\left[\frac{1}{\pi_1'(1-\pi_1')}+\frac{1}{\pi_1'(1-\pi_0')}\right]$ | $\pi_1'$ = anticipated probability of exposure among the cases<br>$\pi_0'$ = anticipated probability of exposure among the controls<br>$\varepsilon$ = relative precision in terms of proportion of OR |
| (d) Hypothesis testing for OR = 1<br>–One control per case | $\frac{1}{(\pi_1'-\pi_0')^2}\left[z_{1-\alpha/2}\sqrt{2\pi_0'(1-\pi_0')}+z_{1-\beta}\sqrt{\pi_1'(1-\pi_1')+\pi_0'(1-\pi_0')}\right]^2$ | $\pi_1', \pi_0'$ as above<br>Medically relevant least OR is $[\pi_1'/(1-\pi_1')]/[\pi_0'/(1-\pi_0')]$ |
| –C controls per case | $\frac{1}{(\pi_1'-\pi_0')^2}\left[z_{1-\alpha/2}\sqrt{\left(1+\frac{1}{C}\right)\pi'(1-\pi')}+z_{1-\beta}\sqrt{\pi_1'(1-\pi_1')+\pi_0'(1-\pi_0')}\right]^2$ | $\pi'=\frac{\pi_1'+C\pi_0'}{C+1}$; $\pi_1', \pi_0'$ as above |

*Note:* Two independent samples; same $n$ for each group.
$\alpha$ is the significance level and $(1-\beta)$ is the statistical power; $z_a$ is such that $P(Z \geq z_a) = a$. For $\alpha = 0.05$, $z_{\alpha/2} = 1.96$; for $\beta = 0.05$, $z_\beta = 1.645$; for $\beta = 0.10$, $z_\beta = 1.28$; for $\beta = 0.20$, $z_\beta = 0.84$.

With significance level $\alpha = 0.05$ and power $(1-\beta) = 0.80$, $z_{1-\alpha/2} = 1.96$ and $z_{1-\beta} = 0.84$, and we get from formula (d) in Table S.2 that

$$n = \frac{1}{(0.25-0.10)^2} \times \left(1.96\sqrt{2\times 0.10\times 0.90}+0.84\sqrt{0.25\times .0.75+0.10\times 0.90}\right)^2 = \frac{1}{0.0225}(0.8316+0.4425)^2 = 73$$

with upward approximation.

Thus, only 73 cases and 73 controls are required for this study. The second formula in (d) in Table S.2 for $C$ controls per case can be approximated as

$$n \approx \frac{(1+1/C)\bar{\pi}(1-\bar{\pi})(z_{1-\alpha}+z_{1-\beta})^2}{\delta^2}.$$

For $C = 1$, this becomes $n \approx \frac{2\bar{\pi}(1-\bar{\pi})(z_{1-\alpha/2}+z_{1-\beta})^2}{\delta^2}$.

With $z_{1-\alpha/2} = 1.96$ for 5% level of significance for a two-tailed test and $z_{1-\beta} = 0.84$ for $C = 1$, the formula can be further approximated as

$$n \approx 8\bar{\pi}(1-\bar{\pi})/\delta^2 \text{ for power} = 80\%$$

and

$$n \approx 11\bar{\pi}(1-\bar{\pi})/\delta^2 \text{ for power} = 90\%.$$

These are simple to remember.

If OR is specified in place of $\pi_1$, the exposure rate in the case group can be obtained as $\pi_1 = \frac{\pi_0(OR)}{1+\pi_0(OR-1)}$. For example, for $\pi_0 = 0.30$ and OR = 3, $\pi_1 = (3\times 0.30)/[1+0.30(3-1)] = 0.5625$. Going further, these values give $\bar{\pi}$ = ½(0.30 + 0.5625) = 0.43125, and for $\alpha = 0.05$, power = 0.90, and from formula (d) of Table S.2,

$$n = \frac{\left[1.96\sqrt{2\times 0.43125\times 0.56875}+1.28\sqrt{0.30\times 0.70+0.5625\times 0.4375}\right]^2}{(0.30-0.5625)^2} = 73.$$

With an approximate formula, for $\delta = 0.5625 - 0.30 = 0.2625$, this becomes $n = (2\times 0.43125\times 0.56875)(1.96+1.28)^2/(0.2625)^2 = 75$, and further an approximate formula for 90% power gives $n = (21\times 0.43125\times 0.56875)/(0.2625)^2 = 75$, which is not much different from 73 arrived at earlier from a more exact formula.

Consider a study on maternal anemia in low birth weight infants versus in normal weight infants. Suppose the anticipated prevalence of maternal anemia in normal birth weight group is 5% ($\pi_0 = 0.05$) and in low birth weight group is OR = 3.0. This is the "exposure" of interest in this example. How big a sample size is required to get a width of 1.5 of 95% CI for OR? A width of CI is the double of $\delta$,

i.e., $\delta = 1.5/2 = 0.75$. With these values, we get $\pi_1 = (0.05 \times 3.0)/[1 + 0.05(3.0 - 1)] = 0.136$, and formula (c) in Table S.2 gives, for $\varepsilon = 0.5$,

$$n = \left(\frac{1.96}{\ln(1-0.5)}\right)^2 \left(\frac{1}{0.136\times 0.864} + \frac{1}{0.05\times 0.95}\right) = 236.38.$$

Thus, 237 cases and similar number of controls are required for this study in a one control per case setup.

For one-to-one **matched** (pair-matched) studies, the sample size is obtained in terms of the number of discordant pairs. This is because only the discordant pairs contribute to the decision in such studies. According to Schlesselman [1], the number of discordant pairs required to detect minimum OR of $\lambda$ with a probability (power) of $(1 - \beta)$ and level of significance $\alpha$ is

$$m = \frac{\left[\frac{1}{2} z_{1-\alpha/2} + z_{1-\beta}\sqrt{\pi(1-\pi)}\right]^2}{(\pi - 1/2)^2},$$

where $\pi = \frac{\lambda}{1+\lambda}$ (under the null, $\lambda = 1$ and thus $\pi = ½$). One strategy to achieve this number could be to go on recruiting subjects until you get these many discordant pairs. This is not a good strategy since no effective planning can be made. To plan, see if you can have some estimate of the proportion of pairs with discordance. If this is $p$, the total number of pairs required is $n = m/p$. Otherwise, an approximation that requires independence of the exposure in the cases and controls is $p \approx \pi_0(1 - \pi_1) + \pi_1(1 - \pi_0)$ [1]. For example, to detect an OR = 2 (this gives $\pi = 1/3$ for $\lambda = 2$) with a power of 90% and two-tailed significance level 5%, the number of discordant pairs required is $m = [1.96/2 + 1.28 \times \sqrt{(⅓ \times ⅔)}]^2/(⅓ - ½)^2 = 91$. If control group exposure $\pi_0 = 0.30$ and OR = 2.0, $\pi_1 = (0.30 \times 2.0)/[1 + 0.30(2.0 - 1)] = 0.4615$, and $p \approx 0.30 \times 0.5385 + 0.4615 \times 0.70 = 0.4846$, assuming independence. These values give $n = 91/0.4846 = 188$. Thus, a total of 188 pairs are required in this study. When the exposure in the case group is not independent of the control group, adjustment suggested by Fleiss and Levin [2] can be used.


1. Schlesselman JJ. *Case-Control Studies: Design, Conduct, Analysis*. Oxford, 1982: pp. 145–61.
2. Fleiss JL, Levin B. Sample size determination in studies with matched pairs. *J Clin Epidemiol* 1988;41(8):727–30. http://www.ncbi.nlm.nih.gov/pubmed/3418361


## sample size for simple situations (mean, proportion, and differences), *see also* sample sizes for study formats, sample sizes for statistical analysis methods, sample sizes for odds ratio and relative risk

The sample size formulas for simple situations are given in Table S.3 for estimation and in Table S.4 for testing of hypothesis. These are for mean, proportion, and their difference in two populations, and do not include formulas for ratios such as odds ratio and relative risk (for OR and RR, see **sample size for odds ratio and relative risk**). These formulas are based on the procedure explained under the topic **sample size determination (general principles)**. The notations are also explained in the tables, and the formulas are valid only for situations in which Gaussian approximation is applicable. The parameter values such as $\pi$ and $\sigma$ that appear in these formulas have to be replaced by their anticipated values that could be either those previously reported, based on a pilot study, or just educated

**TABLE S.3**
**Sample Size Calculation for Simple Estimations (Valid for Large *n* Only)***

| Problem | Formula for Computing *n* | Description of the Notations |
|---|---|---|
| (a) Population proportion with specified absolute precision | $\frac{z^2_{1-\alpha/2}\pi(1-\pi)}{L^2}$ | $\pi$ = Anticipated value of the proportion in the population<br>$L$ = Absolute precision required on either side of the proportion |
| (b) Population proportion with specified relative precision | $\frac{z^2_{1-\alpha/2}\pi(1-\pi)}{(\varepsilon\pi)^2}$ | $\pi$ = Anticipated value of the proportion in the population<br>$\varepsilon$ = Relative precision in terms of fraction |
| (c) Population mean with specified precision | $\frac{z^2_{1-\alpha/2}\sigma^2}{L^2}$ | $\sigma$ = Population SD (can be estimated from a pilot study)<br>$L$ = Specified precision of the estimate on either side of the mean |
| (d) Difference between two population proportions with specified absolute precision—equal $n$ in the two groups | $\frac{z^2_{1-\alpha/2}[\pi_1(1-\pi_1)+\pi_2(1-\pi_2)]}{L^2}$ | $\pi_1$, $\pi_2$ = Anticipated proportions in the two populations<br>$L$ = Absolute precision required on either side of the difference in proportions |
| (e) Difference between means of two populations with specified precision—equal $n$ in the two groups | $\frac{z^2_{1-\alpha/2}\left(\sigma_1^2+\sigma_2^2\right)}{L^2}$ | $\sigma_1$, $\sigma_2$ = Population SD of the two populations (can be estimated from a pilot study)<br>$L$ = Specified precision of the estimated difference on either side of the mean difference |

*Source:* Lwanga SK and Lemeshow S. *Sample Size Determination in Health Studies: A Practical Manual*. World Health Organization, 1991.

*Note:* $1 - \alpha$ is the level of confidence; $z_{1-\alpha}$ is such that $P(Z \le z_{1-\alpha}) = 1 - \alpha$.

For $\alpha = 0.05$, $z_{1-\alpha/2} = 1.96$, and $z_{1-\alpha} = 1.645$. For power = 0.90, $z_{1-\beta} = 1.28$; for power = 0.80, $z_{1-\beta} = 0.84$. For other values, consult Gaussian distribution. If the alternative hypothesis is one-sided, replace $z_{1-\alpha/2}$ by $z_{1-\alpha}$.

*Large $n$ is needed so that the distribution of $p$ can be approximated by Gaussian form, and so that the distribution of sample mean is Gaussian if the distribution of values is not Gaussian.

S

**TABLE S.4**
**Sample Size Calculation for Simple Testing of Hypothesis Situations (Valid for Large *n* Only)*—Two-Sided $H_1$**

| Problem | Formula for Computing *n* per Group | Description of the Notations |
|---|---|---|
| (a) For a population proportion | $\frac{\left[z_{1-\alpha/2}\sqrt{\pi_0(1-\pi_0)}+z_{1-\beta}\sqrt{\pi_a(1-\pi_a)}\right]^2}{\delta^2}$ | $\pi_0$ = Value of $\pi$ under $H_0$<br>$\pi_a$ = Medically important value of population proportion under $H_1$, that is, $\delta = (\pi_0 - \pi_a)$ is the difference proposed to be detected |
| (b) For a population mean | $\frac{\sigma^2\left(z_{1-\alpha/2}+z_{1-\beta}\right)^2}{\delta^2}$ | $\sigma$ = Population SD (can be estimated from a pilot study).<br>$\delta$ = Minimum medically important difference between means under $H_1$ that is proposed to be detected |
| (c) For difference between two population proportions—equal *n* in the two groups | $\frac{\left[z_{1-\alpha/2}\sqrt{2\bar{\pi}(1-\bar{\pi})}+z_{1-\beta}\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}\right]^2}{\delta^2}$ | $\pi_1$, $\pi_2$ = Anticipated proportions in the two populations, that is, $\delta = (\pi_1 - \pi_2)$ is the difference proposed to be detected.<br>$\bar{\pi} = (\pi_1 + \pi_2)/2$, $H_0$: $\pi_1 = \pi_2$ |
| (d) For difference between two population means—equal *n* in the two groups | $\frac{(\sigma_1 z_{1-\alpha/2}+\sigma_2 z_{1-\beta})^2}{\delta^2}$ | $\sigma_1$, $\sigma_2$ = Population SD of the two populations (can be estimated from a pilot study)<br>They will be equal in most situations<br>$\delta$ = Minimum medically important difference between means under $H_1$ that is proposed to be detected |
| (e) For independence in matched pairs | $\frac{\left[z_{1-\alpha/2}\sqrt{\pi_{10}+\pi_{01}}+z_{1-\beta}\sqrt{\pi_{10}+\pi_{01}-(\pi_{10}-\pi_{01})^2}\right]^2}{\delta^2}$ | $H_0$: $\pi_{10} = \pi_{01}$<br>$\pi_{10}$ and $\pi_{01}$ are the discordant probabilities<br>$\delta = \pi_{10} - \pi_{01}$ is the difference proposed to be detected |

*Note:* $\alpha$ is the level of significance and $(1 - \beta)$ is the statistical power; $z_{1-\alpha}$ is such that $P(Z \le z_{1-\alpha}) = 1 - \alpha$.
For $\alpha = 0.05$, $z_{1-\alpha/2} = 1.96$, and $z_{1-\alpha} = 1.645$. For power = 0.90, $z_{1-\beta} = 1.28$; for power = 0.80, $z_{1-\beta} = 0.84$. For other values, consult Gaussian distribution. If the alternative hypothesis is one-sided, replace $z_{1-\alpha/2}$ by $z_{1-\alpha}$.
*Large *n* is needed so that the distribution of *p* can be approximated by Gaussian form, and so that the distribution of sample mean is Gaussian if the distribution of values is not Gaussian.

guesses, as always done in sample size calculation. All formulas assume simple random sampling and a univariate setup.

In many medical situations, the standard deviations (SDs) in the two groups under comparison are equal, that is, $\sigma_1 = \sigma_2$. If the SD in one group is very different from the SD in the other group, sample sizes are optimal if $n_1/n_2 = \sigma_1/\sigma_2$. This helps in minimizing the standard error (SE) of difference in means. For the sample size for estimating proportions with specified precision, if educated guess of $\pi$ is not available, use $\pi = 0.5$, which will give the maximum *n*. The following example illustrates the calculation.

To plan a study on the difference in the prevalence of filariasis in agricultural and nonagricultural workers in an endemic area, a pilot study was carried out with 30 workers of each type. The prevalences, respectively, were 33% and 20%. What sample size is needed if the difference is to be estimated within 3 percentage points with 90% confidence?

This problem is for difference between proportions and would require formula (d) of Table S.3. The calculation requires anticipation that the proportions in the population would be nearly the same as those in the pilot study. Thus, $\pi_1 = 0.33$ and $\pi_2 = 0.20$. Confidence $100(1 - \alpha) = 90\%$ or $\alpha = 0.10$. Since $L = 0.03$, from formula (d) in Table S.3, we get

$$n = \frac{z_{0.05}^2(0.33\times 0.67 + 0.20\times 0.80)}{(0.03)^2}$$
$$= \frac{(1.645)^2 \times (0.3811)}{(0.03)^2}$$
$$= 1146.$$

Thus, a sample of nearly 1150 is required in each group of workers.

The previous example was on estimation. Now consider an example on testing of hypothesis. Suppose a decline of at least 10 mg/dL in triglyceride level (TGL) after a therapy is considered clinically important. It is proposed that a group of nonvegetarian obese and nonobese subjects will be kept on vegetarian diet to see if their TGL declines by this clinically important magnitude. The SD of their TGL is anticipated to be 15.7 and 12.5 mg/dL, respectively. The researcher wishes to detect a 10 mg/dL difference with probability 0.80. What sample size should be chosen if the level of significance is 0.10?

In this case, $\delta = 10$ mg/dL, $\sigma_1 = 15.7$, and $\sigma_2 = 12.5$ (all are estimates, though). The interest is only in decline, so $H_1$ is one sided. (In this case, a rise cannot be ruled out, but keep that aside for this example.) Now, $\alpha = 0.10$ and $(1 - \beta) = 0.80$ give $z_{1-\alpha} = 1.28$ and $z_{1-\beta} = 0.84$. With these values, formula (d) in Table S.4 gives

$$n = \frac{(15.7\times 1.28 + 12.25\times 0.84)^2}{10^2}$$
$$= (20.8 + 10.5)^2/100$$
$$= 10 \text{ when approximated upward.}$$

A sample of 10 obese and 10 nonobese subjects is enough for this study. A relatively small sample is required because the difference to be detected, 10 mg/dL, is quite large. If this is 5 mg/dL, then this gives $n = 40$ in each group. Conversely, if resources permit $n = 100$, then, for $\delta = 5$,

$$100 = \frac{\left(15.7 \times 1.28 + 12.5 \times z_{1-\beta}\right)^2}{5^2},$$

or

$$50 = 20.8 + 12.5 \times z_{1-\beta},$$

or $z_{1-\beta} = 2.384$ and the Gaussian distribution gives $(1 - \beta) = 0.9914$. Thus, a size of $n = 100$ will have a probability of more than 99% of detecting a difference of 5 mg/dL if present.

In case of **sensitivity and specificity**, which are also proportions, obtain the sample size as usual for proportions and divide by the prevalence rate of disease for sensitivity and by (1 – prevalence) for specificity. For more details, see the topic **sample sizes for statistical analyses**.

## sample sizes for statistical analysis methods, *see also* sample sizes for odds ratio and relative risk, sample sizes for simple situations (mean, proportion, and differences), sample sizes for study formats

For basic information on sample size determination, see the topic **sample size determination (general principles)**. This is a prerequisite to fully understand the current section. In this section, we present methods for determining sample size for different statistical analysis setups, namely, ANOVA, associations, correlations, hazard ratios and survival analysis, regressions (logistic and ordinary quantitative regression), and sensitivity–specificity.

### *Sample Size for One-Way ANOVA*

**Analysis of variance (ANOVA)** method is basically used to compare means of three or more groups. Because of multiple groups in this setup, the sample size determination is not as straightforward as in the case of two groups. Although planning of sample size in this setup can be approached in terms of controlling the width of desired confidence intervals (for estimation), we restrict our discussion in this section to **power** approach for testing of hypotheses as this is the dominant reason for using ANOVA. Also, for simplicity, we restrict to **one-way ANOVA** that incidentally also illustrates the kind of issues that arise in this context.

Let us consider that all factor levels are equally important and the sample sizes accordingly are also planned to be equal. If the major interest is in pairwise comparisons of factor level means, it is known that equal sample sizes maximize the power of these comparisons for a fixed combined sample size. Equal sizes are also helpful in ignoring slight departure from **homoscedasticity**. However, we do not deny that unequal sample sizes may be appropriate in some situations such as for comparing the effect of several different doses with that of a control—a situation where a larger sample in the control group is of definite help. Only that our presentation becomes simple if we restrict to the equal sample size situation, and could be understood well by medical professionals.

As always, for determining the sample size for testing of hypothesis situation, you need to specify (i) the significance level, generally fixed at 0.05; (ii) the minimum **medically important effect** as difference between group means that you would want to detect; and (iii) the **power** with which this difference is to be detected if present. Other considerations listed under the topic **sample size determination (general principles)** will continue to have a role as described therein.

Suppose there are a total of $K$ groups in one-way ANOVA setup, and the total number of pairwise comparisons is $K(K - 1)/2$. It is not necessary that all pairwise groups are of our interest—let $H$ of them be of our interest. The significance level for calculation of sample size may be taken as $\alpha/H$ as per the **Bonferroni procedure**, where $H$ is the number of comparisons. Then the sample size required in each group under **Gaussian conditions** for a two-tailed test is

$$n = \frac{(\sigma^2)(z_{1-\alpha/(2H)} + z_{1-\beta})^2}{\delta^2},$$

where the $z$-value comes from the Gaussian distribution corresponding to the prefixed level of significance $\alpha$, $(1 - \beta)$ is the prefixed power to be able to detect a difference of $\delta$ in the means, and $\sigma^2$ is the anticipated common variance that can be estimated by the **mean square error** (MSE) in ANOVA in any earlier study. This is valid when (i) the effects are fixed and not random; (ii) the null hypothesis is that means in all groups are equal and the alternative hypothesis is that at least one mean is different; (iii) all the underlying requirements of ANOVA are fulfilled such as independence of values, homoscedasticity, and Gaussian conditions; and (iv) each subject in each group is either randomly selected or randomly allocated to the groups. Statisticians may advise suitable adjustments in case of other sampling such as stratified or cluster is adopted, and in case the Gaussian conditions do not hold. The effect size for this formula is in terms of mean difference, which is the most straight way of specifying the effect size to be detected in this setup. If the effect size is specified in some other manner such as in terms of the ratio of between and within variance, the formula will change accordingly. Many software and online packages are available that calculate sample size when the effect size is specified in some other manner.

A major problem in adopting this formula, as in all sample size formulas, is obtaining the values of the $\sigma$'s on one hand, and determining the maximum clinically unimportant (or minimum medically important) difference $\delta$ on the other. For the former, the usual practice is to use values reported in earlier studies carried out in similar setups, or to conduct a pilot study and use the results if nothing is available in the literature. To be safe, it is better to inflate this $n$ slightly to compensate for possible error in the guessed value of $\sigma$. The other, possibly more valid, option is to calculate sample size for a range of values of parameters whose values are uncertain, and choose the one that looks feasible without compromising the scientific requirement of the study. This is easy these days as repeated sample size calculations can be done with many readily available online calculators.

For the clinically important difference $\delta$, you need to garner your experience, brainstorm with your colleagues, and consult literature to come up with either a maximum effect that can be considered clinically unimportant or a minimum effect that is clinically important to convince others to change their practice and use your results.

In some rare situations, the alternative hypothesis is not that at least one mean unequal but is that they follow a specified order. This is sometimes termed as *constrained hypothesis*, and the sample size for such alternatives is discussed by Vanbrabant et al. [1]; they have also provided tables of sample sizes for different powers and different alternatives.

### *Sample Size for Associations*

Association between binary attributes in two independent samples is assessed by calculating chi-square: by calculating the odds ratio in a case–control setup, by calculating the relative risk in a prospective setup, and by calculating either in a cross-sectional setup when the

**TABLE S.5**
**Probabilities in a Matched Pairs Setup**

| First Partner of the Pair | Second Partner of the Pair: Positive | Second Partner of the Pair: Negative | Total |
|---|---|---|---|
| Positive | $\pi_{11}$ | $\pi_{10}$ | $\boldsymbol{\pi_{1\cdot}}$ |
| Negative | $\pi_{01}$ | $\pi_{00}$ | $\boldsymbol{\pi_{0\cdot}}$ |
| **Total** | $\boldsymbol{\pi_{\cdot 1}}$ | $\boldsymbol{\pi_{\cdot 0}}$ | **1** |

sample drawn is random. For these, the method given under the topic **sample sizes for odds ratio and relative risk** may be used. For a paired binary setup (e.g., matched case–controls, crossover trials) where the **McNemar test** is used, the number of pairs required is

$$n = \left( \frac{z_{1-\alpha/2}\sqrt{\pi_{10}+\pi_{01}} + z_{1-\beta}\sqrt{\left(\pi_{10}+\pi_{01}\right) - \left(\pi_{10}-\pi_{01}\right)^2}}{\pi_{10}-\pi_{01}} \right),$$

where $\pi$'s are the anticipated probabilities as given in Table S.5 and the null hypothesis under test is $H_0$: $\pi_{10} = \pi_{01}$ (i.e., the probability of discordance of both the types is the same) with two-sided alternative $H_1$: $\pi_{10} \neq \pi_{01}$.

### *Sample Size for Correlations*

Correlation is obtained for two quantitative measurements linked in some manner such as they belong to the same subjects (e.g., correlation between systolic and diastolic blood pressure measured for 75 subjects), to the same time point (e.g., between neonatal and maternal mortality rates in a country from 1985 to 2015), to the same area (e.g., between incidence of cancer and incidence of heart disease in 30 cities), etc. The property that **Fisher *z* transformation** ($z = \frac{1}{2} \ln[(1 + r)/(1 - r)]$) of the product–moment **correlation coefficient** $r$ follows a nearly Gaussian (normal) distribution for large samples can be used to calculate sample size. For significance level $\alpha$ and power $(1 - \beta)$ to detect a correlation of at least $\rho$ with power $(1 - \beta)$, this is given by

$$n = \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{c^2} + 3,$$

where $c$ is the Fisher $z$ transformation of the minimum correlation $\rho$ to be detected. (We are using notation $c$ instead of $z$ to avoid confusion with Gaussian probabilities $z_{1-\alpha/2}$ and $z_{1-\beta}$.) This sample size has a high chance to detect the specified correlation if present. However, the correlation coefficient, just as most other estimates, is unstable for a small sample size, and Schönbrodt and Perugini [2] found by Monte Carlo simulations that it stabilizes when the sample size is nearly 250. You can guess from this how big the sample size requirement can be in some situations.

For comparison of two correlations, the required sample size is given by

$$n = \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\left(c_0 - c_1\right)^2} + 3,$$

where $c_0$ and $c_1$ are the Fisher $z$ transformation of the anticipated correlations $\rho_0$ and $\rho_1$ in the two groups, which means that the minimum difference in correlations to be detected is $\rho_1 - \rho_0$. Practically all the parameter values are replaced by the corresponding sample values.

For obtaining the confidence interval on the correlation coefficient with specified precision, Moinester and Gottfried [3] have provided a procedure and a table of sample sizes.

### *Sample Size for Hazards Ratios and Survival Analysis*

**Hazard ratio** (HR) is the ratio of hazard of occurrence of an event at a particular point in time in the two groups under comparison. Since HR = 1 implies the same hazard in the two groups, the statistical significance is tested against HR = 1 (this is the same as ln(HR) = 0). The required sample size for detecting a particular hazard ratio $\theta$ when present with power $(1 - \beta)$ is obtained in this case by the number of events in place of the total sample. This is given by

$$\text{number of events} : d = \left( \frac{(\theta+1) * \left(z_{1-\alpha/2} + z_{1-\beta}\right)}{\theta - 1} \right)^2,$$

where $\theta$ = minimum hazard ratio to be detected if present, and the level of significance is $\alpha$. For detecting hazard ratio $\theta = 1.5$ with 90% power at 5% two-sided level of significance, this gives $d = 263$. If the event under study is death, you need as big a sample as to have at least 263 deaths. This may turn out to be an enormous number; for example, if 10% are anticipated deaths, a sample of 2630 subjects is needed.

An approximate but simplified and more popular version is

$$d = \frac{4\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{[\ln \theta]^2}.$$

With this formula, to detect HR = 1.5 with power 90% at a two-sided significance level of $\alpha = 0.05$,

$$d = \frac{4(1.96 + 1.282)^2}{[\ln(1.5)]^2} = 256,$$

which is slightly different from the more exact $d = 263$ arrived at earlier. According to this formula, the number of events required to detect different HRs with power 80% and 90% is as given in Table S.6.

Table S.6 shows that the sample size requirement decreases as the HR to detect increases. Thus, one of the ways to reduce the requirement of sample size is to increase the follow-up time that will raise the HR, but this strategy will work when the follow-up is not expensive relative to a study on a larger sample.

**TABLE S.6**
**Number of Events Required for Detecting Specified HR and Power**

| HR | Power: 80% | Power: 90% |
|---|---|---|
| 1.5 | 191 | 256 |
| 2.0 | 66 | 88 |
| 2.5 | 38 | 50 |
| 3.0 | 26 | 35 |

For comparing hazards in two populations,

$$n = \frac{1}{p_A p_B \pi}\left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\ln\theta_1 - \ln\theta_0}\right)^2,$$

where $p_A$ and $p_B$ are the proportions of sample size allotted to the two groups ($p_A + p_B = 1$), $\pi$ is the anticipated overall probability of the event occurring within the study period, and $\theta_1$ and $\theta_0$ are the anticipated HRs in the case group and the control group, respectively. If samples from both the groups are equal, $p_A = ½$ and $p_B = ½$, then

$$n = \frac{4}{\pi}\left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\ln\theta_1 - \ln\theta_0}\right)^2.$$

The $\pi$ in the denominator takes care of the prevalence, and thus the sample size is total $n$ and not the number of events.

In addition to hazards, two other kinds of similar analyses are done in survival studies. The first is simple comparison of survival rate in two groups at a fixed point in time when there are no censored values. This could be like a 5-year survival rate in cancer patients. The sample size for this can be calculated by the usual formula for comparing proportions as given under the topic **sample size for simple situations**. The second is comparing the whole survival pattern by log-rank test where the null hypothesis is $S_1(t) = S_2(t)$ for all $t$, where $t$ is for time point and $S_k(t)$ is the probability of survival in the $k$th group till time $t$. This is the same as $H_1(t) = H_2(t)$ for all $t$, where $H$ is for hazard rate, or the hazard ratio $\theta = H_1(t)/H_2(t) = 1$, or $\ln\theta = 0$. This situation is the same as just discussed and the sample size formula is also the same, which is in terms of the number of events. For survival analyses, the number of events such as deaths and recoveries is important, and the sample size depends on the anticipated rate of occurrence of such events.

In case of a **Cox regression**, where the effect of regressors on the outcome is studied, Schoenfeld [4] showed that the sample size requirement remains the same whether or not an account is taken of such regressors.

### *Sample Size for Logistic and Ordinary Linear Regressions*

Very few biostatistics textbooks deal with sample size considerations for estimation and tests of hypothesis for a regression coefficient (slope) in a simple or multiple ordinary or logistic regression. **Logistic regression** case is even more complicated owing to nonlinearity nature of the logistic function. The formulas are available for relatively simple cases, and they are mostly approximate and valid for large samples only.

In a simple logistic regression model, if the covariate $x$ is continuous with a Gaussian distribution, the sample size formula for detecting the logistic coefficient $\beta_0$ as given by Hsieh et al. [5] is

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\pi(1-\pi)\beta_0^2},$$

where $\pi$ is the anticipated event rate at the mean of $x$. When the covariate $x$ is binary, this becomes

$$n = \frac{\pi(1-\pi)(z_{1-\alpha/2} + z_{1-\beta})^2}{B(1-B)(\pi_1 - \pi_2)},$$

where $B$ is the proportion of the population with $x = 1$, $\pi_1$ is the event rate at $x = 0$, $\pi_2$ is the event rate at $x = 1$, and $\pi$ is the overall event rate. Suppose in an underdeveloped country, the incidence of low birth weight births is 15%, i.e., $B = 0.15$. As many as 60% mothers of low birth weight babies have anemia, i.e., $\pi_1 = 0.60$. The health administration decides that a program of iron supplementation for pregnant women can be launched if it is able to bring this down to 40% or less. Thus, $\pi_2 = 0.40$. How big a sample should be chosen to check that this decline is detected with power 0.90 and significance level 0.025 (one-sided since supplementation program cannot increase the anemia level)? The average prevalence of anemia in this case is $\pi = (0.60 + 0.40)/2 = 0.50$. Substituting these values in the formula just mentioned gives

$$n = \frac{0.5(1-0.5)(1.96+1.28^2}{0.15(1-0.15)(0.60-0.40)} = 103.$$

A field trial with a sample of 103 pregnant women would give a convincing evidence that the supplement is able to bring down the anemia level to 40% or less in this population.

For case–control studies with continuous $x$, Lubin et al. [6] derived the following for $C$ controls per case:

$$n = \frac{C+1}{C} * \frac{\left[z_{1-\alpha}\sigma_x + z_{1-\beta}\sqrt{\frac{C\sigma_1^2 + \sigma_0^2}{C+1}}\right]^2}{\delta^2}.$$

The total sample size is $Cn + n$; $\sigma_1^2$ and $\sigma_0^2$ are the variances of the variable $x$ in the case and control groups, respectively; and $\sigma_x^2$ is the variance under the null hypothesis. For equal samples, just plug in $C = 1$.

For an ordinary simple linear regression, when both $x$ and $y$ are standardized (mean 0 and variance 1 by subtracting mean and dividing by the standard deviation), the hypothesis $\beta = 0$ is the same as the correlation coefficient $\rho = 0$; thus, the sample size requirement is the same, and the formula stated earlier for correlation coefficient is valid for simple linear regression too.

For multiple regression (logistic or ordinary), multiply these sample sizes by the **variance inflation factor** (VIF) = $1/(1 - \eta^2)$, where $\eta$ is the **coefficient of determination**. This can be replaced with the **multiple correlation** coefficient $R$ or its equivalent for linear regression. Same VIF applies to both the setups [5]. As always with sample size formulas, the parameter values are replaced by the anticipated values as estimated from previous studies.

### *Sample Sizes for Sensitivity and Specificity*

**Sensitivity and specificity** measure inherent validity of a diagnostic test against a gold standard. An adequate sample size is necessary to precisely estimate the validity of a diagnostic test and to test statistical significance. For their estimation with specified precision, the formulas are the same as for proportions but have to be adjusted for prevalence of the disease. The formulas when both diagnostic test and gold standard have dichotomous categories are as follows:

$$\text{For estimating sensitivity, } n = \frac{z_{1-\alpha/2}^2 * S_N(1-S_N)}{L^2 * \text{Prevalence}},$$

$$\text{and for estimating specificity, } n = \frac{z_{1-\alpha/2}^2 * S_P(1-S_P)}{L^2 * (1-\text{Prevalence})},$$

where

$S_N$ = anticipated sensitivity
$S_P$ = anticipated specificity
$1 - \alpha$ is the confidence level
$L$ = absolute precision desired on either side (half-width of the confidence interval) of sensitivity or specificity

If the prevalence of disease you are investigating is 12%, divide the usual sample size by 0.12 to obtain the sample size for estimating sensitivity, and divide by 0.88 to obtain the sample size for estimating specificity. This almost ensures that the requisite number of disease positives and negatives, respectively, are available in the sample.

Malhotra and Indrayan [7] have devised a simple nomogram that yields a statistically valid sample size for anticipated sensitivity or specificity. This nomogram could be easily used to determine the sample size for estimating the sensitivity or specificity of a diagnostic test with required precision and 95% confidence level. A nomogram instantly provides the required number of subjects by just moving the ruler and can be repeatedly used without redoing the calculations; it can also be applied for reverse calculations to find the confidence level corresponding to the available sample size, but it is not applicable for testing of the hypothesis setup.


1. Vanbrabant L, Van De Schoot R, Rosseel Y. Constrained statistical inference: Sample-size tables for ANOVA and regression. *Front Psychol* 2015 Jan 13;5:1565. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4292225/
2. Schönbrodt FD, Perugini M. At what sample size do correlations stabilize? *J Res Personality* 2013;47: 609–12. http://www.psy.lmu.de/allg2/download/schoenbrodt/pub/stable_correlations.pdf
3. Moinester M, Gottfried R. Sample size estimation for correlations with pre-specified confidence interval. *Quant Methods Psychol* 2014;10(2):124–30. http://www.tqmp.org/RegularArticles/vol10-2/p124/p124.pdf
4. Schoenfeld DA. Sample size formula for the proportional hazards regression model. *Biometrics* 1983;39:499–503. http://links.jstor.org/sici?sici=0006-341X%28198306%2939%3A2%3C499%3ASFFTPR%3E2.0.CO%3B2-2
5. Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. *Stat Med* 1998 Jul 30;17(14):1623–34. http://personal.health.usf.edu/ywu/logistic.pdf
6. Lubin JH, Gail MH, Ershow AG. Sample size and power for case-control studies when exposure is continuous. *Stat Med* 1988 Mar;7(3): 363–76. http://onlinelibrary.wiley.com/doi/10.1002/sim.4780070302/abstract
7. Malhotra RK, Indrayan A. A simple nomogram for sample size for estimating sensitivity and specificity of medical tests. *Indian J Ophthalmol* 2010; 58:519–22. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2993983/


## sample sizes for study formats, *see also* sample sizes for odds ratio and relative risk, sample sizes for simple situations (mean, proportion, and differences), sample sizes for statistical analysis methods

Whereas sample size is primarily determined on reliability and power considerations as discussed under the topic **sample size determination (general principles)**, the study formats also play a role in providing it a context, and modifications in some situations can be made as needed. This section is devoted to the relevant features of various formats such as case–control studies, clinical trials, cross-sectional studies, surveys, medical experiments, and prospective studies, as applicable to sample size determination. Some essential aspects of sampling methods used in each setup are also included.

### *Sample Size for Case–Control Studies*

Although random sampling is not essential for analytical studies such as case–control, it too involves confidence interval estimation and tests of hypotheses on parameters such as odds ratio (OR) and prevalence rate of various antecedents. Thus, a reasonable sample size is required for these studies also. A **nested case–control study** may involve stratification, and each stratum must have an adequate number of subjects for any reliable conclusion. Control group is a necessary ingredient in case–control setup, and controls are sometimes oversamples (such as 1:2 ratio) in situations where they are easy to obtain and elicit.

Whereas broad methods of determining sample size in case–control studies remain the same, some new issues may come up. For example, in a case–control study on risk factors for breast cancer, the interest could be in age at menarche, age at first live birth, total duration of breast-feeding of all children put together, and the number of first-order blood relatives that have positive history. Cancer prevalence would be different in various categories of different variables. If there are many variables such as in this example, which one should be used for determining the sample size? If the focus is on only one of them, say, age at menarche, and others are just concomitants to remove their confounding effect, use the cancer prevalence rate in different categories of this variable only. If two or more variables have nearly the same importance, calculate the sample size for each of them separately and take the largest. Since there are two groups, case and control in this setup, OR may be a more appropriate parameter of interest. In any case, mean, proportion, OR, etc., and their standard errors, as required for calculation of sample size, would have to be mostly based on a previous study in a similar setting, and the precision in case of estimation, or the medically relevant least difference and power in case of testing of hypothesis, must be specified.

For one control per case, the sample size formula remains the same as for the OR (see Table S.2 in **sample size for odds ratio and relative risk**), where relevant proportions in case–control studies are the proportion of subjects found to have the antecedent of interest (exposure) in the cases and the controls, respectively. The formula for $C$ controls per case is also given at the same place. For continuous exposure where logistic regression is used, Lubin et al. [1] derived the following formula for sample size.

For $C$ controls per case (exposure a continuous variable):

$$n = \frac{C+1}{C} * \frac{\left[ z_{1-\alpha}\sigma_x + z_{1-\beta}\sqrt{\frac{C\sigma_1^2 + \sigma_0^2}{C+1}} \right]^2}{\delta^2},$$

where $\sigma_1^2$ and $\sigma_0^2$ are the variances of the exposure variable $x$ in the case and control groups, respectively, and $\sigma_x^2$ is the variance under the null. The total sample size in this case is $Cn + n$. For one control per case, just plug in $C = 1$ and get the sample size.

### *Sample Size for Clinical Trials*

Sample size is just about the first thing that comes to mind when planning a clinical trial, and it is also among the most important considerations that determine its utility. Since trials are conducted

on human subjects, they require extreme care in all aspects of execution including choosing a sample size. Ethical requirements of clinical trials are fairly well strict and allow neither a larger sample than required because it could mean unnecessary exposure to an unproven treatment, nor a smaller sample as that also could be unnecessary exposure as the results are not likely to be conclusive. In the latter case, a promising regimen that could have provided relief to many patients is missed and subjects needlessly exposed to a trial that was not properly planned. Besides risk of exposure, clinical trials are expensive; thus, they need to be conducted with precision, and a proper sample size has added importance.

In view of more importance of right sample size for clinical trials, let us reiterate that a large *n* could mean a waste of resources—if a study of only 200 subjects can give a reliable answer, why spend resources to study 250? A small sample, on the other hand, may not give evidence one way or the other, and the trial may fail to achieve its objectives. Thus, this also is a waste of resources. An unduly large sample is unethical in trials because it means that some subjects are unnecessarily exposed to an intervention whose utility is in doubt. An unduly small sample, too, is unethical because then the subjects are unnecessarily subjected to a trial of an unproven intervention that is not going to yield a result one way or the other. Having said that, the number of subjects should be reasonably large in each group so that the full clinical spectrum is represented and a trend, if present, can clearly emerge; also the reliability of the results is ensured. It should have adequate statistical **power** to not miss a minimum medically relevant difference when present.

Failure of many **randomized controlled trials** (RCTs) in detecting a medically relevant difference because of insufficient **power** is a matter of concern for medical community. During the 1970s, Freiman et al. [2] observed after studying 71 negative RCTs on new therapeutic procedures that the sample size was too small in most of them for power of 90% to detect a 25% improvement in outcome. In nearly three-fourths of these trials, the sample was inadequate to detect a 50% improvement. In other words, even if there is a 50% improvement in efficacy of the new treatment compared with the old, the trial results were still statistically not significant. It would have most likely become significant had there been a bigger trial. A quarter of a century later, Dimick et al. [3] reported the same for surgical trials. These once again underscore the need to be careful about the size of a trial—the number of subjects must be adequate to inspire confidence that a medically relevant difference would not go undetected.

In addition, a small sample has a high likelihood of being not fully representative and thus of producing biased results. Statistical methods have an in-built provision to take care of the larger sampling errors in small samples, but they are not equipped to take care of the lack-of-representativeness bias that is more likely to creep in small samples.

Sample size becomes especially critical for rare outcomes. With thorough aseptic conditions now maintained in operation theaters, chance of infection is negligible in routine surgeries such as for hernia. However, antibiotics are still given in some settings, albeit for a short duration. If a trial on the effect of antibiotics is done on 100 cases in a test group and another 100 controls, the trial is doomed from the start since no group may develop infection, or just one or two infections may occur in one group. This cannot lead to any reliable conclusion. All of this needs to be taken into account when calculating the sample size in this sort of scenario.

**Logistic regression** can be the key statistical method for analyzing of an RCT. There is a greater need to correctly perform the sample size calculation for this setup. In a logistic regression with *K* independents, conservatives generally advocate that at least 10**K* subjects should be observed for the rarest outcome in the analysis. If $K = 4$, and mortality is the outcome of interest, there must be at least 40 deaths in your sample according to this rule. If mortality is 8%, the total sample must be 500. However, this looks too strict for many situations as the experience suggests that the results are fairly reliable if no category has frequency less than 5.

The primary aim of most clinical trials is testing of hypothesis and not obtaining a confidence interval for the effect size. Thus, the sample size also is mostly calculated to achieve a prespecified power to detect a prespecified effect at a prespecified level of significance (see the topic **sample size determination [general principles]**). For two groups such as test and control, and a two-sided test for quantitative outcome, the sample size in each group is given by the formula provided in Table S.4 in **sample sizes for simple situations (mean, proportion, and differences)**.

Many clinical trials focus on more than one outcome. In this situation, the sample size is either calculated on the basis of the most important outcome if that can be identified, or calculated for all the outcomes and maximum taken. However, in case of multiple outcomes, the chance of false statistical significance increases. If the standard deviations (SDs) of the two groups are very different, it is advisable to choose the samples such that $n_0/n_1 = \sigma_0/\sigma_1$, where $n_0$ is the sample size in the control group and $n_1$ in the treatment group and so are the σ's. If there are more than one active treatment and all the treatments are to be compared with control, the power is maximized by a larger sample of control subjects.

If the response is binary in place of quantitative such as adequate relief/no adequate relief and survived/died, the formula based on proportions is the same as given in Table S.4 in **sample sizes for simple situations (mean, proportion, and differences)**. The formula for means of continuous responses is also given in the same table. For **equivalence trials**, the formulas are slightly modified as follows. For quantitative outcome with Gaussian conditions, if the equivalence margin is from $-\Delta$ to $+\Delta$,

$$n = \frac{(\sigma_0 z_{1-\alpha/2} + \sigma_1 z_{1-\beta/2})^2}{(\Delta - \delta)^2},$$

where δ is the expected difference in means, generally set as $\delta = 0$. For proportions, this becomes

$$n = \frac{\left[z_{1-\alpha/2}\sqrt{2\bar{\pi}(1-\bar{\pi})} + z_{1-\beta/2}\sqrt{\pi_1(1-\pi_1)+\pi_0(1-\pi_0)}\right]^2}{(\Delta - \delta)^2},$$

where $\delta = \pi_1 - \pi_0$. Note how the denominator changes in case of equivalence trials, and β becomes β/2 because **two one-sided tests (TOSTs)** are used for this setup. For details, see Lin [4]. If the outcome is duration, the formulas mentioned for survival analysis would apply.

### *Sample Size for Cross-Sectional Studies*

The term *cross-sectional studies* is used for those observational studies that elicit information on the antecedent and the outcome together on a sample of subjects, which is generally but not always randomly drawn from the target population. In contrast, note that in the case of prospective studies, the samples of the exposed and nonexposed are chosen separately, and in the case of retrospective studies, samples are chosen from the diseased and nondiseased subjects.

This kind of preclassification is not done for cross-sectional studies, although the purpose remains to investigate the relationship between the antecedents and the outcomes.

Validity of conclusions regarding the association of two or more factors depends on their proper representation in the sample in proportion of their presence in the target population. Random sampling serves as a great facilitator to achieve such representation. Assessment of other parameters such as sensitivity–specificity and predictivities also is adversely affected if these proportions are biased in any manner. Thus, random sampling is especially important for cross-sectional studies.

Sample size determination in cross-sectional studies does not pose any additional issues, and all the formulas mentioned in Table S.2 under the topic **sample size for odds ratio and relative risk** can be used depending on whether the objective is to test the hypothesis on OR or on relative risk (RR). All these sample sizes are valid for simple random sampling; suitable adjustments may be needed if the sampling method is different.

### *Sample Size for Descriptive Studies and Surveys*

The primary objective in **descriptive studies** and surveys is not investigating associations but finding the best estimates of the prevalence of the conditions of interest in different segments of a population. Investigation of association lends them to analytical framework, and the study is no longer strictly descriptive. Associations might be investigated in descriptive studies on post-hoc basis to gather additional information for future studies, but the planning, including sample size calculation, is done for estimation. Thus, the procedure and the formulas for determining the sample size for estimations apply, and the desired precision of the estimates is specified in this case in place of the medically important effect and the power needed for calculation of sample size for testing of hypothesis setup. Random sampling is doubly important in descriptive studies so that a sample representing different cross sections of the population is obtained.

Sample size for estimating the parameters such as mean and proportion depends primarily on variance of the estimate and the precision required. This may be intuitively appealing to those who have quantitative thinking—a larger sample is required for smaller margin of error. Depending upon the parameter of interest, any of the formulas given in Table S.3 in **sample sizes for simple situations (mean, proportion, and differences)** can be used to calculate sample size.

A sample size calculation for estimating a proportion requires some explanation. This could be the proportion of subjects of a particular type or a chance of occurrence of a particular event. Two kinds of formulas are available for this parameter: one uses absolute precision, and the other relative precision. When absolute precision is used, the sample size is maximum for anticipated $\pi = 0.50$, where $\pi$ is the proportion in the target population. Absolute precision is specified, for example, as ±5% or ±0.05, which for $\pi = 0.50$ implies that the sample providing an estimate between 0.45 and 0.55 is acceptable. If $\pi = 0.06$, precision ±0.05 gives a range of 0.01–0.11. Relative to such a small value of $\pi$, this range is too large. If prevalence of tuberculosis in a particular population is 6% and the sample estimate is anywhere between 1% and 11%, practically nothing is achieved. Such low precision can be a result of a small sample. Thus, precision for a proportion should be stated in terms of a relative value such as 10% of $\pi$. Ten percent of $\pi = 0.50$ is 0.05 and that of $\pi = 0.06$ is 0.006. A much larger sample size is required for such precision when the anticipated is $\pi$ small.

### *Sample Size for Medical Experiments*

Experiments on animals such as mice are done on small numbers per group, yet they provide reliable results. One might wonder why the statistical formulas do not apply to this setup. There are two reasons: (i) experimental animals are much more homogenous such as those that are of the same strain (Sprague–Dawley, Wistar, Donru, etc.)—thus inter-individual variability is small, and (ii) laboratory conditions are fairly well standardized, and the factors that can influence the results are under good control. Thus, any difference found between groups can be legitimately ascribed to the treatment. For this reason, experiments on five to six mice per group are norm than exception. This cannot be said about clinical trials or observational studies. When data from experiments are analyzed, if the data do not follow a Gaussian pattern, exact or nonparametric methods may have to be used because of a small sample per group.

Laboratory experiments on animals do not raise much concern about sampling methods because the experimental animals of one particular species are generally chosen for this purpose, and they can be easily considered as representative of their "population." However, whenever factors such as age, gender, and weight can affect findings, either the animals should be stratified for independent experiment on each stratum, or only one particular stratum of animals should be investigated. For example, to study the effect of middle turbination resection on facial growth of rabbits, the animals must be of the same age. For telomerase activity, which is implicated in all immortalization and carcinogenesis, age and gender of rats are important because they affect the level of this activity. Also, in this case, the strain of rats can affect the findings. Thus, separate experiments should be done in samples of rats of different strains.

Experiments on biological material such as blood specimens, vaginal swabs, and antigens are also common. If each specimen is identified and linked with a known person, the situation is back to sampling of individuals. For these, the same methods as described for humans should be followed. For experiments on anonymous unlinked biological specimen, think about the possible sources of bias that can inhibit generalization to a larger group. The specimen must represent the full spectrum of material in the target population. Whenever a large number of specimens are available, a random sample should be chosen, although this is rarely done in practice. Because an experiment necessarily involves an intervention, baseline equivalence of the test and control group is generally considered sufficient for validity of the results with a rider that the results are valid under laboratory conditions. Randomness of sampling in such situations has a limited role. Instead, **random allocation** is required.



### *Sample Size for Prospective Studies*

Prospective studies are those observational analytical studies that explore subjects with and without antecedent for developing an outcome of interest with the objective of finding how, if at all, the incidence (risk) of the outcome is affected by the presence or absence of the antecedent. The statistical parameter of interest in most such studies is the **relative risk** of developing the outcome in those with antecedent compared with those without the antecedent. It is easier to talk about the exposed and the nonexposed, and disease and no disease, in place of the antecedent and the outcome. We will come to repeated measures later in this section; for the time being, let the interest be in incidence of the disease at the end of a specified period of follow-up.

For simplicity, which incidentally is also the most common setup, consider a dichotomous disease with present and absent categories. The RR in this case is simply $\pi_1/\pi_0$, where $\pi_1$ is the incidence of

disease in exposed and $\pi_0$ is the incidence among nonexposed. Any such ratio is relatively easy to handle when a logarithm is taken since then $\ln(RR) = \ln \pi_1 - \ln \pi_0$, which has a linear from. Thus, **central limit theorem** can be applied for large $n$, and the Gaussian (normal) distribution can be used for calculation of the sample size. With this background, you may feel comfortable using the formulas given for sample size for the OR and RR. As always, these would require a preliminary guess of the incidences from previous studies, specification of the required precision for estimation, and specification of the medically important RR and associated power for calculation of sample size.

For an example, consider a prospective study of workers in the United States who file claims for work-related musculoskeletal disorders reported by Turner et al. [6]. The primary outcome of interest was duration of work disability in 1 year after filing the claim. The purpose was to develop statistical models that could predict the duration of chronic work disability after the initial suffering from the disorder. The required sample size was calculated as 1800 workers for low back injuries and 1200 for workers with carpal tunnel syndrome (CTS). Values of statistical power used for these calculations were 0.96 for low back and 0.85 for CTS, and the significance level chosen was $\alpha = 0.05$ (two-tailed) for both. The example illustrates that the sample size could be quite large for prospective studies when power and significance level are considered. The concept of power is related to the minimum medically relevant difference, but the abstract of this article does not mention this difference. A large sample size was feasible in this case because the study is mostly based on administrative database and follow-up interviews were conducted over telephone.

For comparison of two groups with respect to the mean of the quantitative outcome, the required sample size in prospective studies is

$$n = \left( \frac{z_{1-\alpha}\sigma_0 + z_{1-\beta}\sigma_1}{\mu_1 - \mu_0} \right)^2 \text{ for one-tailed test}$$

(replace $\alpha$ by $\alpha/2$ for two-tailed test),

where $\alpha$ is the level of significance, $(1 - \beta)$ is the statistical power, $\sigma_0$ and $\sigma_1$ are the respective SDs in the groups under comparison, and the minimum difference to be detected is $\delta = \mu_1 - \mu_0$. If the minimum detectable difference is to be obtained for specific $n$, this is given by

$$\delta = \mu_1 - \mu_0 = \frac{z_{1-\alpha}\sigma_0 + z_{1-\beta}\sigma_1}{\sqrt{n}}.$$

For estimating the incidence rate with $100*(1 - \alpha)\%$ confidence in one group,

$$n = \frac{z^2_{1-\alpha/2}}{\varepsilon^2},$$

where $\varepsilon$ = the required relative precision as a fraction of the anticipated incidence. For testing of hypothesis of equality of incidences in the two groups,

$$n = \frac{\left(z_{1-\alpha/2}\pi_0 + z_{1-\beta}\pi_1\right)^2}{(\pi_0 - \pi_1)^2},$$

where $\pi_0$ = incidence rate (in terms of proportion) under the null hypothesis, and $\pi_a$ = incidence rate in terms of proportion under the alternative hypothesis (the minimum difference to be detected is $\pi_0 - \pi_1$).

In many prospective studies, the follow-up period is not uniform and the incidence per unit of time is calculated after working out person-time (person-years, person-weeks, etc.). This is valid only when the risk of disease at, say, the 15th week is the same as that at the 2nd week of follow-up. Let the subjects either have a common date of entry into the study and are followed up until they develop the outcome or censored, or let them be included in the study in order of their arrival but followed up only until a specified duration. When this is done, the sample size formula for testing the significance of difference in risks (called **attributable risk**) at $\alpha$-level of significance, according to Lwanga and Lemeshow [7], is

$$n = \frac{\left(z_{1-\alpha/2}\sqrt{2f(\bar{\pi})} + z_{1-\beta}\sqrt{f(\pi_0) + f(\pi_1)}\right)^2}{(\pi_0 - \pi_1)^2},$$

where $\pi_1$ and $\pi_0$ are the incidences per unit time in the two groups in the sense that $\pi_1 - \pi_0$ is the minimum medically important difference to be detected with power $(1 - \beta)$; $\bar{\pi} = (\pi_1 + \pi_0)/2$; $f(\pi) = \pi^3 T/(\pi T - 1 + e^{-\pi T})$ if the duration of the study is fixed as $T$ (censored observations) (for $f(\bar{\pi})$, $f(\pi_1)$, and $f(\pi_0)$, replace $\pi$ by $\bar{\pi}$, $\pi_1$, and $\pi_0$, respectively, in these expressions); and $f(\pi) = \pi^2$ if $T$ is not fixed.

**Repeated measures study** requires consideration of the correlation among repeated measures while planning the sample size. If the average difference over time in continuous measure is relevant for calculation of sample size, Hedeker et al. [8] give the following formula for quantitative outcomes:

$$n = \left[ \frac{(z_{1-\alpha} + z_{1-\beta})\sigma}{\delta} \right]^2 * \frac{\left[1 + (K-1)\rho\right]}{K},$$

where $K$ is the number of time points for which repeated measures were taken, $\rho$ is the correlation coefficient between values obtained by repeated measurement (assumed the same for all time points and equal in both the groups), $\sigma^2$ is the variance of the response values (also assumed the same for all time points), and $\delta$ is the time-average difference proposed to be detected. This says that the requirement of sample size will increase as $\rho$ increases when positive, as is likely the case in practically all situations of repeated measures. For example, if $K = 5$ time points, $\sigma = 80$, $\rho = 0.5$, $\alpha = 0.05$, power = 0.90 to detect $\delta = 20$ point average difference, the required sample size is $n = 101$ per group.

For dichotomous outcomes, the formula under the usual notations becomes

$$n = \left[ \frac{z_{1-\alpha/2}\sqrt{2\bar{\pi}(1-\bar{\pi})} + z_{1-\beta}\sqrt{\pi_1(1-\pi_1) + \pi_0(1-\pi_0)}}{\pi_0 - \pi_1} \right]^2 * \frac{\left[1 + (K-1)\rho\right]}{K}.$$


1. Lubin JH, Gail MH, Ershow AG. Sample size and power for case–control studies when exposure is continuous. *Stat Med* 1980 Mar;7(3):363–76. http://onlinelibrary.wiley.com/doi/10.1002/sim.4780070302/abstract
2. Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error, and sample size in the design and interpretation of the randomized control trial: A survey of 71 "negative" trials. *N Engl J Med* 1978;299:690–4. http://www.nejm.org/doi/full/10.1056/NEJM197809282991304

3. Dimick JB, Diener-West M, Lipsett PA. Negative results of randomized clinical trials published in the surgical literature: Equivalency or error? *Arch Surg* 2001; 136:796–800. http://archsurg.jamanetwork.com/article.aspx?articleid=391750, last accessed October 12, 2015.
4. Lin SC. Sample size for therapeutic equivalence based on confidence interval. *Drug Inf J* 1995;29:45–50. http://dij.sagepub.com/content/29/1/45.full.pdf+html
5. Schnitzer TJ, Kong SX, Mitchell JH et al. An observational, retrospective cohort study of dosing pattern of rofecoxib and celecoxib in the treatment of arthritis. *Clin Ther* 2003;25:3162–72. http://www.ncbi.nlm.nih.gov/pubmed/14749154
6. Turner JA, Franklin G, Fulton-Kehoe D, Egan K, Wickizer TM, Lymp JF, Sheppard L, Kaufman JD. Prediction of chronic disability in work-related musculoskeletal disorders: A prospective population based study. *BMC Musculoskeletal Disord* 2004;5:14. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC428578/
7. Lwanga SK and Lemeshow S. *Sample Size Determination in Health Studies: A Practical Manual*. World Health Organization, 1991.
8. Hedeker D, Gibbons RD, Waternaux C. Sample size estimation for longitudinal designs with attrition. *J Educ Behav Stat* 1999;24:70–93. http://jeb.sagepub.com/content/24/1/70.abstract

## sample size reestimation, *see* reestimation of sample size

## sample surveys

The term *sample surveys* is generally used for large-scale **descriptive studies** to estimate the prevalence of certain specified conditions, although it is loosely used for many other types of studies as well. Demographic and Health Surveys [1], periodically carried out in many developing countries to get a snapshot of health and fertility in the country at a specified point in time, and Global Adult Tobacco Surveys [2] for assessing tobacco use in various segments of population in different countries are examples of such surveys. Periodic National Health and Nutrition Examination Surveys in the United States [3] are important source of information on the health status of the population. This also is a sample survey. Associations or correlations and even a cause–effect relationship may come out of such surveys as a by-product, but they are not among the objectives of the surveys. Surveys generally produce results applicable to one specific area, one specific time, etc., as opposed to the results of the experimental studies that tend to transcend time and space. When the surveys are periodic, as in the examples just cited, they provide information on trend also. Prevalence in different segments of population can be statistically compared to find which group has higher prevalence. Reference level of medical parameters such as of cardiac functions in healthy subjects is also obtained by sample surveys, and similarly sample surveys can also be conducted in diseased subjects to find what and how much aberrations in health occur in different segments of population.

Since the objective is estimating the parameters in a survey, the sampling methods that can provide a representative sample become especially important. Most large-scale surveys do not use **simple random sampling** because that can provide scattered subjects; instead use a **multistage sampling** with appropriate stratification to ensure that no important section of the population is left out. Many large-scale surveys use **cluster sampling** so that many subjects are available at one place that can drastically reduce the cost of the survey, and make it easier to administer the survey. Some sections of a population that are not substantial in number but are crucial, such as people of old age, may have to be overrepresented in survey sampling so that reliable estimates can be generated for such segments also. Elaborate calculations of sample size are made for the surveys that can provide precise information on target parameters, keeping the sampling methodology in mind and the corresponding **design effect**. Estimation methods too are adjusted to provide unbiased estimates with the desired precision. **Standard errors** of the estimates are a necessary accompaniment of such estimation that provide an assessment of the reliability of the estimates, and sometimes confidence intervals are worked out to get a range of not implausible values.

Descriptive studies conducted at a local level such as on patients visiting a clinic or a set of clinics are also called a survey as long as the focus is on prevalence and averages rather than associations and cause–effect. Profile of cases arriving, say, for kidney stones, is technically a survey.

Among many examples of sample surveys in health and medicine, a recent one is that conducted by Chen et al. [4] in Wuhan (China) for comparing the number of men who have sex with men among rural-to-urban migrants. Kenny et al. [5] reported a population-based survey on maternal and child health service utilization in Liberia.

1. USAID. *The DHS Program*. http://www.dhsprogram.com/
2. WHO. *GATS (Global Adult Tobacco Survey)*. http://www.who.int/tobacco/surveillance/gats/en/
3. CDC. *National Health and Nutrition Examination Survey*. http://www.cdc.gov/nchs/nhanes.htm
4. Chen X, Yu B, Zhou D, Zhou W, Gong J, Li S, Stanton B. A comparison of the number of men who have sex with men among rural-to-urban migrants with non-migrant rural and urban residents in Wuhan, China: A GIS/GPS-assisted random sample survey study. *PLoS One* 2015 Aug 4;10(8):e0134712. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4524597/
5. Kenny A, Basu G, Ballard M, Griffiths T, Kentoffio K, Niyonzima JB, Sechler GA et al. Remoteness and maternal and child health service utilization in rural Liberia: A population-based survey. *J Glob Health* 2015 Dec;5(2):020401. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4512264/

## SAMPL guidelines

SAMPL stands for "Statistical Analyses and Methods in the Published Literature," guidelines prepared by Lang and Altman [1] to address the issue of errors in reporting of statistical analyses in published literature in health and medicine. This issue has been raked up for a long time by a large number of workers; many have suggested remedies, but SAMPL guidelines put them together. They specify what exactly should be reported with different analyses such as hypothesis testing, association analyses, correlation analyses, regression analyses, analyses of variance and analyses of covariance, etc., besides of course for preliminary and primary analyses. The basic emphasis is on describing statistical methods with sufficient details to enable a knowledgeable reader with access to the original data to verify the reported results.

The guidelines are still in the making and require further improvement. For example, for preliminary analyses, the guidelines mention about collapsing data into categories but do not say that the reason for doing so should also be mentioned and the categories so formed must be justified with the implications of how different categorizations can lead to different results, or whether we should present analysis with and without categories. They also do not mention anything about missing data and outliers. Similarly for primary analyses, the guidelines require that the method for analysis should be fully described, but it does not say that a reference can also be given for the source

where this is described and only modification, if any, needs to be described. The guidelines require specification of the nonparametric tests used to analyze skewed data, but they do not say about recourse to central limit theorem for large samples when the data are marginally skewed. Perhaps it is not possible to include every bit of detail in such a guideline, but a beginning has been definitely made that can improve statistical reporting in medical journals.

1. Lang TA Altman DG. Statistical analyses and methods in the published literature: The SAMPL Guidelines, in: *Guidelines for Reporting Health Research: A User's Manual* (eds. Moher D, Altman DG, Schulz KF, Simera I, Wager E). Wiley, 2014. http://onlinelibrary.wiley.com/doi/10.1002/9781118715598.ch25/summary

## sampling (advantages and limitations)

The concept of sampling is neither new nor unfamiliar in everyday life. A cook examines a few grains of rice to find whether nearly all of them are properly cooked. In medicine, study of blood, urine, stool, and sputum samples and biopsy specimens is common.

The advantages and limitations of sampling are better appreciated when the meaning of the term *population* as used in statistics is clear. As explained under the topic **population (the concept of)**, this term has special meaning in statistics and can be understood as the target group from which sample subjects are chosen, or the larger group to which the findings would be extrapolated. In a descriptive study of acute respiratory infection (ARI) in a country, the target population could be all existing cases of ARI in that country, and for a cervical cancer control program, the target population could be all married women of age above 40 years. For studying risk factors of enlarged prostate, the population of interest could be all men of age 50 years and above in an area. Thus, statistical population depends on the context and the objective. Cost and logistic considerations seldom allow the study of all the subjects in a population, and sampling becomes a natural choice. Nevertheless, all subjects in the target population can also be investigated, in which case the study is called complete enumeration or **census**.

Even if all existing cases are surveyed, there is no guarantee that the results would apply to future cases. Thus, the concept of population in the context of medicine is more hypothetical than real. When all existing cases are indeed included, it still remains a sample considering that future cases are not included. Medical empiricism implies that the findings on the existing cases are used for future cases. Sampling is a prerequisite for this paradigm. However, if the objective is to find the prevalence of diabetes mellitus in the year 2008 among females of age 40 years and above residing in a particular city, complete enumeration is possible. Similarly, if a complete registry of cancer cases in a defined population is available, perhaps sampling is not needed for assessing the existing situation.

A futuristic perspective brings in the concept of **universe**. A universe could be larger than the target population that has implications for the result. Future cases are part of the universe but not of the population.

Although there are many advantages of sampling, sometimes there are also limitations, both of which are described next.

### *Advantages of Sampling*

The advantages of sampling can be listed as follows:

1. Sampling may be the only feasible method for collecting relevant data in some situations. Samples of blood, urine, semen, and biopsies are everyday examples in medicine. Complete enumeration is not feasible for such items.
2. Lower cost and less demand on personnel because of smaller coverage in samples relative to the total population.
3. Higher speed with which the results can be obtained.
4. Because of the relatively small number of subjects in the sample, more reliable information can often be collected by deploying better trained personnel, by adopting technologically more accurate methods, and by close supervision. This is a significant advantage but many times not realized.

### *Limitations of Sampling*

Despite these clear advantages, the results obtained from a sample cannot always be accepted. The following limitations may be noted:

1. Sampling necessarily entails an argument from the fraction to the whole. The validity of this argument depends on the representativeness of the sample. Not all samples are representative, although methods are available that make it likely to happen. But the choice of the method of sampling may be difficult in many situations as it is not always known what kind of sampling would give a truly representative sample to provide results applicable to the target population. At the same time, it is true that representativeness may be essential for descriptive studies where the objective is to generate valid and reliable estimates of prevalence; it is not necessary for an analytical study. As argued by Rothman et al. [1], representativeness by itself may not deliver valid scientific inference. The overall associations observed in the study sample may not apply to every subgroup. The overall effect is merely an average effect that has been weighted by the distribution of people across these subgroups. Thus, if there is a sample that is representative of the sex distribution in the source population, the results do not necessarily apply either to males or to females, but only to a hypothetical person of average sex. In addition, if the objective is to establish a relationship in a specified type of cases, as in most analytical studies including clinical trials, a representative sample is a secondary consideration.
2. When information is required for small segments containing a few individuals, sampling may fail to provide sufficiently precise information on them.
3. Sometimes a complete count is needed anyway, such as for a diagnosis and outcome profile of cases admitted to a hospital each year. In such cases, sampling is unnecessary.
4. Inclusion of some subjects in the sample and exclusion of others for a study may cause a feeling of discrimination among the subjects. Caution, some groundwork, and persuasion may be required.
5. All samples suffer from what is called **sampling fluctuations**, and we need to be on guard that the inference drawn is based on sound statistical considerations. This may require expertise that may not be easily available in some setups.

1. Rothman KJ, Gallacher JEJ, Hatch EE. Why representativeness should be avoided. *Int J Epidemiol* 2013;42 (4):1012–4. http://ije.oxfordjournals.org/content/42/4/1012.full?sid=1488b059-40cf-44d8-aef2-d2e67cf49f63

## sampling distribution of proportion *p* and mean $\bar{x}$

Samples by themselves are a great source of uncertainty. Yet sampling is considered a preferred strategy in most situations because of its overwhelming advantages enumerated in the topic **sampling (advantages and limitations)**. Statistical methods help in arriving at a conclusion regarding a population parameter based on just one sample. The basic tool used for this purpose is the sampling distribution. This is the distribution of the sample estimate seen in repeated samples as explained later in this section. Before discussing sampling distributions, understanding of some basic concepts is necessary.

### *Point Estimate*

Depending upon whether the variable is qualitative or quantitative, the basic parameter generally of interest is proportion (or probability) $\pi$ or mean $\mu$, respectively. In the case of anemia, the interest could be in either the proportion $\pi$ (or percentage) of subjects responding to iron–folic acid supplementation or the mean $\mu$ of rise in hemoglobin (Hb) or hematocrit (Hct) level. In a sample of 200 subjects, if the proportion responding is $p = 0.30$ (or 30%) or the mean rise in Hb level is $\bar{x} = 1.6$ g/dL after supplementation for 100 days, then it is generally expected that the entire target population would also show a similar picture, provided that the sample is adequately representative. The sample values $p$ and $\bar{x}$ are considered to be **point estimates** of $\pi$ and $\mu$, respectively. Such sample values are also called **statistics** where the usage is in the plural sense. When a large number of samples are actually available, the mean of $p$'s over such samples will approach $\pi$ and the mean of $x$'s will approach $\mu$. Much of the uncertainty generated by the **sampling error** can be satisfactorily resolved by studying the behavior of the statistics such as $p$ and $\bar{x}$ from sample to sample. Their intersample variability is measured by their respective standard errors.

### *Standard Error of* p *and* $\bar{x}$

As in the case of individuals, the variability in $p$ and $\bar{x}$ from sample to sample is measured in terms of their standard deviation (SD), but it is now called the standard error (SE). Thus, SE is the measure of variability of estimates from sample to sample. This special term helps to distinguish interindividual variability (measured by the SD) from intersample variability (measured by the SE). The SE of $p$ measures the variability in proportion $p$ from sample to sample, and the SE of $\bar{x}$ measures the variability in mean $\bar{x}$ from sample to sample. The larger the SE, the greater the variability and the lesser the confidence in the sample results. Thus, the inverse of SE is also a measure of reliability. You can intuitively realize that in the sample, some values are small and some are large so that the sample variance is relatively large. For sample mean though, large values average out with small values, and mean tends to be relatively stable across samples—more so as $n$ increases. Thus, SE is always less than SD.

The SE is calculated on the basis of all possible samples of particular size $n$ from the specified target population. However, these samples are not actually drawn. Statistical theory helps to obtain SE on the basis of just one sample, provided it is randomly drawn at least at one stage. This underscores the need to work with random samples. In the case of simple random sampling (SRS), the SEs are computed as follows:

$$\text{standard errors: } \mathrm{SE}(p) = \sqrt{\frac{\pi(1-\pi)}{n}}, \text{ and } \mathrm{SE}(\bar{x}) = \frac{\sigma}{\sqrt{n}}, \qquad \text{(S.1)}$$

where

$n$ is the size of the sample.
$\pi$ is the proportion in the population.
$\sigma$ is the SD of individual measurements in the population.

If measurements of all $N$ subjects in the entire target population are really available, the value of $\sigma$ is computed for all these values. In practice, though, the population parameters are seldom known and cannot be calculated (if they are known, there is no need of a sample). They are *estimated* from the corresponding values in the sample. Because the estimator of $\pi$ is the sample proportion $p$ and that of $\sigma$ is the sample SD $s$,

estimated standard errors:

$$\widehat{\mathrm{SE}}(p) = \sqrt{\frac{p(1-p)}{n}} \text{ and } \widehat{\mathrm{SE}}(\bar{x}) = \frac{s}{\sqrt{n}}, \qquad \text{(S.2)}$$

where $s$ now is computed with denominator $(n - 1)$, which helps to achieve a more accurate estimate in the long run (called **unbiased** in statistical parlance). The following comments in this context are useful:

1. SE($p$) is maximum when $\pi = 0.5$ and smaller when $\pi$ is either small or large. But its interpretation requires additional care. For $n = 100$ and $\pi = 0.05$, SE($p$) = $\sqrt{0.05 \times 0.95/100} = 0.022$, but for same $n$, $\pi = 0.25$, SE($p$) = $\sqrt{0.25 \times 0.75/100} = 0.043$. In absolute terms, the SE in the first case is nearly half of what it is in the second case. In relative terms, though, the first SE is 44% of $\pi$, while the second is only 17% of $\pi$. Thus, the first is higher in a relative sense. These are referred to as the **absolute precision** and **relative precision**, respectively, of $p$ and have serious repercussion on the confidence intervals.
2. Estimate of SE($p$) fails when $p = 0$ or $p = 1$ because then SE = 0. This case requires a separate consideration. The actual SE uses $\pi$, and $\pi = 0$ implies an impossible event. If you get $p = 0$, increase the sample size as much as to get at least one positive response, preferably more.
3. SE($\bar{x}$) is large when $\sigma$ is large. That is, if the individuals vary too much from one another, the sample means too will exhibit a large variation from sample to sample. Conversely, if the individual measurements are nearly alike or homogeneous, the sample mean in one sample will be nearly the same as in another sample.
4. Both SEs are inversely proportional to the square root of sample size $n$ and decrease as $n$ increases. This implies that two samples containing 100 subjects each from the same population would not differ as much with respect to $p$ or $\bar{x}$ as two samples containing only 20 subjects each would. Thus, a larger $n$ increases the confidence in the sample results. It is this feature that propels statisticians to suggest a larger sample. But this increase can be counterproductive if the cost becomes prohibitive. Thus, a balanced approach is needed.
5. Note that $p$ is the proportion of cases possessing a specific attribute in a group of subjects. For example, one may say that 11% ($p = 0.11$) in a sample of $n = 70$ normal births have a gestation period less than 265 days. In this case, SE($p$) = $\sqrt{(0.11 \times 0.89)/70} = 0.037$ or 3.7%. This result should be stated as "the percentage of births with gestation less than 265 days in a sample of 70 normal births is 11, with SE = 3.7%." This SE measures the expected variation in the percentage from one sample of size 70 to another sample of the same size from the same population.

6. In case of the mean, there is a special need to maintain the distinction between SE and SD. SE is for the aggregated picture obtained from the sample, and SD is for individual values. If the average duration of the gestation period in 30 normal births is 280 days with SD = 5 days, then SE is only $5/\sqrt{30} = 0.91$ days. It would be inappropriate to say that the average gestation period is 280 ± 0.91 days. The correct statement is that the average gestation period is 280 with SD = 5 days, that is, the mean should generally be accompanied by SD and not by SE. Also the usage of the ± sign should be discouraged for stating SD or SE since it gives the erroneous impression that adding and subtracting SD or SE gives the entire range of values.
7. Formulas for actual SEs are as given earlier in Equation S.1, but $\pi$ and $\sigma$ would be rarely known and it would be necessary to use the estimates. This would yield Equation S.2. The hat (^) sign indicates that these are estimates and not actual SEs. But this makes the notation very complex, and we ignore the hat sign and use Equation S.2 as though the equations are the actual SEs.

### *Sampling Distribution of* p *and* $\bar{x}$

As already mentioned, the values of proportion $p$ and mean $\bar{x}$ tend to vary from sample to sample. Just as individual measurements have their distribution pattern such as Gaussian, skewed, or J-shaped, the sample proportion and sample mean have a specific distribution pattern when many samples are available. This is called the sampling distribution. This distribution actually can be complex, depending on the form of the underlying distribution of the individual values, but it can be approximated by Gaussian (normal) under **Gaussian conditions** that basically require a large sample if the underlying distribution is not Gaussian. The mean of this distribution would be the same as of the original values, but the SD would be what we call the SE as explained earlier. If Gaussian conditions are not met, the sampling distribution of $p$ would be similar to a binomial distribution, but the distribution of $\bar{x}$ for small $n$ can be weird depending on the distribution of values in the parent population.

## sampling error/fluctuations

One sample from a population in all probability will be different from the second sample, as some or all individuals in the second sample may be new. Thus, the results obtained from one sample may not match with those from another sample. This variability is included in the sources of uncertainties and is called *sampling fluctuation*. If a sample of 300 healthy males of age 60–64 years from a population has an average systolic blood pressure (BP) of 142 mmHg, it is quite possible that another sample of 300 from the same population yields an average of 139 mmHg. But how likely is it that the average in a new sample will be as low as 128 mmHg or as high as 155 mmHg? The magnitude of this "error" depends primarily on three factors:

i. *The method of sampling*. The subjects should be selected in such a manner that a wide spectrum has adequate representation and repeated samples give nearly the same picture.
ii. *The size of the sample*. When the sample includes a large number of subjects, the picture obtained from one sample is not likely to be very different from that of another sample of the same size because both tend to be fair representatives of the population. This cannot be said for small samples.
iii. *Variability among the subjects in the population*. If the cholesterol level differs widely from person to person, then obviously the samples would reflect the same variability.

*Sampling error* is a term used for the difference between sample estimates and the actual value in the population. Thus, the difference $(\bar{x} - \mu)$ in the case of mean and the difference $(p - \pi)$ in the case of proportion are sampling errors. Hence, *sampling error actually is not an error*; it is just a variation arising due to sampling. Since the estimate varies from sample to sample, sampling error will also vary from sample to sample. As the sample size increases, the chance of a large sampling error decreases provided that the sample is randomly drawn. This error is endogenous to the investigation in contrast with nonsampling error that indeed is an error arising from inappropriate design, misreporting, misjudgment, incorrect recording, nonresponse, etc. The nonsampling error is exogenous in nature. The core of statistical methods deals with sampling fluctuations and sampling errors; but nonsampling errors are not ignored either and control of these also is attempted at each stage of investigation.

## sampling fraction, *see* sampling terms

## sampling frame, *see* sampling terms

## sampling techniques (overall)

Sampling technique is the method to draw a sample from the target population. This is primarily divided into random and nonrandom methods. The topic **random sampling methods** provides an overview, and the individual methods such as **simple random**, **stratified random**, **systematic random**, **cluster random**, **multistage random**, and **probability proportional to size sampling** are discussed in this volume under the respective topics. This section presents an overview primarily to distinguish random from nonrandom methods of sampling.

A sample is called **random** when the inclusion or exclusion of a particular eligible subject depends on chance and cannot be predicted in advance. However, the chances are not necessarily equal for all subjects for inclusion in the sample. For this reason, it is sometimes prudent to call it a **probability sample**. The methods we just listed are all probability sampling methods.

Random selection should not be axiomatic but should be regarded as just a strategy to get a representative sample. If representativeness can be achieved by any other method, the same can be adopted; but experience suggests that nothing better than random is available. However, the larger the sample relative to the size of the population, the better the chance of it being representative—random or not.

There are situations where neither random nor representative sample is important. A sample of volunteers for phase I of a clinical trial is a nonrandom sample. Nonrandom samples inhibit generalizability but can still be useful in some situations as in phase I trial since the primary objective there is to find the tolerability of the regimen. This is one of the methods to draw what is called a convenient sample or a **purposive sample**. Such methods are collectively called **nonrandom sampling** methods.

As the name implies, **convenience sample** is the one that is drawn according to convenience. For example, whosoever is available and agrees to participate among those that you could contact

is convenience sampling. Volunteers are an everyday example of convenience sample. This kind of sampling is adopted particularly when a large number of people of the type actually required for the study are not available. Since this is not a random sample at any stage, it may be heavily biased. The findings from such a sample cannot be generalized to the population from where this sample is taken, yet the findings may turn out to be useful in some situations. McCann et al. [1] took a convenience sample of clinical staff to study their attitude toward the causes and management of aggression in acute old-age psychiatry inpatient units. Hwang et al. [2] took a convenience sample of ethnic minorities in South Korea for studying their utilization of complementary and alternative medicine. In both these cases, the convenience sample has not been a wastage of resources.

Another nonrandom method is **snowball sampling** that is used to spot obscure groups of people. Yet another is quota sampling where subjects are sequentially included without any random method to fill a prefixed quota. One can also conceptualize another method of sampling, which can be called **haphazard sampling**. This would be so when some subjects are selected by one or more random methods and others by one or more nonrandom methods, or when one stage is random and the second stage is nonrandom. This method does not follow any scheme.

1. McCann TV, Baird J, Muir-Cochrane E. Attitudes of clinical staff toward the causes and management of aggression in acute old age psychiatry inpatient units. *BMC Psychiatry* 2014 Mar 19;14(1):80. http://www.biomedcentral.com/content/pdf/1471-244X-14-80.pdf
2. Hwang JH, Han DW, Yoo EK, Kim WY. The utilisation of Complementary and Alternative Medicine (CAM) among ethnic minorities in South Korea. *BMC Complement Altern Med* 2014 Mar 19;14(1):103. http://www.biomedcentral.com/content/pdf/1472-6882-14-103.pdf

## sampling terms

Sampling uses several special terms that must be clear to understand and adopt various sampling methods. Some of these are explained in this section.

### *Unit of Inquiry and Sampling Unit*

Different kinds of units are used in sampling. The **unit of inquiry** is the ultimate unit on which the information is obtained. A **sampling unit** is the one that is used for selection of the subjects. In a community survey on protein energy malnutrition, the sampling unit could be a family, but the unit of inquiry could be a child younger than 5 years. One sampling unit can have multiple units or may have no unit to inquire on. Sometimes the sampling is done in stages, such as selection of some hospitals in the first stage, then wards or departments within the selected hospitals in the second stage, and then patients in the selected wards in the third stage. The unit of inquiry could be a patient, but sampling units are multiple in this kind of sampling.

### *Sampling Frame*

The list of all sampling units in the target population is called the *sampling frame*. The units are chosen from this frame. The units must be mutually exclusive, and the frame should be an exhaustive list. If there is a study on hypertensives and diabetics, a person who has both cannot be listed twice. The list may separately include those who have both diseases, those who have only one of the two diseases, and those who have neither of these two diseases. Inclusion and exclusion criteria must be fully known to the sampler. Preparation of the frame requires precise definition of the unit as well as of the population. For example, if the unit is a case of sexually transmitted disease (STD), all the STDs should be specified. State whether a case of hepatitis B infection would be included or not. Also, state whether the cases would be those that have frank disease or cases with subclinical infection would also be included. The criteria of diagnosis of each should be stated, i.e., the diagnosis is based on clinical evidence alone or a laboratory confirmation is also required. In some situations, a map of the study area serves as the frame. In this case, geographical entities are selected instead of individuals or families. Units available in the selected areas constitute the sample.

The requirements of sampling frame are intense for **simple random sampling** since it would include all sampling units. In contrast to this, in **multistage sampling**, for example, sampling frame is required only for next-stage units within the selected previous-stage units. For example, for selecting districts within a state, frame of districts is needed; but once, say, 5 districts are selected where the next stage is census block, only the frame of blocks for the selected districts is needed. If the next stage is selection of families, the frame of families for only the selected blocks is needed. This obviates the need to prepare sampling frame of all the families in the state, which would be needed if the sampling of families is directly done by simple random sampling. This could be a huge convenience in some situations.

### *Sampling Fraction*

This is the proportion of subjects in the target population to be included in the sample. For example, you may decide to include 1 out of 5, which means 20% sample and the sampling fraction is 1/5. This fraction gives an idea to the reader how big is the sample relative to the size of the population, and can be important consideration in assessing the credibility of observational studies since high sampling fraction implies less **sampling error**. Medel-Herrero et al. [1] used sampling fraction to assess the reliability of one hospital statistics resource versus another one in Spain.

Sampling fraction has special application to **systematic sampling**, where one unit is randomly selected as per the sampling fraction and all others follow a systematic pattern in sequence using the same sampling fraction.

### *Sampling with and without Replacement*

Consider a finite population of size $N$ units. If one is selected by random or nonrandom method, the remaining available for sampling now is $N - 1$ units. In the next step, only $N - 2$ units remain. This is sampling without replacement. In case of sampling with replacement, the selected unit goes back to the population so that the size of the population remains the same $N$ for selection of the second and all subsequent units. Under this sampling, the same unit can be selected twice or many times since the unit is available again for selection. This allows infinite samples from a finite population. If $N = 500$ and the sample size is 20, select 20, replace them so that we again have 500 units to choose from, select another 20 (some of these could be the same units as in the first sample), and so on. Without replacement, you can select at most 500/20 = 25 samples, whereas with replacement, you can select as many as desired.

In a random sampling without replacement, the probability of selection of any particular unit is $1/N$ at the first selection but changes to $1/(N - 1)$ at the second selection for the unit not selected, then to $1/(N - 2)$, etc. These changing probabilities essentially mean that the independence is lost—the chance of selection of any particular unit at subsequent sampling is affected by what was selected

earlier. This can complicate the analysis. In case of sampling with replacement, what we get on the first step does not affect what we get on the second, and thus the independence is preserved. Most statistical methods require independence, although we realize that the same unit can rarely be allowed to be selected two or more times in the same sample. In case of a large population, the chance of the same unit being selected twice is anyway small and negligible.

When the population is very large, the probabilities for units selected subsequently without replacement are not much affected and the independence can be assumed. This is not so when the population size is small. Thus, samples chosen from a small population require exact analysis with relatively much more complex calculations. However, this is rarely required in practice. For example, in a simple random sampling, if one wishes to select 20 out of 500, the method such as that of 500 identical chits in a hat with numbers 1 to 500 and selecting 20 at the same time gives the same chance of selection to each unit even when the sampling is without replacement.

In methods such as **bootstrap**, where the sampling is with replacement, any number of samples can be drawn from the available sample. For example, if a sample of just 10 units is available, an infinite number of samples of, say, size 6, can be selected because of replacement. This is what gives strength to the bootstrap method.

1. Medel-Herrero A, Gómez-Beneyto M, Saz-Parkinson Z, Bravo-Ortiz MF, Amate JM. Discordance between two national health statistics sources (EMH and EESCRI, 1990-2009): Analysis of psychiatric morbidity. *Rev Psiquiatr Salud Ment* 2014 Jul 3. pii: S1888–9891(14)00061-5. http://www.elsevier.es/es-revista-revista-psiquiatria-salud-mental-286-linkresolver-discordancia-entre-fuentes-estadisticas-sanitarias-S1888989114000615

## sampling unit, *see* sampling terms

## sampling with and without replacement, *see* sampling terms

## scales of measurement (statistical)

A scale is an instrument on which the characteristics are measured. It can be quantitatively calibrated in the usual sense or can be qualitative. Blood glucose level and urea clearance are measured in terms of quantities, whereas the blood group of a person is a quality recorded as either O, A, B, or AB. Age can be measured in days and hours but is often categorized in years as (0–4), (5–14), (15–49), etc. Disease severity and extent of malnutrition are quantities but are generally measured as none, mild, moderate, and serious. Site of malignancy is also a measurement in a statistical sense but is recorded as oral, lung, abdomen, breast, etc. Thus, there are a large variety of measurements. Statistical methods to study variation depend on the scale of measurements. These scales can be grouped in a variety of ways. A majority of them are described here.

### *Nominal Scale*

Not all measurements are necessarily in terms of quantities. Complaints and site of cancer are qualities but are still considered *measurement* in a statistical sense. The categories in such cases are only names, and the scale of such a measurement is called nominal. All space-related measurements such as the organ affected, site of lesion, and place of occurrence of disease are nominal, and so are attributes such as race and blood group. These names do not have any specific order. Thus, there is no notion of less than or more than in this kind of scale, and the only valid comparison is of equality or inequality. Gender is either male or female, and none is higher or more than the other. Diagnosis of liver disease as hepatitis, cirrhosis, or malignancy is nominal, and so is the nature of a handicap such as visual, speech, orthopedic, mental, etc. These variables are genuinely categorical, and the only way to associate numbers with them is by way of assigning a **code** to each category. Note, however, that codes for nominal categories are not metric scores—a category receiving code 4 is not twice the category receiving code 2. These codes cannot be treated as quantities and cannot be added, subtracted, multiplied, or divided. Codes are not the same as scores.

When the assessment of a characteristic is in terms of only two categories such as yes/no, present/absent, or favorable/unfavorable, these are called **dichotomous categories**. The corresponding variable is called a **binary**. Recording gender as male or female is the most glaring example of such a variable. If the number of categories is more than two, such as cirrhosis, hepatitis, and malignancy for liver disorders, these are called **polytomous categories**. Statistical methods are simpler for dichotomous than for polytomous categories. If a statistician advises you to collapse three or more categories into two for the sake of convenience during calculations, you should agree only if this does not reduce relevance and the operational value of the conclusion is not compromised. Clinical relevance should not be sacrificed for statistical expediency. Computers have largely obviated the need to compromise on this aspect.

Statistically, all measurements on the nominal scale are discrete **variables** since the number of categories is generally small, and the data are generally expressed in **contingency tables** and analyzed by **chi-square test**, particularly for large samples.

### *Metric Scale*

At the other end of the spectrum are characteristics that can be exactly measured in terms of a quantity. Duration of a disease, hemoglobin (Hb) level, heart rate, and parity are examples of such characteristics. These are said to be measured on a metric scale. Often these are recorded in categories such as years of age in (0–4), (5–14), (15–49), (50–69), and (70+) groups. Such categorization tends to ordinalize the metric scale and results in loss of information, yet it is preferred in some situations, as discussed in a short while.

There is a tendency in health and medicine to develop a **scoring system** for **soft data** such as the degree of severity of disease. This introduces a metric scale for soft data and definitely helps in achieving a better exactitude. Such scores are statistically satisfying but sometimes lose clinical relevance. If you are using a scoring system, ensure that it adequately represents varying grades of the observations. Assigning a number should not be at the cost of unjustified air of accuracy.

Sometimes a metric scale is divided further into interval and ratio scales. In an **interval scale**, there is no absolute zero, for example, body temperature. A temperature of 105°F cannot be interpreted as only 5% higher than 100°F. Similarly, it is incorrect to say that a person with an intelligence quotient (IQ) of 160 is twice as intelligent as a person with an IQ of 80. Differences matter but ratios are irrelevant in this kind of scale. This is so for many measurements in medicine that do not start from zero. Plasma glucose level, blood pressure, and heart rate are all examples of this kind of measurement. On the other hand, in a **ratio scale**, a zero point can be meaningfully

designated. It is correct to say that the duration of survival of 6 years is twice as much as 3 years, and parity 3 is thrice as much as parity 1. However, in this case also, their medical interpretation may not be based entirely on a proportional factor. The fine distinction between interval and ratio scales in most cases is not required for managing uncertainties through statistical methods.

The disadvantages of measurements on the metric scale are that, in many cases, such measurements are relatively more difficult, more time consuming, and more expensive. A large number of parameters may have to be considered together to say that the extent of burns is 78%, while it is easy to say on visual inspection alone that the burns are "extensive." However, such shortcuts invariably lack objectivity. Whenever feasible, prefer the metric scale to the ordinal scale. It is always wise to ensure that the metric measurements you are using are indeed valid and reliable.

Metric measurements can also be divided as continuous and discrete. For this, see the topic **variables** where these and other types of variables are described.

### *Ordinal Scale*

There are certain characteristics that should be measured in terms of quantity, but the nonavailability of a good instrument compels measurement in terms of what is called an ordinal scale. Disease severity when measured as none, mild, moderate, serious, or critical; likelihood of the presence of a particular disease when measured as ruled out, unlikely, doubtful, likely, or confirmed; and self-perception of health from very bad to very good on, say, a seven-point scale are examples of such characteristics. These are inherently metric, but an ordinal scale is more convenient in such measurements. The main reason for nonavailability of a metric scale is that many of these characteristics are multifactorial, for example, disease severity depends on signs and symptoms, measurements such as blood pressure and plasma glucose levels, and radiological assessment, etc., which makes assigning a single quantity extremely difficult.

Sometimes a device to measure a characteristic is easily available but is not adopted because such a level of accuracy is not needed. Smoking can be measured as the number of cigarettes smoked, but categories such as none, light, moderate, and heavy seem to serve the purpose sufficiently well in many clinical situations. Age can be measured in terms of years, but categorization into child, adult, and old may be adequate in some situations.

The basic advantage of using an ordinal scale rather than a metric scale is convenience in eliciting, recording, and reporting, and there is no need of any sophisticated device for measurement. Ordinal categories are often easier to comprehend than metric categories, although valuable and accurate information is lost in the process, and the analysis of data is rendered less efficient. Metric measurements are amenable to a host of mathematical manipulations that are not possible with ordinal measurements. Thus, prefer using hard measurements such as blood pressure level instead of grade of hypertension, and prostate volume instead of grade of enlargement. If the efforts required for hard measurements are enormous such as in measuring the size of the brain, or when no metric scale is available, use an ordinal scale. However, beware of anomalies in some ordinal categories. What is mild for one may be moderate for another, since these terms are rarely strictly defined.

Between nominal and ordinal, there might be measurements that are *semiordered*. Classification of malignancy as definitely absent, probably absent, uncertain, probably present, and definitely present is an example of a semiordered scale. These categories are partly nominal and partly ordinal.

**TABLE S.7**
**Comparative Features of Nominal, Ordinal, Interval, and Ratio Scales of Measurement**

| Type of Scale | Does It Measure Magnitude? | Is This Equally Spaced? | Does It Have Absolute Zero? |
|---|---|---|---|
| Nominal | No | No | No |
| Ordinal | Yes | No | No |
| Interval | Yes | Yes | No |
| Ratio | Yes | Yes | Yes |

Quite often, numerals are associated with ordinal categories such as 0 for none, 1 for mild, 2 for moderate, and 3 for serious. These numbers are then subjected to all sorts of algebraic calculations. Such calculations are valid only when the moderate degree is considered two times the mild degree, and the serious degree is considered three times the mild. These numerals also assume that the difference between mild and no disease is the same as that between serious and moderate disease. In practice, this may not be so, and sufficient caution is required in assigning numerals to ordinal categories and in drawing conclusions when based on calculations involving such numbers. Note that these numerals are not codes but are used as scores.

Main features of nominal, ordinal, and metric (interval and ratio) scales are summarized in Table S.7, which may give you a snapshot of their comparison.

### *Other Types of Scales of Measurement*

Several other types of scales are used in statistical data, and these are discussed elsewhere in this volume. Among these are **logarithmic scale** as opposed to the usual linear scale, **Likert scale**, and **Guttman scale**—the latter two are used for items in a questionnaire.

## scalogram analysis, *see* Guttman scale

## scatter diagrams

A scatter diagram aims to show the variation in the values of one variable in relation to another. Simple examples are blood pressure (BP) values in subjects of different ages and cholesterol levels in subjects with different body mass indexes. Different symbols can be used to distinguish between male and female subjects, or any such groups. Preferably both horizontal ($x$-axis) and vertical ($y$-axis) should be metric for a scatter diagram, but it is necessary for at least the $y$-axis to be metric ($x$-axis can be nominal or ordinal). If one variable is dependent on the other, the dependent variable is shown on the vertical axis. In Figure S.2a, the percent cholesterol esters is shown as dependent on total bilirubin in cases of cirrhosis, hepatitis, and common bile duct (CBD) obstruction. In the presence of regurgitation jaundice, the flow of bile into the duodenum may decrease, causing a reduction in the esterified form of cholesterol. Note an **outlier** in this figure. Scatter diagrams provide a great opportunity to spot such outliers.

When the scale on the $x$-axis is ordinal or nominal, the scatter takes the shape of stacking of symbols over one another as in Figure S.2b. Different symbols are used for different severities of jaundice in this diagram. Already it is difficult to untangle the points and identify the pattern for cirrhosis, hepatitis, and CBD obstruction

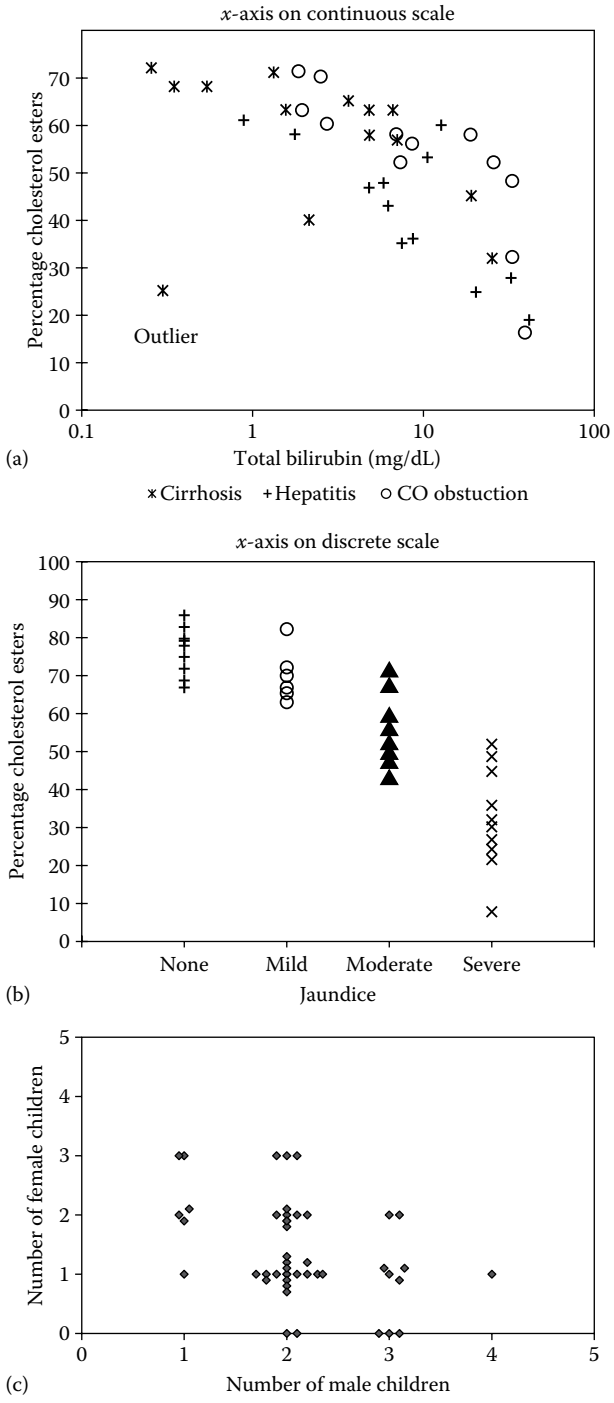



**FIGURE S.2** Three types of scatter diagram. (a) Dependence of percentage cholesterol esters on degree of jaundice measured by total bilirubin. (b) Percentage cholesterol esters in cirrhosis patients with different degrees of jaundice. (c) Use of jitter for overlapping points—male and female children at the time of sterilization in families preferring male children.

in Figure S.2a. The problem accentuates when measurements with different units are shown on the right and left axes (not shown). The next problem occurs due to too many points at one place in a scatter diagram that may look like a vague blob. For such overlapping points, try **jitter**, which is a minor perturbation for plotting purposes only, and use small size symbols. See Figure S.2c for an example.

## scoring systems (methods for developing), *see also* scoring systems for diagnosis and for gradation of severity

In our context, scoring systems give scores to various background characteristics of the people (age, sex, weight, occupation, etc.), signs–symptoms (pallor, spleen, pain, etc.), medical measurements (blood pressure, pulse rate, etc.), laboratory and radiological investigations (electrocardiogram [ECG], glucose level, chest radiograph, etc.), and other features, with the following objective: first to try to give them exactitude in terms of quantities, and second to combine them to provide a holistic picture. In some difficult-to-comprehend situations, such scoring systems can help address part of the problem arising from knowledge limitation (**epistemic uncertainties**) and can help to establish diagnosis and to grade the severity of the condition.

Developing a good scoring system has not been easy because of gaps in our knowledge to be exact and due to the need to have expertise for a realistic assessment. Yet many of them are available and used extensively. Developing a scoring system requires that the final status of the patients (either the diagnosis or the gradation of severity) is available with a widely acceptable set of criteria for at least some of them. Once the system is validated for its reliability and validity, it can be used on new cases.

Characteristics that have gradients are relatively easy to score than measurements on the nominal scale. If the gradient is already numeric such as pain on the visual analog scale, the score is immediately obtained for this variable; but the problem arises for ordinal factors that have gradients but no numeric scale is available. Examples are severity of disease, which is categorized as mild, moderate, serious, or critical, and degree of satisfaction categorized as completely dissatisfied to fully satisfied. Characteristics on the nominal scale such as presence or absence of signs and symptoms in any case defy quantitation at the individual level. The method of assigning numerics to ordinal and nominal characteristics should be such that it can reduce uncertainties and does not introduce additional questions regarding validity of the scoring system. A more general method based on **logistic regression** is discussed later in this section, but consider the following for the time being.

### *Method of Scoring for Graded Characteristics*

The simplest scoring for ordinal characteristic is linear, for example, 0 for no disease, 1 for mild disease, 2 for moderate disease, 3 for serious disease, and 4 for critical condition. As already mentioned, such a scoring assumes that the difference between mild and no disease is the same as that between critical and serious disease. It is legitimate to ask for such a scoring why scores 0, 3, 6, 9, and 12 are not more appropriate for severity of disease, which also are linear (3 added each time), or even why geometric scoring (twice of the previous) such as 0, 1, 2, 4, 8 is not better. Very few studies have been carried out to investigate various alternatives, and thus nothing can be stated with confidence. Nevertheless, the 0, 1, 2, 3, and 4 type of linear scoring remains the most widely used scores because of their simplicity. They go unnoticed. No explanation is generally required when such simple scores are used, while any other scores are expected to be accompanied by justification. However, it is better to investigate a set of patients with different severities to find how severe this disease is for them and assign a score accordingly. A more complex procedure to generate scores is illustrated in the following example. Similar innovative ways can be devised to develop a scoring system.

Functionality in multiple sclerosis patients has a gradient, and its metric measurement has always been a challenge. Various scales are used to assess functionality in the patients; among those relatively easy to perform are the 10 m timed walk (TMTW) and nine-hole peg test (NHPT) for the right and left hands. Vaney et al. [1] developed a short and graphic ability score (SaGAS) as (2 × TMTW + NHPTright + NHPTleft) after taking the logarithm of these timed values. The authors demonstrated good correlation of SaGAS with established tests such as Multiple Sclerosis Functional Composite, Expanded Disability Status Scale, and Rivermead Mobility Index. Features of SaGAS such as simplicity, intuitiveness, and nonphysician-based measurability were enumerated as advantages for its use in multiple sclerosis patients. The authors developed an easy-to-implement scoring system, and its good correlation with other more complex scoring methods is one way to demonstrate its validity. For a real assessment, comparison with a gold standard could be more convincing. Such a standard may not exist in this case; in most cases, the gold standard might be too cumbersome and expensive. Equivalence of the new scoring system with the gold standard when demonstrated would establish its real validity.

Consider measuring nursing workload in an intensive care unit (ICU) by a scoring system. Yamase [2] assessed this workload by 88 items relating to (i) the number of nurses required, (ii) muscular exertion, (iii) mental stress, (iv) skill, and (v) intensity. A three-round **Delphi** survey among 20 skilled ICU nurses assigned a consensus four-grade (0 to 3) score to each of the 73 items after excluding 15 items considered unnecessary. These scores were confirmed by surveying 118 nurses in other ICUs. The "comprehensive nursing intervention score" is the sum of all these individual item scores. The scoring system was confirmed as reflecting the true workload by applying it to the daily care of 107 patients. This example illustrates how a simple method can be used to develop a scoring system. Validation of the individual scores by another group of nurses and of the scoring system by using it in actual conditions tends to enhance the confidence in the scoring system.

### *Method of Scoring for Diagnosis*

Scoring for establishing diagnosis might help in situations where diagnosis or differential diagnosis is difficult and requires a lot of expertise that may not be immediately available. See the topic **scoring systems for diagnosis and for gradation of severity** for an example on the diagnosis of hypothyroidism that uses scores on clinical signs and symptoms. Here, the scoring helps to establish or rule out hypothyroidism in nearly half the cases. Thus, the need for further investigations is reduced to one-half. For another example, see Guo et al. [3] who provided a scoring system for the diagnosis of Hirschsprung's disease in neonates. Rosen et al. [4] proposed a smaller 5-item index as a diagnostic tool for erectile dysfunction compared with the larger 15-item scoring. Their index uses linear scores for each item.

Signs and symptoms are qualitative and play a significant role in establishing diagnosis. Converting such qualities to a numerical score has always been a challenge, and no widely acceptable method is available yet. The following examples and their discussion are based on the methods used by some workers who ventured into this area.

There are examples of assigning arbitrary scores. Obviously, they provide tentative results. A common and acceptable method of assigning scores to individual characteristics is based on the **regression coefficients** that are estimated using a multiple regression method when the outcome is quantitative. For qualitative outcomes, particularly those that are dichotomous, such as the presence or absence of a disease, **logistic regression** coefficients are used. The factors surmised to determine the outcome are antecedents whose coefficients in the logistic regression equations are significant. Since they can be interpreted as the log of odds ratio (OR), the larger the OR, the better the predictive utility of the factor. Thus, the score in proportion of the OR can be assigned to those factors that turn out to be significant predictors. The method is illustrated in the following example on unstable angina. This example also is more on the gradation of a disease than on diagnosis, although the outcome is dichotomous: either death or survival. For a more rigorous example, see **APACHE score**.

Unstable angina is a complex syndrome prognosticated by a host of factors such as age, hypertension, diabetes, hypercholesterolemia, smoking, previous myocardial infarction (MI), ST segment elevation in ECG, troponin test, etc. Piombo et al. [5] studied a large number of such factors in 473 patients and found four of them to be significant predictors in a multivariate logistic regression of in-hospital occurrence of refractory ischemia, acute MI, or death. These three together formed an unfavorable outcome. The ORs were 4.03, 2.29, 2.21, and 2.0 for ST segment deviation, age ≥ 70 years, previous coronary artery bypass grafting, and positive troponin test (T ≥ 0.1 mg/mL), respectively. The scores assigned were 4, 2, 2, and 2 corresponding to the respective ORs, and the highest possible score was 10. It was divided into three categories: 0 or 2 for low risk, 4 or 6 for intermediate risk, and 8 or 10 for high risk (scores 1, 3, 5, 7, and 9 are not possible in this scoring system). This scoring system was validated in another group of 242 patients that provided similar results. Nearly 63% of patients were assigned to the low-risk group, 31% to the intermediary-risk group, and 6% to the high-risk group. The predictive power of the scoring system assessed by the ***C*-statistic** (area under the **ROC curve**) was 0.72. The *C*-statistic is one of the important criteria that determine the validity of a scoring system.

This example on unstable angina was chosen not because it provides a valid scoring system but because it uses an appropriate method for selection of factors and for assigning them a proper score. There are many other examples of this type. Purasiri et al. [6] combined the results of clinical examination, mammogram, ultrasonograph, and fine needle aspiration cytology by assigning them weighted scores using stepwise logistic regression. This combined score performed better than any of the others individually in differentiating malignancy from benign lesions in suspected cases of breast cancer. In this study, the confirmed diagnosis was later available so that the "gold" was present. However, the authors used the term index and not score. Another example is the scoring system developed by Chiu et al. [7] for early detection of oral submucous fibrosis based on a self-administered questionnaire. This had a *C*-statistic of 0.90. Rassi et al. [8] also assigned scores proportional to the regression coefficients to the independent significant factors for death in Chagas disease. The *C*-statistic was 0.84 for this scoring system.

The method of assigning scores proportional to the regression coefficients is valid only when the values of the predictive factors are standardized to mean 0 and variance 1 before the regression is run. The regression coefficients without this standardization are not comparable and are severely affected by the unit of measurement. Since computation is not a limitation these days, regression equation itself can be used as the scoring system without any standardization.

Since a regression coefficient after standardization measures the contribution of the factor to the outcome, this method of scoring looks at least face valid as it assigns higher score to the factor that contributes more to the outcome. Also since the regression is able to identify the factors that are significant independent contributors and need to be included in the scoring system, this method has some desirable properties. However, it may fail to provide a valid scoring system in some situations as discussed next.

### *Validity and Reliability of a Scoring System*

Although qualities such as ease of understanding and ease of implementation can be cited, basic statistical qualities of scoring systems are **validity** and **reliability**. Of these two, validity is more important and difficult to assess too. If a scoring system were not valid, its good reliability would seldom be useful.

*Validity of scoring system.* How does one assess that a scoring system is providing the right result? First, it should look just about right (called face validity) and should correspond well to the knowledge of experts. If a scoring system surprises you and the experts, reconsider the elements that cause this surprise and make necessary modifications. But the most important assessment of validity is against a gold standard. The basic difficulty in assessing this is in identifying and implementing the gold standard against which the validity is checked. If the gold standard is easy to perform, there is no need for a scoring system. In the case of pregnancy-induced hypertension, Thurnau et al. [9] compared the scoring results with clinical manifestation. If clinical manifestation is to be considered the gold standard and if clinical assessment is relatively easy, nothing additional is gained by the scoring system. A scoring system is useful only when it really adds to the clinical picture or when it replaces a complicated procedure. The latter can happen when a final diagnosis is based on consensus of experts or when it emerges later in the course of disease. An explicit advantage is the objectivization the scores introduce, which clinical assessment may lack.

When the results from the gold standard are not available, the worth of a scoring system is assessed using alternatives. One is to see whether the scoring system gives results that are consistent with undisputable outcomes such as death. Rassi et al. [8] reported for their scoring system for predicting death in Chagas disease that patients with a low (0–6 points) score had 10% 10-year mortality rate, those with a medium (7–11 points) score had 44%, and those with a high (12–20 points) score had 84%. This provides an indirect evidence of validity of the scoring. Note in this case that the gold standard is death, and there is no way to assess it in advance except by prognostic factors summarized by the total score.

The second aspect of validity is establishing it in a different sample. This, in fact, testifies repeatability. If another sample of similar nature gives similar results, it is safe to conclude that the scoring system is not sample-specific and has at least some generalizability. In almost all examples discussed in this section, the validation sample is different from the development sample, and the results were shown to replicate adequately.

The third type of validation of a scoring system is its comparison with an established and more cumbersome system. This is to check if an easier version provides the same results. Both could be excellent or both could be poor, but that is not the issue in this kind of comparison. The comparison is not with a gold standard in this case. Such concurrent validity provides evidence that the easier version can replace the cumbersome procedure. Moreno and Morais [10] compared a 28-question simplified therapeutic intervention scoring system (TISS) with the standard 72-question TISS for nursing workload in ICUs and came up with the conclusion that the simplified version is just as good. Evans et al. [11] developed a scoring system for identifying mutation and found that it outperforms existing models.

*Reliability of a scoring system.* In addition to being valid, any medical assessment tool should also be reliable with respect to repeatability and reproducibility. Reliability may suffer if the wordings are not precise and instructions are not explicit, so that there is room for subjective interpretation either by the assessor or by the assessee or by both.

The reliability of a scoring system is assessed interobserver (also called an interrater) as well as intraobserver. For both these assessments, **intraclass correlation** is used. This correlation must be in excess of 0.90 for good reliability provided that there are no factual repetitions, and a value between 0.80 and 0.89 is considered acceptable. Any tool with intraclass correlation less than 0.80 is suspect and used with caution.

Another measure of reliability is the narrow width of the confidence interval (CI). This can be calculated for sensitivity, specificity, predictivities, and the area under the ROC curve for the derived scoring system as illustrated in the following example. If the width is sufficiently narrow, the system can be considered reliable.

Xu et al. [12] studied 129 patients with suspicion of prostate cancer who underwent transrectal sonography-guided repeat biopsies. They devised a scoring system based on the results of multivariate logistic regression. ROC analysis found that the best cutoff is 2.5 at which the area under the ROC curve was 0.816. Sensitivity was 76.5% and specificity 74.7% at this cutoff. Relatively low values of sensitivity–specificity and not particularly high area under the concentration curve show just about satisfactory validity of this scoring system but not high validity. CIs were not calculated; had they been, they would also show that the reliability also was not adequate.


1. Vaney C, Vaney S, Wade DT. SaGAS, the Short and Graphic Ability Score: An alternative scoring method for the motor components of the Multiple Sclerosis Functional Composite. *Mult Scler* 2004;10:231–42. http://msj.sagepub.com/content/10/2/231.abstract
2. Yamase H. Development of a comprehensive scoring system to measure multifaceted nursing workloads in ICU. *Nurs Health Sci* 2003;5:299–308. http://www.ncbi.nlm.nih.gov/pubmed/14622382
3. Guo W, Zhang Q, Chen Y, Hou D. Diagnostic scoring system of Hirschsprug's disease in the neonatal period. *Asian J Surg* 2006;29:176–9. http://www.ncbi.nlm.nih.gov/pubmed/16877220
4. Rosen RC, Cappelleri JC, Smith MD, Lipsky J, Peña BM. Development and evaluation of an abridged 5-item version of the International Index for Erectile Function (IIEF-5) as a diagnostic tool for erectile dysfunction. *Int J Imp Res* 1999;11:319–26. http://patientreportedoutcomes.ca/files/2014/04/IIEF-5-Rosen-1999.pdf, last accessed November 30, 2015.
5. Piombo AC, Gagliardi JA, Guelta J, Fuselli J, Salzberg S, Fairman E, Bertolasi C. A new scoring system to stratify risk in unstable angina. *BMC Cardiovasc Disord* 2003;3:8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC194644/
6. Purasiri P, Abdalla M, Heys SD, Ah-See AK, McKean ME, Gilbert FJ, Needham G, Deans HE, Eremin O. A novel diagnostic index for use in the breast clinic. *J R Coll Surg Edinb* 1996;41:30–4. http://www.ncbi.nlm.nih.gov/pubmed/8930039
7. Chiu CJ, Lee WC, Chiang CP, Hahn LJ, Kuo YS, Chen CJ. A scoring system for the early detection of oral submucous fibrosis based on a self-administered questionnaire. *J Public Health Dent* 2002;62:28–31. http://www.researchgate.net/publication/227704354_A_Scoring_System_for_the_Early_Detection_of_Oral_Submucous_Fibrosis_Based_on_a_Selfadministered_Questionnaire, last accessed November 30, 2015.
8. Rassi A Jr, Rassi A, Little WC, Xavier SS, Rassi SG, Rassi AG, Rassi GG, Hasslocher-Moreno A, Sousa AS, Scanavacca MI. Development and validation of a risk score for predicting death in Chagas heart disease. *N Engl J Med* 2006;355:799–808. http://www.nejm.org/doi/full/10.1056/NEJMoa053241
9. Thurnau GR, Dyer A, Depp OR III, Martin AO. The development of a profile scoring system for early identification and severity assessment of pregnancy-induced hypertension. *Am J Obstet Gynecol* 1983;146:406–16. http://www.ncbi.nlm.nih.gov/pubmed/6859162

10. Moreno R, Morais P. Validation of the simplified therapeutic intervention scoring system on an independent database. *Intensive Care Med* 1997;23:640–4. http://link.springer.com/article/10.1007/s001340050387#page-1
11. Evans DG, Eccles DM, Rahman N, Young K, Bulman M, Amir E, Shenton A, Howell A, Lalloo F. A new scoring system for the chances of identifying a BRCA1/2 mutation outperforms existing models including BRCAPRO. *J Med Genet* 2004;41:474–0. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1735807/
12. Xu B, Min Z, Cheng G, Mi Y, Tong N, Feng N, Song N et al. Evaluating possible predictor of prostate cancer to establish a scoring system for repeat biopsies in Chinese men. *J Ultrasound Med* 2011;30:503–8. http://www.jultrasoundmed.org/content/30/4/503.full

## scoring systems for diagnosis and for gradation of severity, *see also* scoring systems (methods for developing)

Medical professionals can be statistically dichotomized as those preferring qualities and subjective assessments using clinical acumen, and those who prefer quantities that bring in exactitude. Some of those belonging to the latter group may like to quantify even signs and symptoms of a disease. How does one convert qualities into quantities at an individual level? The answer is through the scoring system, whereby signs and symptoms unique to a disease or more commonly seen are assigned a higher score and the others a low score. Thus, measurements on a nominal scale are ordinalized and finally converted to a metric scale. **APACHE** and **Apgar scores** are everyday examples of such conversions, although in Apgar, each sign gets an equal score. Pain on a visual analog scale measures pain intensity, but the McGill pain score comprises qualities like throbbing pain, shooting pain, and stabbing pain. Scoring systems also reduce a multivariate entity into a univariate quantity, thus increasing comprehension and utility. They also help in reducing some of the **epistemic uncertainties** that can arise from inadequate realization of how much weight is to be given to various pieces of information.

Realize that the scores are the quantitative values assigned to **ordinal data**. This increases the statistical power of the methods of analysis, if, and this is a big if, the scores truly reflect the "importance or medical significance" of the ordinal categories. In case the data are already quantitative and metric categories had to be formed for convenience, the midpoints of the intervals (called *midranks*) are a fairly good approximation of the category scores. If the categories were unequal such as age categorized into 0–1, 1–4, 5–14, etc., these scores would not be equally spaced. The temptation to give equally spaced scores such as 1, 2, 3, etc. (called *linear scores*) to such categories could provide misleading results. In the case of predictor variables, midranks and linear scores may not be able to capture the right importance. For example, the number of cigarettes smoked per day may be categorized as 0, 1–4, 5–9, 10–19, and 20+; however, their linear score as 0, 1, 2, 3, and 4, or midranks as 0, 2.5, 7, 14.5, and, say, 29, may not be able to capture the intensity of smoking for hurting the health, and therefore a more rational scoring may be needed. The same applies to scores 1, 2, 3, etc., assigned to mild, moderate, serious, etc., pertaining to the severity of diseases. However, the challenge is to ordinalize or assign quantities to signs and symptoms. As already mentioned, this is done by assigning a higher score to those that are more commonly seen or those that are specific to the disease.

Scoring systems are gradually gaining importance similar to the laboratory and radiological investigations. There was a time when these were considered below the dignity of a clinician, but now hardly any medical decision is taken without recourse to such investigations. A large number of scoring systems on various aspects are being developed: the list already runs into hundreds. These systems have found useful applications in gradation of severity of a disease as they consider several individual assessments together. Scoring systems have also been applied for establishing a diagnosis, particularly for differentiating one condition from another similar looking condition. Both types of scoring systems—for diagnosis and for severity of a disease—can add to the knowledge with which a medical condition is managed.

### *Scoring System for Diagnosis*

Attempts have been made periodically to quantify the field of medicine and developing scoring systems that can help in diagnosis. The following example on scoring for diagnosing thyroidism illustrates the procedure.

Hypothyroidism poses a challenge to physicians at the time of diagnosis. It is not easy to distinguish hyperthyroidism, euthyroidism, and hypothyroidism from each other based on clinical signs and symptoms. Assays for measuring $T_3$, $T_4$, and thyroid-stimulating hormone (TSH) levels help in reaching a diagnosis, but these tests are slightly expensive, and some hospitals in developing countries may not have the facility for such testing. Clinically clear cases of hyperthyroidism and hypothyroidism can be treated without taking into account laboratory results, and the tests are ordered only in doubtful cases. This can also reduce the load on laboratory services. A diagnostic index for hypothyroidism was devised by Billewicz et al. [1] taking into account the frequency of various symptoms and signs in this disease. They developed a scoring system excluding four features with zero score (Table S.8).

**TABLE S.8**
**Scoring System for Thyroidism**

| | Score | |
|---|---|---|
| **Signs and Symptoms** | **Present** | **Absent** |
| Physical tiredness | 0 | +2 |
| Slow cerebration | –3 | +2 |
| Diminished sweating | +6 | –2 |
| Dry skin | +3 | –6 |
| Cold intolerance | +4 | –5 |
| Dry hair | –2 | +2 |
| Weight increase | +1 | –1 |
| Constipation | +2 | –1 |
| Hoarseness of voice | +5 | –6 |
| Paresthesia | +5 | –4 |
| Deafness | +11 | 0 |
| Slow movements | +4 | –3 |
| Coarse skin | +7 | –7 |
| Cold skin | +3 | –2 |
| Periorbital puffiness | +4 | –6 |
| Pulse rate <75/min | +4 | –4 |
| Ankle jerk | +15 | –6 |
| **Total score** | **≤ –30** | **Euthyroid** |
| | **–29 to +24** | **Doubtful** |
| | **≥ +25** | **Hypothyroid** |

The scores are based on the logarithm of the ratio of frequency of presence and absence of symptoms in established hypothyroid and euthyroid cases. According to this scoring system, a clear diagnosis can be made based on signs and symptoms alone if the score is 25 or more (hypothyroid) or –30 or less (euthyroid). Laboratory assistance is required only when the total score is between –29 and +24. This can reduce the load on the laboratory by more than 50%. This scoring system was later simplified by Zulewski et al. [2]. Although this is a score, the authors liked to call it an *index*. Many such examples of mixed use of the terms exist in the literature.

Not many researchers look for a simple scoring system as illustrated in our example. Instead, a **logistic regression** equation is used as a scoring system (see **scoring system [methods for developing]**). Examples are a scoring system for repeat biopsies in suspected patients of prostate cancer [3] and for initial treatment failure in suppurative kidney infections [4]. These are some examples of various scoring systems available for diagnostic purposes; many others are available. Almost all such scoring systems are based on data from developed countries, which may not directly apply to the subjects in developing countries because of different nutritional and environmental factors. These scoring systems have to be appropriately modified before using them for patients in such countries.

Such systems should be used only when a valid diagnosis is difficult to establish or when the diagnosis depends upon physicians' preferences, their expertise, or results of laboratory or radiological investigations that lack credibility. Such inadequacies in the diagnostic process are more common than are otherwise apparent. A useful strategy is to use these scores as just additional adjunct to the clinical and laboratory evidence, and take a decision in a holistic manner.

### *Scoring for Gradation of Severity*

Prognostic assessment and the management of a patient depend to a large extent on the severity of the disease. There is considerable **epistemic uncertainty** about how to assess this severity. Different professionals use different methods. For uniformity and exactitude, at times, scoring is considered desirable.

The Glasgow Coma Scale [5] is used to grade coma patients by using a numeric scale for eye, motor, and verbal response. The **APACHE score** is used to assess severity in critically ill hospitalized adults [6]. Various variations of this score are available. The **Apgar score** is used to assess prognosis in a neonate. In case you are stuck with a problem of grading the severity of patients, see if a valid scoring system is available. If not, you may like to devise a scoring system yourself. One example of a scoring system for gradation of severity is as follows.

Chagas disease is an important health problem in Latin America, and cardiac involvement in this disease increases the severity and risk of death. Rassi et al. [7] developed a risk score based on the evaluation of 424 outpatients from a regional Brazilian cohort as given in Table S.9. The score is obtained as a sum of these points. They divided the score into three groups: low risk, 0–6 points; intermediate risk, 7–11 points; and high risk, 12–20 points. The risk pertains to the risk of death. The 10-year mortality rates for these three groups were 10%, 44%, and 84%, respectively. For an example of clinical risk-scoring algorithm to forecast scrub typhus severity, see Sriwongpan et al. [8].

As already stated, all scoring systems try to convert multiple measurements into a single unified but meaningful index. They transform multivariate data into a univariate score. Sometimes several scoring systems are available for the same condition, and it would be difficult to choose the right system. For example, severity in peritonitis cases can be assessed by APACHE score, Peritonitis Severity Score, Mannheim Peritonitis Index, Hacettepe Score, American Society of Anesthesiologists Score, etc. Choose a scoring system that looks more appropriate for your patients.

**TABLE S.9**
**Scoring System for Death in Chagas Disease**

| Risk Factor | Points |
|---|---|
| New York Heart Association Class III or IV | 5 |
| Evidence of cardiomegaly on radiography | 5 |
| Left ventricular systolic dysfunction on echocardiography | 3 |
| Nonsustained ventricular tachycardia on 24 h Holter monitoring | 3 |
| Low QRS voltage on electrocardiography | 2 |
| Male sex | 2 |


1. Billewicz WZ, Chapman RS, Crooks J, Day ME, Gossage J, Wayne E, Young JA. Statistical methods applied to the diagnosis of hypothyroidism. *Q J Med* 1969; 38:255–66. http://qjmed.oxfordjournals.org/content/38/2/255
2. Zulewski HK, Muller B, Exer P, Miserez AR, Staub J. Estimation of tissue thyroidism by a new clinical evaluation of patients with various grades of hypothyroidism and controls. *J Clin Endocrinol Metab* 1997; 82:771–6. http://press.endocrine.org/doi/full/10.1210/jcem.82.3.3810, last accessed November 3, 2015.
3. Xu B, Min Z, Cheng G, Mi Y, Tong N, Feng N, Song N et al. Evaluating possible predictor of prostate cancer to establish a scoring system for repeat biopsies in Chinese men. *J Ultrasound Med* 2011; 30:503–8. http://www.jultrasoundmed.org/content/30/4/503.full
4. Stojadinovic MM, Milovanovic DR, Gajic BS. Scoring system development and validation for initial treatment failure in suppurative kidney infections. *Surg Infect (Larchmat)* 2011; 12:119–25. http://www.ncbi.nlm.nih.gov/pubmed/21545280
5. Jennnett B, Teasdale G, Braakman R, Minderhoud J, Heiden J, Kurze T. Prognosis of patients with severe head injury. *Neurosurgery* 1979; 4:283–9. http://www.ncbi.nlm.nih.gov/pubmed/450225
6. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, Sirio CA et al. The APACHE III prognostic system: Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* 1991; 100:1619–136. https://apachefoundations.cernerworks.com/apachefoundations/resources/APACHE%20III%20Chest%201991.pdf
7. Rassi A Jr, Rassi A, Little WC, Xavier SS, Rassi SG, Rassi AG, Rassi GG, Hasslocher-Moreno A, Sousa AS, Scanavacca MI. Development and validation of a risk score for predicting death in Chagas heart disease. *N Engl J Med* 2006; 355:799–808. http://www.nejm.org/doi/pdf/10.1056/NEJMoa053241
8. Sriwongpan P, Krittigamas P, Tantipong H, Patumanond J, Tawichasri C, Namwongprom S. Clinical risk-scoring algorithm to forecast scrub typhus severity. *Risk Manag Healthc Policy* 2013 Dec 16;7:11–7. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3872011/


## screening trials, *see* clinical trials

## scree plot/test

Proposed by Cattel [1] in 1966, a scree plot is a depiction of cumulative contribution of the factors or principal components in a multivariate

data to the total variance. In **principal component analysis** (PCA) and **factor analysis**, the important components and the factors are identified by studying how much of the total variation they explain. Scree plot is used as a visual test to decide how many factors or principal components have substantial contribution, and these are then studied in detail. This is done by identifying natural breaks as revealed by sudden flattening in the plot. Sometimes the fall is gradual, and no clear answer to the number of appropriate factors is available; but in many situations, a scree plot has proved to be of substantial help. The name *scree* comes from the rubble that accumulates at the foot of a mountain as is evident from Figure S.3.

Suppose the percent contribution of the various factors in a factor analysis is as shown in Table S.10. The scree plot for this is shown in Figure S.3. Nearly 35% of the variance in this example can be considered uncovered as the contribution of the sixth factor itself is just 2%. There is no sudden leveling, and it is difficult to say how many factors would be sufficient in this case.

An alternative is to examine the numerical values of the percentage contribution of the factors. When this percentage suddenly falls or becomes less than any prespecified threshold, stop looking for other factors. This could be quite subjective. In Table S.10, the first two factors together are able to account for 47% variation, and the contribution of others looks minor. The unexplained part of variation is also depicted by this plot.

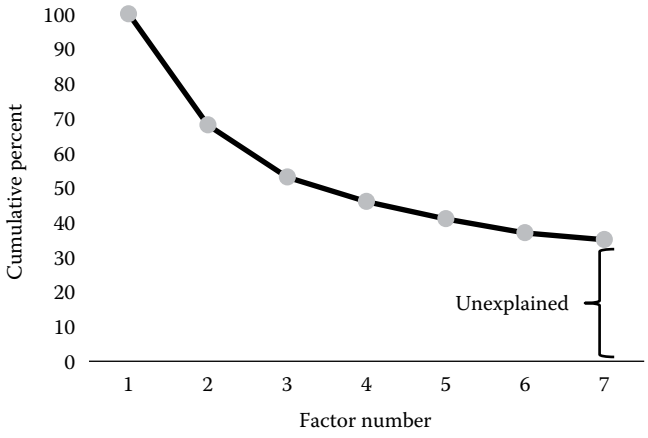



**FIGURE S.3** Scree plot for the data in Table S.10.

**TABLE S.10**
**Percent Contribution of Various Factors in a Factor Analysis**

| Factor | Percent Contribution | Cumulative Percent |
|---|---|---|
| 1 | 32 | 100 |
| 2 | 15 | 68 |
| 3 | 7 | 53 |
| 4 | 5 | 46 |
| 5 | 4 | 41 |
| 6 | 2 | 37 |
| 7+ | 35 | 35 |

1. Cattel RB. The scree test for the number of factors. *Mult Behav Res* 1966;1:245–76. http://garfield.library.upenn.edu/classics1983/A1983PY50000001.pdf, last accessed September 14, 2015.

## seasonal trend, *see* time series

## secondary data, *see* data sources (primary and secondary)

## secular trend, *see* time series

## semi-interquartile range, *see* variation (measures of)

## sensitivity and specificity

Tools such as laboratory tests are an integral part of modern medicine. Fine-needle aspiration cytology is done to detect breast cancer, Pap smear for cervical cancer, Western blot for human immunodeficiency virus infection, and chest x-ray for tuberculosis. Often, the sign–symptom syndrome functions as a *test* to form the basis for establishing a diagnosis. Generalized maculopapular rashes and fever with cough are considered indicative of measles, and prolonged acute chest pain indicates myocardial infarction (MI).

The tools used for evaluation and management of health and disease are seldom perfect. All tests are flawed to a degree. These produce correct results in many cases but fail, fully or partially, to perform well in some cases. Healthy individuals are occasionally classified wrongly as ill, sending false alarm, and some individuals who are really ill may not be detected, raising complacency. All such errors need to be controlled because there is a cost involved—cost of unnecessary treatment, cost of side effects, inconvenience, progression of disease, and even death of some patients. The ability of a tool or of a procedure to perform its assigned function correctly is called its **validity**.

A valid diagnostic test would correctly detect the presence as well as the absence of the disease. Some tests are more valid than others, although they may be more expensive. Rectal sonography is considered more efficacious than prostatic specific antigen values for detecting prostate cancer. Scintigraphy gives better results than electrocardiogram (ECG) for MI. However, **gold standards** that give perfect results all the time are rare, and most of them are expensive in time and effort. No medical test is valid in absolute sense. Errors in classification such as misdiagnosis and missed diagnosis occur no matter what test is used.

Malaria is characterized by high fever with chills and rigors, splenomegaly, and a positive blood smear. How valid is this set of criteria? Can it correctly identify all the cases of malaria, and can it correctly exclude all the nonmalarial cases? These two aspects are discussed next.

The ability of a test to give a positive result in true cases of a disease is called *sensitivity*. *Specificity* is the ability to give a negative result in cases where the disease is absent. Both are probabilities, though many times expressed in percentage terms by multiplying by 100. These are two components of validity of a test that measure its inherent goodness. Denote the presence of disease by *D*+ and absence by *D*–, and test positivity by *T*+ and test negativity by *T*–. Also let true positives be abbreviated as TP, false positives as FP, true negatives as TN, and false negatives as FN. The ability of a test to be positive when the disease is present is

$$\text{sensitivity } S(+): P(T+ \mid D+) = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

The ability of a test to give a negative result when the disease is absent is

$$\text{sensitivity } S(-): P(T- \mid D-) = \frac{\text{TN}}{\text{TN} + \text{FP}}.$$

These are best illustrated with the help of an example. Suppose ECG was performed on a total of 700 subjects with complaints of prolonged acute chest pain. Of these, 520 cases were earlier confirmed for the presence of MI and 180 for its absence. The results obtained are given in Table S.11.

Note the following for ECG as a test for the diagnosis of MI in this example:

TP = 416
FP = 9
TN = 171
FN = 104

It is known that 520 subjects had MI and the other 180 did not. Thus,

$$\text{sensitivity of ECG, } S(+) = \frac{416}{520} \times 100 = 80\%,$$

and

$$\text{sensitivity of ECG, } S(-) = \frac{171}{180} \times 100 = 95\%.$$

Thus, in this example, a negative ECG seems sufficiently specific in non-MI cases but not so sensitive in the cases with MI. Many cases of MI, in this example 20%, are missed by ECG, possibly because the required elevation in ST segment does not appear.

The following measures the combination of sensitivity and specificity, when both are equally important:

$$\text{inherent validity of a test} = (\text{TP} + \text{TN})/n,$$

where $n$ is the total number of subjects. In our example, the inherent validity of ECG is $(416 + 171)/700 \times 100 = 84\%$.

**TABLE S.11**
**Myocardial Infarction (MI) and ECG in Cases of Acute Chest Pain**

| | MI | | |
|---|---|---|---|
| **ECG** | **Present** | **Absent** | **Total** |
| Positive | 416 (TP) | 9 (FP) | **425** |
| Negative | 104 (FN) | 171 (TN) | **275** |
| **Total** | **520** | **180** | **700** |

### *Features of Sensitivity and Specificity*

A difficulty with the concepts of sensitivity and specificity is that they can be evaluated only when the presence or absence of the disease is known. If the disease status is already known, where is the need for a test? The concepts are still useful as shortly explained. The following points deserve attention:

1. The true diagnosis is evaluated on the basis of more refined methods, a gold standard, that may be far more difficult to adopt. Often, the real diagnosis emerges after the passage of time, for instance, on response to therapy or by autopsy. Many times, a surrogate is used as a gold standard such as histological evidence for cancer. If the gold itself is a bit shoddy, a good sensitivity or specificity may give a false sense of security.
2. The values of sensitivity and specificity can be changed by altering the criterion for positivity. In our MI example, if the creatine phosphokinase (CPK) level is also considered in addition to ECG, then some other values of sensitivity and specificity are obtained. The exact values will depend on what threshold of the CPK level is chosen to indicate MI. Different thresholds will give different values of sensitivity and specificity.
3. Generally speaking, an alteration in the criteria will either increase or decrease sensitivity but will affect the specificity in an opposite manner. That is, an increase in sensitivity is generally accompanied by a decrease in specificity, and vice versa.
4. It is assumed in these calculations that the disease and the test can be clearly classified as present/absent or negative/positive. In practice, you can have a category "may be" or "indeterminate" or "±." The calculations then become extremely complex and raise many additional issues that are outside the scope of the present book. Genuine cases with equivocal or uninterpretable test results can be excluded. In practice though, such results are seen to have some association with disease state, and the bias can creep in even when they are excluded.
5. As for almost any other criterion, values of sensitivity and specificity are valid for the type of cases actually used for calculation. If only mild cases are included, the values would not be valid for severe cases. For example, large or advanced tumors are easily picked up. You may like to include appropriate proportion of cases with different spectrum of disease to get more representative estimates of the sensitivity and the specificity. But the number in each category must be sufficient for estimates to be reliable.
6. Comorbidities may also affect the values. For example, NESTROFT (naked eye single tube red cell osmotic fragility test), used for screening thalassemia in children, shows good sensitivity in patients without any other hemoglobin disorder but also produces a positive result when other types of hemoglobinopathies are present.
7. The subjects without disease should be similar to those encountered in practice. They should be subjects with initial suspicion because otherwise the question of testing does not arise.
8. Another condition is that prior knowledge of presence or absence of disease should not affect the performance of the test. That is, the person interpreting the test should not consider the disease status. Better still if the person is kept blind to the disease status.

9. Guard against possible interobserver variation if more than one observer is used. This is particularly so when observer abilities are important in diagnosis such as for bone density assessment through magnetic resonance imaging by an experienced radiologist versus the one by a junior radiologist.
10. If a really rare (e.g., one in a million) disease is being studied, a positive result even with a good test is likely to be false on somebody in the remaining 999,999, and the calculation of sensitivity and specificity may be misleading.

Relation between sensitivity and specificity at different values of a quantitative test is studied by **receiver operating characteristic (ROC) curve**.

## sensitivity analysis, *see also* uncertainty analysis

The sensitivity analysis refers to the study of effect of changes in the basic premise such as individual and societal preferences or assumptions made at the time of model development. These assumptions are made to plug the knowledge gaps, which we call **epistemic uncertainties**. Thus, sensitivity analysis establishes or refutes robustness of the model. It is sometimes confused with **uncertainty analysis** and **confidence intervals**. The difference can be explained as follows.

In almost all practical situations, parameters of a statistical distribution of a medical measurement are not known and have to be estimated based on sample values. These sample estimates are point estimates and are subject to **sampling fluctuations**. This is measured by their standard error and reflected in the corresponding confidence interval (CI). This is the **aleatory** part of what is called the **parameter uncertainty**. In addition, there is epistemic uncertainty arising from our incomplete knowledge about the processes, which mostly remains guesswork and is determined by the experience of the experts, although some experts are able to postulate a whole distribution of parameter values that would incorporate both aleatory and epistemic parts. Including uncertainty to model parameters provides a range of prediction error that would be substantially wider than the statistical CIs. For studying the effect of parameter uncertainty on model outputs, consider both systematic increments as well as random perturbations.

Models, particularly statistical models, are not unique. Different sets of parameters may reasonably reproduce the same sample values. Thus, agreement between the model and the observed data does not imply that the model assumptions accurately describe the underlying process. Only that it is one of the plausible explanations and is empirically adequate. Thus, the model needs to be checked under varying conditions as stipulated under sensitivity analysis.

In contrast to this, uncertainty analysis is the process of measuring the impact on the result of changing the values of one or more key inputs about which there is uncertainty. Thus, this mostly arises due to the use of sample estimates. The inputs are varied over a reasonable range that can practically occur. In essence, sensitivity analysis is primarily for epistemic uncertainties, whereas uncertainty analysis is mostly for aleatory uncertainties.

For example, in the calculation of **disability adjusted life years (DALYs)** done initially by the World Health Organization, the assumption that death in young age is much more important than death in childhood or old age is a value choice and a suitable candidate to be examined by sensitivity analysis. In addition to such assumptions, variables used for developing a model also are basic to the model. These also depend on the choice of the investigator. A simple example of a model is systolic blood pressure studied to depend on age, sex, and body mass index in healthy subjects. Another choice could be age, sex, and socioeconomic status. These choices are deterministic rather than stochastic. More accurate measurement or scientifically sound methodology does not help alleviate this uncertainty. Sensitivity analysis varies these basic conditions and verifies that the broad conclusion still remains the same. Thus, this may be called an exercise in external **validation**.

The following example illustrates another aspect of sensitivity analysis. Risk of coronary disease can be modeled to depend upon the presence or absence of diabetes, hypertension, and dyslipidemia. This model might be able to correctly predict 10-year risk in 62% of the cases. However, addition of smoking and obesity can increase this to 70%. This addition of 8% is substantial. Thus, predictivity of coronary disease is *sensitive* to the choice of risk factors. The first model is based on three risk factors and the second on five factors. Had the contribution of smoking and obesity been only 2% or 3%, the conclusion would be that prediction of coronary disease is insensitive to smoking and obesity when diabetes, hypertension, and dyslipidemia are known. Then robustness of the smaller model would have established.

The sensitivity analysis investigates how the result is affected when the basic premise is altered. Unlike uncertainty analysis, this has nothing to do with the variation in the input values. Input factors themselves are changed. Sensitivity analysis deals with uncertainty in model structure, assumption, and specification. Thus, it pertains to uncertainty arising from what parameters are included and what are excluded.

The purpose of sensitivity analysis is to examine whether the key result continues to point to the same direction when the underlying structure is altered within a plausible range. The process involves identifying key outcome in the first place, and then the basic inputs that can affect this outcome. For example, in a clinical trial setup, the outcome could be mortality or the length of hospital stay. Both should generally lead to nearly the same conclusion. Thus, the outcome measure can also be changed to see if the results are still the same. Patients' inclusion and exclusion criteria can be relaxed, the method of assessment can be altered, and even the data analysis methods can be changed to see if this affects the final result. Intention-to-treat analysis can be tried in addition to the regular per protocol analysis that excludes the missing or distorted data. If the results do not materially change, the confidence in the results strengthens. The following example further illustrates this concept.

Estenssoro et al. [1] compared Argentinean patients on mechanical ventilation in intensive care unit (ICU) for more than 21 days ($n = 79$) with those with less ventilation ($n = 110$) for severity score, worst $PaO_2/FIO_2$ fraction, presence of shock on ICU admission day, length of stay in ICU, and length of stay in the hospital. Logistic regression identified shock on ICU admission day as the only significant predictor with an OR = 3.10. This analysis excluded patients who died early. It was not known that this result would or would not hold for patients dying early. To bridge this epistemic gap, the authors conducted a sensitivity analysis by including 130 patients who died early. Shock remained a powerful predictor that gives the conclusion that the only prognostic factor for prolonged mechanical ventilation is shock on ICU admission day irrespective of early or late death (or survival). Perhaps shock itself is a by-product of severity of illness and hypoxemia.

1. Estenssoro E, Gonzalez F, Laffaire E, Canales H, Sáenz G, Reina R, Dubin A. Shock on admission day is the best predictor of prolonged mechanical ventilation in ICU. *Chest* 2005;127:598–603. http://journal.publications.chestnet.org/data/Journals/CHEST/22021/598.pdf

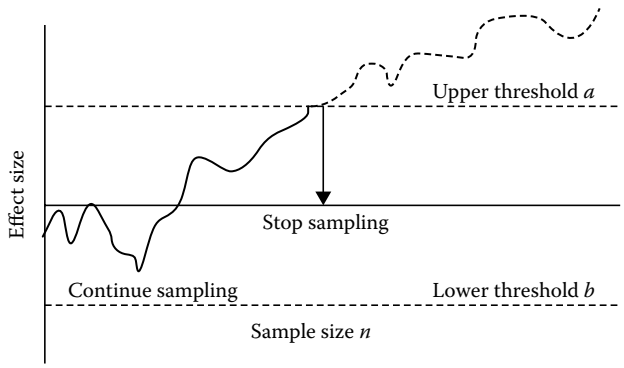



**FIGURE S.4** Illustration of sequential analysis.

## sequential analysis

This is the method of analysis of data generated by **sequential sampling**; that is, units or groups of units are selected one after the other, and the null hypothesis is tested every time on the basis of the accumulated data. The procedure stops either when the null is rejected or when a predetermined size of sample is reached. The procedure looks at medium- to long-term trend and may not be affected by a few outliers. Because of repeated testing of hypothesis, this procedure raises questions about controlling the **Type I error**. The method of sequential analysis, developed by Abraham Wald in 1945 [1], helps in this respect.

Sequential analysis tries to minimize the sample size for a fixed level of significance α and statistical power (1 – β). Wald worked out that the best approximation for the lower threshold is $a = \ln[\beta/(1-\alpha)]$, and that for the upper threshold is $b = \ln[\beta/(1-\alpha)]$. Incidentally, these give $\alpha = \frac{1-e^{a}}{e^{b}-e^{a}}$ and $\beta = \frac{e^{-b}-1}{e^{-b}-e^{-a}}$, just in case you already have thresholds $a$ and $b$, and want to see what will be the values of α and β for those thresholds.

The procedure is illustrated in Figure S.4 where the effect size obtained at each stage of sampling is plotted along with the upper and lower thresholds. Further sampling is stopped as soon as the effect size crosses any of these limits. In this figure, the effect size obtained after crossing the upper threshold is shown by the dotted line if the sampling continues.

For example, a clinician wishing to compare the effects of a regimen with a placebo may like to stop the trial if, at any stage, a convincing difference can be demonstrated on the basis of the available data. The procedure can also be used to find if the available data are adequate for a valid decision, as investigated by Egerup et al. [2] after a systematic review of randomized trials on the effect of intravenous immunoglobulins in women with recurrent miscarriages. They observed that insufficient information has been accrued so far for any convincing evidence in favor of the effect.

For further details of sequential analysis, see Lai [3] and Bakeman and Quera [4].

1. Wald A. Sequential tests of statistical hypotheses. *Ann Math Stat* 1945;16 (2):117–86. http://www.jstor.org/stable/2235829
2. Egerup P, Lindschou J, Gluud C, Christiansen OB, ImmuReM IPD Study Group. The effects of intravenous immunoglobulins in women with recurrent miscarriages: A systematic review of randomised trials with meta-analyses and trial sequential analyses including individual patient data. *PLoS One* 2015 Oct 30;10(10):e0141588. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0141588
3. Lai TL. Sequential analysis: Some classical problems and challenges. *Stat Sin* 2001;11:303–408. http://www3.stat.sinica.edu.tw/statistica/oldpdf/A11n21.pdf
4. Bakeman R, Quera V. *Sequential Analysis and Observational Methods for the Behavioral Sciences*. Cambridge University Press, 2011.

## sequential sampling

This kind of sampling is generally used for accepting or rejecting a lot on the basis of the number or percentage of "defectives" in the sample. One unit or a group of units is selected at a time from the lot, and the decision regarding rejecting or accepting it is made with reference to prespecified criteria based on the analysis of the accumulated values available at each stage; the sampling continues if no decision one way or the other can be taken. In most situations, a random sample is taken at each stage, although sequential sampling is more convenient when the units are coming in a queue such as patients in a hospital ward. Compare this with the usual procedure where a fixed sample is available and a decision regarding rejecting or accepting a lot is made with no third option. In case of sequential sampling, the third option to continue sampling is available. When one unit at a time is selected, this is called *unit-by-unit sequential sampling*, and when a group of units is selected each time, it is called *group sequential sampling*.

The advantage of sequential sampling is that a start can be made with a small sample, and a decision is made without wasting time and effort as soon as the adequate evidence accumulates one way or the other. In addition, in some situations, the researcher can do minor changes in subsequent sampling if needed to improve the research method and analysis. In most situations, this kind of sampling results in a smaller sample than the fixed sample approach. However, the big problem is the inconvenience of choosing a sample each time and undergoing a repetitive process that could be frustrating in some situations. The sample size in this case is a random variable as it is not fixed and depends on chance. Another problem is that the time elapsed between observations may alter the values due to natural changes or some unanticipated sudden changes. Since the sample size is not fixed, a proper planning of time schedule and resources required cannot be done. Despite such severe drawbacks, sequential sampling can still be recommended, where getting observations is very expensive, provided they can be quickly collected so that the elapsed period will not be able to make any difference in the values.

Consider a lot of testing kits, some of which can be defective; suppose that the lot would be accepted if not more than 1% are defective, that the lot would be rejected if more than 2% are defective, and that sampling would continue if the defective is between 1% and 2%. Statistically, the limits that give 95% assurance in large samples are $p_L - 1.645*\sqrt{(p_L q_L/n)}$ and $p_U + 1.645*\sqrt{(p_U q_U/n)}$, where $p_L$ and $p_U$ are the lower and upper limits, respectively, which are fixed in advance. These are $0.01 - 1.645 \times \sqrt{(0.01 \times 0.99/n)}$ and $0.02 + \sqrt{(0.02 \times 0.98/n)}$, which will change with $n$, and a plot of something like Figure S.5 will guide about acceptance, rejection, or continuation of sampling.

Sequential sampling has applications to the test of hypothesis setup also. However, in this case, the null hypothesis is never accepted, and the sampling continues until such time that we are able to reject the null or reach our predefined maximum sample. If the null cannot be rejected even with our maximum sample, the conclusion is that there is no sufficient evidence against the null. This method is sometimes used for **clinical trials** where the participants

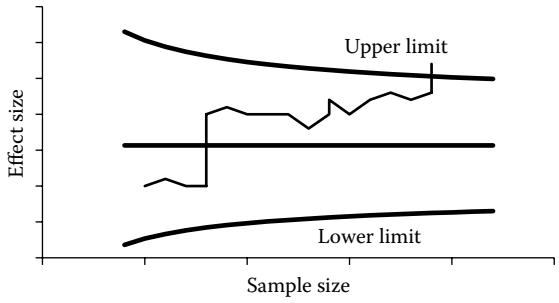



**FIGURE S.5** Decision in sequential sampling.

are sequentially added and the hypothesis regarding the effect of the regimen is tested. This is also used in **meta-analysis** where studies are sequentially added to the analysis to come to a firm conclusion. The difficulty in this procedure is that special methods are required to control the Type I error in view of repeated testing of hypothesis. For this purpose, the method of **sequential analysis** is used.

The term *sequential sampling* has been used in medical literature with varying meaning; there are not many examples of its use in the statistical sense we have presented in this section. Among some of the examples is that by Hirschfeld et al. [1], who used the sequential sampling approach to find a sample size that provides stable estimates of the odds ratios (ORs) in multiple logistic regression. For this, they added participants one by one to the dataset and computed ORs by running logistic regression every time. They had 10 regressors and found that all the ORs tend to stabilize at $n = 1000$ in their data.

1. Hirschfeld G, Wager J, Zernikow B. Physician consultation in young children with recurrent pain-a population-based study. *Peer J* 2015 Apr 28;3:e916. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419529/

## sequential trials, *see* group sequential designs

## serial correlation, *see* autocorrelation

## sex ratio

Sex ratio is conventionally the number of males per 100 females in a population or its segment. This innocuous-looking indicator has extensive implications, particularly for humans because humans tinker with the naturally determined ratio. Theoretically, the ratio should be 1:1 in a natural setting such as random mating and equal exposures, but it varies widely from area to area.

The sex ratio at birth in humans in the absence of any intervention has been observed 105 males per 100 females. This is widely observed across nations and considered "natural." The reason for higher male births is speculated to be the higher motility of X chromosomes that is relatively quicker to fertilize the female ovum compared with the Y chromosome, and is surmised to be the nature's response to higher mortality of males later in life due to greater exposure to environmental hazards. There are unfortunate human interventions that disturb the sex ratio at birth, primarily sex-selective abortions because of preference for sons and prejudices against females in some societies such as in India and China. According to an estimate, the ratio at birth in China in the 1990s was 120:100 and was nearly 110:100 in 2015 [1], and in India it was 110:100 in 2011 [2].

Despite the higher ratio for males at birth, this steeply declines in old age because of higher life expectancy of females almost everywhere in the world. The ratio declines to as much as 45:100 in old age people of Russia [1]. The overall sex ratio in the population is fairly even in most developed countries but is adverse to females in many developing countries. For example, in India, it was 106 males per 100 females in 2011.

Many people consider sex ratio as an important indicator of social health of a population, and an equal ratio is considered to indicate a good social health. Sex-selective abortions are decried not just because they skew the sex ratio but also because this practice jeopardizes the health of the women. Gender prejudices affect the health of the women in several other subtle ways.

Sex ratio is also used to describe the relative incidence of some diseases such as bladder cancer, which is more common in males, or possible skewed sex ratio in spontaneous abortions.

1. About Education. *Sex Ratio*. http://geography.about.com/od/populationgeography/a/sexratio.htm, last accessed November 30, 2015.
2. Census of India. *SRS Statistical Report 2013*. http://www.censusindia.gov.in/vital_statistics/SRS_Reports_2013.html

## Shapiro–Francia test

This is a modification of the **Shapiro–Wilk test** to find if a dataset has a non-Gaussian (non-normal) pattern. Gaussianity is required for many statistical tests such as the Student *t*-test, the analysis of variance *F*-test, and tests for regression coefficients, and these procedures cannot be used straightaway on non-Gaussian data. Although the Shapiro–Wilk test is considered the best omnibus test for this purpose, this test requires coefficients that are difficult to compute. The Shapiro–Francia test is relatively simple and has almost the same **power** as the Shapiro–Wilk test. This is given by

$$\text{Shapiro–Francia test: } W' = \frac{(\Sigma_i z_i x_{[i]})^2}{\Sigma z_i^2 \times \Sigma_i (x_i - \bar{x})^2}, \; i = 1, 2, \ldots, n,$$

where $x_{[i]}$ are the ordered values of $x$ ($x_{[1]}$ being the lowest), and $z_i$ is the ***Z*-score** of the $i$th **order statistics** [1]. *Z*-score is the expected standardized value if the distribution is Gaussian. *W*' can be shown as equal to the square of the Pearsonian correlation coefficient between $x_{[i]}$ and $z_i$, which measures the degree of linear relationship—thus *W*' indicates the straightness of the **normal probability plot**. Small values of *W*' are evidence of departure from Gaussianity, and the *P*-value can be obtained by using an appropriate statistical package. The test was developed by Shapiro and Francia in 1972 [2], and Royston [1] has provided a simple method to evaluate the Shapiro–Francia test. Although Royston [3] later also developed a pocket-calculator algorithm for this test with an application to medicine, the actual usage of this test is still rare, as simplification has lost relevance ever since all calculations are done by computer software packages. Nonetheless, deSouza et al. [4] used this test to check Gaussianity of urethral paravaginal fascial volume and of retropubic urethral length in females with urinary genuine stress incontinence, if you want to see its application.

1. Royston JP. A simple method for evaluating the Shapiro–Francia W' test of non-normality. *The Statistician* 1983;32(3):297–300. http://www.jstor.org/stable/2987935
2. Shapiro SS, Francia RS. An approximate analysis of variance test for normality. *J Amer Stat Assoc* 1972;67:215–6. http://www.tandfonline.com/doi/abs/10.1080/01621459.1972.10481232

3. Royston P. A pocket-calculator algorithm for the Shapiro-Francia test for non-normality: An application to medicine. *Stat Med* 1993 Jan 30;12(2):181–4. http://www.ncbi.nlm.nih.gov/pubmed/8446812
4. deSouza NM, Daniels OJ, Williams AD, Gilderdale DJ, Abel PD. Female urinary genuine stress incontinence: Anatomic considerations at MR imaging of the paravaginal fascia and urethra initial observations. *Radiology* 2002 Nov;225(2):433–9. http://www.ncbi.nlm.nih.gov/pubmed/12409577

## Shapiro–Wilk test

The Shapiro–Wilk test is considered the best test so far for testing non-Gaussian distribution of the data in hand and is quite commonly used. This was designed by Shapiro and Wilk in 1965 [1] and is calculated as follows:

$$\text{Shapiro–Wilk test: } W = \frac{\left(\Sigma_i a_i x_{[i]}\right)^2}{\Sigma(x_i - \bar{x})^2},$$

where $x_{[i]}$ are the ordered sample values from the lowest to the highest, and the $a_i$'s are constants generated from the means, variances, and covariances of the order statistics of a sample of size $n$ from a normal distribution. Small values of $W$ indicate that the distribution is non-Gaussian, and the $P$-value can be obtained by using an appropriate statistical package.

Royston [2] showed via Monte Carlo simulation that the transformed variable $y = (1 - W)^{\lambda}$, where $\lambda$ is determined by sample size $n$, is approximately Gaussian. He also gave a method of evaluating $W$ and its statistical significance for sample sizes between 3 and 2000. As almost any other statistical test, this test can be used to detect non-Gaussianity but not for confirming Gaussianity, although many researchers do not make any distinction between failure to detect non-Gaussianity and confirming Gaussianity. The Shapiro–Wilk test focuses on lack of symmetry, particularly around the mean, and is not much sensitive to differences toward the tails of the distributions. Tails may not be as important either because of extremely low frequencies.

Tiwari et al. [3] used this test for checking non-Gaussianity of fear assessment picture scale and modified facial affective scale in 60 children of age 6 to 8 years who were visiting a dental hospital in India and needed pulpectomy treatment. De Lima et al. [4] used this for testing non-Gaussianity of plasma concentrations of triglycerides, lactate, and glucose in adults in Brazil.

1. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika* 1965 December;52(3/4):591–611. http://links.jstor.org/sici?sici=0006-3444%28196512%2952%3A3%2F4%3C591%3AAAOVTF%3E2.0.CO%3B2-B
2. Royston, P. An extension of Shapiro and Wilks's W test for normality to large samples. *Appl Stat* 1982;31(2):115–24. http://www.jstor.org/stable/2347973?seq=1#page_scan_tab_contents
3. Tiwari N, Tiwari S, Thakur R, Agrawal N, Shashikiran ND, Singla S. Evaluation of treatment related fear using a newly developed fear scale for children: "Fear assessment picture scale" and its association with physiological response. *Contemp Clin Dent* 2015 Jul–Sep;6(3):327–31. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4549982/
4. de Lima FD, Correia AL, Teixeira Dda S, da Silva Neto DV, Fernandes ÍS, Viana MB, Petitto M et al. Acute metabolic response to fasted and postprandial exercise. *Int J Gen Med* 2015;8:255–60. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4540134/

## side effects, *see* clinical trials (overview)

## sigmoid curve, *see* logit

## signed-ranks test, *see* Wilcoxon signed-ranks test

## significance (statistical)

Significant normally means something of consequence that affects our thoughts or practices. However, the term *statistical significance* is used in the context of **tests of hypotheses** where a statistical test is said to provide a significant result when the corresponding ***P*-value** is less than the predetermined threshold, called the **level of significance**. The procedure generally adopted to check statistical significance is to set up a null hypothesis and an alternative hypothesis, use a test criterion such as chi-square depending on the type of data, and calculate the $P$-value. This is the probability of the sample coming from the hypothesized distribution. If this $P$-value is less than the predetermined level of significance, the null hypothesis is rejected and the result is declared to have achieved statistical significance. It only indicates that the chances of the sample coming from the hypothesized null are too small—consequently, the null is not admissible and rejected. Generally, a threshold of 5% is used as the level of significance, meaning thereby that if the chance of sample values coming from a distribution with a parameter value as specified in the null hypothesis is less than 5%, the result is said to be statistically significant. It is customary to call a result highly significant when such a chance is less than 1%. Failure to get significance does not mean that the effect is not present—only that the amount of data available is not enough to show that statistically significant effect is present.

There is a considerable debate about the actual utility of the concept of statistical significance. See the topic **statistical tests versus confidence intervals** that emphasizes that a statistically significant result does not necessarily imply important or useful result. Note that failure to achieve significance such as of an independent predictor of risk in a multivariable model does not necessarily mean the variable has no clinical importance. Reasons a variable may fail to achieve such significance include insufficient sample size, correlation with some other variable in the model, etc., besides poor design or poor methodology. Many researchers unnecessarily get excited when they achieve statistical significance. Some even would resort to what is called **data dredging** to achieve significance. Statistical significance can always be legitimately achieved by increasing the sample size. For example, the difference of 1 mmHg in mean blood pressure of two groups would be statistically significant when the sample size is, say, 2000 in each group. Similarly, an exceedingly weak correlation coefficient of 0.1 may be statistically significant when the sample size is large. In most situations, statistical significance only says that some effect is present but does not say how much. For "how much," or to test "at least this much," a modification of the test is needed. Also, consider the statistical **power** of the test to be able to detect a specified minimum **medically important effect** when present.

The term *significance* has pervaded so much in empirical research that some journals have reserved it for statistical significance. Since the term *significance* is not the exclusive domain of statistics, and can be of various types, including one being the medical significance, we would suggest that significance based on $P$-values be specified as *statistical* significance. This helps to distinguish it from other types of significance.

One more caution: Statistical significance of a result can occur due to an honest sampling error in the sense that by chance you happen to have a sample that rejects the null hypothesis. When the level of significance is 5%, as is mostly the case, on average, 1 out of 20 results could be significant in error. This can particularly affect results when several statistical tests have been carried out for reaching to a holistic conclusion. Many researchers ignore this aspect of statistical significance. In addition, statistical significance can also occur in error due to incorrect data, faulty design, miscalculations, wrong method of analysis, etc.

Our advice is to consider statistical significance as just an aid to decision and not to use this as the sole criterion. Try to find a plausible biological reason for such significance, and use your common sense. If the results do not appeal to your conscience, there is a high chance that nonsampling error of the types just mentioned may have occurred. Believe in yourself more than the data, but report what you actually obtained—in some cases, a plausible reason is discovered at some time in the future.

## significance level, *see* level of significance, and also significance (statistical)

## significant digits

There is a tendency to overuse decimals in the reporting of results, creating a false sense of accuracy. Perhaps the higher number of decimals is considered to give a scientific look. The fact is that the final result should have only an appropriate number of decimal places, although for this a large number of decimal places should be used for intermediary calculations (a computer will automatically do that).

What is the appropriate number of decimals? A rule for percentages is to have only as many decimal places as are needed to retrieve the original number. As an extra precaution, one more decimal place can be used. For the percentage of subjects, this can be obtained by the following rules.

In percentages,

Maximum one decimal place if $n \leq 99$
Maximum two decimal places if $100 \leq n \leq 999$
Maximum three decimal places if $1000 \leq n \leq 9999$
And so on

It is sometimes desirable to use the same number of decimal places uniformly throughout a report even if $n$ varies from one section of the report to another. The number of decimal places in this case would depend on the highest $n$. If there are 260 subjects belonging to 80 families, the percentages for subjects as well as for families can go up to two decimal places each. In addition, there is a concept of significant digits. This is applicable particularly to the reporting of calculated values with an extremely large denominator. The leading zeroes are ignored while counting the significant digits. The value 0.00720 and the value 0.458 both have three significant digits and have the same accuracy. The values 3.06 and 0.069 have two significant digits each. This concept is particularly useful when the reported values are meant to be used for further calculations. Although we have advised that all reporting should have the same number of significant digits, many times varying digits due to leading zeroes can create the confusion in the minds of ordinary readers.

In expressing quantities other than percentages, such as mean and standard deviation (SD), the following rule is generally adequate: Report one decimal place more than in the original measurements from which mean and SD are calculated. For example, if hemoglobin level (in g/dL) is measured with one decimal place as 11.7, 13.8, etc., the mean and SD should be reported with two decimal places. The number of decimal places in a coefficient, such as in a regression equation, would depend on the numerical magnitude of the quantity with which it is multiplied. The coefficient 1.07 for birth weight in kilograms has the same accuracy as 0.00107 for birth weight in grams.

Sometimes original measurements are also made with excess accuracy. It would be a waste of resources to measure survival time in cancer patients in days and report it as 3.7534 years, or insulin level in a subject to two decimal places. A minute difference in such readings does not affect the validity of the measurement. The same sort of logic can be applied to percentages and means also. It does not matter whether the mean diastolic blood pressure level of a group is 86.73 or 86.54 mmHg. Both could be rounded off to 87 for medical interpretation. Thus, the decimal places in this case do not serve any useful purpose. Instead, they complicate the presentation and interpretation. However, they might be important for trend to show that one number is smaller than the other.

Many rates can be expressed as percentage, but for others, a different multiplier is used for convenience. For example, the death rate due to cervical cancer is stated as 8 per 100,000 women. This is the same as 0.00008 per woman or 0.008%. Siegrist [1], however, found that rates expressed as a frequency (8 per 100,000) are perceived differently than rates expressed as a probability (0.00008).

For rounding off, the following rule may be helpful: If the last digit to be rounded off is 5, make the previous digit even, i.e., 1.15 is rounded off as 1.2 and 3.45 as 3.4. Or, you can decide to stick to the odd digit in place of even digit. The idea is that 5 should go up half the time and down half the time because it is exactly midway. When the last digit is other than 5, rounding off is the nearest previous digit. But something like age in adults is conventionally reported in completed years, ignoring the digit after the decimal even if it is more than 5.

1. Siegrist M. Communicating low risk magnitudes: Incidence rates expressed as frequency versus rates expressed as probability. *Risk Analysis* 1997; 17:507–10. http://onlinelibrary.wiley.com/doi/10.1111/j.1539-6924.1997.tb00891.x/abstract

## sign test

This is a nonparametric test for checking that the **location** (generally the median) of a distribution matches with the one hypothesized, or whether the locations in paired observations are the same. Thus, this test does not talk about the mean, but any location, and median is the most commonly used location for this purpose. The only information required for this test is whether values are higher or lower than the hypothesized value, and the actual values are not needed. This is what makes it a nonparametric test, but this simplification is also responsible for rendering it less powerful compared with the tests that use actual values with known distributions. If quantities are to be considered, a **Student *t*-test** does a much better job under Gaussian conditions, and even if just ranks are available, a **Wilcoxon signed-ranks test** will be more powerful in detecting a difference if present than a sign test. Independence of observations is required for this test just as for most other tests.

A sign test is basically a small sample test since the Student *t*-test can be used taking advantage of the **central limit theorem** for large samples. This test counts the number of positive and negative signs, and uses a **binomial distribution** to find the *P*-value since the probability of positive and negative sign is known under the null

hypothesis. For example, if the null is for median, this probability is $\pi = ½$. If the null is for first quartile, $\pi = ¼$. In case of paired observations, the null is that the values are equally likely to be smaller or larger than the other, that is, $H_0$: $\pi = ½$ for median. If the paired values are $(x_i, y_i)$ $(i = 1, 2, \ldots, n)$, then count the number of pairs $s$ for which the difference $x_i - y_i > 0$ (positive) and find the probability of so many or more positives using a binomial distribution with $\pi = ½$ and given $n$. This will be the $P$-value for the right-tailed test. For the left-tailed test, find $P(S \le s)$, and for two-tailed, the $P$-value is twice the smaller tail value. The critical value can be approximated by $s = \frac{n+1}{2} + 0.98\sqrt{n}$ for small samples at 5% level of significance.

This can also be used to test if the median (or any other quantile) of a population is a particular value $\tilde{\mu}$. Consider an example of selenium intake measured for $n = 9$ male subjects of prostate cancer. Suppose these are (μg/day) 43, 62, 58, 32, 54, 50, 67, 47, and 51. Can the median be taken as 55 μg/day? This null hypothesis can be checked by a Student $t$-test assuming that the distribution is nearly Gaussian. A software package gives $t = 0.988$ and $P = 0.35$—thus, the null cannot be rejected. Now, for sign test, note that out of 9 subjects, 6 have intake less than this hypothesized value of 55 and 3 have more. If the null is true, we expect nearly half of the values greater than this and the other half less than this. From the binomial distribution with $n = 9$ and $\pi = 1/2$, $P(S \le 3) = 0.25$. For a two-tailed test, $P$-value = 2×0.25 = 0.50, and the null cannot be rejected by this method also; but the $P$-values are very different because the sign test is far too approximate in the sense that it does not consider actual values. The approximate critical value is $s = \frac{9+1}{2} + 0.98\sqrt{9} = 2.06$. Neither the number of positives nor the number of negatives is less than this critical value—thus the null cannot be rejected this way also.

Now consider the hypothesis that the third quartile is at least 60 μg/day. This gives $H_0$: $\pi = 0.75$. The number of observations less than or equal to 60 is 7 out of 9 in the same example. The binomial distribution for $n = 9$ and $\pi = 0.75$ gives $P(S \le 7) = 0.10$. This $P$-value also is not small and we concede the null.

## similarity (statistical measures of), *see* measures of dissimilarity and similarity

## simple linear regression, *see also* regression models (basics of), multiple linear regression

A regression is called simple linear when there is only one regressor and one dependent variable, both are quantitative, and the relationship can be graphically expressed by a line. The regression equation in this case is

$$\text{simple linear regression: } y = \alpha + \beta x + \varepsilon,$$

where α is the value of $y$ at $x = 0$, called **intercept**; β is the rate of change in the value of $y$ per unit change in the value of $x$, called **slope**; and ε is the remainder not accounted for by the regression, called error. Linear regression implies that the relationship under investigation is such that $y$ always changes by the same quantity β as $x$ increases from $x$ to $x + 1$, and that is almost always valid for the range of values of $x$ and $y$ under study, although slight extrapolation is allowed. For example, suppose the regression of systolic blood pressure (BP) in mmHg ($y$) on age in years ($x$) is SysBP = 110 + ½*Age that may be valid for healthy adults of age, say, from 20 to 80 years, and not applicable to children or persons of age above 80 years since the trend outside 20–80 years is not known. This regression implies that for each year of age, the systolic level of BP in healthy adults increases by ½ mmHg (or by 5 mmHg every 10 years), say, due to accumulation of atheromatous plaque, and this change is uniform between the age 20 and 80 years. If the rise in BP is faster in old age, the linear regression will not be adequate. If a healthy person of age 60 years has SysBP = 138 mmHg against the expected 140 mmHg as per this regression, the difference –2 mmHg is what we called error. This is not a mistake but a fluctuation around the line. This kind of ordinary regression requires that $x$ is considered fixed, and $y$ is the only **stochastic** variable. In our example, this would mean that there are, say, 4 adults of age 22 years, 7 of age 26 years, 1 of age 30 years, etc., and their systolic level is measured, so that the age is known and the BP is obtained. At each value of $x$, there may be one or more values of $y$; in fact, the theory requires that there is a whole distribution of the values of $y$ for each value of $x$. The value of $x$ is our choice, and the $y$ on the left side is the average for any fixed value of $x$.

Linear regression in most cases is a simplified version of the actual relationship, which is generally more complex and nonlinear. To reiterate, the linear version gives satisfactory results when the dependent and regressor variables have a straight relationship, i.e., where a unit change in the level of the regressor variable *over its entire range* in the target group is accompanied by nearly the same change in the level of the dependent variable. A straight line can then depict the trend over that range. In many situations, this does not hold. The relationship of total leukocyte count and hemoglobin (Hb) level with age in healthy subjects is nonlinear over the age range 0–20 years because these two decline up to age 6–8 years (physiological anemia) and then increase. In case the relationship is nonlinear, the linear regression would depict the linear part of the relationship.

### *Intercept and Slope*

Suppose $x_i$ is the age of the $i$th child in months and $y_i$ is his or her weight (in kg). A simple linear regression for children up to the age of 1 year could be $y_i = 3.4 + 1.1x_i$. This is shown in Figure S.6a. According to this regression equation, if $x_i = 0$, i.e., when age = 0 months, weight $y_i = 3.4$ kg. This is the intercept and, in this example, this can be interpreted as the average birth weight.

Graphically, the intercept is the distance from the origin to the point where the regression line intersects the $y$-axis (Figure S.6). The intercept has a biological meaning in our example, but in many situations, this is just imaginary. In the case of regression of weight on height in children, one cannot think of height = 0. In this case, the regression line would be extended to the $y$-axis to see the intercept. Algebraically, as mentioned earlier, just plug in $x = 0$ in the equation and get the value of the intercept when the regression is simple linear. If you have two or more independent variables (multiple regression), the value of all the $x$'s will be substituted as zero to get the value of the intercept. In case of sample values, this will give the estimate of the corresponding intercept in the population.

The slope is the gradient of the line. The higher the value of β, the steeper the slope, and its negative value implies that the slope is declining ($y$ decreasing with increasing $x$). In our example on weight of children, the value of β is 1.1 kg, which says that as age increases by 1 month, the weight increases on average by 1.1 kg. Since the regression has been obtained only for age 0–12 months, this statement is applicable to only this age group. Had the slope been 1.4,

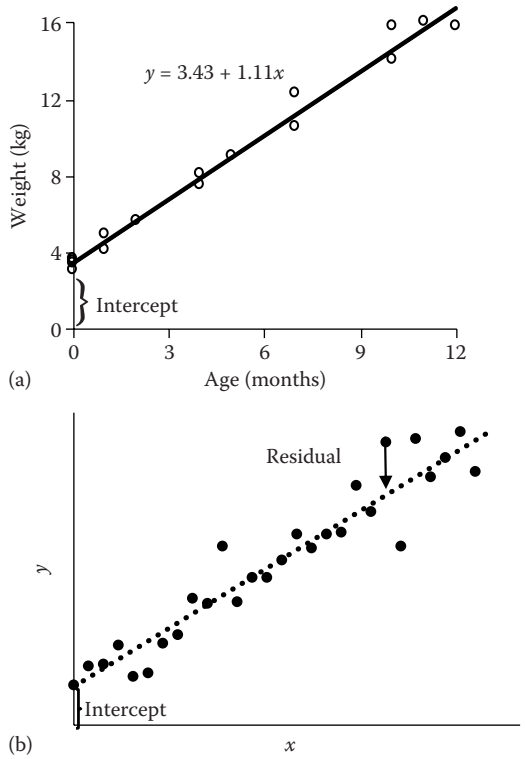



**FIGURE S.6** (a) Simple linear regression of weight on age of children less than 1 year. (b) Simple linear regression of $y$ on $x$ showing scatter plot, regression line, residual, and intercept.

the slope would be steeper, and if it were 0.9, the slope would be gradual. Also note that the interpretation of the slope depends on the units of measurement. In our example, age is in months and weight is in kilograms. If age is in weeks or in years, the value of the slope parameter will accordingly change.

### *Estimation*

The preceding discussion assumes as though the values of α and β are known. This will almost never be the case, and these parameters have to be estimated from the data obtained in a sample of subjects. A **least squares method** is generally used to estimate the values of α and β that finds those estimates that minimize the sum of square of differences between the expected and actual values. Some of these differences are positive and some negative, and the least squares method assures that the sum of these differences is always zero. Expected are those that are determined by the equation thus obtained. This method is nonparametric and does not require Gaussian (normal) distribution of any variable. When these estimates are substituted, the simple linear regression becomes

$$\text{estimated simple linear regression: } y = a + bx + e,$$

where $a$ and $b$ are the point estimates of α and β, respectively, and $e$ is now called the **residual** such that $\Sigma e = 0$. See Figure S.6b for the regression line, intercept, and residuals. For a given set of data, estimates $a$ and $b$ can be calculated as follows:

$$b = \frac{s_{xy}}{s_x^2} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} \quad \text{and} \quad a = \bar{y} - b\bar{x}.$$

Thus, the estimate of the regression coefficient is the covariance between $x$ and $y$ divided by the variance of $x$. In terms of the correlation coefficient, this can be written as $b = rs_x/s_y$, where $s_x$ and $s_y$ are the sample standard deviations of $x$ and $y$, respectively, and $r$ is the correlation coefficient. The regression coefficient and the correlation coefficient have one-to-one correspondence in the sense that if one is negative, the other is also negative; but the correlation coefficient is bound between –1 and +1, while the regression coefficient is unbounded.

### *Confidence Intervals and Tests of Hypotheses for Simple Linear Regression*

So far, we have not used any Gaussian condition, but for finding the confidence intervals (CIs) and tests of hypotheses on the intercept and the slope, we need Gaussian (normal) distribution of the residuals. Since their sum is always zero, we say that $e \sim N(0, \sigma_e^2)$. The lesser the variance, the more reliable the estimates. The condition of Gaussianity can be relaxed for large $n$ because of the **central limit theorem**; otherwise, think of an appropriate transformation to convert the distribution to Gaussian. Independence of values is required anyway. Also, the variance of $y$ at different values of $x$ should be nearly the same, which is called **homoscedasticity**. When these conditions are met, it has been established that

$$b \sim N\left(\beta, \frac{\sigma_e^2}{\Sigma(x-\bar{x})^2}\right) \text{ and } a \sim N\left(\alpha, \sigma_e^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x-\bar{x})^2}\right]\right).$$

Because of these properties, we can find the CI and test the hypothesis using a Student $t$-test after replacing $\sigma_e^2$ with its estimate given by $s_e^2 = \Sigma(y-\hat{y})^2/(n-2)$, where $\hat{y}$ is the regression-predicted value of $y$ given by $\hat{y} = a + bx$. This is popularly known as the mean square error (MSE). For details, see the topic **confidence interval (CI) for regression coefficient and intercept in simple linear regression**. A similar CI can also be obtained for the predicted mean and the predicted value of for any given $x$ within the admissible range. For this, see the topic **confidence interval (CI) for predicted *y* in simple linear regression**, where the standard errors (SEs) are also given. **Confidence band for the regression line** can also be obtained. It is helpful to note that the SEs of $a$ and $b$ have $\Sigma(x-\bar{x})^2$ in the denominator—thus, larger spacing of $x$'s would decrease the SE, and the estimates and conclusions would be more reliable. Wherever possible, widely different values of $x$ should be chosen—this can be done since the choice of $x$ values is with the researcher.

Under the Gaussian and other conditions just stated, statistical significance of the intercept and the slope can be tested by a Student $t$-test using the corresponding SE. This will have the form $t$ = (estimate – hypothesized value)/(SE of the estimate) and will have df = $(n-2)$. For illustration, consider the data in Table S.12 on duration of hospitalization of 10 patients in critical condition and their APACHE score at admission.

**TABLE S.12**
**Duration of Hospitalization and APACHE Score at Admission of Critically Ill Patients**

| Patient No: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| APACHE Score ($x$) | 42 | 21 | 29 | 32 | 26 | 34 | 37 | 32 | 33 | 38 |
| Duration of Hospitalization (Days) ($y$) | 10 | 3 | 5 | 8 | 5 | 8 | 7 | 6 | 8 | 9 |

S

A software package gives the following for the data in Table S.12. These can be manually verified (there might be some minor variations due to rounding off):

$$\text{mean APACHE score } \bar{x} = 32.40, \text{ and}$$

$$\text{mean duration of hospitalization } \bar{y} = 6.90 \text{ days}$$

$$\Sigma(x - \bar{x})^2 = 330.00,\ \Sigma(y - \bar{y})^2 = 40.90,\ \Sigma(x - \bar{x})\,(y - \bar{y}) = 107.40$$

$$b = \frac{107.40}{330.00} - = 0.325, \text{ and } a = 6.90 - (0.325)\times 32.40 = -3.632$$

so that the regression equation is

$$\text{duration of hospitalization (days)} = -3.632 + 0.325^{*}\text{APACHE}.$$

This says that as the APACHE score increases by 1, the duration of hospital stay increases by 0.325 days on average (or nearly 1 day for every rise of nearly 3 in APACHE score) if the relation is really linear.

$$\text{MSE } s_e^2 = \Sigma(y - \hat{y})^2/(n - 2) = 0.749,$$

and this gives

$$SE^2(b) = 0.749/330.00 = 0.00227, \text{ and } SE^2(a) = 0.749\times[1/10 + (32.40)^2/330.00] = 2.547.$$

Thus, with Student $t$ at 8 df = 2.306,

$$95\%\text{ CI for the regression coefficient } \beta \text{ is } [0.325 \pm 2.306 \times \sqrt{(0.00227)}], \text{ or } (0.215, 0.435)$$

and 95% CI for the intercept $\alpha$ is $[-3.632 \pm 2.306 \times \sqrt{(2.547)}]$, or (–7.244, –0.020).

The SEs can also be used to test the hypotheses on $\alpha$ and $\beta$, although the former has very little relevance. Just for completeness, $t = 6.829$ ($P = 0.049$) for $H_0$: $\beta = 0$, and $t = -2.319$ ($P < 0.001$) for $H_0$: $\alpha = 0$. Both the nulls are individually rejected in this example.

## simple matching dichotomy coefficient, *see* association between dichotomous categories (degree of)

S

## simple random sampling

This is the basic method of selecting a sample of specified number of units from a given population. When the scheme is such that each unit of the population has the same chance of being included in the sample, it is called simple random sampling (SRS). A rigorous definition is that all possible samples of the same size have an equal chance of selection. It is customary in sampling to denote the size of the sample by $n$ and the size of the population by $N$. SRS is like picking $n$ slips from a lot containing $N$ lookalike slips numbered 1 through $N$. A more scientific method is to use random numbers, which can be very easily generated on a computer.

SRS seems easy but can turn out to be difficult to implement. A prerequisite is the availability of the **sampling frame**, preparation of which can be a very expensive exercise in many cases. If a study involves several hospitals, each hospital will have a list of its own patients, but a joint list of the patients in all participating hospitals may not be available at one place. In a domiciliary study, it may be difficult to prepare a list of families in an area. The next problem is that the randomly selected units can be physically very far apart residing in different areas, admitted to different hospitals, or being attended to in clinics in different locations. Note that such physical divergence may or may not add to the representativeness of the sample. There is no guarantee that SRS will adequately represent different segments of the target population. Being random, it is possible that the sample happens to include cases of serious or moderate severity but none of mild severity. Or all cases can be adults with no or very little representation of children. If adequate representation of such segments is required, the method of selection should be **stratified random sampling**.

Consider waist–hip ratio (WHR) and triglyceride (TG) data in a population of 100 male hypertensive subjects (Table S.13 on next page). This is the total population in this illustration. An SRS of size 16 may be the subject numbers 3, 18, 21, 31, 33, 45, 49, 53, 59, 62, 66, 71, 78, 79, 91, and 96 as marked by an asterisk in the table. For this sample, the mean ($\bar{x}$) TG level is 153.9 mg/dL, and the sample SD ($s$) is 15.8 mg/dL, whereas the population mean ($\mu$) for all the 100 subjects is 155.5 mg/dL and population SD ($\sigma$) is 17.4 mg/dL. The estimates from this sample are slightly lower than the values of the corresponding population parameters. This can happen with any sample, but in the long run, SRS provides **unbiased estimates**.

In this sample of 16 subjects, only 3 (nearly 19%) have WHR ≥ 1.10, whereas in the population, there are 48% in this category of WHR. Similarly, only 5 (31%) have TG ≥ 160 mg/dL against 43% in the population. There is no guarantee in the SRS that the sample will correctly represent the entire spectrum of subjects. Another sample may provide different values and different representation, called **sampling fluctuation**. However, as the sample size increases, the chances that the sample will be more representative of the features of population steeply increase, including the distribution, and the values of sample mean and SD become more representative. Theoretically, this is true for the average of the repeated samples also, although repeated samples are almost never actually taken. When sampling from a relatively small population, as in our example where $N = 100$, the estimates may require **finite population correction** as discussed under that topic.

## Simpson paradox

Simpson paradox occurs when the results obtained by the whole are different from what you get from the subgroups. This can happen when, for example, the case mix in the two groups is very different—one group has more serious cases and the other group has more mild cases. This imbalance in the subgroups tends to provide a distorted picture of the whole. In Figure S.7, pain score decreases with dose of the drug in both males and females, but when a combined scatter

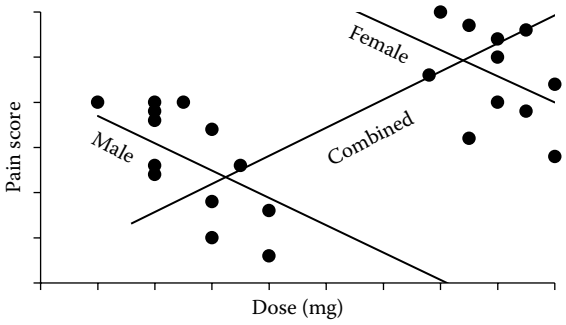



**FIGURE S.7** Relation of pain score with dose of a drug in male and female patients illustrating Simpson paradox.

**TABLE S.13**
**Serum Triglycerides (TG) in mg/dL and Waist–Hip Ratio (WHR) in a Population of 100 Male Hypertensive Subjects**

| Subject Number | WHR | TG | Subject Number | WHR | TG | Subject Number | WHR | TG |
|---|---|---|---|---|---|---|---|---|
| 1 | 1.21 | 152 | 35 | 1.17 | 151 | 68 | 0.93 | 148 |
| 2 | 0.97 | 145 | 36 | 1.41 | 167 | 69 | 1.27 | 162 |
| 3* | 1.52 | 183 | 37 | 0.80 | 131 | 70 | 1.33 | 178 |
| 4 | 0.83 | 138 | 38 | 0.99 | 128 | | | |
| 5 | 1.06 | 147 | 39 | 1.24 | 142 | 71* | 1.09 | 140 |
| 6 | 1.15 | 167 | 40 | 1.07 | 162 | 72 | 1.07 | 138 |
| 7 | 1.44 | 178 | | | | 73 | 1.14 | 169 |
| 8 | 1.23 | 160 | 41 | 1.13 | 148 | 74 | 1.20 | 128 |
| 9 | 1.17 | 180 | 42 | 0.95 | 152 | 75 | 1.27 | 139 |
| 10 | 0.98 | 128 | 43 | 1.26 | 158 | 76 | 1.07 | 172 |
| | | | 44 | 1.06 | 150 | 77 | 0.92 | 136 |
| 11 | 1.89 | 197 | 45* | 1.17 | 157 | 78* | 0.97 | 148 |
| 12 | 1.41 | 162 | 46 | 0.84 | 136 | 79* | 0.98 | 169 |
| 13 | 1.32 | 158 | 47 | 0.91 | 147 | 80 | 1.32 | 173 |
| 14 | 1.04 | 150 | 48 | 0.90 | 162 | | | |
| 15 | 1.35 | 175 | 49* | 1.07 | 157 | 81 | 1.48 | 185 |
| 16 | 0.76 | 161 | 50 | 1.04 | 149 | 82 | 1.27 | 196 |
| 17 | 0.81 | 151 | | | | 83 | 1.08 | 147 |
| 18* | 0.90 | 148 | 51 | 1.73 | 195 | 84 | 1.08 | 173 |
| 19 | 1.44 | 159 | 52 | 1.39 | 187 | 85 | 1.38 | 175 |
| 20 | 1.20 | 167 | 53* | 1.08 | 151 | 86 | 1.40 | 184 |
| | | | 54 | 1.17 | 162 | 87 | 1.01 | 160 |
| 21* | 1.00 | 142 | 55 | 1.10 | 148 | 88 | 1.02 | 151 |
| 22 | 1.37 | 138 | 56 | 1.12 | 138 | 89 | 1.07 | 145 |
| 23 | 1.12 | 149 | 57 | 1.57 | 167 | 90 | 1.15 | 162 |
| 24 | 1.29 | 162 | 58 | 1.31 | 175 | | | |
| 25 | 1.08 | 158 | 59* | 0.94 | 167 | 91* | 1.18 | 171 |
| 26 | 1.34 | 149 | 60 | 0.85 | 145 | 92 | 0.90 | 162 |
| 27 | 1.61 | 182 | | | | 93 | 0.87 | 128 |
| 28 | 0.73 | 120 | 61 | 0.75 | 132 | 94 | 0.95 | 149 |
| 29 | 0.88 | 138 | 62* | 0.66 | 125 | 95 | 0.96 | 160 |
| 30 | 1.17 | 129 | 63 | 1.53 | 192 | 96* | 0.84 | 134 |
| | | | 64 | 1.91 | 169 | 97 | 0.88 | 158 |
| 31* | 1.01 | 138 | 65 | 1.02 | 152 | 98 | 1.15 | 167 |
| 32 | 0.84 | 141 | 66* | 1.07 | 165 | 99 | 1.20 | 162 |
| 33* | 0.93 | 167 | 67 | 0.94 | 129 | 100 | 1.17 | 139 |
| 34 | 0.92 | 129 | | | | | | |

| WHR | | TG | |
|---|---|---|---|
| ≤0.89 | 14 subjects | ≤159 | 57 subjects |
| 0.90–1.09 | 38 subjects | ≥160 | 43 subjects |
| ≥1.10 | 48 subjects | | |
| Population mean ($n$ = 100) 1.127 | | Population mean ($n$ = 100) 155.5 mg/dL | |
| Population SD ($n$ = 100) 0.238 | | Population SD ($n$ = 100) 17.4 mg/dL | |

*In the sample.

is considered, the conclusion would be that the pain score increases with the dose of the drug. This seems like an improbable thing to occur, but this paradox is occurring in this case because females have much higher pain score than males.

Edward Simpson was the first to formally address this phenomenon in 1951 [1], although the phenomenon was known earlier. This paradox has posed a dilemma of whether aggregated results or the disaggregated ones should be considered for action. Our example illustrates that the disaggregated results provide more correct assessment. In many situations, disaggregated data are not available, and the aggregated results are considered true, particularly when it is not known what stratifying characteristic should be used for disaggregation. The paradox also highlights why some statistical associations cannot be taken on face value.

Marang-van de Mheen and Shojania [2] have discussed this phenomenon in the context of standardized mortality ratios. **Standardization** with respect to the "culprit" characteristic of the two groups is the remedy in such situations. For further details of Simpson paradox, see Pearl [3].


1. Simpson EH. The interpretation of interaction in contingency tables. *J Royal Stat Soc Series B* 1951;13: 238–41. http://www.jstor.org/stable/2984065
2. Marang-van de Mheen PJ, Shojania KG. Simpson's paradox: How performance measurement can fail even with perfect risk adjustment. *BMJ Qual Saf* 2014 Sep;23(9):701–5. http://qualitysafety.bmj.com/content/23/9/701.full.pdf+html
3. Pearl J. Understanding Simpson's paradox. Technical Report UCLA R-414 December 2013. http://ftp.cs.ucla.edu/pub/stat_ser/r414.pdf


## simulation studies, *see* Monte Carlo methods

## single linkage method of clustering

See **cluster analysis** for basic details of this kind of analysis. Single linkage is one of the several methods of hierarchical clustering. In hierarchical clustering, a measurement of dissimilarity such as **Euclidean distance** is used to classify the units into various groups using one of the several possible algorithms. Two units (or subjects) that are most similar (or least distant) are grouped together in the first step to form one group of two units. This group is now considered as one entity. Now the distance of this entity from other units is compared with the other distances between various pairs of units. Again, the closest are joined together. This hierarchical agglomerative process goes on in stages, reducing the number of entities by one each time. The process is continued until all units are clustered together as one big entity. See **hierarchical clustering** for the method to decide when to stop the agglomerative process so that natural clusters are obtained.

The primary problem in clustering is in computing the distance between two entities containing, say, $n_1$ and $n_2$ units, respectively. Several methods are available. The first is to consider all units in an entity centered on their average. Another method is to compute the distance of the units that are farthest in the two entities. A third method is to base it on the nearest units. There are several others. Single linkage is one such method.

In the single linkage method of clustering, the distance between two clustered entities is measured by the distance between the nearest located units of the entities. If one entity has 7 units with values denoted by $(x_1, x_2, \ldots, x_7)$ and the other has 4 units with values denoted by $(y_1, y_2, y_3, y_4)$, the first step would be to calculate the distance between $x_1$ and $y_1$, $x_1$ and $y_2$, …, $x_2$ and $y_1$, …, $x_7$ and $y_4$. The minimum of these 28 distances will be considered as the distance between these two entities (Figure S.8). If entity A contains $I$ units $(a_1, a_2, \ldots, a_I)$ and entity B contains $J$ units $(b_1, b_2, \ldots, b_J)$, then, under the single linkage method, the distance between these two entities is measured as $d_{AB} = \min_{ij}(dij)$, where $d_{ij}$ is the distance between the $i$th unit of the first entity and the $j$th unit of the second entity. This method can be easily extended to a multivariate setup.

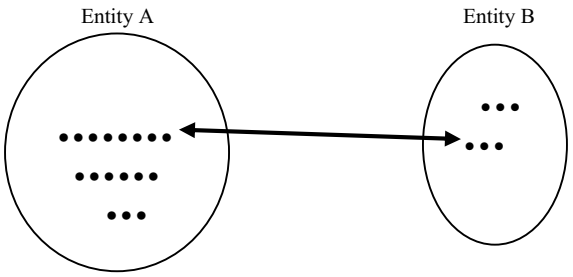



**FIGURE S.8** Single linkage distance between entity A and entity B.

A large distance indicates that the entities are really very different from each other and thus should not be clustered together. If this distance is small, the entities can be considered similar, and you can merge these two entities together to form a bigger entity.

This type of method seems more appropriate for clustering genomes as advised by DeLuca et al. [1] since the objective is to look for the closest similar genome, and for protein sequencing as mentioned by Petryszak et al. [2].


1. DeLuca TF, Cui J, Jung JY, St Gabriel KC, Wall DP. Roundup 2.0: Enabling comparative genomics for over 1800 genomes. *Bioinformatics* 2012 Mar 1;28(5):715–6. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3289913/
2. Petryszak R, Kretschmann E, Wieser D, Apweiler R. The predictive power of the CluSTr database. *Bioinformatics* 2005 Sep 15;21(18):3604–9. http://bioinformatics.oxfordjournals.org/content/21/18/3604.long


## six-sigma methodology

This methodology is used for **quality control** of products including medical products where quality issues are in any case paramount as they deal with life and death. Sigma (σ) is the notation for standard deviation (SD), and it is well known that three-sigma on either side of the mean in case of Gaussian (normal) distribution covers 99.7% of the probability and leaves out only 0.3% probability. Use it for errors and realize that these limits leave out only 0.3% chance of error as uncovered. Extend it to six-sigma on either side of the mean and that tolerates only 3.4 errors per a million opportunities. Notice how intolerant this is compared to the usual 95% confidence intervals defined by ±1.96SE that leaves out 5% chances of error.

In fact, only 4.5σ leaves out a chance of 3.4 errors per million opportunities, but an additional 1.5σ shift on either side (Figure S.9) is allowed as an abundant precaution for possible shift in the process in the future that can unknowingly happen due to subtle changes in the conditions under which a process operates.

Health care is a big industry, and quality issues cannot be sidelined. Medication errors, diagnostic errors, prognostic errors, etc., are quite common as we deal with a highly variable entity called human beings. Any manager in any industry can easily realize how difficult it is to achieve an error rate of less than 3.4 per million, more so in health care. A motivated team is required to identify areas of improvement and devise strategies to tackle them. Patient satisfaction surveys would be an integral part of this exercise. Statistically, this may require studies on cause and effect, and development of tree diagrams and flow charts, besides study of error distributions with its mean and variance, confidence intervals, and tests of significance. The process requires comprehensive and coordinated efforts from all the concerned departments such as administrators, managers, physicians, and technicians right from the stage when a product or a service is conceived. Six-sigma is generally implemented as a DMADV (define, measure, analyze, design, verify) process that can improve the system and can produce products and services with such high-quality levels.

Six-sigma is not just a statistical concept, and a caution is required when nonexperts use this methodology. For example, they may not realize the importance of a large sample size and instead may have unacceptably low numbers for monitoring that could jeopardize

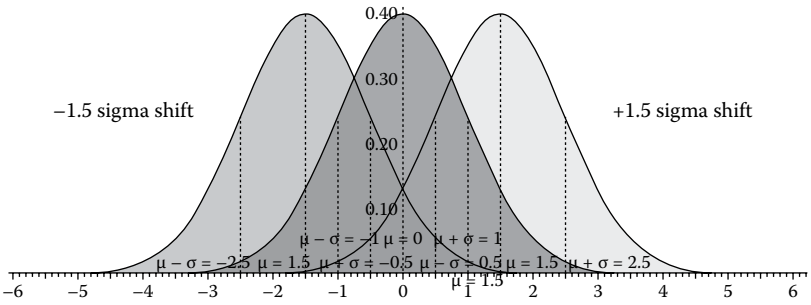



**FIGURE S.9** 4.5σ limits of a Gaussian distribution and 1.5σ shift on either side.

Gaussian application. In addition, errors such as ill-placed control limits (see **quality control**) and wrongly chosen factors can provide misleading results and may push some manager to consider six-sigma methodology as a nuisance. When properly used with well-meaning intentions, the methodology has enormous potential to improve various aspects of health care.

Six-sigma started with engineering products (Motorola in 1986) but has profound medical applications. Ortiz et al. [1] compared several six-sigma projects in health care industry in Colombia and reported that obstetric outpatients were the most suitable six-sigma project for improving care opportunities. Bedi et al. [2] used DMAIC (define, measure, analyze, improve and control) methodology of six-sigma strategy to assess the musculoskeletal disorders in India.

1. Ortíz MA, Felizzola HA, Isaza S. A contrast between DEMATEL-ANP and ANP methods for six sigma project selection: A case study in healthcare industry. *BMC Med Inform Decis Mak*. 2015;15 Suppl 3:S3. http://www.biomedcentral.com/1472-6947/15/S3/S3
2. Bedi HS, Moon NJ, Bhatia V, Sidhu GK, Khan N. Evaluation of musculoskeletal disorders in dentists and application of DMAIC technique to improve the ergonomics at dental clinics and meta-analysis of literature. *J Clin Diagn Res* 2015 Jun;9(6):ZC01–3. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4525595/

## skewness and the coefficient of skewness

Gaussian distribution pervades statistical thoughts so much that it is seen afflicted by ghost of Gauss. Although many quantitative medical measurements in healthy subjects do indeed follow a Gaussian pattern (Figure S.10a), not all do. Gaussian is a **symmetric distribution** with the left half being the exact replica of the right half. Many distributions do not have this feature. Figure S.10b shows the distribution of serum total homocysteine in healthy subjects in Norway [1]. This distribution has a longer tail on to the right because higher-than-mode levels are more common in healthy subjects. This is called a *right-skewed distribution*. The homocysteine level in any population is likely to follow the same pattern. Triglyceride levels in children are also found to be highly skewed to the right in Luxembourg [2]. On the other hand, the distribution of hemoglobin level is generally *left-skewed* (Figure 10.c) because lower values are commonly seen in healthy subjects.

The distribution of most medical measurements in healthy subjects is Gaussian and skewed in sick subjects. Examples of measurements with remarkable skewness among patients are tumor markers such as carcinoembryonic antigen (median in lung cancer 3.5 ng/mL but range 1–7580). In cancer particularly, the aberration increases in a multiplicative manner rather than additive such as becoming twice as much per unit of time. Logarithms transform multiplications to sums and tend to give the distribution a symmetric shape. Most right-skewed distributions can be converted to a Gaussian form by a suitable logarithmic or square-root transformation. For a left-skewed distribution, a reciprocal transformation can be examined; sometimes it works, sometimes not.

### *Checking Skewness—Simple But Approximate Methods*

Since most statistical methods are based on Gaussian distribution, this should be checked when in doubt. Sometimes the biological process underlying a medical measurement provides sufficient clue whether the distribution of a particular measurement is Gaussian or not as seen in the examples given in preceding paragraphs. In all other cases, a judgment must be made on the basis of the available data on a sample of subjects. In this situation, any of the following methods can be used for checking skewness in this situation. These methods work well for large $n$ but may fail for small $n$, and subjective judgment may be needed for small $n$.

If you are calculation oriented, just calculate the coefficient of skewness. Actual procedure for calculating this coefficient is complex as it requires the sum of the cubes of deviations from the mean, but a simple procedure is to calculate

$$\text{coefficient of skewness } I \approx \frac{\text{mean} - \text{mode}}{\text{SD}}.$$

This works reasonably well for unimodal (single-peak) distributions. A negative value of this coefficient indicates left skewness, and a positive value indicates right skewness. For a symmetric distribution, this coefficient is zero since then, in a unimodal distribution, mean = mode. A value <–1 or >+1 indicates highly skewed distribution.

In a Gaussian distribution—in fact in all symmetric unimodal distributions—mean, median, and mode are equal (Figure S.10a). In sample values, this could be approximately so. For other distributions, note the following:

Right-skewed distribution: mode < median < mean
Left-skewed distribution: mean < median < mode.

These are also shown in Figure S.10b and c. Incidentally, these words appear in a dictionary in the order seen for left-skewed distribution and reverse in the right-skewed distribution. Also the distance between the mean and median in a dictionary is small relative to the distance between median and mode as in these figures. The coefficient of skewness I is also based on such considerations. Thus, the first method to find that a distribution is symmetric or not is to calculate mean, median, and mode, and see if they follow any of the above-mentioned patterns. In samples, the difference between mean, median, and mode must be substantial for the distribution to be considered skewed.

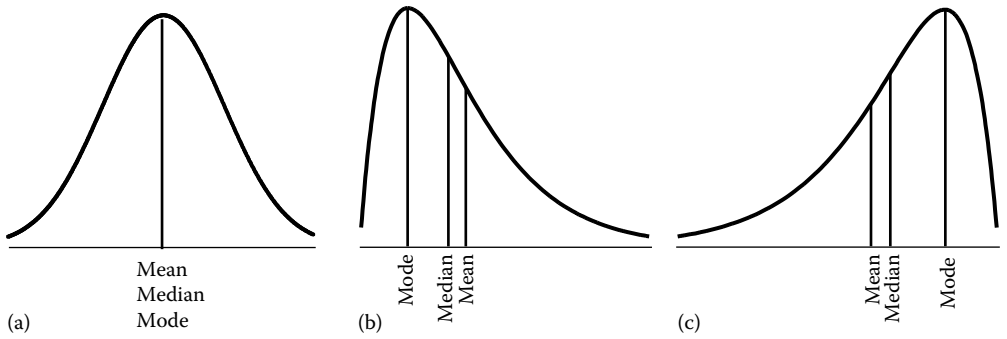



**FIGURE S.10** Location of mean, median, and mode in (a) symmetric; (b) right-skewed; and (c) left-skewed distributions.

If you are graph oriented, the most basic way to check Gaussianity is by **histogram**. Draw it for frequencies in different class intervals and see if it largely follows a bell shape or not. An alternative to histogram is **stem-and-leaf plot**. These are well known, and we are not giving any further details. A second approximate method is *quartile plot* of the type shown in Figure S.11a–c. For this, compute the first, second, and third **quartiles** $Q_1$, $Q_2$, and $Q_3$ and plot them on the Minimum to Maximum axis. If the distance between $Q_1$ and $Q_2$ is nearly the same as that between $Q_2$ and $Q_3$, it is safe to assume that the distribution is symmetric and possibly Gaussian. If the pattern is different as in Figure S.11b and c, the distribution is either left-skewed or right-skewed. If the sample size $n$ is really large, try this type of plot with **deciles** instead of quartiles.

You can see from Figure S.11b and c that $Q_3 - Q_2 > Q_2 - Q_1$ for positive skewness and $Q_3 - Q_2 < Q_2 - Q_1$ for negative skewness. Thus, another measure of skewness is

$$\text{coefficient of skewness II} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}.$$

An alternative to the quartile plot is the **box-and-whiskers plot** as presented in that topic. For symmetry, the boxes above and below median as well as the whiskers on both the sides should be nearly equal. A third graphic method is **ogive**, which is a plot of cumulative frequencies against the $x$-values. In a Gaussian distribution, this takes the shape of a **sigmoid**. If the shape is substantially different, take it as an indication of non-Gaussian shape.

Since most statistical methods work best for Gaussian distribution, there is a tendency to use transformation that can correct skewness in a skewed distribution. For positive skewness, logarithmic transformation is common. This rapidly shrinks large values to small values—thus the right tail of the distribution shifts toward the center, making it more symmetrical. The other popular transformation is square root ($x^{0.5}$), which also shrinks the right tail. Both of these are applicable to only positive values. In case all values are not positive, add slightly more than the highest negative so that none is zero or negative, and keep track of this while interpreting the results. For left skewness, square transformation ($x^2$) may work reasonably—square of 3 is 9 and square of 16 is 256—thus left tail and the negative skewness are reduced. It may be evident that the power of $x$ is less than 1 to correct right skewness and greater than 1 for left skewness. Procedures are available to find exactly which power of $x$ is best to minimize the skewness. For this, see the topic **power transformation**.

1. Arnesen E, Refsum H, Bonaa KH, Ueland PM, Forde OH, Nordrehaug JE. Serum total homocysteine and coronary heart disease. *Int J Epidemiol* 1995;24:704–9. http://www.ncbi.nlm.nih.gov/pubmed/8550266
2. Guillaume M, Lapideus L, Beckers F, Lamert A, Bjorntorp P. Cardiovascular risk factors in children from the Belgian province of Luxembourg: The Belgian Luxembourg Study. *Am J Epidemiol* 1996;144:867–80. http://aje.oxfordjournals.org/content/144/9/867.full.pdf

## slope (in regression), *see* simple linear regression

## slope–ratio assays, *see also* bioassays

Assays are experiments where we intentionally do something in different intensities to see what the effect is and compare it with the effect of a standard intervention (see **bioassays**). The effect

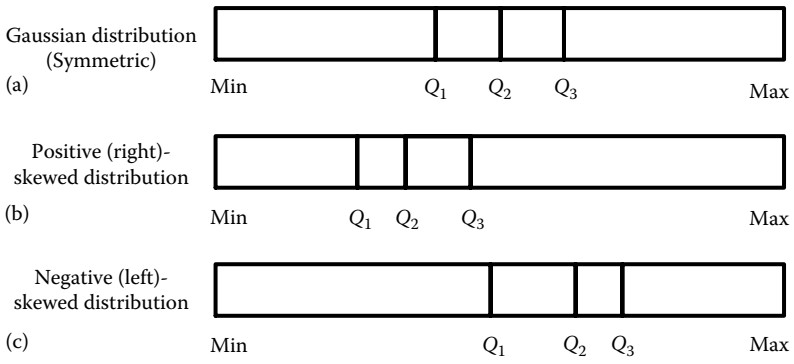



**FIGURE S.11** Pattern of quartiles in (a) symmetric (Gaussian); (b) right-skewed; and (c) left-skewed distributions.

of an intervention is technically called *response*, and the same intervention at different intensities is applied in the hope that as the intensity increases, the response will be higher. This particularly applies to poisons such as insecticides and pesticides, where larger dose results in higher response. In this case, the response is death, but in general, response could be any other outcome. This feature is seen in anesthetic agents, but the outcome of interest there is relief from pain. The primary objective of assays is to estimate the relative potency of the test intervention compared with the standard intervention for achieving a particular outcome. The assays that expect multiplicative response such as $2y$ at dose $2x$ but $4y$ at dose $3x$ are called slope–ratio assays. The basic setup in slope–ratio assays is that the response is the same for test and standard preparation at some baseline but increases (or decreases) at a faster rate for one preparation than the other as the dose increases. However, the trend in both should be linear for the slope–ratio assay to be valid. This is illustrated with the help of an example later in this section.

Death is called a quantal response (see **quantal assays**) as it occurs or does not occur, and there is nothing in between. In place of death, consider the time elapsed before death. If you give 2 mg of poison to rats, how much time does it take for a rat to die, and if you give 5 mg then how much time is taken? For the same dose, say 5 mg, some rats will take 4 h and some will take 5 h to die. Different rats will take different time, but the response is not death, instead is measured in quantitative terms in hours taken to die. When the comparison of such quantitative response is with a standard poison, this is called *quantitative assay.*

Slope–ratio assays are many times contrasted with parallel-line assays. In **parallel-line assays**, the plot of response versus dose falls in a straight line for both the test preparation and the standard preparation, and these lines are parallel to each other. This means that the difference in response remains constant as the dose escalates. In our example, the difference in time to death can be 4 h between test preparation and standard preparation at dose 2 mg, and the difference remains same 4 h at 3 mg, at 4 mg, etc. in the case of parallel-line assays. In place of actual dose $x$, one can have a transformation such as $x^{\lambda}$, called the *dose metameter.* Sometimes this kind of transformation is used to achieve a response that looks like that in parallel lines.

Quite often, the difference in response at different doses of standards and test preparation will not be constant. The difference in time to death could be 4 h at dose 2 mg but 6 h at dose 3 mg and 8 h at dose 4 mg. As the dose is increased by 1 mg, the difference increases by 2 h on average each time in this example and does not remain constant. This kind of response gives rise to slope–ratio assays.

For illustration, consider an experiment to find how the **bioavailability** (measured by area under the concentration curve) of drug B compares with the bioavailability of standard drug B in male adults.

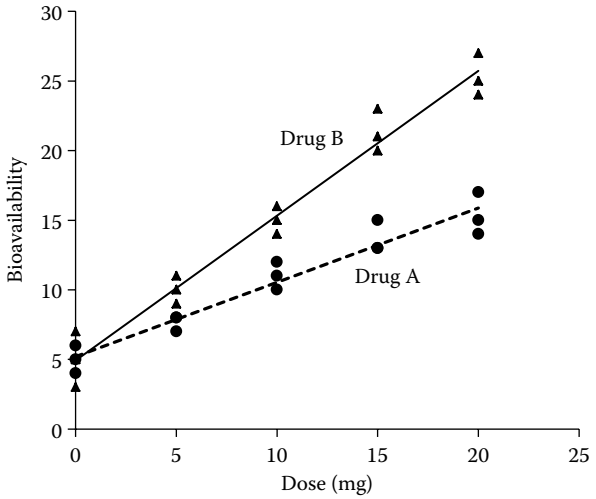



**FIGURE S.12** Regression plots for the data in Table S.14.

Both were given in doses of 0, 5, 10, 15, and 20 mg to three persons each; that is, a group of 30 male adults was randomly divided into two groups of 15 to receive drug A and drug B. The data obtained are shown in Table S.14.

If bioavailability is considered $y$ and dose is $x$, the linear regression of $y$ on $x$ is as shown in Figure S.12. Besides linearity, an important feature of regressions in this figure is that both the lines have the same intercept, but the slope for drug B is steeper than that for drug A. Let drug A be the standard and drug B the test preparation. The dose–response regression equations can be stated after ignoring the error term as

$$y_s = \alpha + \beta_s x_s \text{ for the standard preparation}$$

$$y_t = \alpha + \beta_t x_t \text{ for the test preparation,}$$

where $y_s$ is the response of standard preparation and $y_t$ of the test preparation. Different doses of standard and test preparations under trial are denoted by $x_s$ and $x_t$. $\alpha$ and $\beta_s$ and $\beta_t$ are the intercepts and slope of regression lines, respectively. Note that both have the same intercept—thus $\alpha_s = \alpha_t$, which is denoted by $\alpha$ in these equations. In the case of parallel-line assays, the intercepts are different and the slope is the same, but in the case of slope–ratio assays, the intercepts are the same and the slope differs. **Relative potency** in this case is the ratio of these slopes, that is,

$$\text{relative potency (slope–ratio assays) } \rho = \frac{\beta_t}{\beta_s},$$
$$\text{and its sample estimate is } R = \frac{b_t}{b_s},$$

S

**TABLE S.14**
**Bioavailability of Different Doses of Drug A and Drug B**

| | Receiving Drug A (Standard) | | | | | Receiving Drug B (Test) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Person** | **Blank 0 mg** | **Dose 1 5 mg** | **Dose 2 10 mg** | **Dose 3 15 mg** | **Dose 4 20 mg** | **Blank 0 mg** | **Dose 1 5 mg** | **Dose 2 10 mg** | **Dose 3 15 mg** | **Dose 4 20 mg** |
| 1 | 5 | 8 | 10 | 13 | 14 | 7 | 11 | 16 | 23 | 25 |
| 2 | 6 | 8 | 12 | 15 | 17 | 3 | 9 | 15 | 20 | 24 |
| 3 | 4 | 7 | 11 | 13 | 15 | 5 | 10 | 14 | 21 | 27 |
| **Mean bioavailability** | **5.0** | **7.7** | **11.0** | **13.7** | **15.3** | **5.0** | **10.0** | **15.0** | **21.3** | **25.3** |

where $b_s$ and $b_t$ are the estimates of $\beta_s$ and $\beta_t$, respectively. These slopes are readily estimated with the conventional **regression** techniques.

A careful look at Figure S.12 reveals that the estimated regression equations are approximately $y_t = 5 + 1.0x_t$ for the test preparation, and $y_s = 5 + 0.5x_s$ for the standard preparation. This gives relative potency = 1.0/0.5 = 2.0. This is an approximation, but assume for the time being that this is correct. Then drug B is twice as potent with regard to bioavailability as drug A in this example. In other words, twice as much drug A is needed to get the same bioavailability as drug B. This example is extraordinarily simple in many ways as noted below:

1. In this example, the dose itself is being used in the regression and not its logarithm or any other transformation. Thus, the dose metameter is the dose itself. The same goes for response (bioavailability). In general, you may have to use a dose metameter and in some cases response metameter,first to achieve Gaussian distribution and second to get linear regression.
2. Doses are equispaced 0, 5, 10, 15, and 20 mg in this example. In actual experiments, this may not be so. Somebody may receive 12 mg, some 14 mg, some 17.5 mg, etc. This does not change the setup as the usual regression can be run even when the doses are not equally spaced.
3. Doses for the test and standard preparations are the same in this example, but, in general, they could be different.
4. Dose 0 mg is called the blank dose, and this factually is given to 6 persons in this example, whereas other doses are given to 3 persons each. A higher number of subjects for blank dose help to get a more reliable estimate of the intercept.
5. The number of doses under trial is the same for both the preparations in this example, yielding to what is called a *symmetric assay*. These are 4 doses for each plus one blank (although blank appears in both the groups). Thus, this is a 9-point assay. In general, the number of doses can differ; for example, standard can be tried for 3 dose levels and test for 5 dose levels. At least 3 doses of standard and 3 doses of test preparation are needed to validate linearity (see point 8), and a blank is needed to check for equality of the intercept (see point 7).
6. In this example, each dose has been given to 3 persons. In practice, one dose may be given to 4 persons, another dose to 7 persons, third dose to only 2 persons, etc., as these numbers can vary in an experiment. That also does not change the basic structure of the experiment, and the regressions can be obtained as usual. However, an equal number of subjects for each dose level minimize the requirement of the total number of subjects, and the computations are relatively simple. When the number of subjects are the same for each dose, this is called **balanced design**.
7. Figure S.12 shows for these data that the intercepts for the two regressions are the same. In practice, a statistical test based on analysis of variance (**ANOVA**) is conducted to check whether the intercepts are significanty different or not. Once they are found to be not different, then only the ratio of slope will provide the correct estimate of the relative potency.
8. In Figure S.12, the regressions are clearly linear. In practice, check that they are really linear, and deviation from linearity is not statistically significant. This test also is based on ANOVA and would be valid when the requirements of ANOVA are fulfilled.
9. Fiducial limits for relative potency, being a ratio, can be obtained by **Fieller theorem**.

For further details of slope–ratio assays, see Hewitt [1].

1. Hewitt W. *Microbiological Assay for Pharmaceutical Analysis: A Rational Approach*. CRC Press, Boca Raton, FL, 2003.

## Slutzky–Yule effect

Discovered by Eugen Slutzky [1] and Udny Yule [2] independently in 1920s, this is the correlation induced by a method such as **moving averages** in a **time series** of uncorrelated values as is apparent from the following example:

$$\text{moving average of order 3 at time } t\text{: } y_t = \frac{x_{t-1} + x_t + x_{t+1}}{3}.$$

The values in series $y$ created by moving averages would be correlated since $x_t$ would appear in $y_{t-1}$, $y_t$, and $y_{t+1}$, and even when the $x$-values are uncorrelated, **autocorrelation** of lag 1 could be 2/3 and 1/3 at lag 2 because of repeated appearance of the same values. If any particular value is large, all moving averages involving that value will be inflated. Generally, the inflation will be the greatest when that value is in the middle, and will progressively reduce on either side. The Slutzky–Yule effect is the spurious cyclical effect that autoregressive series can produce when there is none. In fact, it is because of such correlations that moving averages are able to provide a relatively smooth pattern.

Smoothing a time series by forming moving averages is a commonly deployed approach—thus, this effect is commonly seen in such cases. **ARMA and ARIMA models** of time series also tend to induce such correlations. Conversely, if a time series is exhibiting an oscillatory movement, examine if it is due to summation (or averaging) of values over time.

1. Slutzky EE. Slozhenie sluchainykh prichin, kak istochnik tsiklicheskikh prosessov. *Voprosy kon "yunktury* 1927;3:34–64.
2. Yule GU. Why do we sometimes get nonsense correlations between time series?—A study is sampling and nature of time series. *J Royal Stat Soc* 1926;89(1):1–64. http://links.jstor.org/sici?sici=0952-8385%28192601%2989%3A1%3C1%3AWDWSGN%3E2.0.CO%3B2-L

S

## small area estimation

Under this methodology, the results of a survey for a big population are sought to be used to generate estimates for small subareas, for which the sample size is inadequate to provide any reliable estimates, by using auxiliary information (such as from census or administrative records) already available for all units of the small area. Use of auxiliary information is basic to this method, and thus the method is also called the *synthetic estimation*. Small area can be a geographical area such as a county but can also be any specific group of interest such as a particular age–sex group. It is better understood as a domain, which is a more general term, and encompasses the spatial, demographic, disease, and such other groups. The method can be used when an increased sample size for each small area yields an unusually large sample that exceeds the resources

available for the survey, and the subareas have an inadequate sample for providing a reliable estimate.

As you can see, small area estimation is an indirect method of estimation as compared with a direct method that uses the collected data on the area. Its success depends on proper choice of the auxiliary variable since this method derives strength from this variable to simulate a large sample size. This auxiliary variable must be measured uniformly over the total area so that the information can be effectively used. If there is some nonresponse, this should be small and random, and not biased (selective). The link between the main variable of interest and the auxiliary variable is exploited to improve the reliability of the estimates, and this link should be clearly established and fully understood so that it can be properly used. If the auxiliary variable is not chosen properly, this may introduce bias into the estimates.

Small area estimation has a lot of relevance in assessing health status (such as disabilities) and health practices (such as drug use) at county or regional level that helps in developing adequate plans to meet the local needs. The National Center for Health Statistics in the United States pioneered the use of such synthetic estimation for developing state estimates of disability and other health characteristics from the National Health Interview Surveys. Sample sizes in most states were too small to provide reliable direct estimates for the states.

The method generally used for small area estimation is similar to what is used for **ratio estimators**. The link between the auxiliary variable and the variable of interest may be complex involving **hierarchical models** with **random effects**. Suppose $y_{ij}$ is the value of the study variable for the $i$th unit ($i = 1, 2, \ldots, n_j$) of the $j$th small area ($j = 1, 2, \ldots, J$), where $n_j$ is the number of population units in the $j$th area. Let the auxiliary variable be $x$ whose value $x_{ij}$ is available for each unit in the small area. The basic relationship between $x$ and $y$ is the link model in this case and can be expressed through a nested **regression model** as

$$y_{ij} = \beta x_{ij} + u_j + \varepsilon_{ij},$$

where $u_j$ is the area-specific random effect of area $j$, $\beta$ is the regression coefficient, and $\varepsilon_{ij}$ is the random error. For simplicity, we have taken only one auxiliary variable and are considering only the linear model, but there is no such restriction in small area estimation. As mentioned earlier, the validity of the estimation certainly depends on the suitable choice of the auxiliary variable and also on the suitable link model. In case needed, more auxiliary variables and curvilinear relationships can be included. This link model allows estimation of the parameters of $y$ (such as mean) for the small areas with relatively high precision. For details, see Rao [1].

1. Rao JNK. *Small Area Estimation*. Wiley, 2003. http://samples.sainsburysebooks.co.uk/9780471431626_sample_386961.pdf

## smoke-pipe distribution

This is the name proposed by Indrayan [1] for distribution of deaths by age in a developing country where infant mortality is high. Thus, this distribution has two modes—one around age 6 months (or even early) and the other around 75 years (Figure S.13). After the age of 80 years, the number of people remaining alive is small, and consequently, the number of deaths is also small, although the death rate in old age groups is extremely high (Table S.15). This feature gives it a shape of smoke pipe.

Since this shape is entirely different from Gaussian, it highlights the importance to be careful while dealing with age distributions of deaths in a population such as in building up of confidence intervals and testing of hypothesis. Statistical properties of smoke-pipe distribution are yet to be investigated.

1. Indrayan A. *Medical Biostatistics*, Third Edition. Chapman & Hall/CRC Press, 2012.

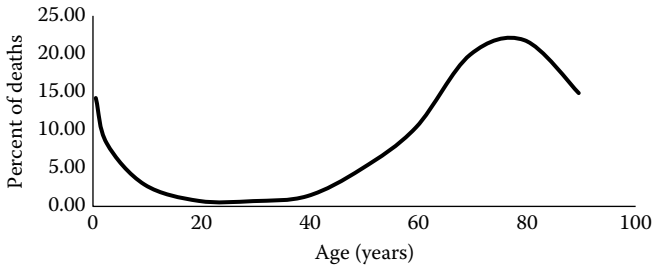



**FIGURE S.13** Smoke pipe distribution of deaths by age in a developing country.

**TABLE S.15**
**Distribution of Deaths by Age in a Developing Country**

| Age Group (Years) | Population ('000) | Age-Specific Death Rate (per 1000) | Number of Deaths | Number of Deaths per Year of Age | Percent of Deaths per Year of Age |
|---|---|---|---|---|---|
| 0–1 | 1.5 | 20 | 30 | 30 | 14.22 |
| 1–4 | 7.0 | 10 | 70 | 17.5 | 8.30 |
| 5–14 | 15.0 | 4 | 60 | 6 | 2.84 |
| 15–24 | 15.0 | 1 | 15 | 1.5 | 0.71 |
| 25–34 | 14.5 | 1 | 14.5 | 1.45 | 0.69 |
| 35–44 | 14.2 | 2 | 28.4 | 2.84 | 1.35 |
| 45–54 | 13.0 | 8 | 104 | 10.4 | 4.93 |
| 55–64 | 9.5 | 23 | 218.5 | 21.85 | 10.36 |
| 65–74 | 7.0 | 60 | 420 | 42 | 19.91 |
| 75–84 | 2.0 | 230 | 460 | 46 | 21.81 |
| 85–94 | 0.7 | 448 | 313.6 | 31.36 | 14.87 |

## smoking index, *see* Indrayan smoking index

## smoothing methods, *see also* spline regression, cubic splines

The term *smoothing* in statistics is used for extracting a trend from noisy data in such a way that important features of peaks and troughs remain preserved. Thus, this is similar to curve fitting by regression, but here, values adjacent to any point are given more weightage—low values adjacent to a high value are raised, and high values adjacent to low values are reduced so that they are aligned and produce a smooth trend. See Figure S.14 for an example. It is difficult to get this kind of trend by regression equation.

One example of such smoothing is in obtaining reference centile curves for growth parameters of children (see **growth charts**) where actual data at individual time points may show an irregular pattern, but curves are obtained after smoothing without losing the periods of slow growth and fast growth. Smoothing is done in this case in the realization that there is no reason for sudden jumps or falls in growth from age to age, and whatever changes occurring are gradual, particularly when the average for the entire "population" is considered. Similar reasons exist for other trends.

A large number of methods are available for smoothing, but all of them are intricate and require considerable expertise. An easy method is **moving averages** as generally used for time series data. The method generally used for centile curve smoothing is **cubic splines** as illustrated in that topic.

Another prominent use of smoothing is in spatial studies as done by Szonyi et al. [1] for incidence of Lyme disease in the United States. The objective here is identifying directional trends such as waning occurrence as we move from south to north. The same is the case with imaging as discussed by Patané [2] for different anatomical sites. The method they used is based on kernel smoothing and is different from what we have described.

Further details of smoothing methods have been provided by Simonoff [3].


1. Szonyi B, Srinath I, Esteve-Gassent M, Lupiani B, Ivanek R. Exploratory spatial analysis of Lyme disease in Texas—What can we learn from the reported cases? *BMC Public Health* 2015 Sep 19; 15(1):924. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4575478/
2. Patané G. Diffusive smoothing of 3D segmented medical data. *J Adv Res* 2015 May;6(3):425–31. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4522549/
3. Simonoff JS. *Smoothing Methods in Statistics*. Springer, 1996.

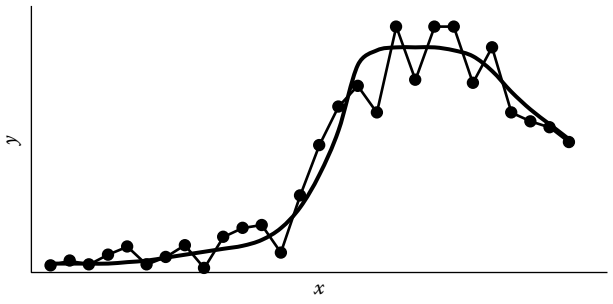




**FIGURE S.14** Noisy data and smooth trend.

## snowball sampling

This is a nonrandom method of sampling that tries to reach an obscure and possibly closed group by forming a chain of eligible subjects. A popular example is drug users who rarely identify themselves as such a user, and are hard to locate, and similarly sex workers. If a researcher is able to somehow find one drug user, this one can reveal some others who are in his or her contact, and those in turn can reveal more of their contacts. For this reason, this is also referred to as *chain referral sampling*. Some would reveal their contacts; others would not. If each person is able to identify at least 3 more persons, the sample size could become 81 after 4 stages. This can help in finally recruiting a reasonably sized sample for an empirical study, although it is futile to expect a large sample by this method.

Despite being nonrandom, some idea of what is happening can be obtained by studying this kind of sample. Since this method is not based on probability sampling, statistical methods such as confidence intervals cannot be used. However, in **analytical studies** such as in experiments where the subjects are randomly divided, it should be possible to make a generalized statement regarding, say, feasibility of a regimen in such a group, similar to the exercise we do in phase I of a clinical trial.

Snowball sampling is relatively simple and cost-efficient, in some cases the only way out to reach people of the type one wishes to study, but can seldom provide a representative sample of the target population. Since the subjects identify others of their own type, a strong clustering effect may be present that can produce biased estimates. There is a possibility that after some links, the process becomes repetitive; that is, you start getting the same subjects again. Since this method is used for people who are mostly in illegal activities, special efforts may be needed to address their sensitivities and to get their cooperation to participate in the research.

## social classification

Social classification is an indicator of the social health of a person and is generally defined by his or her education, income, and occupation. These can be combined to devise a system of classification of people into various socioeconomic classes. There are many classifications available, but our preference is for the one devised by Indrayan [1] and given in Table S.16 because of its international applicability. In case of family, this applies to the head of the household, and the entire family is considered to belong to the same class.

Table S.16 suggests a scoring system for the three components, and a classification based on the aggregate score. The categories of schooling years and of occupation are absolute in this table and can probably be used anywhere in the world without alteration. But the income level in one country can seldom be compared with that in another country except possibly in terms of the **purchasing power parity**. But this is far too complex to calculate. Thus, a classification of income is on the basis of percentile, which will be specific to the area. There will always be 20% of the population below the 20th percentile and another 20% between the 20th and 40th percentiles, etc., which in fact are **quintiles**. Yet, the classification may be valid for comparing one subpopulation within a country to another subpopulation, and to some extent for an international comparison also. Such quintiles are generally available from the distribution of per capita income of the nation. If not available, it is relatively easy to assess that the person's income belongs to which quintile than exact income.

**TABLE S.16**
**Scoring for Social Classification**

| Score | Years of Schooling | Income[a] | Occupation |
|---|---|---|---|
| 0 | Nil (illiterate) | Below poverty line[b] | Unproductive or burden on society (e.g., begging), including unemployed |
| 1 | <5 | <20th percentile[c] | Unskilled labor |
| 2 | 5–10 | 20th–40th percentile | Skilled labor, artisan, small business, student, small farmer, soldier |
| 3 | 10–15 including vocational | 40th–60th percentile | Clerk, medium business, medium farmer, technician, salesperson |
| 4 | 15+ but nonprofessional nontechnical | 60th–80th percentile | Teacher, researcher, industrialist, big farmer, big business, government officer, manager |
| 5 | 15+ some of which is professional or technical | ≥80th percentile | Executive, doctor, attorney, consultant, engineer |

| Aggregate Score | Social Class |
|---|---|
| 0–3 | V |
| 4–6 | IV |
| 7–9 | III |
| 10–12 | II |
| 13–15 | I |

[a] In case of a family, calculate per capita income in the family.
[b] Income required to purchase low-cost balanced food to provide 2400 calories.
[c] Excludes those that are below poverty line.

A large number of common occupations are listed in Table S.16, but certainly not all. An indication of the score for other occupations not included in this table should be available from the pattern.

The minimum possible aggregate score is 0 and the maximum is 15. A classification is suggested in the bottom of Table S.16. The percentage of the population in different classes would give an indication of the social health of people. Many diseases and other health conditions have been observed to be associated with such social classification of the subjects. See **social health (indicators of)** for details.

For official classification followed in the United Kingdom, see the ONS document [2].

1. Indrayan A. *Medical Biostatistics*, Third Edition. Chapman & Hall/CRC Press, 2012.
2. ONS. SOC2010 volume 3: The National Statistics Socio-economic classification (NS-SEC rebased on SOC2010). http://www.ons.gov.uk/ons/guide-method/classifications/current-standard-classifications/soc2010/soc2010-volume-3-ns-sec-rebased-on-soc2010-user-manual/index.html

## social health (indicators of)

Social health is generally measured in terms of social status comprising education, income, and occupation. These, independently as well as in combination, are believed to have a deep impact on the well-being of the people. Education increases awareness and adds to the productivity of people. Income causes a sense of well-being by allowing access to a wider set of choices. Occupation helps to attain satisfaction of doing something worthwhile for society. These together are sometimes considered distal factors of health that affect proximal factors such as diet, smoking, and hygiene. The proximal factors affect physiological and pathophysiological processes that lead to disease and infirmity. Genetic factors and environmental factors such as pollution and water supply contribute to this process.

See **education indicators** for a large number of indicators used to measure education level in a community. We are not giving details here.

The most widely used measure of the level of income of a community is the per capita **gross domestic product (GDP)**, popularly called per capita income. Almost all countries compute this for the nation as a whole and usually for each of their states separately. Further subdivisions may not be easily available. Income per se in the national currencies is seldom comparable across countries because of differential purchasing power. A Singapore dollar cannot buy as much in Singapore as the Yuan equivalent can in China. For international comparison, the World Bank has worked out **purchasing power parity dollars**. Note, however, that income is mostly a family trait rather than an individual because children are not supposed to earn. If the average size of family is four, the average family income is four times the per capita income. When income data are not available, an indirect way to assess the economic status of a community is by the presence of amenities such as housing, telephones, televisions, computers, and automobiles per family or per 1000 population.

Ideally speaking, all jobs have dignity and deserve full respect. Practically, though, professionals such as attorneys, doctors, chartered accountants, and consultants are considered to enjoy better social health. Industrial laborers and coalmine workers do not generally enjoy the same status. Thus, there is an inbuilt hierarchy in occupation in most societies. Except for the percentage in different occupations, no worthy measure is available to measure occupational distribution.

Singh-Manoux et al. [1] compared three models for exploring the links between different measures of what they called adult socioeconomic position (SEP)—education, occupation, income—and psychosocial health. Their focus was predicting psychosocial health based on social position. Model I was a basic regression model with a measure of SEP as the predictor, and model II was a multiple regression model with all three measures of SEP considered as predictors. Model III treated education as a distal measure of SEP, and as an antecedent to the proximal measures of SEP in the prediction equations linking SEP to psychosocial health. The three models lead to completely different conclusions, but the authors preferred model III that showed education to have a stronger *indirect* effect on psychosocial health when compared to its direct effect. They concluded that proximal measures of social position might discriminate better as they portray the current and accumulated socioeconomic circumstances of the individual more accurately.

1. Singh-Manoux A, Clarke P, Marmot M. Multiple measures of socioeconomic position and psychosocial health: Proximal and distal measures. *Int J Epidemiol* 2002;31:1192–99. http://ije.oxfordjournals.org/content/31/6/1192.long

## Somer *d*, *see* association between ordinal characteristics (degree of)

## Spearman rank correlation, *see* rank correlation

## Spearman–Brown formula

Spearman–Brown formula is used to reassess the **reliability** of a multi-item test or instrument when its length is increased or decreased, and the formula can also be inversely used to determine the length for fixed reliability. If the new test is $K$ times longer than the original, the reliability of the new test is given by

$$\text{Spearman–Brown formula: } \rho_K = \frac{K\rho_1}{1+(K-1)\rho_1},$$

where $\rho_K$ is the reliability of the new test, which is $K$ times as long as the original test, and $\rho_1$ is the reliability of the original test. For example, if the original test has reliability $\rho_1 = 0.62$, doubling ($K = 2$) the length of the test will give reliability $\rho_2 = 2 \times 0.62/[1 + (2 - 1) \times 0.62] = 0.77$. Inversely, for desired reliability $\rho_K$,

$$K = \frac{\rho_K(1-\rho_1)}{\rho_1(1-\rho_K)}.$$

If a test already has reliability of $\rho_1 = 0.75$ and it is proposed to be increased to $\rho_K = 0.85$, this gives $K = [0.85(1 - 0.75)/[0.75(1 - 0.85)] = 1.89$. If the original test had 10 items, the new test should have 19 items to increase its reliability from 0.75 to 0.85.

The formula was independently proposed by Charles Spearman [1] and William Brown [2] in 1910. It is applicable only when the new test has items of similar difficulty—they could be, for example, more indicators of the same trait—and is commonly used for psychometric evaluation of the test items in social research including health. Alves et al. [3] used this formula to determine the number of items in a linguistic adaptation of the Neonatal Intensive Care Units Family Needs Inventory in Portugal. Aadland and Ylvisåker [4] had an interesting application of the formula for determining the number of monitoring days to achieve an intraclass correlation coefficient of 0.80 for reliable measurement of sedentary time and physical activity in adults in Norway.


1. Spearman CC. Correlation calculated from faulty data. *Br J Psychol* 1910;3:271–95. http://onlinelibrary.wiley.com/doi/10.1111/j.2044-8295.1910.tb00206.x/abstract
2. Brown W. Some experimental results in the correlation of mental abilities. *Br J Psychol* 1910;3:296–322. http://onlinelibrary.wiley.com/doi/10.1111/j.2044-8295.1910.tb00207.x/abstract
3. Alves E, Severo M, Amorim M, Grande C, Silva S. A short form of the neonatal intensive care unit family needs inventory. *J Pediatr (Rio J)* 2015 Oct 9. pii: S0021-7557(15)00144–8. http://www.sciencedirect.com/science/article/pii/S0021755715001448
4. Aadland E, Ylvisåker E. Reliability of objectively measured sedentary time and physical activity in adults. *PLoS One* 2015 Jul 20;10(7):e0133296. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4508000/


## specificity, *see* sensitivity and specificity

## sphericity, *see* Mauchly test for sphericity

## spline regression, *see also* cubic splines

Splines are projections of shaft that join another shaft through grooves or any other mechanism so that both can move together. In statistics, these are functions, mostly piecewise **polynomials**, which graphically fit into one another to provide a continuous though irregular shape. The joints are called *knots*, although these may not be visible in a finished graph. Polynomials are algebraic functions of any specified degree such as linear, quadratic, cubic, and quartic. When these are lines, the knots would be easily visible as the points joining lines with different slopes. Figure S.145a has five linear splines and four knots where these splines are tied. In **cubic splines**, for example, each one of the splines would be a cubic function that can give a shape of ∾ with no break, and they will smoothly join with one another on the knots (Figure S.15b). This figure is based on a different set of data and shows linear splines as well.

To understand splines further, consider two quadratic polynomials $y_1 = 3x^2$ and $y_2 = x^2 + 4x - 2$, the latter of which can also be written as $y_2 = 3x^2 - 2(x - 1)^2$. The reason that we want to give this alternative version of $y_2$ is to show that the two functions have the same value $3x^2$ at $x = 1$, and this is the knot where these two quadratic "splines" can be joined. We can say that $y$ is $3x^2$ for $x < 1$ and is $x^2 + 4x - 2$ for $x > 1$, both being the same at $x = 1$. Both polynomials are truncated. If functions do not have exactly the same value at any value of $x$, the closest value is searched and smoothing is applied so that they can be joined. This is done through an appropriate software package.

Spline regression is the process of joining many splines of the desired degree (linear, quadratic, cubic, etc.). This is used when you expect many different patterns of trends when the value of $x$ increases—the pattern for $x$ in the interval $(0–\theta_1)$ different from the pattern for $x$ in the interval $(\theta_1–\theta_2)$, and so on, for $x$ in the interval

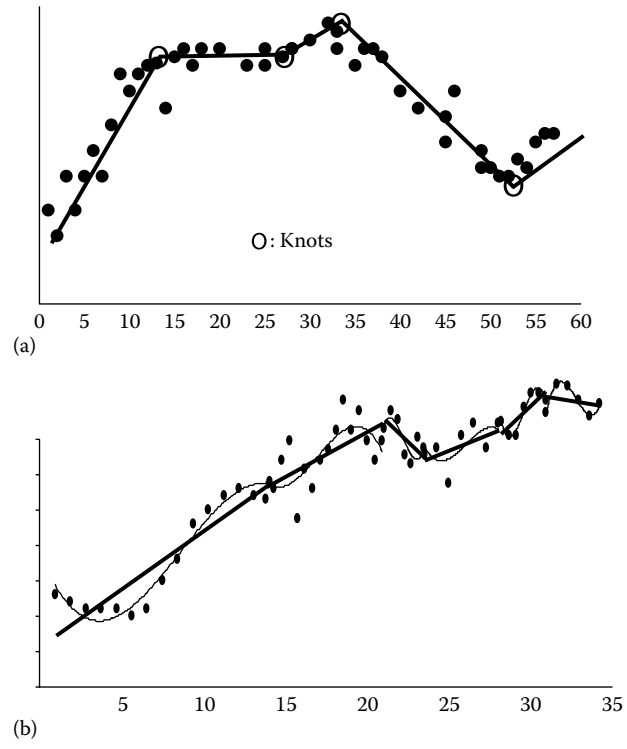



**FIGURE S.15** (a) Linear splines. (b) Cubic splines (snaky curve) imposed on linear splines.

($\theta_{K-1}$–$\theta_K$), where $\theta_k$'s are the $K$ knots. A common example is the construction of growth charts for children that may have phases of slow growth for some age intervals and of fast growth in some other intervals. In case of linear splines, this reduces to changing slopes up or down at some specific values of the regressor (Figure S.15a). A general spline regression takes the following form.

Regression equation for splines of degree $D$ with $K$ knots:

$$y = \beta_0 + \Sigma_d \beta_{1d} x^d + \Sigma_k \beta_{2k} (x - \theta_k)_+^D ;\ d = 1, 2, \ldots, D;\ k = 1, 2, \ldots, K$$

where the first part of the equation is the regular polynomial regression, and the second part is the one that is for knots and also for different polynomials for different splines, although each spline has the same degree $D$. The notation $(x - \theta_k)_+^D$ is for $[\max(0, x - \theta_k)]^D$ so that only values greater than or equal to $\theta_k$ are considered. The number of $\beta_{1d}$'s is $D$, the number of $\beta_{2k}$'s is $K$, the number of $\theta$'s is also $K$, and there is one $\beta_0$—the total of $(D + 2K + 1)$ parameters for estimation. The number of knots and their location mostly will not be known, and their choice is a challenge—a software package is used for this purpose also. Too few knots may not be able to detect all the points of slow and fast movements, and too many knots may show spurious features with unnecessarily wiggly regression. Obviously, reasonable estimation of so many parameters is possible only when the number of $x$ values is much larger so that overfitting is avoided. For this, the method of penalized least squares or **penalized likelihood** is used that can reduce the number of parameters for estimation by imposing penalty on the higher number of parameters. It is important to realize that this method allows flexibility since different types of polynomials can be fitted to different segments of $x$-values, thus removing the restriction of one regression equation to the entire range of $x$ as done in the usual regression. The challenge is also to ensure that the regressions in different segments exactly join at the knots with no jump or step. A good statistical software package and sufficient statistical expertise are required for fitting spline regression to the data.

Besides child growth curves, splines have been used in other setups also. For example, Zhang et al. [1] have studied the relationship between dose of smoking and difference in methylation intensity by cubic splines in older adults in Germany, and Ray et al. [2] demonstrated for subjects in a sleep clinic in the United States that spline regression method has the potential to improve sleep estimates using wrist actigraphy.

For further details of spline regression, see Marsh and Cormier [3].


1. Zhang Y, Schöttker B, Florath I, Stock C, Butterbach K, Holleczek B, Mons U, Brenner H. Smoking-associated DNA methylation biomarkers and their predictive value for all-cause and cardiovascular mortality. *Environ Health Perspect* 2015 May 27. http://ehp.niehs.nih.gov/wp-content/uploads/advpub/2015/5/ehp.1409020.acco.pdf
2. Ray MA, Youngstedt SD, Zhang H Robb SW, Harmon BE, Jean-Louis G, Cai B et al. Examination of wrist and hip actigraphy using a novel sleep estimation procedure. *Sleep Sci* 2014 Jun;7(2):74–81. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4286157/
3. Marsh LC, Cormier DR. *Spline Regression Models*. Sage, 2001.


## split-half consistency, *see also* consistency

Consider a test in the form of a questionnaire containing several questions or items, say, on a scale of 0 to 5. An example is the assessment of ability to perform activities of daily living (ADL) by a geriatric population, where items such as ability to dress, to bathe, to walk around, and to eat are scored from 0 for complete inability (full dependence) to 5 for complete independence requiring no assistance. It is expected that the ability score on one item would correspond to the score on other items. If there were a difference, it would generally persist across subjects. In fact, all items are different facets of the same entity—the ADL in this case. Uniformity of the underlying construct across all items of a test is an important prerequisite for measuring internal consistency. If that is so, the responses will be consistent with one another provided the items are properly framed and the questions appropriately asked.

One way to measure internal consistency is to split the test into ostensible equal halves where feasible. This is generally possible for a questionnaire by randomly dividing questions into two parts. The other method is to put odd-numbered questions in one-half and even-numbered questions in the other half. The product–moment correlation coefficient between total scores in the two parts across several subjects is a measure of internal consistency. This is called split-half consistency. Note that there is no one-to-one correspondence between one item in one-half of the test and an item in the other half of the test. Thus, intraclass correlation cannot be used in this case, and the usual product–moment correlation between total scores is used to assess split-half consistency.

Split-half implies that the number of items available for calculating the total score is one-half of the items in the whole test. Such reduction in the number of items adversely affects reliability assessment. An adjustment, called the **Spearman–Brown formula**, can be used to estimate the reliability if the test were to consist of all the items. For details, see Groth-Marnat [1].

Split-half is commonly used for consistency measure in questionnaire development. Wang et al. [2] used this for psychometric evaluation of the questionnaire of health-related quality of life among rural-to-urban migrants in China, and Yoshii et al. [3] used it for assessing reliability of the workplace social distance scale in Japan.


1. Groth-Marnat G. *Handbook of Psychological Assessment*, Fifth Edition. Wiley, 2009.
2. Wang P, Chen C, Yang R, Wu Y. Psychometric evaluation of health related quality of life among rural-to-urban migrants in China. *Health Qual Life Outcomes* 2015 Sep 24;13(1):155. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4581414/
3. Yoshii H, Mandai N, Saito H, Akazawa K. Reliability and validity of the workplace social distance scale. *Glob J Health Sci* 2014 Oct 29;7(3):46–51.




## spot map

Maps are a powerful medium for showing the spatial distribution of a disease or of a health condition. A dot can represent one or many cases (such as 1 dot = 10 cases). A concentration of dots in any area indicates that the incidence of disease is high in that area. Figure S.16a is a map of India for human immunodeficiency virus prevalence rate in persons with sexually transmitted disease in 2005. Areas with higher incidences have many dots. Such a map is called a spot map and can be used to investigate a localized outbreak of a disease. Figure S.16b shows the spot map of tularemia case in the United States in 1990–2000.

Ward et al. [2] developed a spot map of their catchment area in the United States and identified areas where nonaccidental trauma occurs with increased rate. Jent et al. [3] provided a spot map showing areas requiring close monitoring for fecal contamination in Puerto Rico.

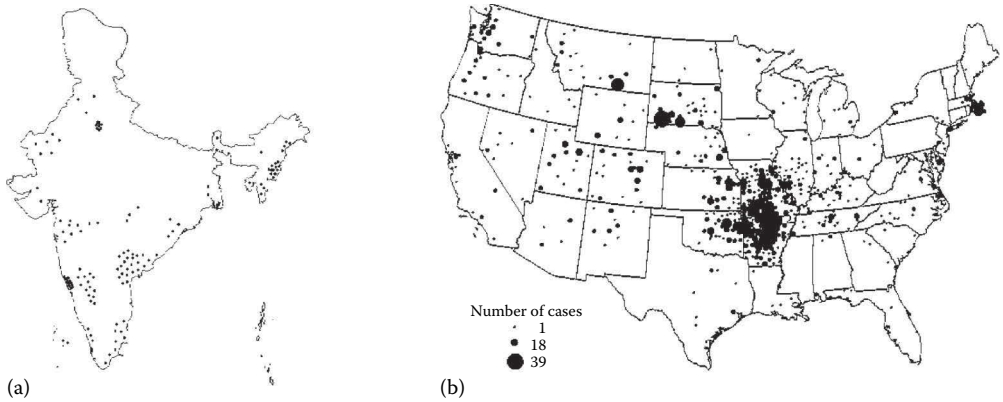



**FIGURE S.16** Spot map: (a) HIV prevalence in STD cases in India, 2005. (b) Reported cases of tularemia in the United States—1990–2000. (From Hunt R. *Parasitology Chapter Seven Part Two: Ticks. Microbiology and Immunology Online*. http://www.microbiologybook.org/parasitology/ticks.htm.)


1. Hunt R. *Parasitology Chapter Seven Part Two: Ticks. Microbiology and Immunology Online*. http://www.microbiologybook.org/parasitology/ticks.htm
2. Ward A, Iocono JA, Brown S, Ashley P, Draus JM Jr. Non-accidental trauma injury patterns and outcomes: A single institutional experience. *Am Surg* 2015 Sep;81(9):835–8. http://www.ncbi.nlm.nih.gov/pubmed/26350656
3. Jent JR, Ryu H, Toledo-Hernández C, Santo Domingo JW, Yeghiazarian L. Determining hot spots of fecal contamination in a tropical watershed by combining land-use information and meteorological data with source-specific assays. *Environ Sci Technol* 2013 Jun 4;47(11):5794–802. http://www.ncbi.nlm.nih.gov/pubmed/23590856


## spurious correlation, *see* correlation (the concept of)

## square table, *see* contingency table

## standard deviation

Dispersion, scatter, and variability all connote the same phenomenon. Standard deviation (SD) is the most widely acceptable and most commonly used measure of **variation** in quantitative values. Its magnitude depends on the extent of the difference between a value and that of the others. Instead of calculating so many differences, it is convenient to compute the difference of each from a central value, namely the mean. Let the sample observations be $x_1$, $x_2$, …, $x_n$ and the mean $\bar{x}$. The difference $(x_i - \bar{x})$, where $i = 1, 2, \ldots, n$, is called the deviation of the $i$th value from the mean. Some of these deviations would be positive and some negative. An average of these deviations could be a measure of dispersion, but this would always be zero. One way out is to ignore the sign, get the absolute values, and calculate the average $\Sigma|x_i - \bar{x}|/n$. This is called the **mean deviation**. Absolute values are mathematically difficult to handle. It is easy to get rid of the minus sign by squaring the deviations. The average of squared deviations $(x_i - \bar{x})^2$ is $s^2 = \Sigma(x_i - \bar{x})/n$. This is called the sample **variance** and is a very useful and popular measure of dispersion. For population variance, use $\mu$ in place of $\bar{x}$ and denote by $\sigma^2$. Thus,

$$\text{population variance: } \sigma^2 = \Sigma(x_i - \mu)/N,$$

where $x_i$ now refers to the $i$th value in the population, and the sum is over all $N$ values in the population. It is seen in the case of samples that the denominator $(n - 1)$ in place of $n$ gives a better estimate of the population variance in the long run (called **unbiased**), but the population variance is calculated with $N$ in the denominator.

Variance has been extensively studied and found to be a very adequate measure. The only difficulty is that its calculation is such that the unit of the variable values is also squared (e.g., square meters for area). The original unit is retrieved by taking the square root. The quantity obtained as the positive square root of the variance is called the **standard deviation** (SD). Thus,

$$\text{population SD: } \sigma = \sqrt{\Sigma(x_i - \mu)/N},$$

$$\text{and sample SD: } s = \sqrt{\Sigma(x_i - \bar{x})/(n-1)}.$$

The term "standard deviation" was first proposed by Karl Pearson in 1894 as a substitute for cumbersome "root mean square deviation" [1].

See Figure S.17 to get an idea of what variance measures, where scatter of three medical measurements is shown. The serum iron varies in this figure from 50 to 170 μg/L in healthy subjects and has a larger variance relative to systolic blood pressure (BP) that varies between 115 and 140 mmHg; BP has a larger variance than serum sodium since it varies within a narrow range of 135–144 mEq/L.

### *Calculation of Variance and SD in Ungrouped Data*

Ungrouped data are the original values as measured in exact quantities. As already explained, the formulas for ungrouped data can be stated as follows:

$$\text{sample variance: } s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} \text{ and sample SD: } s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$$

Calculation of the variance and SD of the systolic BP for the two groups of subjects is shown in Table S.17 where one group has compact values from 128 to 132 mmHg, and the other group has widely scattered values from 100 to 150 mmHg. SD in group 2 is more than four times the SD in group 1. The value of SD can be legitimately used to conclude that the variation in group 2 is nearly four times than that in group 1.

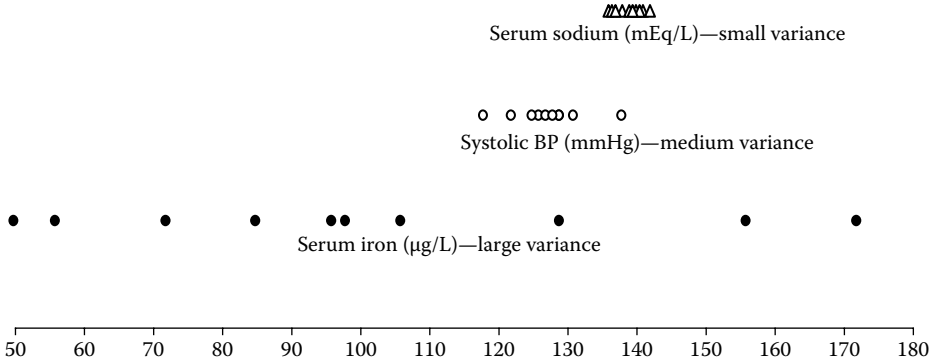



**FIGURE S.17** Schematic of differential variance in three measurements.

**TABLE S.17**
**Calculation of Variance and Standard Deviation (SD) in Two Disparate Groups**

| Group 1 | | | Group 2 | | |
|---|---|---|---|---|---|
| SysBP ($x$) | Deviation ($x-\bar{x}$) | Squared Deviation $(x-\bar{x})^2$ | SysBP ($y$) | Deviation ($y-\bar{y}$) | Squared Deviation $(y-\bar{y})^2$ |
| 134 | 4 | 16 | 110 | –20 | 400 |
| 132 | 2 | 4 | 140 | 10 | 100 |
| 124 | –6 | 36 | 118 | –12 | 144 |
| 132 | 2 | 4 | 150 | 20 | 400 |
| 128 | –2 | 4 | 132 | 2 | 4 |
| **Total** $\Sigma(x_i-\bar{x})=0;\Sigma(x_i-\bar{x})^2=64$ | | | **Total** $\Sigma(y_i-\bar{y})=0;\Sigma(y_i-\bar{y})^2=1048$ | | |
| Sample variance $\frac{\Sigma(x_i-\bar{x})^2}{n-1}=\frac{64}{4}=16$ | | | $\frac{\Sigma(y_i-\bar{y})^2}{n-1}=\frac{1048}{4}=262$ | | |
| Sample SD $\sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n-1}}=\sqrt{16}=4$ | | | $\sqrt{\frac{\Sigma(y_i-\bar{y})^2}{n-1}}=\sqrt{262}=16.2$ | | |

SD is invariant under the change of origin, that is, it is not affected by addition or subtraction of each value by a constant. This implies that the SD of 5, 10, 15, and 20 is the same as that of 105, 110, 115, and 120. However, it is affected by scale. If you divide each value of $x$ by 4, the SD also will be one-fourth. In general,

$$\text{SD}(a+bx)=b^*\text{SD}(x).$$

Since SD is heavily dependent on mean $\bar{x}$, it can give a weird result when the mean is not an adequate central value. Thus, avoid using SD for data where mean is not appropriate such as when outliers are present or when the distribution is highly **skewed**.

The value of the variance (or of SD), just as of the mean, depends on the unit of measurement. It would be high if the unit was micrograms per liter instead of milligrams per liter. The SD of serum iron values in healthy subjects may be 25 when measured in micrograms per liter but would be only 0.025 when measured in milligrams per liter.

### *Variance and SD in Grouped Data*

Let there be $K$ intervals $(a_0, a_1)$, $(a_1, a_2)$, …, $(a_{K-1}, a_K)$ and their midpoints $x_k=(a_{k-1}+a_k)/2$, $k=1, 2, \ldots, K$. Let the frequency in the $k$th interval be $f_k$. Then, for

$$\text{grouped data: sample variance } s^2=\frac{\Sigma f_k(x_k-\bar{x})^2}{n-1},$$

and for

$$\text{grouped data: sample SD } s=\sqrt{\frac{\Sigma f_k(x_k-\bar{x})^2}{n-1}},$$

where $\bar{x}=\Sigma f_k/n$ and $n=\Sigma f_k$. As in the case of the mean for grouped data, SD also involves approximation in that all observations in an interval are assumed centered on the midpoint. This works fairly well when the width of the intervals is not large.

Sometimes, as when the data do not have a Gaussian distribution, interquartile range or range is used as a measure of dispersion in place of SD. However, SD may still be needed for specific applications such as for meta-analysis. For estimating the sample mean and SD from the sample size, median, range, and/or interquartile range, see Wan et al. [1].

### *Variance of Sum or Difference of Two Measurements*

Mean has the property that mean$(x+y)$ = mean$(x)$ + mean$(y)$. But variance or the SD does not have this feature. The SD of sum or difference of the values is not the sum or difference of SDs. In a specific situation when $x$ and $y$ are linearly independent, variance$(x+y)$ = variance$(x)$ + variance$(y)$. In general,

$$\text{variance}(x+y)=\text{variance}(x)+\text{variance}(y)+2^*\text{covariance}(x,y),$$

**TABLE S.18**
**Illustration of Relation of Variance($x - y$) with Variance($x$) and Variance($y$)**

| | SysBP | DiasBP | Pulse Pressure |
|---|---|---|---|
| | $x$ | $y$ | $x - y$ |
| | 116 | 72 | 44 |
| | 124 | 82 | 42 |
| | 145 | 89 | 56 |
| | 139 | 87 | 52 |
| | 140 | 93 | 47 |
| | 118 | 77 | 41 |
| | 136 | 85 | 51 |
| | 127 | 79 | 48 |
| | 129 | 84 | 45 |
| | 110 | 67 | 43 |
| | 115 | 79 | 36 |
| | 128 | 83 | 45 |
| **Mean** | **127.25** | **81.42** | **45.83** |
| **Variance** | **124.20** | **52.08** | **29.24** |
| **Covariance($x$, $y$)** | **73.52** | | |

where **covariance**$(x, y) = \Sigma(x_i - \bar{x})(y_i - \bar{y})/(n - 1)$. This is the sum of the product of deviations and is applicable when each value of $x$ has correspondingly one value of $y$. Note that $(x + y)$ can be calculated only in such paired case. For variance of $(x - y)$, the sign before the covariance becomes negative. For linearly independent $x$ and $y$, covariance becomes zero. The following example illustrates a situation where the difference of values is used, and such equations could be useful.

Pulse pressure is the difference between systolic pressure (SysBP) and diastolic pressure (DiasBP). A value around 45 mmHg is considered normal. A genuine low value such as less than 25 mmHg may occur in shock or aortic stenosis. A high value such as more than 60 mmHg may indicate stiffness of major arteries or a leak in the aortic valve. Thus, the pulse pressure itself is an important medical parameter irrespective of actual values of systolic and diastolic levels.

Suppose 12 persons have systolic and diastolic levels as shown in Table S.18. Pulse pressure is also shown, as well as the mean and SDs of $x$ and $y$. Note that mean$(x - y)$ = mean$(x)$ – mean$(y)$, but variance$(x - y) \neq$ variance$(x)$ – variance$(y)$. In this case, covariance$(x, y) = 73.52$—thus, the variance of pulse pressure = $124.20 + 52.08 - 2 \times 73.52 = 29.24$. This is the same as obtained in Table S.18 by directly computing the variance of differences.

In case you are interested, see the topic **confidence interval (CI) for variance**. The CI for the SD is obtained by taking the square root of the lower and upper limits of the CI for variance.

1. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Medical Research Methodology* 2014, **14**:135. http://www.biomedcentral.com/1471-2288/14/135/

## standard error (SE), *see also* sampling distribution of proportion $p$ and mean $\bar{x}$

The term *standard error (SE)* is used for the **standard deviation** (SD) of the sample summaries such as mean, proportion, correlation coefficient, median, and odds ratio. This presumes that a large number of samples are available, and this kind of summaries can be computed for each sample, which would differ from sample to sample, giving rise to the need to measure intersample variation. However, in practice, a large number of samples are not needed, and the SE can be worked out in most cases by theoretical derivations. For example, the SE of mean is generally smaller than that of sample median, which implies that means do not differ as much from sample to sample as medians do. Let there be no confusion: the SD is a measure of variation of individual values from subject to subject, and the SE is a measure of variation of sample summaries, called sample "statistics" in plural.

SEs are needed to assess the **reliability** of the estimates, for constructing **confidence intervals** (CIs), and for constructing **test criterion** for statistical tests of hypotheses. Reliability of the estimation is the inverse of the SE—the higher the SE, the lesser the reliability. As will be seen shortly, SEs heavily depend on the sample size, and for any fixed sample size, the statistics with lower SEs are considered better relative to those with higher SEs since a lower SE implies more reliability—the estimates with low SEs do not differ as much from sample to sample as do the estimates with higher SEs. For example, the SE of the mean is $\sigma/\sqrt{n}$, and that of the median from a Gaussian distribution is nearly 1.25 times the SE of the mean—thus, the mean is preferred over the median in most situations.

The most commonly used SEs are for sample mean $(\bar{x})$ and sample proportion $(p)$, which are described in the topic **sampling distribution**. For SE of attributable risk, correlation coefficient, difference between means and proportions, median, odds ratio, predicted $y$ in simple linear regression, regression coefficient and intercept in simple linear regression, relative risk, and variance, see the concerned topic on **confidence interval**. The SEs mentioned therein are based on **Gaussian conditions** that are generally fulfilled with large samples; small sample counterparts for most of these SEs are intricate.

An important feature of the SEs is that they heavily depend on sample size, and $n$ is invariably in the denominator, meaning thereby that the SE decreases as $n$ increases. This sounds intuitive also as we expect that the reliability of an estimate increases as the sample size increases. Higher sample size and consequent lower SE imply narrower CIs (which means more confidence in the results) and higher statistical **power** to detect a specified difference by tests of hypothesis.

## standardization of values, *see* *Z*-scores

## standardized death rates

Standardized death rate is the one that is "standardized" for age–sex structure of the population so that the effect of differential age–sex structure is eliminated and the rate becomes comparable across populations. Standardization is a method of adjustment for increasing the comparability. This is generally done for age differentials as just stated but can also be done for other factors. The objective is to remove the effect of differential structure of the subgroups in the two populations under comparison. The rates are then brought to a common base and thus made comparable.

Different types of **death rates** are computed depending upon groups of interest, for example, groups classified by age and gender. The most common of these is crude death rate (CDR) computed as deaths per 1000 population per year irrespective of the age and sex structure of the population. It is called *crude* as it disregards the age–sex structure of the population. If people in an area are predominantly old, a high CDR is not as bad as in an area where

the population is predominantly young. Thus, a CDR of 8 per 1000 population in Sweden should not be construed to mean that the health status is nearly the same as in India, where also the CDR is nearly the same. The CDR in this case is misleading since India has only 8% of its population with age 60 years or above, whereas Sweden has more than 20%. The death rate among old people is naturally high, and therefore the CDR becomes high. A valid comparison is obtained when the rate is recomputed by assuming the same age structure in the two countries. This is one form of standardization. The other brings the age-wise mortality pattern to a common base. Both require age-specific death rates (ASDRs) and also assumption of a standard or a reference population. This could be real or hypothetical, but in both cases, the standard is arbitrary (see the example in this section). The same can be done for sex differentials also, but let us keep that aside for our discussion. In case of age standardization, the standard may have a predefined age structure or a predefined ASDR. These two methods of standardization can give entirely different results.

### *Direct and Indirect Standardized Death Rate*

In the direct method, actual ASDRs in the study population are used on the standard population that has a predefined age structure. Thus,

$$\text{directly standardized death rate} = \frac{\Sigma_k P_{ks} d_k}{\Sigma_k P_{ks}};\ k = 1,2,\ldots,K,$$

where the population is divided into $K$ age groups.

$P_{ks}$ is the predefined standard population (in percent or count) in the $k$th age group.
$d_k$ is the ASDR in the $k$th age group of the study population.

Thus, the age structure is standardized. When such a standardization is done for two populations, any difference between the two can be ascribed to the difference in their ASDRs.

Direct standardization is possible only when ASDRs in the study population are known. If they are not known, the **indirect method** is used. In this method, predefined values of ASDRs in the standard population are used on the actual age structure of the study population. Thus,

$$\text{indirectly standardized death rate} = \frac{\Sigma_k p_k D_{ks}}{\Sigma_k p_k};\ k = 1,2,\ldots,K,$$

where

$p_k$ is the actual *study* population (in percent or count) in the $k$th age group.
$D_{ks}$ is the predefined ASDR in the $k$th age group of the standard population.

When this is done for two populations, both are then based on the same ASDRs, and any difference between the two would be due to their differential age structure.

### *Comparison between Direct and Indirect Methods*

Comparison between these two methods of standardization can be easily studied by means of an example. Consider the data in Table S.19 on age distribution of population of the United States and Venezuela in 2002, and the age-specific death rates [1].

The standard population given in Table S.19 is as suggested by the World Health Organization (WHO) [2]. We proposed the standard ASDRs in the last column. No widely acceptable standard is available for ASDRs.

Note that the CDR in the United States is nearly twice of the CDR in Venezuela. Also, note that the United States has more people in the old age group. From the formula given earlier, the directly standardized death rate is given as follows:

$$\text{for the United States} = \frac{1.7\times 8.9 + 0.2\times 17.3 + \cdots + 148.3\times 0.6}{8.9 + 17.3 + \cdots + 0.6}$$

$$= 5.5 \text{ per 1000 population,}$$

and

**TABLE S.19**
**Population and Death Rates in Different Age Groups in the United States and Venezuela, and the WHO World Standard**

| | United States 2002 | | Venezuela 2002[a] | | Standard | |
|---|---|---|---|---|---|---|
| Age Group (Years) | Population (%) | ASDR per 1000 | Population (%) | ASDR per 1000 | Population (%) | ASDR per 1000 |
| A | B | C | D | E | F | G |
| 0–4 | 6.8 | 1.7 | 11.1 | 3.5 | 8.9 | 10.0 |
| 5–14 | 14.2 | 0.2 | 21.6 | 0.3 | 17.3 | 1.0 |
| 15–24 | 14.1 | 0.8 | 19.4 | 1.8 | 16.7 | 0.5 |
| 25–34 | 13.8 | 1.0 | 15.6 | 2.1 | 15.5 | 0.5 |
| 35–44 | 15.6 | 2.0 | 12.9 | 2.4 | 13.7 | 1.0 |
| 45–54 | 13.9 | 4.3 | 9.2 | 4.5 | 11.4 | 5.0 |
| 55–64 | 9.2 | 9.5 | 5.5 | 8.8 | 8.3 | 10.0 |
| 65–74 | 6.3 | 23.1 | 3.2 | 21.5 | 5.2 | 15.0 |
| 75–84 | 4.4 | 55.6 | 1.3 | 52.4 | 2.4 | 35.0 |
| 85+ | 1.6 | 148.3 | 0.3 | 140.0 | 0.6 | 100.0 |
| **Total** | **100.0** | **8.5 CDR** | **100.0** | **4.2 CDR** | **100.0** | **5.0 CDR** |

[a] Data from United Nations Statistics Database.

$$\text{for Venezuela} = \frac{3.5\times 8.9 + 0.3\times 17.3 + \cdots + 140.0\times 0.6}{8.9 + 17.3 + \cdots + 0.6}$$

$$= 5.8 \text{ per } 1000 \text{ population.}$$

The difference has now almost vanished after standardization.

The indirectly standardized death rate is as follows:

$$\text{for the United States} = \frac{10.0\times 6.8 + 1.0\times 14.2 + \cdots + 100.0\times 1.6}{6.8 + 14.2 + \cdots + 1.6}$$

$$= 6.8 \text{ per } 1000 \text{ population,}$$

and

$$\text{for Venezuela} = \frac{10.0\times 11.1 + 1.0\times 21.6 + \cdots + 100.0\times 0.3}{11.1 + 21.6 + \cdots + 0.3}$$

$$= 3.9 \text{ per } 1000 \text{ population.}$$

This death rate for Venezuela is much lower.

Note the following for this example:

1. Venezuela has a higher population in younger age groups, and the ASDRs are also higher in younger age groups.
2. For direct standardization of death rate in the United States, ASDRs in column C are multiplied by the standard population in column F, added, and divided by the total of the standard population in column F. The numerator so obtained is the expected number of deaths that would have occurred if the age structure were standard. For direct standardization of death rate in Venezuela, its ASDRs in column E are multiplied by the standard population in column F.
3. For indirect standardization, the population (column B for the United States and column D for Venezuela) is multiplied by the ASDRs in the standard population (column G), added, and divided by the total population in the country, i.e., the total of column B for the United States and of column D for Venezuela. The numerator in this case is the expected number of deaths that would have occurred if the ASDRs were the same as in the standard population.
4. Directly standardized DR is less than the CDR in the United States because the standard population is less in the age groups where the United States ASDR is high. When the standard population is greater in higher mortality groups as in Venezuela, the directly standardized DR is greater than the CDR.
5. Directly standardized rate brings the age structure of the two populations to the same pattern. When this is done, the death rate in the United States becomes nearly the same as that in Venezuela.
6. Indirect standardization has a different effect in this case. The indirectly standardized DR is higher in the United States. The difference between the indirectly standardized rates in the United States and Venezuela is mostly due to the difference in age structure in the two countries.
7. In this example, the age structure is given in terms of percentages, but the actual population can also be used in the same formulas.
8. In this example, the two methods of standardization can give entirely opposite results. Thus, it is important that the method be correctly chosen in accordance with your objective of doing the standardization. In addition, the interpretation of the standardized rate should be proper for the method used.

The standardized rate depends heavily on the standard chosen. No universal standard is available, and this is arbitrarily chosen. If a desirable structure exists or can be constructed, then that can be chosen as the standard. If not, a structure that is of middling type or commonly seen or easy to implement can be chosen as the standard. Column F of Table S.19 gives the WHO World Standard population [2] as mentioned earlier. This actually extends to age 100 years and above and is given in 5-year intervals, but in this table, an abridged version with 85 years and above as the last group and age intervals of 10 years is given.

Since the standard is arbitrary anyway, it should be simple. This is not the case with the WHO World Standard population. The ASDRs in column G are our own standards and are simple. For interregional comparison within a country, the age structure of the total country or its ASDRs can be used as the standard.

The direct method seems more appropriate in most cases because this gives the death rate that is expected for standard age structure. But this method cannot be used when the ASDRs are not known. Also, this method should not be used when ASDRs are unstable and are based on a small number of subjects. Indirect standardization is generally used for disease mortality because of unstable ASDRs. This, when used for small groups instead of the general population, is called the standardized or adjusted mortality rate and leads to a popular measure known as the **standardized mortality ratio**. Such an adjustment can be done not only for age but also for any other factor that might influence the mortality pattern.

The illustration in our example is for population mortality as it is the most common application of standardized death rate. However, this can also be used for disease-specific mortality. For example, if age-specific death rates for circulatory diseases are known and if the purpose is to compare two age-wise diverse populations, the comparison should be based on standardized rates. For emphasizing the importance of the choice of the standard in this context, Ahmad et al. [2] show for the year 1995 that the standardized death rate due to circulatory disease in males in the United States using WHO World Standard is 285 per 100,000 population but is 372 when a Scandinavian standard is used. This reflects a large difference of 23% and underscores the heavy dependence of standardized death rate on the standard chosen for this purpose.

1. Hoyert DL, Kung H-C, Smith BL. Deaths: Preliminary data for 2003. *Natl Vital Stat Rep* 2005;53(15):7. http://www.cdc.gov/nchs/data/nvsr/nvsr53/nvsr53_15.pdf
2. Ahmad OB, Boschi-Pinto C, Lopez AD, Murray CJL, Lozano R, Inoue M. Age standardization of rates: A new WHO standard. *GPE Discussion Paper Series: No. 31*. World Health Organization. http://www.who.int/healthinfo/paper31.pdf

## standardized deviate, *see* *Z*-score

## standardized mortality ratio (SMR)

The standardized mortality ratio (SMR) is the ratio of the number of observed deaths in a study group to the number that would be

expected if the study group had the same specific rates as in a standard group (see **standardized death rates**). This procedure is the same as the indirect method of standardization. The ratio is sometimes multiplied by 100 for expressing it in terms of percentage. Thus,

$$\text{SMR} = \frac{\text{observed number of deaths}}{\text{expected number of deaths}} * 100,$$

where the denominator is based on the specific rates in the chosen standard population. These are not necessarily age-specific but could be gender-specific, exposure-specific, or specific for any other categorization. An SMR greater than 100 is interpreted as indicating that the study group has excess mortality relative to the standard. In this method, more stable rates of the larger population are applied to the smaller study group to obtain the expected number of deaths. The SMR gives a measure of the likely excess or reduction in mortality in the study group. Note that the SMRs in two groups can be compared only when they are based on the same standard population. The study and the standard group can be disease and control groups, respectively, in a case–control study or general populations of two types or any other groups of interest. The following example illustrates its application.

Stress of work, exposure to fiber dust, and other factory environmental factors are known to cause excess mortality in the staff of a textile mill. Their age distribution and calculation of SMR are shown in Table S.20.

Expected deaths are obtained in this example by applying the national age-specific death rates (ASDRs) to the number of textile workers in different age groups. Since the national rate for the age group 25–34 years is 3.0 per 1000 population, the expected deaths in 400 workers of this age group are 3.0 × 400/1000 = 1.2; the same goes for other age groups. If the total number of observed deaths is 10 and the expected deaths based on the national rates are 6.8, then from the formula given earlier, SMR = (10/6.8)100 = 147. This shows that the mortality level of textile workers (in this mill) was 147% of the national average. This is 47% higher than that experienced by the national population. Excess mortality in textile workers was never in doubt, but SMR delineates the exact magnitude of this excess.

For a similar but more extensive application of SMR, see Johansen and Olsen [1]. They investigated mortality from amyotrophic lateral sclerosis (ALS) and other chronic disorders among male employees of electric supply companies in Denmark. The 21,236 men included in the study accrued nearly 303,000 person-years of follow-up. They observed 14 deaths from ALS when only 6.9 deaths were expected on the basis of the national ASDRs. This yielded an SMR of nearly 200. This means that these employees are dying from ALS nearly twice as much as the general population. These excess deaths can be attributed to the risks present in the company environment.

**TABLE S.20**
**Calculation of Standardized Mortality Ratio (SMR) in Textile Workers**

| Age Group (Years) | ASDR[a] in the National Population per 1000 | Textile Workers | |
|---|---|---|---|
| | | Number of Workers | Expected Deaths |
| 25–34 | 3.0 | 400 | 1.2 |
| 35–44 | 5.0 | 300 | 1.5 |
| 45–54 | 8.0 | 200 | 1.6 |
| 55–64 | 25.0 | 100 | 2.5 |
| **Total** | | **1000** | **6.8** |

[a] Age-specific death rate.

1. Johansen C, Olsen JH. Mortality from amyotrophic lateral sclerosis, other chronic disorders, and electric shocks among utility workers. *Am J Epidemiol* 1998;148:362–8. http://aje.oxfordjournals.org/content/148/4/362.full.pdf

## standard normal distribution, *see* Gaussian probability (how to obtain)

## STARD statement

An acronym for Statement for Reporting Studies of Diagnostic Accuracy, STARD consists of a checklist of 25 items and a flow chart that authors can use to ensure that all relevant information is present in their report on research on diagnostic accuracy of a medical test [1]. This is one of the resources used to improve reporting of diagnostic accuracy studies, and it is realized that complete and informative reporting can only lead to better decisions in health care.

Medical tests are quite commonly used in clinical setups for providing a certain degree of assurance and to confirm the diagnosis for a disease or any other health condition. The term *medical test* is generic and refers to any method of obtaining additional information on the patient, and includes not just laboratory and imaging tests but also history and examination. Thus, it can be only signs and symptoms. These tests help in initiating a treatment and in deciding when to modify or stop the treatment. However, the problem is that these tests rarely, if ever, provide infallible results, and the validity of the results depends on a host of factors such as condition of the patient, method of administering the test, the inherent qualities of the test, quality of reagents, care in following the recommended procedure, etc.

Whenever a new test is developed (call it *index test*), it is expected that its performance will be evaluated against the existing procedure (or a reference standard) and its superiority would be established. The superiority could be either in terms of efficacy as measured by sensitivity and specificity (or predictivities) or could be efficiency in terms of convenience, cost, and time. The test is tried on a series of subjects and results are reported. It has been observed that sometimes this reporting is incomplete, and the reader is not able to comprehend all aspects. STARD statement is an attempt to standardize the reporting so that nothing is missed and all steps of the study are properly described. Its first version appeared in 2003. The latest version (2015) is given in Table S.21, and the flow chart is in Figure S.18.

Ever since STARD statement has been issued, it is being increasingly used for reporting of diagnostic studies, and the reporting quality seems to be improving [2]. Cheung et al. [3] used this format to report assessment of breast specimens with or without calcifications for diagnosing malignant and atypia for mammographic breast microcalcifications without mass, and Alcoba et al. [4] used this to report results of their study on proadrenomedullin and copeptin in pediatric pneumonia.

## TABLE S.21
## STARD Statement

| Section and Topic | No | Item |
|---|---|---|
| | | **Title or Abstract** |
| | 1 | Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity/specificity, predictive values, or area under the curve) |
| | | **Abstract** |
| | 2 | Structured summary of study design, methods, results, and conclusions |
| | | **Introduction** |
| | 3 | Scientific and clinical background, including the intended use and clinical role of the index test |
| | 4 | Study objectives and hypotheses |
| | | **Methods** |
| Study design | 5 | Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study) |
| Participants | 6 | Eligibility criteria |
| | 7 | On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry) |
| | 8 | Where and when potentially eligible participants were identified (setting, location, and dates) |
| | 9 | Whether participants formed a consecutive, random, or convenience series |
| Test methods | 10a | Index test, in sufficient detail to allow replication |
| | 10b | Reference standard, in sufficient detail to allow replication |
| | 11 | Rationale for choosing the reference standard (if alternatives exist) |
| | 12a | Definition of and rationale for test positivity cutoffs or result categories of the index test, distinguishing prespecified from exploratory |
| | 12b | Definition of and rationale for test positivity cutoffs or result categories of the reference standard, distinguishing prespecified from exploratory |
| | 13a | Whether clinical information and reference standard results were available to the performers or readers of the index test |
| | 13b | Whether clinical information and index test results were available to the assessors of the reference standard |
| Analysis | 14 | Methods for estimating or comparing measures of diagnostic accuracy |
| | 15 | How indeterminate index test or reference standard results were handled |
| | 16 | How missing data on the index test and reference standard were handled |
| | 17 | Any analyses of variability in diagnostic accuracy, distinguishing prespecified from exploratory |
| | | **Results** |
| Participants | 18 | Intended sample size and how it was determined |
| | 19 | Flow of participants, using a diagram (Figure S.18) |
| | 20 | Baseline demographic and clinical characteristics of participants |
| | 21a | Distribution of severity of disease in those with the target condition |
| | 21b | Distribution of alternative diagnoses in those without the target condition |
| | 22 | Time interval and any clinical interventions between index test and reference standard |
| Test results | 23 | Cross-tabulation of the index test results (or their distribution) by the results of the reference standard |
| | 24 | Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals) |
| | 25 | Any adverse events from performing the index test or the reference standard |
| | | **Discussion** |
| | 26 | Study limitations, including sources of potential bias, statistical uncertainty, and generalizability |
| | 27 | Implications for practice, including the intended use and clinical role of the index test |
| | | **Other Information** |
| | 28 | Registration number and name of registry |
| | 29 | Where the full study protocol can be accessed |
| | 30 | Sources of funding and other support; role of funders |

*Source:* Bossuyt PM et al., *BMJ* 2015;351:h5527. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4623764/.
*Note:* At the start of each item row, authors should specify the page number of the manuscript where the item can be found.

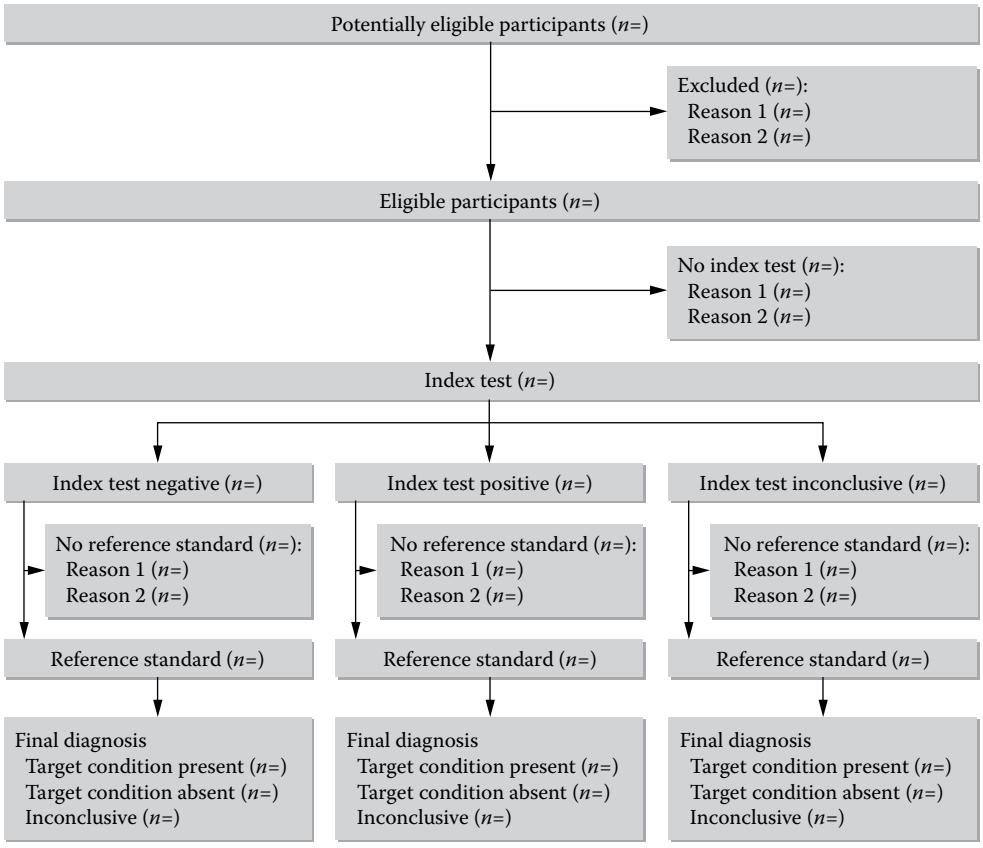



**FIGURE S.18** STARD flow chart. (From Bossuyt PM et al., *BMJ* 2015;351:h5527. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4623764/.)


1. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG et al. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 2015;351:h5527. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4623764/
2. Selman TJ, Morris RK, Zamora J, Khan KS. The quality of reporting of primary test accuracy studies in obstetrics and gynaecology: Application of the STARD criteria. *BMC Women's Health* 2011;11:8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3072919/
3. Cheung YC, Juan YH, Ueng SH, Lo YF, Huang PC, Lin YC, Chen SC. Assessment of breast specimens with or without calcifications in diagnosing malignant and atypia for mammographic breast microcalcifications without mass: A STARD-compliant diagnostic accuracy article. *Medicine (Baltimore)* 2015 Oct;94(42):e1832. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4620838/
4. Alcoba G, Manzano S, Lacroix L, Galetto-Lacour A, Gervaix A. Proadrenomedullin and copeptin in pediatric pneumonia: A prospective diagnostic accuracy study. *BMC Infect Dis* 2015 Aug 19;15:347. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4543464/


## statistical analysis, *see* analysis (statistical)

## statistical fallacies, *see* fallacies (statistical)

## statistical inference, *see* inference (statistical)

## statistical models, *see* models (statistical), and also parametric models

## statistical reviews

As is true for almost all **empirical research**, nearly all medical research articles are full of numbers. Proper designing and execution, satisfactory collection of data, adequate processing of data, and correct interpretation are the key for appropriate conclusions, as well as for assessment of the confidence in those conclusions. In an editorial in a reputed magazine *Science*, McNutt [1] stated that there have been far too many cases where the quantitative analysis of those numbers has been flawed, causing doubts about the authors' interpretation and uncertainty about the result. Since subject matter experts generally are not in data analysis, statistical reviews are gaining increasing attention. Effective 1 July 2014, *Science* has established a Statistical Board of Reviewing Editors, consisting of experts in various aspects of statistics and data analysis, to provide better oversight of the interpretation of observational data. Many medical journals have also realized the importance of statistical reviews and have expert biostatisticians in their panel.

There are a large number of studies that have established that statistical quality, and hence of conclusions, is rather poor in many articles. This happens because the spread of knowledge of statistical methods among medical professionals has not kept pace with the speed with which statistical applications are done due to wide availability of software packages. This availability seems to have propelled many researchers to use these methods as a **black box** without understanding the intricacies. This has resulted in an urgency for statistical reviews.

In a comparison of articles in two top medical journals, namely, the *New England Journal of Medicine* and the *Nature Medicine*, Strasak et al. [2] observed that while nearly 90% of the articles they reviewed contained inferential statistics, the documentation of

statistical methods applied in the articles was generally poor or insufficient, and recommended that closer attention to statistical issues should be considered to raise the standards. In another study by the same authors [3] on Austrian journals, the findings were even worse.

The problem is compounded by the availability of huge data that are now stored online for various interactions and transactions. According to EMC International Data Corporation Digital Universe Studies [4], there were close to 4 zetabytes ($10^{21}$) of information stored in 2013 with doubling every 2 years. This, however, lacks quality but is almost indiscriminately used to draw inferences. Expert data analysts that can filter meaningful messages out of such noisy data are not many yet. In the context of medicine, results of laboratory and imaging investigations and genotyping may be of good quality, but there is a question mark on the quality of research articles that are now so vociferously published across the world. It is not just the analysis and interpretation, but also the quality of data that go into empirical studies can be questionable and seems to have been rarely assessed. It may be a good idea if each journal periodically undertakes audit of what it has been publishing, and implement the lessons learned.

1. McNutt M. Raising the bar. *Science* 4 July 2014;345(6192):9. https://www.sciencemag.org/content/345/6192/9.full
2. Strasak AM, Zaman Q, Marinell G, Pfeiffer KP, Ulmer H. The use of statistics in medical research: A comparison of *The New England Journal of Medicine* and *Nature Medicine*. *The Amer Stat* 2007;61(1):47–55. http://www.tandfonline.com/doi/abs/10.1198/000313007X170242?journalCode=utas20
3. Strasak AM, Zaman Q, Marinell G, Pfeiffer KP, Ulmer H. The use of statistics in medical research: A comparison of Wiener Klinische Wochenschrift and Wiener Medizinische Wochenschrift. *Austrian J Stat* 2007;36(2):141–52. http://www.stat.tugraz.at/AJS/ausg072/072Strasak.pdf
4. IDC. *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*. http://www.emc.com/leadership/digital-universe/2014iview/index.htm

## statistical significance, *see* significance (statistical)

## statistics, *see also* biostatistics

The term *statistics* is used in two senses—(i) as plural as in day-to-day language for data and (ii) as singular for the science dealing with data. As plural, statistics are just plain facts, mostly numerical measurements or counts, and it is in this sense that we talk about industry statistics, geographical statistics, and even health statistics such as how many people are dying of what causes and how many doctors of various types we have in a community. These statistics are collected for reference or for information. They can be primary (directly collected from the source) or secondary (collected by someone else and put to use by somebody else). For example, the government may collect data on a number of cases admitted of various diseases in different hospitals so that they are primary for them, and researchers use them for analysis as secondary data. In addition, different data can have different **scales of measurement** such as nominal, ordinal, and metric. These characteristics are utilized for deciding the method of summarizing the data and their presentation for easy understanding and to be able to use them later for analysis.

As singular, statistics is a science, a discipline that extracts meaning out of data. This can be viewed as comprising two distinct entities (Figure S.19). The first is *data science* where there is no probability component, such as data mining from enormous data that are now generated online or otherwise, which is done to explore for patterns and configurations that can help, for example, in marketing or in assessing health sector demands by different segments of population. This can also be termed as quantitative information focused on specific groups. Collection and presentation of data in terms of tables, graphs, and maps are part of this exercise.

Statistics is best defined as the science of managing uncertainties [1], particularly their measurement, control, and communication. Basically it is an inferential science where probabilities play a vital role, and all conclusions are attached with a measure of uncertainty. Processing and analysis of data are the main tools that propel this activity. Models that express a complex phenomenon in simple terms, such as regressions and structural equations, are built up based on probability theory so that most features of the data are captured and the model can be effectively utilized in practice. Many times extrapolations are done such as forecasting of future events, and sometimes interpolations for intermediary unobserved values are done using these models. However, the main strategy requiring probabilities is inferring about a "population" from a "sample." When successful, making inferences becomes much less expensive and possibly more accurate with this strategy as better methods can be used to obtain data on a fraction of subjects instead of all the subjects. Sample **designs** and design of experiments such as clinical trials help in controlling uncertainties, and quantities like ***P*-value** measure the uncertainties and are used for communicating the results.

Inferential statistics too can be subdivided into two distinct components—one that deals with estimation and the other that is concerned with testing of hypothesis. These are discussed in detail under the topics **estimate** and **confidence intervals (the concept of)** for estimation, and the topic **testing of hypothesis (philosophy of)** for testing. Both these methods are based on samples and are heavily based on probabilities.

**Uncertainties** pervade so much in medical sciences that statistics in this discipline has acquired a new identity called medical **biostatistics**. This "branch" of statistics is more medical than mathematical, whereas statistics by itself is mostly a mathematical science.

1. Indrayan A. *Medical Biostatistics*, Third Edition. Chapman & Hall/CRC Press, 2012.

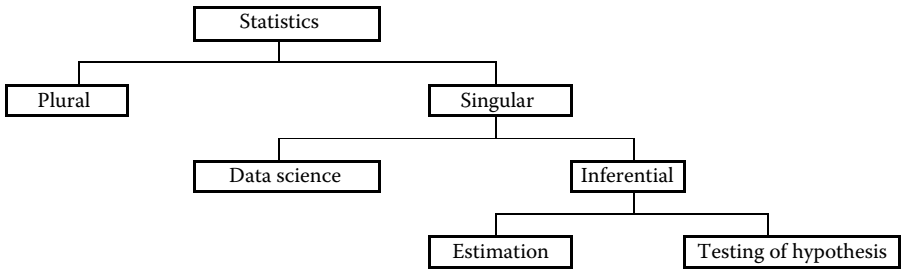



**FIGURE S.19** Usage of the term *statistics*.

## stem-and-leaf plot, *see* histogram

## stepwise methods, *see also* best subset method of choosing predictors

In a **regression model** setup, including logistic regression, stepwise methods are used for variable selection when a large number of candidate variables are available as regressors. Including all these regressors together (called the "enter" method) (i) may compromise the applicability of the model and lose its parsimony; (ii) may require an inordinately large sample that may not be feasible; (iii) may cause the cost of obtaining data on so many variables to become prohibitive; and (iv) may increase the chance of multicollinearity that adversely affects the efficiency of the regression models. For example, for predicting thyroid status, more than 100 signs and symptoms can be identified such as tremors, hoarseness of voice, altered skin texture, and brittle nails. Including all of them in a regression will require an enormous sample, and even if this kind of sample is available, the model so obtained based on so many regressors can hardly be used. Thus, it is imperative to look for a subset of the regressors that can still predict or explain a phenomenon with a reasonable degree of validity. The best option is to select the variables based on their biological relevance for the outcome—include ones that are more relevant and exclude the others. When the biological picture is not clear, or when all the variables seem to have nearly equal relevance, methods based on the **statistical significance** of the variables in explaining the outcome can be used. In many situations, statistical selection of variables tends to overfit—more variables would be added than required. Thus, it is advisable to use a strict criterion for addition of the variables.

Variable selection by itself is not an end. Ultimately, the model finally arrived at must serve the purpose. Beware that the model so chosen may have no biological relevance—that is not the consideration in this method and this must be separately assessed.

Stepwise is one of the many methods available for variable selection. This is one of a set of statistical methods for automatically selecting a "better" (although not necessarily the best) subset of **regressors** for regression analysis. The dependent (outcome) in the regression can be quantitative (ordinary regression), qualitative (logistic regression), or hazard (Cox regression), and the method is applicable to all types of regressions. According to Morozova et al. [1], such automatic stepwise methods in regression often perform poorly in terms of both the variable selection and estimation of the coefficients, especially when the number of independent variables is large and multicollinearity is present. The model arrived at from these methods tends to work well on the data used for their development but works poorly on a new set of data. If a subset of variables is to be treated as a group because of biological relevance or for any other reason, this method does not allow that because it is blind to such configurations unless specific provision for this is made at the time of analysis. Yet, stepwise algorithms remain the dominant method in medical and epidemiological research for variable selection. Also, these automatic selection methods can render an explanatory model ineffective since the objective of explanatory models is to find the overall adjusted effect, and the adjustment in case the case of automatic selection is only for the variables that remain after such selection. All these are large sample methods; small samples relative to the number of variables under consideration can give weird results. The most commonly used stepwise methods are forward selection, backward elimination, and a combination of both of these, also known as stepwise method. For alternative methods, see Morozova et al. [1].

### *Forward Selection*

In the forward selection method, the start is made with no variable in the model, and sequentially one is added at a time depending on a preset criterion. Once a variable is included, it stays. The criterion used to decide whether to add a particular variable to an existing model could be one of the following in ordinary quantitative regression: (i) The statistical significance of the reduction in the **residual sum of squares** (RSS) resulting from including that variable in the model. Incidentally, this is the same as the increase in the regression sum of squares whose significance is tested by ***F*-test**. This test will give the ***P*-value** for each variable that measures the statistical significance. Add that variable first for which the reduction in RSS has the least *P*-value (i.e., statistically the most significant). The process is repeated for the remaining variables. You can preset another criterion such as any variable with reduction yielding a *P*-value greater than, say, 0.10 will not be added. This procedure is called *P*-value to add. (ii) A similar procedure is when *P*-value is obtained directly for the variable and not for the RSS. Statistical software packages generally have provision for these two procedures. (iii) The third criterion could be the difference in the value of the **coefficient of determination** $\eta^2$ (which becomes **multiple correlation coefficient** $R^2$ in case of **linear regression**) when the variable is added. In this case, statistical significance would take you back to the first criterion, but you can decide on the basis of the absolute addition in place of the statistical significance. Do not add variables that you find are not substantially increasing the value of $\eta^2$. (iv) There are other criteria such as **Akaike information criterion**, **adjusted $R^2$**, and univariate **correlation coefficient** with the outcome variable. All these may give the same result. For logistic and Cox regression, the criterion changes to the **chi-square**–based **deviance**. Further details of forward selection method have been provided by Harrell [2].

For forward selection in action, see Akbar et al. [3] who used this as one of the methods for selecting best predictors of retention time of phenolic compounds in a chromatography. Wang et al. [4] used the forward selection method to confirm their findings regarding best biomarkers of neonatal sepsis identified by other methods.

### *Backward Elimination*

Now consider starting with all the potential regressors and delete one by one as per their nonsignificance assessed by a prefixed criterion, and the regressors once deleted remain outside the purview of inclusion. In the case of forward selection, the regressors are added one at a time based on their significance, but in backward elimination, the regressor with, say, the highest *P*-value is deleted first. This method deletes all nonsignificant regressors, whereas forward selection retains only the significant ones. To understand the difference, just realize that the regressors not significant are not necessarily the same as the complement of the significant ones. This method is more appropriate when the candidate variables are not too many and you want to throw out some that are less relevant. On the other hand, forward selection is better when you have a large number of variables and you want to select a few most relevant ones.

Pina et al. [5] used the backward elimination method for selecting variables related to pain intensity in cancer, and used the criteria of $P = 0.2$ for deletion. Bucsa et al. [6] also used this method for evaluating demographic characteristics, comorbidities, medications, and laboratory monitoring, and quantified the patients with serum creatinine and potassium levels above the upper normal limit in logistic regression to identify significant predictors of combination therapy in hospitalized patients on angiotensin-converting enzyme inhibitors.

### *Real Stepwise Method*

This is a combination of the forward selection and backward elimination in the sense that the regressors once included by forward selection are assessed again at the next stage, and deleted if any of them loses statistical significance. Thus, the option of exclusion of one or more of the existing regressors is exercised after inclusion of a new regressor as per their significance. This can help in fine-tuning the model more than any of the two previously described methods. The significance can be tested either by using *t*-test for the regression coefficients or criteria such as *F*-to-enter and *F*-to-remove based on the entire model at hand at successive stages. These and *P*-values have to be decided beforehand. It is very important not to delegate your responsibility to the computer—you should examine the model for its biological adequacy so that it is medically justified and can be properly explained.


1. Morozova O, Levina O, Uusküla A, Heimer R. Comparison of subset selection methods in linear regression in the context of health-related quality of life and substance abuse in Russia. *BMC Med Res Methodol* 2015 Aug 30;15:71. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4553217/
2. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* Springer, 2010.
3. Akbar J, Iqbal S, Batool F, Karim A, Chan KW. Predicting retention times of naturally occurring phenolic compounds in reversed-phase liquid chromatography: A quantitative structure–retention relationship (QSRR) approach. *Int J Mol Sci* 2012 Nov 20;13(11):15387–400. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3509648/
4. Wang K, Bhandari V, Chepustanova S, Huber G, O'Hara S, O'Hern CS, Shattuck MD, Kirby M. Which biomarkers reveal neonatal sepsis? *PLoS One* 2013 Dec 18;8(12):e82700. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3867385/
5. Pina P, Sabri E, Lawlor PG. Characteristics and associations of pain intensity in patients referred to a specialist cancer pain clinic. *Pain Res Manag* 2015 Sep-Oct;20(5):249–54. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4596632/
6. Bucsa C, Moga DC, Farcas A, Mogosan C, Dumitrascu DL. An investigation of the concomitant use of angiotensin-converting enzyme inhibitors, non-steroidal anti-inflammatory drugs and diuretics. *Eur Rev Med Pharmacol Sci* 2015 Aug;19(15):2938–44. http://www.europeanreview.org/article/9312


## stillbirth rate/ratio

Generally, pregnancy terminations are considered stillbirths when the fetus or the infant is born dead and weighs at least 500 g or has body length (crown–heel) of at least 20 cm. When these measurements are unavailable, the gestational age should be at least 28 weeks. Such deaths are also called late fetal deaths. Volume of stillbirths in a community can be measured by one of the two ways as follows:

$$\text{stillbirth ratio} = \frac{\text{stillbirths}}{\text{live births}} * 1000$$

$$\text{stillbirth rate (SBR)} = \frac{\text{stillbirths}}{\text{still} + \text{live births}} * 1000.$$

The first has only live births in the denominator, whereas the second has live births + stillbirths; thus they are called "ratio" and "rate," respectively.

Stillbirths are a great public health issue as they jeopardize the life and health of the mother besides the moral and ethical questions. After expecting a child, a stillbirth can be a devastating trauma for the family. Some 2.6 million stillbirths occurred worldwide in 2009 according to the first set of comprehensive estimates [1]. Maternal obesity is considered as the dominant risk factor. Intensive efforts are made to control such births such as those outlined by Yoshida et al. [2].


1. The Lancet. *Stillbirths, 2011.* http://www.lancet.com/series/stillbirth
2. Yoshida S, Martines J, Lawn JE, Wall S, Souza JP, Rudan I, Cousens S et al. Setting research priorities to improve global newborn health and prevent stillbirths by 2025. *J Glob Health* 2016 Jun;6(1):010508. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4576458/


## stochastic processes

Stochastic process connotes a system that changes over time in an uncertain manner. Because of uncertainty, probability is an integral part of such a system, hence the name stochastic. Consider a single characteristic such as amount of drug present in the body. This changes over time in an uncertain manner—thus it is a stochastic process. Sex does not change at all over time, and age does not change in an uncertain manner—thus none of these is a stochastic process. Characteristics such as amount of drug in the body can be continuously measured over time that will give rise to continuous time process, or can be measured each morning to give rise to discrete time process.

The characteristic (e.g., amount of drug) at time $t$ can be denoted by $x(t)$, which is now a **random variable**, and $t$ can be any value between, say, 0 and $T$ for continuous time process and would be $t = 1, 2, \ldots, T$ for a discrete time process. On the other hand, a characteristic such as developing a sign of a disease is a discrete event (the sign will develop or not develop), which will give rise to discrete state process, which can also be for continuous or discrete time. Discrete state process is for counts, whereas continuous state process is for continuous variables. Thus, there are four types of stochastic processes: continuous state, continuous time; continuous time, discrete state; discrete time, continuous state; and discrete time, discrete state.

A basic discrete state process generates **Poisson distribution** when (i) the probability of at least one occurrence of the event in a given time interval is proportional to the length of the interval (if the time interval is large, the probability of occurrence is also high); (ii) the probability of two or more occurrences of the event in a very small time interval is negligible; and (iii) the numbers of occurrences of the event in disjoint time intervals are mutually independent. Consider deaths occurring in a hospital where the chance of death occurring in any given 30 h is more than the chance of occurrence in any given 10 h; the chance of occurrence of two deaths within a short span of time is negligible; and the number of deaths occurring on any day does not depend on how many occurred on any other day. Thus, all the conditions are fulfilled for deaths in a hospital in the normal course for this to be called a *Poisson process*. Since deaths occur in an uncertain manner (cannot be predicted in advance how many deaths will occur), it is a stochastic process. This process can be studied with the help of Poisson distribution, for example, to find the probability that more than 4 deaths will occur in the next 24 h. The number of accidents occurring in a county in a day and the number of cases of myocardial infarction admitted in a hospital in a week are other examples of variables with Poisson process.

There are several types of stochastic processes such as Weiner process, Dirichlet process, and Gaussian process, but the most talked about is the Markov process.

### *Markov Process*

This is the process under which change in status (such as from health to disease) depends only on the existing condition, and the past or the history does not play any role. The transition depends only on the current value and is independent of past values. Also, there must be only a few states such as full health, mild disease, severe disease, and death. But the most severe restriction in Markov process is that the probability of transition from one state to another is the same for all the subjects. This is named after the Russian mathematician Andrei Markov who proved this in 1909 [1].

Andrei Markov

When a birth takes place, the population count increases by one from the existing count, and when a death takes place, the population decreases by one from the existing count. If the chance of birth is the same and the chance of death also remains the same whatever the population count, the **birth–death** is a Markov process as far as count of people is concerned. What happens next does not depend on where the process started—whether the population is 1000 or 100,000, a birth will increase it by one and a death will decrease it by one. As long as you know the current count, the number of births or deaths that occurred previously does not matter. The same is true for an *arrival–departure process* such as for patients coming in and going out from emergency in a busy hospital. In this case, departure could be by way of admission to the regular ward from emergency, by patients leaving against medical advice, by death, or by discharge. The only condition is that the probability of arrival and departure of each patient should be the same. This may not be so, particularly for departure, since it depends on the severity of the condition of the patient. If the chance varies from patient to patient, it is not a Markov process.

Realize that Markov is a stochastic process in the sense that the change from one state to another depends on chance. It is not deterministic and cannot be predicted. Second, under the conditions just stated, a Markov process stabilizes—the population in different states will become stable at some time in the future. When stable conditions exist, modeling by Markov process helps to estimate the predictivity of transition from one state to another.

In a clinical setup, the conditions of Markov process may be fulfilled for a patient moving from disease to healthy to disease states where relapse occurs. Recurrent conditions such as epilepsy and angina episodes may also be eligible for modeling by Markov process. Borisenko et al. [2] used this to study four states of inoperable patients with functional mitral value regurgitation in German settings while getting repair of percutaneous mitral valve. These are New York Health Association (NYHA) functional classes reported by them as NYHA-I to IV. Markov process modeling helped them to work out the transition probabilities from one state to another. They used this to do cost–utility analysis.

1. Encyclopaedia Britannica. Andrey Andreyevich Markov. http://www.britannica.com/EBchecked/topic/365793/Andrey-Andreyevich-Markov
2. Borisenko O, Haude M, Hoppe UC, Siminiak T, Lipiecki J, Goldberg SL, Mehta N, Bouknight OV, Bjessmo S, Reuter DG. Cost-utility analysis of percutaneous mitral valve repair in inoperable patients with functional mitral regurgitation in German settings. *BMC Cardiovasc Disord* 2015 May 14;15:43. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4443594/

## stochastic variables, *see* variables

## stopping rules (for trials), *see also* O'Brien–Fleming procedure, Lan–deMets procedure

Stopping rules are used in clinical trials to decide to stop a trial either when enough evidence is available for the desired efficacy or when sufficient evidence is available for the futility of continuing the trial. Suppose you decide a priori that a regimen is not worth pursuing unless it has at least 30% efficacy, and decide to do a trial on $n = 500$ subjects that should give at least $500 \times 0.3 = 150$ positive responses on average in repeated trials. Since the standardized Gaussian (normal) deviate (which can be used in this case in view of large $n$) $z = (x - 150)/\sqrt{(500 \times 0.3 \times 0.7)} > 1.96$ gives $x > 170.084$, at least 171 successes are needed to be reasonably assured that the efficacy is at least 30% at 2.5% level of significance. If after 400 subjects, only 70 respond positive, you know that if all the remaining 100 respond positive, the number will not become 171. Thus, it is futile to continue the trial, and the trial should be stopped. Even if 100 respond in the first 400 subjects that gives estimated efficacy of 25%, the chance that at least 70 will respond out of the remaining 100 at this rate is $P[z > (70 - 25)/\sqrt{(100 \times 0.25 \times 0.75)}] = P(z > 10.4)$, which gives $P < 0.001$. Since the chance is less than 1 in a thousand (in fact, almost zero), it seems futile to continue the trial.

As already stated, an ongoing trial can be stopped for two reasons. One is that adequate evidence emerges of the desired efficacy of the regimen, and the other reason is that the efficacy is found too low to go any further. The first is stopping for efficacy, and the second is stopping for futility. Futility has special appeal for phase II trial because the primary aim of this phase is to provide proof of concept in the sense that the regimen has minimum efficacy. Futility will save it from going through the expensive phase III. Stopping for efficacy has application to both phase II and phase III but has special appeal to phase III as it can save substantial time and resources. For these reasons, stopping rules are relatively relaxed for phase II trials as they allow early stopping for futility if the concept is not found sufficiently rigorous. Phase III trials do not want to stop early unless the evidence of efficacy is firm. This is further explained in the following.



### *Stopping for Futility*

Statistical changes in an ongoing trial are done in a manner that the prefixed **level of significance** and statistical **power** of detecting a prespecified effect remain unaltered. The procedure to decide futility becomes intricate in view of these constraints, and assistance of a good software package is required. Among many procedures, the following, called *Simon method*, is relatively simple to understand [1].

As an illustration of a two-stage design, consider $\pi_0 = 0.20$ as the null, $\pi_a = 0.30$ as the alternative, power $(1 - \beta) = 0.90$, and $\alpha = 0.05$. Software can be asked to find a suitable $n_1$ for an interim look at the data in the first stage and propose a final $n$ in the second stage. The algorithm that minimizes the expected requirement of total $n$ gives $n_1 = 19$ and $n_2 = 35$ for a total of $n = 54$ in this case [2]. This is more than the size obtained for a single-stage trial for the same power.

However, there is a chance that the trial is stopped early in the case of two-stage design. The software will also tell that if the number of positive responses is ≤4 in the first stage out of $n_1 = 19$, it is futile to continue, as the efficacy of 0.30 is extremely unlikely to accomplish at completion when the number of responses is so few at the first stage.

The computer package may take a substantial time to arrive at these numbers, as the computations are enormous. You can see how complicated this procedure could become if there are two arms (case and control) and many stages. For such rules for some other values of $\pi_0$ and $\pi_a$, see Simon [1].

Single-stage design promises a large $n$ upfront, but this can expose such a large number of needy patients to an unproven regimen that may also have side effects. A two-stage **group sequential design** can partially be saved from these vagaries. Stopping rules for futility for two-stage design can also be worked out by following a similar procedure with the help of an appropriate software package, and the method can be extended to multistage sequential design and **multiarm trials**.

### *Stopping for Efficacy*

This requires that the level of significance α is judiciously apportioned at each appraisal such that the total Type I error does not exceed α. A simple though inadequate approach is to spend α in $K$ equal parts if $K$ appraisals are planned including the final analysis of all the data. This means that the hypothesis at the first appraisal is tested at $\alpha/K$ level of significance, the second at $2\alpha/K$ level of significance, and the last at α level of significance. More generally, in this case, level of significance at the $k$th appraisal is $\alpha(\tau_k) = \tau_k\alpha$, where $\tau_k$ is the proportion covered at the $k$th appraisal. If $\alpha = 0.05$ and $K = 2$ (only one midterm appraisal), use $\alpha = 0.025$ level of significance to test the null at the midterm stage. This implies for two-tailed Gaussian $z$-test that the critical value will be ±2.24 (from Gaussian table for $\alpha = 0.025$) in place of the conventional ±1.96. This, in fact, is more stringent than it apparently looks because, at this stage, the sample size available is also one-half of what would be eventually available. If the results meet this criterion, reject the null of equality of the test and the control group, conclude that you have enough evidence available in favor of the alternative hypothesis, and stop the trial.

Statistically, the equal-α spending procedure just stated is too liberal than what it should be to control Type I error, and rejects the null relatively easily. It sounds reasonable to have a procedure that is even more stringent at initial stages. For equally spaced group sequential designs, one such popular procedure is due to **O'Brien–Fleming**. An improvement over this is due to **Lan–deMets**, which is flexible and accommodates unequally spaced appraisals. This can be used for equally spaced sequential designs as well. This preserves the overall Type I error regardless of timing of the appraisals but makes it difficult to stop the trial early unless there is a strong evidence of the desired efficacy.

No matter which method is used, there would be a bias involved in early stopping. If you plan $n = 400$ subjects in a trial and stop it after analysis of data on 150 subjects since the data tell you that strict significance has been reached, the question arises whether these 150 are as representative of the population as 400 would have been. If the first 150 subjects are not random, further bias is apparent. For further details, see Chow et al. [3].

1. Simon R. Optimal two-stage designs for phase II clinical trials. *Cont Clin Trials* 1989;10:1–10. http://www.ncbi.nlm.nih.gov/pubmed/2702835
2. Groulx A, Moon k, Chung SC. Using SAS® to determine sample sizes for traditional 2-stage and adaptive 2-stage phase II cancer clinical trial designs. *SAS Global Forum* 2007. http://www.jazdlifesciences.com/pharmatech/research/Scian-Services-Inc.htm?contentSetId=111471&supplierId=30010143
3. Chow S-C, Wang H, Shao J. *Sample Size Calculations in Clinical Research,* Second Edition. Chapman & Hall/CRC Press, 2007.

## stratification, *see* stratified random sampling

## stratified analysis, *see* Mantel–Haenszel procedure

## stratified randomization, *see* block, cluster, and stratified randomization, and minimization

## stratified random sampling

Stratified random sampling is the random selection of the units after dividing the population into relevant groups. A big drawback of **simple random sampling** (SRS) is that it can fail to give adequate representation to one or more subgroups of interest. For example, in a study on the relationship of maternal complications with parity, it is necessary that women of different parities are included in the sample. The definition of the **sampling unit** in this case could be a currently pregnant woman reporting in a particular group of antenatal clinics. An SRS of size, say, 60 women in this case can yield a sample in which by chance parity 4 is not represented at all or inadequately represented. Note that the objective of the study is not met unless all parities have adequate representation. The procedure therefore should be to first divide the **sampling frame** by parity status such as 1, 2, 3, 4, 5, and 6+ and then draw an independent SRS of size, say, 10 from each division. Such a division of the frame is called **stratification** and each division is called a **stratum**.

Consider the data on waist–hip ratio (WHR) and triglyceride (TG) levels of 100 subjects in the total population in Table S.22. In this table, the strata could be WHR categories as given at the bottom of the table. One may decide how many units to be selected from different strata—they need not be equal. This procedure of choosing the sample is called stratified random sampling (StRS). The characteristic chosen for stratification is either the one suspected to affect the variable under study or the one that makes groups of interest for which separate results are required. After this, the sample would adequately represent the stratifying characteristic but not necessarily other factors that may be of consequence. For example, in a study on mortality in diabetes mellitus, the subjects may be stratified by the maximum level of plasma glucose, but the sample may still not be representative for obesity, age–gender, coexisting diseases, patient cooperation, etc. All these can affect the prognosis or

**TABLE S.22**
**Five-Year Survival Rate of Gastric Cancer Cases with Different Types of Operation**

| Type of Operation | Total Cases | Five-Year Survival | | |
|---|---|---|---|---|
| | | Rate (%) | Survived | Died |
| Distal subtotal gastrectomy | 678 | 37.2 | 252 | 426 |
| Total gastrectomy | 466 | 32.4 | 151 | 315 |
| Proximal subtotal gastrectomy | 354 | 20.1 | 71 | 283 |
| **Total** | **1498** | **31.6** | **474** | **1024** |

the outcome. Thus, care is always needed in extrapolation of results even after adopting stratified sampling.

In Table S.13 in **simple random sampling**, the subjects are divided into thin, normal, and obese according to WHR ≤ 0.89, 0.90 ≤ WHR ≤ 1.09, and WHR ≥ 1.10. These are the strata in this case. With this categorization, the SRS of 16 subjects marked with * in this table happens to contain 2 thin, 11 normal, and 3 obese subjects. However, the actual numbers of thin, normal, and obese in the population according to this categorization are 14, 38, and 48, respectively. The obese are clearly underrepresented in the sample. If you divide the population in the first instance and then take a sample, the result could be very different. A commonly adopted strategy in this case is to take a **proportionate sample** from each stratum. In this case, 16 out of 100 gives a sampling fraction of 0.16. Applying this to the strata gives a sample of 2 from the first stratum, 6 from the second stratum, and 8 from the third stratum. These, when randomly drawn by us, are subject numbers 28 and 62 from the first stratum; 5, 21, 33, 47, 76, and 95 from the second stratum; and 8, 54, 58, 63, 69, 73, 85, and 99 from the third stratum. You may get different subjects. The means are now calculated separately for each stratum. These means the TG levels of subjects in our sample are 122.50, 155.83, and 169.62, respectively. These are multiplied by the respective stratum size and divided by the population size, i.e.,

$$\text{mean in stratified sample: } \bar{x}_{st} = \sum N_k \bar{x}_k / N,$$

where $N_k$ is the size of the $k$th stratum, and $\bar{x}_k$ is the mean obtained for the $k$th stratum. In our sample,

$$\bar{x}_{st} = (122.50 \times 14 + 155.83 \times 38 + 169.62 \times 48) / (14 + 38 + 48) = 157.8.$$

The stratum size is now the "weight" for calculation of the mean. A similar formula is required to calculate the sample standard deviation. This sample mean for StRS is slightly greater than the population mean of TG in this example. It is natural for such a difference to arise in any sampling. Now the sample has adequate representation of each WHR category.

In the case of proportionate samples as in this example, the probability of selection of each subject is nearly the same in different strata. The advantage with this is that the sample becomes self-weighting. This can be explained as follows.

Proportionate sample implies that the sample from each stratum is in the same proportion as in the population. This means

$$\frac{N_k}{N} = \frac{n_k}{n} \text{ for all the } k \text{ (i.e., for each stratum).}$$

When this is substituted, we get

$$\text{mean in stratified sampling: } \bar{x}_{st} = \Sigma n_k \bar{x}_k / n.$$

This can be easily seen to be the same as the usual sample mean. Thus, in the case of proportionate samples, the mean for the stratified sample can be calculated just as a usual unweighted mean is calculated. This property is called *self-weighting* of the sample.

Nonetheless, proportionate samples are not a requirement for stratified sampling. If 5 subjects are selected from each stratum in this example, the probability of selection is 5/14, 5/38, and 5/48 for subjects in the first, second, and third strata, respectively. Such unequal probabilities do not render StRS invalid. Many medical investigations use equal samples from strata of different sizes. It is important to realize that the probabilities for different units can be unequal despite random samples, and they impose a restriction of the type illustrated by the formula of $\bar{x}_{st}$. However, in this case, the estimate of mean in different strata would have different precision, and it is difficult to justify mixing estimates of varying precision for obtaining the combined estimate.

## STROBE statement

Strengthening of reporting of observational studies in epidemiology (STROBE) is a guideline for fully reporting the results from an observational study. This contains a checklist of 22 items on the contents of the article on results of observational studies, and is endorsed by a large number of biomedical journals. Some points are separate for prospective, retrospective, and case–control studies, and others are common for all observational studies. For example, for cohort studies, one item is "summarize follow-up time" (e.g., average and total amount), which is not required for other studies.

The statement requires that scientific background and rationale for the study should be clearly stated and also requires a cautious overall interpretation of results considering the objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence. A discussion of the existing external evidence is particularly important for studies reporting a small increase in risk.

As of now, STROBE does not specifically require that new findings be put into context by conducting a systematic review of other similar studies. This obviously would be a positive feature as "this could alert us about consistency of findings, allow exploration of sources of heterogeneity and would ensure, in the same way as it does for randomized trials, that research effort is not spent on rediscovering the same finding again" [1].

For details and the latest on STROBE, see Ref. [1].

1. STROBE statement. http://www.strobe-statement.org/index.php?id=available-checklists.

## structural equation models

Structural equation models (SEMs) define a set of data analysis tools that allow for the testing of theoretically derived and a priori–specified causal (or just dependence) hypothesis [1]. Most dependence hypothesis exploring methods such as analysis of variance and regression can be regarded as special cases of SEM, but its immediate predecessors are **path analysis** and confirmatory **factor analysis**. SEM can be viewed as a combination of these two methods, where path analysis provides the structure of the relationship and the factor analysis provides the latent variables. **Latent variables** are those that cannot be directly observed such as mental capabilities, courage to face adverse situations, and beliefs. These are also called latent factors or common factors or constructs because they are mostly construed to be explained by a combination of observed variables. The observed variables are also termed as *exogenous variables* and the latent variables as *endogenous variables*. SEMs try to reduce a large set of observed variables to a small set of latent variables based on the covariance structure—thus incorporating the measurement errors by adjusting the correlations and the path coefficients accordingly. Path analysis by itself assumes that all measurements are perfect, while that is not so with SEMs. The development of structural models follows the seminal work of Joreskog [2] in 1973, who presented a general method of estimating a linear structural equation system, and the work is still in progress.

As in all **model** fitting, the process of building up SEM involves stages such as initial model conceptualization, specification of parameters and estimation of these parameters, model fitting and assessment of its adequacy, and model modification as needed after this assessment. Of these, the most challenging is model conceptualization, which requires brainstorming and meticulous thinking so that the specified model represents the theories of the underlying biological process fairly well. Lack of consonance between the two can render the entire exercise futile. The conception of this model can come from the path analysis, which helps to visualize the structural part, and the factor analysis, which is based on the measurement part. The next stage of parameter specification requires that the number of parameters, $K$, must be substantially less than the number of unique covariances of the observed variables. Mostly the method of **maximum likelihood** is used for estimation of the parameters, although distribution-free methods such as generalized least squares also exist. The preceding two steps would provide the model. Whether this is an adequate model or not is the question answered in the third stage. This is generally done by chi-square test that compares the observed values with the values predicted by the model. An improvement over chi-square is provided possibly by **Akaike information criterion** that penalizes for a large number of parameters. If the model so arrives is good to your satisfaction, the work is done, and if not, the exercise may give some leads of what changes are needed.

Since the method itself is in the making, the reporting of the results of SEM too is not standardized yet. McDonald and Ho [3] have presented a detailed account of what should be reported and how. See Lee et al. [4] for SEM in action. They have used the SEM approach for studying the impact of noise on self-rated job satisfaction and health in open-plan offices in China and Korea.

For more details of SEMs, see Loehlin [5].


1. Mueller RO, Hancock GR. Best practices in structural equation modeling, in: *Best Advanced Practices in Quantitative Methods* (ed. Osborne JW). Sage, 2007: pp. 488–508.
2. Joreskog KG. A general method for estimating a linear structural equation system, in: *Structural Equation Models in the Social Sciences* (Goldberger AS, Duncan OD). Seminar Press, 1973: pp. 85–112.
3. McDonald RP, Ho MR. Principles and practice in reporting structural equation analysis. *Psychol Methods* 2002;7(1):64–82. http://www.nyu.edu/classes/shrout/G89-2247/McDonaldMoon-Ho2002.pdf, last accessed September 16, 2015.
4. Lee PJ, Lee BK, Jeon JY, Zhang M, Kang J. Impact of noise on self-rated job satisfaction and health in open-plan offices: A structural equation modelling approach. *Ergonomics* 2015 Sep 14:1–13. http://www.ncbi.nlm.nih.gov/pubmed/26366940
5. Loehlin JC. *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis*, Fourth Edition. Psychology Press, 2003.




## Studentized range, *see* Tukey test for multiple comparisons

## Student *t*-distribution

William Gosset, who wrote under the pseudonym of Student, first proposed *t*-test while optimizing yeast composition in 1908 [1]. This is defined as

$$\text{Student } t\text{: } t_{\upsilon} = \frac{z}{\sqrt{\chi^2/\upsilon}}, \tag{S.3}$$

where $z$ is a standard Gaussian variable, i.e., $z \sim N(0, 1)$, and $\chi^2$ has chi-square with $\upsilon$ degrees of freedom (df's), which is independent of $z$. The statistic $t$ also has the same df's as the chi-square in the denominator. Student $t$ has different distribution for different df's just as people of different age has different distribution of systolic blood pressure. Review the topic on the concept of **degrees of freedom** if that is not clear. The shape of the distribution of Student $t$ is symmetric, which gives a feeling that it is Gaussian but has relatively a larger variance (Figure S.20). For large df's, the distribution is approximately Gaussian, and for infinite df's, it is exactly Gaussian.

William Gosset

Student $t$ has wide applications in statistical testing of hypothesis problems. For example, this is used for testing the statistical significance of difference in means (paired and unpaired), one-sample comparison with prespecified mean, and regression coefficient. Test for **regression coefficient** is separately presented, and others are discussed under the topic **Student *t*-test**. All these can be shown to have the same basic form as given in Equation S.3 under the null hypothesis when **Gaussian conditions** hold. For example,

$$\text{Student } t\text{-test (one sample): } t_{n-1} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

actually is

$$t_{n-1} = \frac{(\bar{x} - \mu_0)/\left(\dfrac{\sigma}{\sqrt{n}}\right)}{\sqrt{\dfrac{\Sigma(x-\bar{x})^2/\sigma^2}{n-1}}},$$

where $\sigma$ cancels out from the numerator and the denominator. If you carefully notice, under Gaussian conditions, the numerator $(\bar{x} - \mu_0)/\left(\dfrac{\sigma}{\sqrt{n}}\right)$ is a standardized Gaussian deviate $z$ =

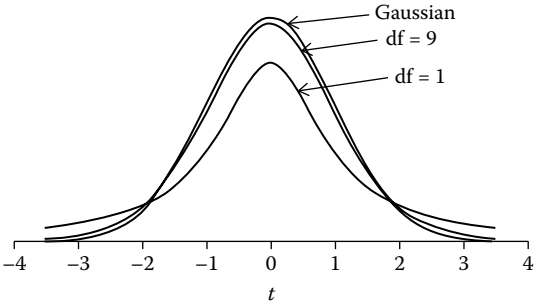



**FIGURE S.20** Student $t$ distribution has larger variance but approaches Gaussian as df's increase.

(estimate – mean)/(SE of the estimate) under the null hypothesis that the population mean is $\mu_0$, and the denominator $\sqrt{\frac{\Sigma(x-\bar{x})^2/o^2}{n-1}}$ is $\sqrt{\chi^2/\upsilon}$ with $\upsilon = n - 1$ since $\Sigma(x-\bar{x})^2/o^2$ is chi-square, being the sum of squares of the standard Gaussian deviates, provided the observed values are independent. Also it is theoretically established that the denominator is independent of the numerator. All Student *t*-tests can be similarly explained.

1. Student. The probable error of mean. *Biometrika* 1908 Mar:g(1):1–25. http://www.york.ac.uk/depts/maths/histstat/student.pdf

## Student *t*-tests

Two common problems handled by Student *t* are discussed in this section. First is comparing the mean in a sample of subjects with a prespecified value, and the second is comparing the mean in one sample with that in another sample. The two samples could be paired or may represent two different groups. **Gaussian conditions** are required for Student *t* to be valid. Among others, it is also used for **comparison of two regression coefficients** as separately explained.

### *Comparison with a Prespecified Mean*

Let the interest be in finding whether patients with chronic diarrhea have the same average hemoglobin (Hb) level as normally seen in healthy subjects in the area. Suppose the normal level of Hb is 14.6 g/dL. This is assumed to be known and fixed for the present example. Since chronic diarrhea can only decrease the Hb level, and not increase it, it is a one-tailed situation. In an unlikely event of sample mean being >14.6 g/dL, the evidence is immediate that Hb level does not decrease in chronic diarrhea patients, and there is no need to proceed further. A higher mean can occur by chance in the sample.

Suppose further that a random sample of 10 patients with chronic diarrhea is investigated, and the average Hb level is found to be 13.8 g/dL. Thus, the sample mean is lower than normal, and this could occur if the sample happens to comprise subjects with a lower level. Such subjects are not uncommon in the healthy population as well. If another sample of 10 patients is studied, the average could well be 14.8 g/dL. Can it be concluded with reasonable confidence based on the previous sample that patients with chronic diarrhea indeed have a lower Hb level on average?

There is only one sample in this example, and the comparison is with the known average in the healthy subjects. It is a one-sample problem, although the comparison is of two means—one observed in the sample and the other known for the healthy population.

The null hypothesis in the preceding example is $H_0$: $\mu = 14.6$ g/dL. Since the possibility of a higher average Hb level in patients with chronic diarrhea is ruled out, the alternative hypothesis is one-sided. That is, $H_1$: $\mu < 14.6$ g/dL. If $H_0$ is rejected, then $H_1$ is considered true.

The first step is to choose an appropriate criterion to test this hypothesis. The value of this criterion is then calculated assuming that $H_0$ is true. Then the probability of the observed or a more extreme value is obtained. This is the ***P*-value**. If this probability is very small, $H_0$ is considered not plausible and rejected. The conclusion reached is that $H_1$ is true. If the *P*-value is not sufficiently small, say not less than 0.05, the null hypothesis is conceded with the rider that this does not imply that $H_0$ is accepted. It is just that it cannot be rejected on the basis of that sample because sampling fluctuation is not adequately ruled out as a likely explanation.

Heuristically, the answer depends on the magnitude of the difference between the sample mean and the known mean of the healthy subjects. In the preceding example, this difference is 13.8 – 14.6 = –0.8 g/dL. This magnitude is assessed relative to the expected variation in means from sample to sample. This variation is measured by the standard error (SE) of the mean, $\sigma/\sqrt{n}$. In a rare case when $\sigma$ is known, the criterion is

$$\text{Gaussian test: } z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

This follows Gaussian distribution provided the underlying distribution of *x*'s is also Gaussian. The value of this criterion is compared with its value in the Gaussian table to find if the *P*-value is sufficiently small for rejecting the null.

In practice, the SD $\sigma$ would be rarely known and is replaced by its estimate *s*. Thus, the criterion for this setup is

$$\text{Student } t\text{-test (one sample): } t_{n-1} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

where $\mu 0$ is the value of the mean under $H_0$. This criterion is valid under mild conditions such as the observations are independent. In the context of this example, this means that the Hb level in one subject should not have any influence on the level in other subjects. This is clearly satisfied in this example, but an example is given later where independence is violated. The other condition is that the distribution of *x*'s themselves is Gaussian. This condition can be relaxed if the sample size is large. The higher the value of *t*, the greater the chance that the sample has not come from a population with mean $\mu 0$. In that case, the decision to reject $H_0$ would have less chance of error. When this probability of Type I error is low, say less than 0.05, $H_0$ can be safely rejected. For *t* in the formula given in the just mentioned equation, df = $(n - 1)$. This is specified as a subscript of *t*.

Consider again the example of the Hb level in chronic diarrhea patients. Suppose that in a random sample of size $n = 10$, the levels in g/dL are as follows:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 11.5 | 12.2 | 14.9 | 14.0 | 15.4 | 13.8 | 15.0 | 11.2 | 16.1 | 13.9 |

These give mean $\bar{x} = 13.8$ g/dL and SD $s = 1.67$ g/dL. The hypothesis under test is that the average Hb level in the patients with chronic diarrhea is the same 14.6 g/dL that is normal in healthy subjects. Thus, $H_0$: $\mu = 14.6$ g/dL. The alternative as already explained is $H_1$: $\mu < 14.6$ g/dL. In this case, under $H_0$,

$$t_9 = \frac{13.8 - 14.6}{167/\sqrt{10}} = -1.51.$$

A statistical package gives $P(t < -1.51) = 0.0827$. This is the probability of getting the sample mean this much or more extreme in favor of $H_1$. Since this $H_1$ is one-sided, the probability required is also one-tailed. The distribution of *t* is also symmetric just as is the Gaussian distribution. Thus, $P(t < -a) = P(t > a)$. In this case, *P* is greater than 0.05 because 1.51 is less than the critical value 1.833 of Student *t* at 9 df. Thus, the chance is more than 5% that $H_0$ is true. Therefore, this $H_0$ cannot be rejected. The difference between the sample mean 13.8 g/dL and the population-normal 14.6 g/dL is not statistically significant. This sample does not provide sufficient

evidence to conclude that the mean Hb level in chronic diarrhea patients is less than normal.

The conclusion in this example is partly the result of the high variability in the Hb level in the sample patients: whereas it was only 11.2 g/dL in one patient, it was 16.1 g/dL in another. Widely scattered values gave a high value of sample SD *s* and led to the expectation of high intersample variability. Consequently, it became difficult to say anything definite about the lower mean Hb in the patients.

### *Difference in Means in Two Samples*

Now let us consider a situation where two samples are available. These could be from two groups such as males and females, patients suffering from disease A and disease B, or patients of age 20–39 years and age 40+ years, or could be from one group only, which is measured before and after a treatment. The latter is called a **paired samples** setup. The general form of the criterion in a two-sample setup is

$$\text{Student } t \text{ (two-sample)} = \frac{\text{sample mean difference} - \text{population mean difference under } H_0}{\text{estimated SE of difference in sample means}}.$$

This takes a different form in paired setup compared with that in independent samples setup.

*Paired samples setup*: The procedure used to calculate the SE of the difference in case of paired samples is different from that of unpaired samples. Observations are said to be paired when they are obtained twice for the same subject, such as blood pressure (BP) before and after treatment or erythrocyte sedimentation rate (ESR) measured by two methods in the same group of subjects. They are also considered paired when the subjects in the two groups are one-to-one matched such as in some case–control studies. This is an example where the observed values are not independent. In the case of paired samples, obtain the difference between the pairs as $d_i = (x_{2i} - x_{1i})$, $i = 1, 2, \ldots, n$, where $n$ is the number of pairs. These differences are independent of one another. Calculate the SD of these $d$'s as usual by

$$s_d = \sqrt{\frac{\sum (d - \bar{d})^2}{n-1}}.$$

The null hypothesis of interest for rejection in this case is $H_0$: $\mu_1 = \mu_2$ (or $\mu_1 - \mu_2 = 0$), where $\mu_1$ is the mean of one population and $\mu_2$ is the mean of the other population. For example, these could be mean diastolic BP before and after treatment, mean ESR obtained by two methods, or mean of any other such population. Under this $H_0$, the criterion, from the formula given earlier for one sample setup, is

$$\text{Student } t_{n-1} \text{ (paired samples)} = \frac{\bar{d}}{s_d/\sqrt{n}}.$$

Basically, this is the same as the one-sample formula for $H_0$: $\mu_d = \mu_1 - \mu_2 = 0$. After the differences are obtained, the paired sample problem reduces to a one-sample problem for these differences with the null hypothesis that the mean difference is zero. When Student *t* is significant, the null is rejected. Since the point estimate of the difference $(\mu_1 - \mu_2)$ is $(\bar{x}_1 - \bar{x}_2)$, this is the estimate of the treatment effect in the case of measurements before and after treatment.

Given below are the serum albumin levels (g/dL) of six randomly chosen patients with dengue hemorrhagic fever before and after treatment. The null hypothesis is that the mean after is the same as the mean before, i.e., the treatment of dengue fever (this treatment is mostly symptomatic) has not altered the albumin level.

| | | | | | | |
|---|---|---|---|---|---|---|
| Before treatment | 4.8 | 4.1 | 5.3 | 3.9 | 4.5 | 3.8 |
| After treatment | 5.2 | 4.9 | 5.2 | 4.8 | 4.6 | 4.4 |
| In this case, difference, $d_i$: | 0.4 | 0.8 | –0.1 | 0.9 | 0.1 | 0.6 |

Mean differences, $\bar{d} = 0.45$, and SD of differences, $s_d = 0.3937$.

Thus, under $H_0$: $\mu_1 = \mu_2$,

$$t_5 = \frac{0.45}{0.3937/\sqrt{6}} = 2.80.$$

Since there is no assertion in this case that the albumin level after the treatment will increase or decrease, the alternative hypothesis is $H_1$: $\mu_1 \neq \mu_2$. For this $H_1$, a two-tailed probability $P(|t| > 2.80)$ is needed. For $(n - 1) = 5$ df, this is $P = 0.038$ from statistical software. Because the probability of a Type I error is sufficiently small ($P < 0.05$), reject $H_0$ and conclude that the mean albumin level after the treatment is different from the mean before the treatment.

*Unpaired (independent) samples setup*: Paired observations are highly desirable in the situation of our example, but it is possible that the albumin investigations could not be done before the treatment started. Suppose that a separate group of six similar but randomly drawn patients was investigated for albumin level before the treatment. Thus, there are a total of 12 patients—6 in each group. The two groups are now independent. In such cases, the first step is to check that the variances in the two groups are not widely different since this is one of the requirements for validity of Student *t*. Generally, a ratio $s_1^2/s_2^2 < 3$ is considered adequate if each $n$ is around 10 or 15. If $n$ is smaller, even $s_1^2/s_2^2 < 4$ can also be tolerated. If $n$ were 30 or greater, a ratio of 3 may be too high. The conventional statistical test for $H_0$: $\sigma_1^2 = \sigma_2^2$ is $F = s_1^2/s_2^2$ with higher SD in the numerator. However, it is necessary for this test that the underlying distribution is Gaussian, at least approximately. One should prefer the **Levene test** for equality of variances because it is valid even in the case of departure from the Gaussian pattern. The Student *t*-test is applicable even when the underlying distribution is very different from Gaussian provided that $n$ is sufficiently large. The Levene test is also applicable in this situation. Now there are two possibilities.

(i) *Population variances are equal*: In this case, the samples can be combined to obtain a more reliable estimate of the variance. This is given by

$$\text{pooled variance: } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)},$$

where $s_1$ is the SD in the first sample and $s_2$ in the second sample, and $n_1$ is the size of the first sample and $n_2$ of the second sample. Now, calculate

$$\text{SE (mean difference)} = s_p\sqrt{1/n_1 + 1/n_2}.$$

S

From the formula given earlier, the criterion for testing $H_0$: $\mu_1 = \mu_2$ is

Student $t$ (two independent samples): $t_{n_1+n_2-2} = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2(1/n_1 + 1/n_2)}}$,

where $\bar{x}_1$ and $\bar{x}_2$ are the respective sample means. The degrees of freedom for $t$ in this formula are $(n_1 + n_2 - 2)$.

(ii) *Population variances are unequal*: You would be rarely interested in equality of variances per se. Statistically, it is kind of a **nuisance parameter** that we have to deal with anyway. Realize though that when means are very different, there is a good likelihood that variances would also differ. A group with higher mean may have higher variance too. In that sense, equality of variances becomes a major issue. If the population variances are known to be unequal or if the sample variances are very different from one another, then the samples cannot be pooled and the formula for pooled variance is not valid. In this case, calculate the separate variance Student $t$ (two independent samples) as

$$t_v = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

where the degrees of freedom

$$v = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\dfrac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \dfrac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}.$$

The case of unequal variances in two samples is popularly known as the **Welch test**. Note that this reduces to the usual Student $t$ in case of equal $n$ and equal variances.

For an example, consider the same data as in the previous example, but now the observations are for 12 different patients in place of pairs of observations on 6 patients. The patients measured before treatment are not the same as those measured after the treatment. In this case,

mean albumin level in before treatment group, $\bar{x}_1 = 4.40$ g/dL
SD in this group, $s_1 = 0.5797$ g/dL
mean albumin level in after treatment group, $\bar{x}_2 = 4.85$ g/dL
SD in this group, $s_2 = 0.3209$ g/dL

In this example, $n_1 = 6$, $n_2 = 6$ so that df $= n_1 + n_2 - 2 = 10$.

For such small samples, the SDs do not differ too much and we can pool them. Thus,

$$s_p^2 = \frac{5 \times 0.5797^2 + 5 \times 0.3209^2}{10} = 0.2195.$$

Therefore,

$$t_{10} = \frac{4.40 - 4.85}{\sqrt{0.2195(1/6 + 1/6)}} = -1.66.$$

The alternative hypothesis continues to be two-sided. The probability of getting these sample values or more extreme in favor of $H_1$ when $H_0$ is true is $P(|t| > 1.66)$. From a statistical package, this $P$-value is 0.1272, which is large relative to the usual $\alpha = 0.05$. Thus, the null hypothesis of equality of means cannot be rejected. The evidence is not strong enough to conclude that the mean albumin level after the treatment is any different from the mean before the treatment.

### *Some Features of Student* t

Our examples very aptly illustrate the following features of Student $t$:

1. The $t$-test is based on the magnitude of the difference and its variance but ignores the fact that five of the six patients in the paired setup in our example showed some rise. If the interest is in the *proportion* of subjects showing a rise and not in the magnitude of the rise, then use the methods for proportions.
2. The same difference is statistically significant in the paired setup in our example but not significant in the unpaired setup. This occurred because the difference in the paired setup is fairly consistent, ranging from –0.1 to +0.9 g/dL. Each patient served as his/her own control. In the unpaired setup, the interindividual variation is large. A paired setup may be a good strategy in many situations. Advice of matching of controls with cases is given since such matching simulates pairing and reduces the effect of at least one major source of uncertainty.
3. The size of the sample is 6 in the paired setup and a total of 12 in the unpaired setup. Both are small. Recall that Student $t$ is valid only when means follow a Gaussian pattern. When $n$ is large, this pattern is nearly always Gaussian due to the **central limit theorem**, whether or not the underlying distribution of the individual measurements is Gaussian. Where $n$ is small (say, less than 30), the $t$-test is valid only if the underlying distribution is Gaussian. This is the assumption made in these examples. If the underlying distribution is far from Gaussian and $n$ is small, other methods such as nonparametric (e.g., **Wilcoxon**) tests are used.
4. One should always examine whether the data meet the requirement of $\sigma_1^2 = \sigma_2^2$. The ratio $s_1^2/s_2^2$ in this example is $0.5797^2/0.3209^2 = 3.26$. For a sample of size 6 each, this may be within the tolerance limit. Many statistical software packages would automatically test this. In case of violation, the $P$-value is obtained by using Welch $t$ in place of the pooled variance $t$. Some statistical software would also do this automatically or would provide $P$-values based on both types of $t$. However, caution is required when using a separate variance estimate. When the variances are different, it is clear that the two populations are different. Then equality of means may not be of much consequence as the distribution pattern is different anyway. If it is known, for example, that body mass index is much more variable in women than in men, equality of their means would rarely help. Thus, use of a separate variance estimate in a two-sample $t$ is rare.
5. In some cases, e.g., in estimation of antibody titers, the conventional (arithmetic) mean is not applicable and the **geometric mean** (GM) is used to measure the central value. The logarithm of GM has the same features as the

usual arithmetic mean. Thus, the Student *t*-test and other means-based tests can be carried out on geometric means after taking the logarithm of the values. However, in this case, the conclusions will be applicable to log values and not to the values themselves. You should carefully examine whether or not the results obtained on log values can be extended to the original values. In many cases, such extension does not cause any problem.

6. All statistical tests give valid conclusions for the groups and not for individuals. The *t*-test is for the *average* values in the groups. Individuals can behave in an unpredictable way even when the difference in means is statistically significant. Thus, use caution in applying results to individual subjects. Clinical features of a person may completely override the means-based conclusions.
7. Statistical inference is not valid for nonrandom samples. In our examples, we may not have mentioned this explicitly, but this is implied.
8. The fundamental requirement for the Student *t*-test is that the sample values are independent of each other. In our unpaired example, the albumin level in one patient is not going to affect the level in any other patient, and thus the values are independent. In the paired example, there is no such independence because of paired setup, but once the difference is obtained, these differences would be independent. Paired *t*-test is based on the differences.

   Consider the blood pressure measured for two or more subjects belonging to the same family. **Familial aggregation** for medical measurements is well known, and thus values belonging to the members of the same family are not independent. The Student *t*-test cannot be used for such values unless family effect is first removed. Ingenious methods may be needed for removing this effect. Most practical situations do not have this constraint and Student *t* can be safely used.

### *Effect of Unequal* n

The Student *t*-test for two independent samples does not have any restriction on $n_1$ and $n_2$—they can be equal or unequal. However, equal *n*'s are preferred because of two reasons: (i) When a total of $2n$ subjects are available, their equal division among the groups maximizes the statistical **power** to detect a specified difference; and (ii) two-sample *t* is not robust to $\sigma_1^2 \neq \sigma_2^2$ unless $n_1 = n_2$. If a smaller sample has larger variance, the problem is aggravated. To reiterate what was stated earlier, if you have prior knowledge or find from the data that $\sigma_1^2 \neq \sigma_2^2$, further testing for equality of means may not be relevant since the distributions are different anyway owing to different SDs. If the interest persists, Welch test should be used in place of Student *t*.

Although equal *n*'s are desirable, in many medical situations, this may not be a prudent allocation. In clinical trials, many times controls are easy to investigate and more than one control per case could be a good strategy. Thus, it is not necessary to have equal *n*, although that is highly desirable.

## study designs, *see* designs of medical studies (overview)

## subjective probability, *see* probability

## sum of squares (types of)

Sum of squares (SS) actually is $\Sigma_i x_i^2$, where $x_1, x_2, \ldots, x_n$ are the observed values for *n* subjects in the sample, but the convention in statistics is to adjust all the values by the mean. Thus,

$$\text{sum of squares} = \Sigma_i (x_i - \bar{x})^2,$$

where $\bar{x}$ is the sample mean. This is the numerator of variance, and many times it is used as its surrogate. For example, in **analysis of variance** (ANOVA), it is the sum of squares that is broken up into its components (and later divided by the degrees of freedom) but is called components of variance. The setup in **one-way ANOVA** is that there are *J* groups and $x_{ij}$ is the *i*th value ($i = 1, 2, \ldots, n_j$) in the *j*th group ($j = 1, 2, \ldots, J$), which actually implies that there are $n_1$ observations from the first group, $n_2$ from the second group, etc.; the mean of the *j*th group is denoted by $\bar{x}_{\bullet j}$ and the overall mean is denoted by $\bar{x}$. With these notations, the basic equation in one-way ANOVA is the following:

$$\begin{aligned}\Sigma_{ij}\left(x_{ij} - \bar{x}_{\bullet\bullet}\right)^2 &= \Sigma_{ij}\left[(x_{ij} - \bar{x}_{\bullet j}) - (\bar{x}_{\bullet j} - \bar{x}_{\bullet\bullet})\right]^2 \\ &= \Sigma_{ij}\left[(x_{ij} - \bar{x}_{\bullet j})^2\right] - 2\Sigma_{ij}\left[(x_{ij} - \bar{x}_{\bullet j})(\bar{x}_{\bullet j} - \bar{x}_{\bullet\bullet})\right] \\ &\quad + \Sigma_{ij}\left[(\bar{x}_{\bullet j} - \bar{x}_{\bullet\bullet})^2\right] \\ &= \Sigma_{ij}\left[(x_{ij} - \bar{x}_{\bullet j})^2\right] + \Sigma_{ij}\left[(\bar{x}_{\bullet j} - \bar{x}_{\bullet\bullet})^2\right]\end{aligned}$$

(the middle term can be shown to be zero for independent values),

where it is evident that the first term on the right side is based on the deviation of values from the mean of their respective group (thus within-groups SS), and the second term is based on the deviation of group means from the overall mean (thus between-groups SS). This gives

$$\text{total SS} = \text{within-groups SS} + \text{between-groups SS}.$$

These SSs are divided by the respective df's to get mean squares. Within-groups SS is also called **error sum of squares** and when divided by the df becomes **mean squares due to error** (MSE). The term on the left side is deviation of values from the overall mean (thus is the total SS).

It can be similarly shown for a **two-way ANOVA** that total SS = Factor 1 SS + Factor 2 SS + Interaction SS + Error SS. Similar decomposition of the total SS can be obtained for multiway ANOVA. Factorwise SS measures how much each factor is contributing to the total variation—thus helps in assessing whether the factors have significant effect or not when compared with error SS. Factor SS should be substantially higher than the error SS for the effect of the factor to be statistically significant. This is done by the ***F*-test**. However, the "devil is in the details" as explained next.

### *Type I, Type II, and Type III Sums of Squares*

When the number of factors is two or more, various SSs can be calculated in at least three different ways. Each has relevance in a particular setup. We are explaining these for two-way ANOVA, but they are similar for higher-way ANOVA.

*Type I SS*: For a two-way ANOVA, Type I SS measures the effect $\alpha$ of factor 1 after adjusting for $\mu$ (overall mean), measures the effect $\beta$ of factor 2 after adjusting for $\alpha$ and $\mu$, and measures the effect of the interaction $\theta$ after adjusting for $\beta$, $\alpha$, and $\mu$. Thus, this assesses the incremental effect of parameters in the order they appear in the

model or in the order as specified by the user while using the software package. If the effect of sex and body mass index (BMI) is being studied on creatinine level, Type I SS would require that you decide which is to be entered first. If sex is the first, Type I SS would assess the significance of sex when nothing is in the model (or just overall mean), would assess the significance of BMI when sex is in the model (called sex-adjusted effect), and would assess the interaction when both sex and BMI are in the model.

*Type II SS*: This is used for assessing the effect of those parameters that do not include it. Since interaction includes both $\alpha$ and $\beta$, Type II SS measures the effect $\alpha$ after adjusting for $\beta$ and $\mu$ but not $\theta$, and the effect $\beta$ after adjusting for $\alpha$ and $\mu$ but not $\theta$. This is rarely used.

*Type III SS*: This assesses the effect of parameters after adjusting for all other terms in the model. Thus, this measures the effect $\alpha$ after adjusting for $\theta$, $\beta$, and $\mu$, and the effect $\beta$ after adjusting for $\theta$, $\alpha$, and $\mu$. This is the same as Type II SS if there is no interaction in the model. This type of SS is used to assess the significance of the effect of each factor when all other parameters are present in the model. In our example, this means the effect of sex when BMI and interaction are in the model, the effect of BMI when sex and interaction are in the model, and the effect of interaction when sex and BMI are in the model.

*Type IV SS:* This is used for missing cells such as in partially factorial designs. It is similar to Type III SS but has different strategy because of missing cells. If there are no missing cells, Type IV SS is the same as Type III SS.

In most situations, Type III SS is computed and it serves the purpose well. This also is the default in most statistical software packages.

## superiority and noninferiority trials, *see* equivalence and noninferiority trials

## surface and internal attributes, *see* factor analysis

## surface chart, *see* response surface

## surrogate measurement/variables

Surrogate measurement is a proxy for something that is either impractical to measure or is impossible but still gives a fairly good idea about the state of the difficult-to-measure characteristic. For example, most people want to keep their income confidential, but an idea of income can be obtained by the type of housing, cars, amenities in the household, etc. Health is almost impossible to measure because of its high personal overtones but can be guessed by finding the number of sick days, severity of disease if any, feeling of comfort, etc. A perfect measurement for many medical conditions is not available, and a surrogate is used, perhaps as much as the surrogate has positioned itself as the actual measurement of interest. Severity of disease is a common example for which no widely acceptable scale is available, and scales such as APACHE and Glasgow coma scale are used as surrogates. In a clinical setup, sometimes color of lips and fingernail beds are used as surrogates for the outcome of heart surgery. Many times cholesterol level is used as a surrogate endpoint, whereas actual interest may be in incidence of cardiac ailments; waiting for these ailments to develop may take an inordinately long time. In all such cases, it is important that the characteristic being used as a surrogate is strongly linked to the actual condition of interest.

The problem arises when conclusions are drawn on the actual variable of interest forgetting that only surrogate has been studied. For example, significant reduction in mean cholesterol level by a regimen should not be construed to mean that cardiac ailments will definitely reduce or the cardiac mortality will reduce. There might be several other intervening factors that determine the final outcome. In addition, it stands to reason that the conclusions based on surrogates will have a larger margin of error than what it apparently looks.

Bárcena et al. [1] found CD94hi/HLADR+ phenotypic profile a useful marker for natural killer cell clonality that otherwise lacks a universal and specific marker, and Moriceau et al. [2] examined whether time-to-treatment can be a good surrogate for cancer quality of treatment in France, and came to a negative conclusion.

1. Bárcena P, Jara-Acevedo M, Tabernero MD, López A, Sánchez ML, García-Montero AC, Muñoz-García N et al. Phenotypic profile of expanded NK cells in chronic lymphoproliferative disorders: A surrogate marker for NK-cell clonality. *Oncotarget* 2015 Nov 6. http://www.impactjournals.com/oncotarget/index.php?journal=oncotarget&page=article&op=view&path[]=5480&pubmed-linkout=1
2. Moriceau G, Bourmaud A, Tinquaut F, Oriol M, Jacquin JP, Fournel P, Magné N, Chauvin F. Social inequalities and cancer: Can the European deprivation index predict patients' difficulties in health care access? A pilot study. *Oncotarget* 2015 Nov 2. http://www.impactjournals.com/oncotarget/index.php?journal=oncotarget&page=article&op=view&path%5B%5D=6274&path%5B%5D=16302

## surveys, *see* sample surveys

## survival analysis, *see also* survival curve/function

Survival is complementary to mortality, but the statistical use of the term is for the duration of survival. The duration of survival of a patient with cancer after detection of malignancy and median survival time in cases of leukemia are common examples. Sometimes the interest is in the percent surviving for a specific duration such as 5-year survival rate of end-disease renal patients on dialysis.

In a generic sense, the method of survival analysis is used to analyze any data on duration. This can also be understood as time-to-event data. It could be duration of stay in the hospital, birth interval, duration of immunity, or any such duration. This could be any failure such as occurrence of metastasis, toxicity, or relapse, or any success such as recovery, discharge from the hospital, or disappearance of a complaint. Although the term *survival* is used generically to measure time to any event under investigation, other terms such as *waiting time*, *failure time*, and *transition time* are also used. A follow-up is an integral part of survival studies.

Why can durations not be analyzed as any other quantitative variable? One difficulty is that most durations do not follow a Gaussian pattern, and the distribution is highly skewed. This is not a big problem since this can be overcome by choosing a sample of sufficiently large size. The real problem arises when the duration for some subjects is not fully observed because the person moves out, dies due to an unrelated cause, refuses to cooperate after some follow-up, etc. The primary cause of such incomplete observation, however, is that the end-point event does not occur during the follow-up period, and the study is terminated after a fixed time. All these result in incomplete segments, called **censored observations**. The method

of survival analysis is geared to meet this contingency, which can rarely be handled in any other manner.

Consider an example of the follow-up of breast cancer cases. If the follow-up is for a period of 6 years, there might be cases that are still alive at the end of the follow-up. Their duration of survival will not be known except that the survival is for *at least* 6 years. Also, in such a follow-up study, there is always a likelihood that some patients withdraw or are lost to follow-up. If a patient was alive at the 24-month visit and lost thereafter, the survival duration again is not known. The statistical technique used to analyze such data is called the survival analysis.

Knowledge of survival pattern helps patients and physicians to decide which treatment or health strategy to prefer and when. For example, short-term survival may be better (in terms of percentage) with one regimen and long-term survival with another. The survival patterns are lessons to health care providers and seekers about what to expect in specific cases. The actual experience in individual cases would be different but not too much if the survival curves are valid and reliable.

Besides modeling the survival pattern over a period of time, the other objectives of survival analysis are (i) to investigate factors that influence the duration of survival, (ii) to compare two or more modalities of treatment for survival pattern, and (iii) to estimate the future survival of individuals or groups with specified features.

It goes without saying that for any duration, there must be a beginning (entry) point and an end (exit) point. In the case of diabetes, for example, the beginning point can be the time when the first symptom was noted, when the disease was first diagnosed, when blood glucose level showed a significant rise from the preexisting level, when a treatment was started, or any other point of time considered appropriate. The end point can be when blood glucose was first noted back to the normal level, when the treatment was discontinued, when the person is able to eat normally, or any other. Because of multiple choices, you can see that both the beginning point and the end point should be sharply defined for the duration to be observed without error. It is also important to decide whether the intervening period between the entry and exit has to be completely free of events and whether the occurrence of other events is to be disregarded.

For convenience of presentation, we are using the terms *survival* and *mortality* in this section, but they really mean change from one state to the other. Survival time is any time-to-event duration.

## Survival Data

Although the duration of survival can be studied in terms of probability of outcome within a specified duration, such as chance of recurrence of sputum positivity in tuberculosis within, say, a 6-month period after the treatment was discontinued in the middle of the course, this is a by-product of the survival analysis that models the entire pattern of survival from the beginning to the end point. When a single summary measure is required, the choice goes to median survival time instead of the mean because of the highly skewed distribution of survival time in most situations. Some other features of survival are described in this section.

In the case of survival studies, not all the subjects of a cohort necessarily begin at the same point of time. The subjects can join and leave at any point. The joining time is considered time zero, and all durations are measured from this point. Other features of survival time can be described as follows.

Survival time data can be obtained in a variety of ways. The method of data collection can have important implications on the interpretation and sometimes on the method of analysis because different collection methods give rise to different kinds of censoring.

The most convenient method is to take a cross-sectional random sample of subjects who have the condition of interest and inquired when the condition started. This can be done when the starting point is known to the subjects, and they are capable of reporting it correctly. Thus, this method can be adopted when the beginning point is well defined, for example, by appearance of complaints. It can be defined by laboratory assessment if the concerned laboratory report is available with each subject. Any other record of interest may also help to identify the time of beginning of the condition. If there is no follow-up, the end point will remain unknown in this method. One alternative in this situation is to consider the duration for all subjects as censored at the time of inquiry. This censoring for *all* subjects is not admissible. In fact, incomplete segments for survival analysis should not be far too many, certainly not more than the complete segments. If there are no or very few subjects with complete segments, the pattern would not be known, and no worthy analysis can be performed.

The only way to assess the end point in case of a cross-sectional sample is to follow up the subjects until the end point is reached. This should be done for at least one-half of the subjects who should be a random subsample so that no bias creeps in. Because of such a follow-up, the strategy does not remain cross-sectional but hopefully would still provide a representative picture of the target population.

The second method of collection of survival time data is to take a general sample from the target population and inquire if they ever (or in the previous few years) had the condition of interest—when it started and when it finished. This entirely depends on proper recall and again can be adopted for those conditions only for which the beginning and the end points are known to the subjects.

The third method is to take a random sample of those who have just experienced the end point and find out when the condition started. For example, this can be adopted for death due to Alzheimer's disease. The time of start of the disease can be obtained by inspecting records. A great advantage of this method is that there are no incomplete segments. See if it is easier for you operationally to assess survival time in this manner.

The most satisfying but perhaps most difficult method is to enroll people at the beginning of the condition and follow them up till the end point. If the end point is death, the total follow-up period could be quite long. Thus, the observation can be terminated after a prefixed reasonable time so that most subjects are able to reach the end point. Some who do not could be considered as censored values.

Before proceeding ahead with survival analysis, assess that there is no discernible pattern in the durations or in reaching to the end point when arranged by enrolment. After reordering by the date of entry, confirm that deaths and durations of survival are randomly distributed (see **censoring of observations**). Early recruiters should not have unusually high (or low) death rates or unusually long (or short) duration of survival.

## Statistical Measures of Survival

Because of censoring of values, computation of mean survival time is problematic. If these are ignored, the mean of the remaining known durations can be unreliable and biased: unreliable because the $n$ available now is smaller whereas actually there are more subjects in this study, and biased because the subjects excluded from this calculation due to incomplete observations could be very different. If this is the duration of survival after a treatment, the persons alive at the end of follow-up are exactly those that may have benefited. If these censored observations were included with truncated values, the means would obviously be an underestimate. Thirdly, survival duration is known to have highly skewed distribution since some people

tend to survive for long duration. And for such highly skewed values, it is known that mean is not a valid representative value.

If mean is not appropriate for survival, which summary measure should we use? The choice immediately falls on median, which is much less affected by such vagaries. **Median survival time** is indeed used as a summary to measure the central value of the duration of survival and to compare two or more groups, although you can see that this too is not completely free of bias because of incomplete segments.

For varying durations, one can think of person-years as the denominator and come up with an average such as 2.7 deaths per person-year. Although person-year can take care of the incomplete segments, this annoyingly considers first year after treatment on the same pedestal as say the fifth year. Clearly this is not true for survival as the risk of death increases as the time passes—one because of advancing age and two because the effect of treatment generally wanes. Thus, a measure such as deaths per person-year (or 100 person-years) is also ruled out for survival studies.

Beside median survival time, the other appropriate measure is survival rate such as 5-year survival rate. This can be adjusted for censored values through the method of survival analysis, and is a commonly used parameter in survival studies.

The most popular measure of duration of survival is **expectation of life**. This is discussed as a separate topic and requires that there are no incomplete segments. Main methods of survival analysis with incomplete segments include the **life table method** and the **Kaplan–Meier method**, which are also separately presented. These are the core methods that delineate the pattern of survival for a group on a particular regimen. Both these methods do not consider prognostic factors affecting the durations. The methods of assessing the role of various factors in affecting the duration are presented in the topic **Cox regression**. For comparing the survival pattern in two groups, **log-rank method** is generally used.

## survival bias, *see* bias in medical studies and their minimization

## survival curve/function, *see also* Kaplan–Meier method

The term *survival curve* is used for graphical representation of the pattern of survival over time, and survival function is its mathematical counterpart. The curve is drawn on the basis of the function. This is obtained as follows.

Consider first the life table method where survival is recorded at intervals of time rather than continuous. Under the **life table method**, where the data are obtained at intervals of time and not continuously,

$$p_k = \frac{n_k - d_k - c_k/2}{n_k - c_k/2};\ k = 1,2,\ldots,K;$$
$$= 1 - \frac{d_k}{n_k - c_k/2},$$

where

- $p_k$ is the proportion surviving the $k$th interval among those who survived the $(k–1)$th interval (the estimated conditional probability).
- $n_k$ is the number of subjects known to be alive at the beginning of the $k$th interval.
- $d_k$ is the number of deaths in the $k$th interval.
- $c_k$ is the number of subjects with incomplete segments in the $k$th interval whose fate is unknown.

The numerator of this formula is the number of survivors and the denominator is the number of subjects at risk. Both contain half of those with incomplete segments. The cumulative probability of surviving the $k$th interval can be estimated using this equation. By product rule for conditional probabilities, this is

$$s_k = p_k p_{k-1} \ldots p_2 p_1;\ k = 1, 2, \ldots, K.$$

This probability of survival can be obtained for each of the $K$ intervals. This is what is called the *survival function*. Its plot versus time $k$ is called the *survival curve*. Since each $p_k \le 1$, the survival probability $s_k$ is nonincreasing and will eventually decline as time passes. Statistical software can easily do these calculations and plot the survival curve. Note that this method is nonparametric as it does not depend on any particular form of distribution of survival period.

For illustration, consider the following data on survival time of 15 patients following radical mastectomy for breast cancer. The study started in January 2011 and continued till December 2014. Thus, the maximum follow-up was 4 years. However, many patients joined the study after January 2011 as and when radical mastectomies were being performed.

Survival time (months):

6 8 20 20 20+ 24+ 25+ 30+ 35+ 37 37 38+ 40+ 42 45+,

where + means that the patients are lost to follow-up or not followed up after this period. They are the incomplete segments. Their exact survival time is not known, but it is at least as many months as shown. The patients are deliberately ordered by survival time. This order is not important but makes the presentation simple. The sample size is small and there are many incomplete segments, which is not desirable for survival analysis, but the data are still adequate to illustrate the method.

The survival curve is shown by the dotted lines in Figure S.21. Median survival time corresponds to survival function = 0.5, which is the same as the time at which 50% survive only if there are no censored observations. Thus, 50% survival and survival function = 0.5 can be different. More exactly, median is calculated as the smallest survival time at which survival probability is at least 0.5. From the figure, this can be worked out to nearly 43 months in this example.

Since the duration of survival time is generally highly skewed, mean is not used. However, if needed, mean survival time is estimated by the area under the survival curve. Some software packages use only uncensored time points for calculation of mean, but that tends to give a biased estimate.

Dispense with the restriction that survival is recorded in time intervals. Assume that continuous monitoring is done and the exact time of death is recorded. Analysis of such survival data is done with

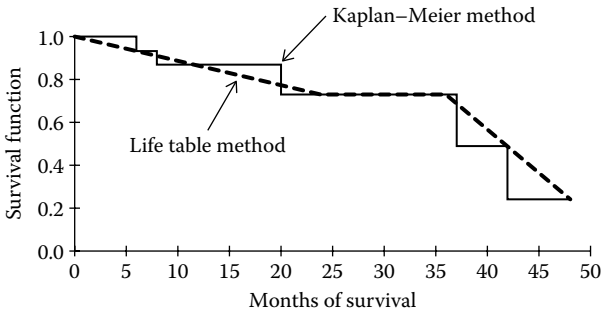



**FIGURE S.21** Survival function of breast cancer cases: life-table method and Kaplan–Meier methods.

the help of the **Kaplan–Meier method**. As explained in that topic, the survival function in this case is

$$s_t = p_t p_{t-1} \ldots p_2 p_1;\ t = 1, 2, \ldots, T,$$

where

$$p_t = \frac{n_t - d_t}{n_t};\ t = 1, 2, \ldots, T.$$

$p_t$ is the proportion surviving the $t$th time point among those who survived the $(t-1)$th time point.
$n_t$ is the number of subjects at risk of death at the $t$th time point, i.e., those who are still being followed up (this excludes those with incomplete segments, i.e., $n_{t+1} = n_t - c_t - d_t$).
$c_t$ is the number of subjects with incomplete segments.
$d_t$ is the number who died at the $t$th time point.

This is the estimated survival function in this case. It is basically the same product as in the formula given earlier but is now computed for each time point instead of the time interval. Only unique time points are considered. If a time point is applicable to two or more subjects, it is counted only once. This method requires the calculation of as many survival rates by the product rule as there are events, unless several events occur at the same time. Hence, the Kaplan–Meier (K-M) method is also called the **product-limit** method. The larger the $T$, the smoother the survival curve. The K-M method for estimating survival probability at the last time point is usually very unreliable because of heavy censoring toward the end of the trial.

Consider the same duration of survival of breast cancer patients after radical mastectomy as in the previous example, but now there is no grouping of survival time. The survival curve is shown in Figure S.21. Since time is continuously observed in this setup, many deaths at one time point are shown by a stepladder. The median duration corresponding to 50% survival (0.5 on $y$-axis) is nearly 37 months. This is very different from the 43 months arrived at by using the life table method.

For other types of survival function, see **hazard functions**.

## survival rate, *see also* survival analysis

This is the proportion of subjects that survive for a particular time during the follow-up in survival studies. Thus, a 5-year survival rate would be different from a 10-year survival rate. Duration is a necessary ingredient of all survival studies, and survival rate is no exception. Computation of this is not straightforward because of **censoring** in survival data but can be easily computed when the **survival function** is available.

Survival rates at any point in time are like proportions, and the comparison of two groups can be easily done by chi-square when the usual validity conditions of this test are fulfilled. These survival rates are calculated by the **life table method** when the observation of survival time is in the intervals, and by the **Kaplan–Meier method** when it is continuously observed. Survival–deaths in the two groups will give a 2 × 2 table, which will provide chi-square with 1 df as usual. The procedure remains essentially the same if there are more than two groups. A weakness of the procedure is that the time points chosen for calculating survival rate can be arbitrary, and different time points can give different results. The procedure to compare the overall survival pattern is the **log-rank**, but this procedure fails to detect if a big and significant difference exists at one or two specific time points, and not at other time points.

Survival in gastric cancer patients depends on a large number of factors, but this example is restricted to the type of surgical operation. Zhang et al. [1] have presented 5-year survivals as shown in Table S.22.

Censored observations are excluded, although, as mentioned earlier, this may cause bias. Now the last two columns of the table are a regular 3×2 **contingency table**, and the statistical significance can be tested by the usual chi-square. In this case, $\chi^2 = 31.66$, which is highly significant ($P < 0.01$) for 2 df. Thus, there is sufficient evidence to conclude that gastric cancer cases with three types of operations have different 5-year survival rates.

With such a large $n$ as in this example, statistical significance is a foregone conclusion. The actual utility of this type of analysis is that the survival in proximal subtotal gastrectomy is nearly half of that in distal gastrectomy. Deaths are 79.9% in proximal and 62.8% in distal. This tends to quantify the hazard of death and says that the hazard ratio for death in proximal operation is nearly 1.3 relative to the distal operation (this is based on deaths and not on survivals). Not many studies of this type aim at this kind of a useful conclusion.

For the confidence interval on survival rate, we need its standard error (SE). If the expression for **survival function** is messy, the SE cannot be simple. However, if deaths are not too many at any time point and if you started with a sufficient number of subjects so that the number of survivors remains reasonably large at the end, the SE is approximated by

$$\mathrm{SE}(st) = \sqrt{\frac{s_t(1-s_t)}{n_t}},$$

where $s_t$ is the survival function at time $t$, and $n_t$ is the total number of subjects at risk at time $t$. This SE is the same as for the usual proportions and is applicable to the individual time points. The SE would be different for different points of time, just as survival $s_t$ and the number of subjects $n_t$. For life-table survival also, the approximate SE is the same—just that the subscript $t$ is replaced by subscript $k$. This is not applicable to small $n$, and even when $n$ is large, note from SE that it increases (the precision declines) with time as the survival reduces, and it is based on fewer and fewer subjects. For cumulative survival till time $T$,

$$\mathrm{SE}(S_T) = s_T \sqrt{\sum_{t \le T} \frac{d_t}{n_t(n_t - d_t)}},$$

where $d_t$ is the number of deaths at time $t$. The interest in survival analysis is mostly in this SE to find the confidence interval for the survival function. The confidence interval is obtained as usual by ±1.96*SE under Gaussian conditions.

1. Zhang XF, Huang CM, Lu HS, Wu X-Y, Wang C, Guang G-X, Zhang J-Z, Zheng C-H. Surgical treatment and prognosis of gastric cancer in 2,613 patients. *World J Gastroenterol* 2004;10:3405–8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4576218/

## symmetric distribution, *see* skewness and the coefficient of skewness

## synergism, *see* interaction

## synoptic chart, *see* Lexis diagram

## synthesis of research, *see* research synthesis, and also meta-analysis

## systematic random sampling

A sampling is called systematic random when the first unit is selected at random from the initial set as per the **sampling fraction** and the remaining automatically as per the same sampling fraction. If 25 subjects are planned to be in the sample out of a total population of 200, the sampling fraction is 1 out of 8, and the first is selected at random out of first 8. Others are selected by an automatic procedure by adding the sampling fraction. If the first randomly selected is number 3, the next in our example will be number 11, then 19, etc. This presumes that the subjects have a defined sequence, that is, you know the serial number of each patient. If the subjects are not in sequence, this sampling scheme requires that each subject of the population be assigned a serial number before the sampling is executed.

Let there be $N$ subjects in the total population, numbered 1 to $N$ in any order. If the size of the sample is $n$, the sampling fraction is $I = N/n$; in fact, the fraction is $n/N$, but for systematic sampling, it is easier to work with $I = N/n$. If $I$ is not an integer, then the integer part is taken. The first unit is selected at random from the first $I$ units. Suppose this is the $r$th unit. Then the subsequent units in the sample are $(r + I)$th, $(r + 2I)$th, etc. Thus, the first selected unit determines the entire sample.

Let there be $N = 101$ subjects in the target population out of which $n = 8$ are proposed to be selected by the systematic method. The sampling interval is $I = 101/8$, and its integer part is 12. If the randomly selected unit (subjects in this case) out of the first 12 is the 9th, then the remaining units have numbers 21, 33, 45, 57, 69, 81, and 93.

Although the systematic method works well when the sampling interval is an integer, two kinds of anomalies can occur in other cases: (i) The last few units may never have a chance to be included in the sample. In the preceding example, if the first selected number is the maximum 12, then the last number in the sample is 96. Thus, unit numbers 97–101 can never be selected. (ii) In some cases, the sample size may increase by one. In this example, if the first randomly selected unit is the 4th, then the last would be the 100th. The sample then would contain nine units instead of the stipulated eight. However, this is not serious when $n$ is large, say greater than 30. A way out of this mess is **circular sampling**, in which the start is made from a random number between 1 and $N$, and every $I$th thereafter is selected in a cyclical manner, i.e., the $(N + 1)$th unit becomes the first unit again. This is done until $n$ units are selected.

Systematic random sampling (SyRS) has the following merits: (i) It is very easy to execute. When the patients are coming to a clinic, every eligible $I$th after the random first can be easily pulled out for inclusion in the study. It is also very quick relative to simple random sampling (SRS), where each unit is to be selected at random. (ii) SyRS does not need a full **sampling frame**—knowledge about the total number of units in the population is enough. If you know that a clinic generally gets $N = 275$ patients a month and a sample of size $n = 30$ is to be chosen from among the patients coming in a particular month, then $I = 9$. If the first randomly selected patient is the 7th, then the others in the sample are the 16th, 25th, etc. These can be chosen according to the sequence of arrival, and no list of patients is required for sampling. (iii) In some situations, it can yield a much more representative sample than SRS since equally spaced selection until the end gives the opportunity to all sections of the population to be represented. If a sex clinic runs 6 days a week with Mondays and Thursdays reserved for teenagers, Tuesdays and Fridays for male adults, and Wednesdays and Saturdays for female adults, then SyRS would automatically give proportional representation to each of the three groups of patients when the sampling includes all the days.

SyRS is not without demerits: (i) The SyRS could yield a biased sample if a periodicity or trend is hidden in the subjects as you go from 1 to $N$. (ii) As in the case of SRS, this method can fail to give adequate representation to some specific groups of interest.

For an example, see Vitolo et al. [1] who describe results of a study in Sao Leopoldo (Brazil) for evaluating risk factors of folate deficiency among adolescents. In the first stage, they selected census regions, then they selected random blocks and street corners where sampling would start, and finally houses were sampled systematically one in three. All individuals of age 10–19 years in the selected houses constituted the sample. In this study, systematic sampling was done at the last stage to select the houses. In most large-scale studies where such multistage sampling is used, systematic sampling can be at one or more stages.

1. Vitolo MR, Canal Q, Campagnolo PD, Gama CM. Factors associated with risk of low folate intake among adolescents. *J Pediatr (Rio J)* 2006;82: 121–6. http://www.scielo.br/scielo.php?pid=S0021-75572006000200008 &script=sci_arttext&tlng=en

## systematic reviews, *see also* Cochrane collaboration/reviews

Aptly called research into research, systematic reviews are focused reviews of the relevant literature on a particular topic or on a defined outcome, and are done with the objective to filter a common message and to highlight the discrepancies as reported by different studies. Being systematic requires that the following steps are listed in the **protocol** and rigorously followed: (i) identify all published and unpublished evidence on the topic of focus through scanning of literature databases, libraries, government documents, and all other sources (for this, it is helpful to list your search criteria upfront); (ii) prepare a set of criteria for inclusion and exclusion of reports in your review in a manner that all relevant documents are included and the irrelevant ones are excluded; (iii) select studies for inclusion that meet this preset criteria; (iv) assess the quality of each study on the basis of some established norms such as representative sample, design that controls the bias, adequate analysis, and clear exposition; (v) prepare a subset of documents that meet the defined quality standard; (vi) read these documents thoroughly, filter the common findings keeping any flaws in mind, and synthesize the varying results without any personal bias (for this, help of **meta-analysis** can be taken but that is not necessary); and (vii) interpret the findings and present a unified conclusion—in case of significant differences among studies, report that prominently. These differences could be due to varying background characteristics of the population under study, difference in design for collection of data, difference in analytical strategies, altered focus of the authors or the agency publishing the report, etc.

Systematic reviews require a lot of critical thinking about the possible bias of the investigators as most authors like to further their own hypothesis, and their report may be silent on issues unfavorable to their hypothesis. In addition, the **publication bias** can completely distort the findings as the negative results rarely find their way into quality journals. Perhaps one way out is to give extra weight, say double, to negative results assuming that only one of the two negative findings are published (the situation though seems to be rapidly changing). Extra care is required if both qualitative and quantitative evidences are to be synthesized because they require special expertise.

Systematic reviews provide high **level of evidence** and are respected in medical sciences because many times wide variation is seen in findings of different workers, and it becomes necessary to scan them closely to come to a unified conclusion. Most respected are **Cochrane reviews** [1]. They do help the medical fraternity to confidently use a new method of prevention, diagnosis, or treatment when such a systematic review gives a positive recommendation. They also reveal the areas where the evidence is nonexistent or week, and thus where the new research ought to be focused.

A large number of systematic reviews are published each year, which are peer-reviewed for credible evidence, but some are not as meticulous and can mislead. Thus, the results of all systematic reviews cannot be blindly accepted.

For further details, see Cochrane Consumer Network [2].

1. Cochrane Library. *Cochrane Reviews*. http://www.cochranelibrary.com
2. Cochrane Consumer Network. *What Is a Systematic Review*? http://consumers.cochrane.org/what-systematic-review

# T

## tabular presentation of data and results

Tabular presentation is a summary of data in the form of a table. Tables are powerful tools to display the data in an intelligible and precise format. They tend to condense and summarize the information and sometimes are able to communicate the intricacies better than text. Tables are also made to highlight specific features of a data set. Generally, only the salient features of the data sets or of the results are presented in the form of tables and the unimportant features are ignored. Data that are not easy to describe in a couple of sentences in text are generally tabulated.

The purpose of a table is to give an overall view of the data or results—thus, it should not be too complex since then it is not able to fulfill its purpose. All row headings and column headings must be unambiguous. Sometimes, it is not feasible to explain all column and row headings in a proper manner, or sometimes, numbers are inconsistent, which requires additional explanation. For readers to understand the table in one shot, add footnotes to the table where necessary. Footnotes should be marked with *, **, †, @ in this order or with superscripts 1, 2, 3, 4, and so on. Also, you may have to explain abbreviation in a footnote.

If you have two or more tables with identical rows or columns, examine if they can be combined. The message that a table is supposed to convey should be obvious at a glance. When writing about the table in the text, it should be straightforward in that it explains the key points of the data in the table in a couple of sentences; preferably, the table should be self-explanatory. The table titles, format, and presentation should be consistent throughout the report. A thorough exposition of the design and usage of tables is given in Ref. [1]

Primarily, a table presents the number of subjects in different categories such as the frequency with different cholesterol levels like 100–149 mg/dL, 150–199 mg/dL, and 200+ mg/dL categories, or cases with cancer at different sites in various years. The purpose is to see the **distribution** of subjects—what values are more common and what values are rare, whether the values are highly concentrated or pretty much evenly scattered, whether the dispersion is large or is it small, and so on. For this purpose, the old method is of tally marks as described in the topic **frequencies** but now almost everybody uses the "bins" for quantitative data, which are defined with the help of a software package. The system of manual tally marks in any case is unsuitable for large sets of data as it is time consuming and prone to errors.

For an adequate tabular presentation of numerical data, it is necessary that the bins are suitably constructed. Customarily, categories beginning with digits 5 or 0 are made such as 97.0–97.4, 97.5–97.9, and so on, for body temperature (in degrees Fahrenheit), and <20.0, 20.0–24.9, 25.0–29.9, and so on, for body mass index (in kilograms per meter squared), but that is being gradually replaced with "natural" categories dictated by the data. Generally, 6 to 10 bins are considered adequate to represent the distribution, but you can have more bins for large data sets. See the topic **class-intervals** for open categories at the beginning and end of the data values. Percentages of frequencies should be based on appropriate total and the number of subjects in each group should be mentioned, preferably in the column heading. Ensure that the numbers in different tables are consistent with one another. Specify the unit of each measurement.

For **qualitative measurements**, there is hardly any opportunity to choose the "bins"—the qualities generally determine the number of categories. For example, the categories for blood groups have to be O, A, B, and AB, and the categories for site of cancer for males have to be lung, prostate, esophagus, brain, blood, and so on. However, in this case, all the uncommon sites can be lumped together in an "others" category so that the number of categories does not become unmanageable and the number of subjects in each category is adequate for statistical analysis. For other details of these kinds of frequency tables, see the topic **contingency tables**.

### *Types of Tables*

Simple tables are the usual tables that we are not discussing but, in contrast, are composite tables that contain detailed information on two or more groups such as the use of smokeless tobacco in males and females of some Southeast Asian countries (Table T.1) and combined for both sexes. Table for each sex is what we are referring to as a simple table.

Table T.1 does not contain frequencies, but instead contains percentage prevalence. The table illustrates the following:

- The title of the table should be as specific as possible without being lengthy. The title of Table T.1 explains that these are prevalences in percentages and are given for different age-groups and sex for some countries. It also says that the data belong to the years 2006–2009.
- All entries are with one decimal. Some are NA and the footnote at the bottom of the table explains that NA means "not available."
- There is another footnote explaining what the asterisk sign stands for. The data for Bhutan are restricted to one city only.
- There are separate blocks for males, females, and persons (where "persons" stands for both the sexes combined).

Data on groups you want to compare should be in columns. For example, if your objective is to compare cases with controls, data on cases should be in one column and data on controls should be in adjacent column. If you want to compare males and females, data on these two groups should be in column (and not two rows). Table T.1 on tobacco use primarily compares prevalence in different countries as the countries are in the columns. When two or more tables present results on the same set of variables, the order of variables should be the same.

Beside frequency tables (or percentages), statistical results are also presented in tabular form as done for logistic results in Table T.2. This is on the factors affecting postoperative pain relief in patients who have had bariatric surgery. The factors considered are (i) the kind of support, if any, that the patients walked-in with; (ii) preoperative continuous positive airway pressure (preop CPAP) done or not; (iii) STOPBANG score; and (iv) body mass index (BMI) at the

**TABLE T.1**
**An Example of a Compound Table: Age- and Sex-Wise Prevalence (%) of Smokeless Tobacco (Current Users) in Some Southeast Asian Countries (2006–2009)**

| Sex | Age-Group (Years) | Country | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Bangladesh | Bhutan[a] | India | Indonesia | Myanmar | Nepal | Thailand | Timor-Leste |
| Males | 15–24 | 9.3 | NA | 23.1 | 1.2 | 45.0 | 23.5 | 0.1 | 13.6 |
| | 25–34 | 22.0 | 24.2 | 39.1 | 3.4 | 62.9 | 28.9 | 0.3 | 27.3 |
| | 35–44 | 32.5 | 22.4 | 39.5 | 2.2 | 54.3 | 38.8 | 0.7 | 35.2 |
| | 45–54 | 41.4 | 17.4 | 33.9 | 1.2 | 49.6 | 49.0 | 1.1 | 43.9 |
| | 55–64 | 38.9 | 17.2 | 33.7 | NA | 41.6 | 29.0 | 2.1 | NA |
| Females | 15–24 | 4.0 | NA | 8.2 | 0.1 | 5.3 | 0.3 | 0.0 | 0.4 |
| | 25–34 | 18.1 | 18.6 | 14.8 | 0.3 | 11.9 | 2.5 | 0.5 | 2.1 |
| | 35–44 | 37.6 | 18.9 | 21.1 | 0.4 | 18.8 | 6.5 | 0.7 | 5.9 |
| | 45–54 | 52.5 | 16.3 | 25.4 | 0.6 | 21.1 | 14.1 | 3.9 | 10.2 |
| | 55–64 | 62.7 | 13.8 | 32.7 | NA | 19.2 | 7.1 | 15.0 | NA |
| Persons | 15–24 | 6.6 | NA | 16.1 | NA | 21.4 | 12.4 | 0.1 | NA |
| | 25–34 | 19.9 | 21.7 | 27.1 | NA | 31.2 | 17.1 | 0.4 | NA |
| | 35–44 | 35.0 | 20.8 | 30.8 | NA | 31.5 | 22.0 | 0.7 | NA |
| | 45–54 | 46.7 | 16.9 | 29.8 | NA | 32.1 | 31.2 | 2.6 | NA |
| | 55–64 | 49.3 | 15.7 | 33.2 | NA | 28.4 | 20.5 | 8.8 | NA |

*Note:* NA, not available.
[a] Available for Thimpu city only.

**TABLE T.2**
**Example of Presentation of the Results of a Logistic Regression: Factors Affecting Postoperative Pain Relief in Patients Undergoing Bariatric Surgery**

| Factor | *n* (820) | Logistic Coefficient (*b*) | SE of *b* | *P*-Value | Odds Ratio (exp*b*) | 95% CI for Odds Ratio | |
|---|---|---|---|---|---|---|---|
| | | | | | | Lower | Upper |
| Support | | | | 0.738 | | | |
| No support | 703 | Reference | | | | | |
| Minor support | 27 | 22.702 | 2.761E4 | 0.999 | 7.234E9 | 0.000 | Very large |
| Major support | 90 | −0.455 | 0.583 | 0.435 | 0.635 | 0.203 | 1.989 |
| Preop CPAP[a] | | | | | | | |
| No | 710 | Reference | | | | | |
| Yes | 106 | 1.721 | 0.351 | 0.000 | 5.593 | 2.812 | 11.124 |
| STOPBANG score | 820 | 0.304 | 0.077 | 0.000 | 1.355 | 1.165 | 1.575 |
| BMI | 820 | 0.071 | 0.016 | 0.000 | 1.073 | 1.041 | 1.107 |
| Constant | | −8.022 | 0.957 | 0.000 | 0.000 | | |

[a] Preoperative continuous positive airway pressure for sleep apnea—information not available for four patients.



time of admission. Support and preop CPAP were categorical and STOPBANG score and BMI were continuous. The table gives the number of subjects in different categories, the logistic coefficients (*b*), their standard error (SE), odds ratio, which is $e^b$, and the 95% confidence interval (CI) for the logistic coefficient. It illustrates that the tables containing statistical results have a format very different from that of the tables containing number and percentage of subjects in different categories.

Such tables give structure to the answer and provide concrete evidence as large set of summary statistics can be meaningfully recorded in the table. However, they can also be a source of confusion when not properly drawn—thus, exercise due care. Tables on statistical significance should contain exact *P*-values that give more insight to the reader than just saying $P < 0.05/P > 0.05$ or significant/not significant, but $P < 0.001$ is acceptable. Customarily, only two or three decimal places are used in values of *P*. Use capital *P* since lowercase *p* is used to denote proportion in the sample. When feasible, also give the name of the statistical test on which *P*-value is based, but do not give a value for the test statistic such as *t*, *F*, and $\chi^2$ in a report unless required by the organization for whom the report is being prepared or by the journal to whom a paper is being sent for publication. State the CIs wherever applicable.

**TABLE T.3**
**Tabular Presentation of Analysis of Variance Results and Multiple Comparisons (Too Much Information Packed in One Table): Scores of Peritonitis Patients in Physical and Mental Domains of Quality of Life at Baseline and 1, 2, and 3 Months after Two Types of Surgical Regimens**

| | | Surgery 1 | | | Surgery 2 | | | Differences in the Two Surgeries | |
|---|---|---|---|---|---|---|---|---|---|
| Domain and Time | n | Mean Score | *P*-Value for Time Differences (*F*-Test) | Tukey Test for Time Differences | Mean Score | *P*-Value for Time Differences (*F*-Test) | Tukey Test for Time Differences | *P*-Value (*F*-Test) | Tukey Test |
| **Physical** | | | | | | | | | |
| Baseline | 20 | 10.25 | <0.001 | Sig from 1, 2, and 3 months | 10.30 | <0.001 | Sig from 1 and 2 months | 0.085 | NS |
| 1 month | 20 | 12.20 | | Sig from 2 and 3 months | 12.35 | | Sig from 2 and 3 months | | NS |
| 2 months | 20 | 13.05 | | Sig from 3 months | 11.15 | | Sig from 3 months | | NS |
| 3 months | 20 | 14.30 | | | 10.70 | | | | NS |
| **Mental** | | | | | | | | | |
| Baseline | 20 | 9.95 | <0.001 | Sig from 1, 2, and 3 months | 10.15 | <0.001 | Sig from 1, and 2 months | 0.167 | NS |
| 1 month | 20 | 11.65 | | Sig from 2 months | 11.60 | | Sig from 3 months | | NS |
| 2 months | 20 | 11.70 | | Sig from 3 months | 11.25 | | Sig from 3 months | | NS |
| 3 months | 20 | 13.65 | | | 9.90 | | | | NS |

*Note:* All Tukey tests at the 5% level of significance. NS: not significant.

Presentation of statistical results for repeated measures and for multiple comparisons can be very difficult. Table T.3 illustrates the problems encountered in this presentation and suggests one way of presentation. The table contains the means and SDs of quality-of-life scores in physical and mental domains in peritonitis patients at baseline and at 1, 2, and 3 months after two types of surgical intervention. Although this provides an overview of the results, this looks too congested for the reader. You might be able to devise a more ingenious way of presenting these kinds of results such as splitting the table to have one for the physical domain and another for the mental domain and having a separate table for difference between two surgeries.

A statistical table can be prepared in a large number of other formats also depending on the data they contain. We can have a table of random numbers (see the topic **random numbers**), a table of features of studies included in a meta-analysis, a life table (see the topic **expectation of life and life table**), and several other types of tables.

1. NC State University *LabWrite Resources.* Revised 16th May 2005. http://www.ncsu.edu/labwrite/res/gh/gh-tables.html, last accessed December 10, 2015.

## *t*-distribution, *see* Student *t*-distribution

## Taguchi designs

Developed by Genichi Taguchi in 1960 and explained fully by him in 1987 [1], Taguchi designs seek to integrate **quality control** into product design when he identified that the loss is minimized by operating on target with minimum variance. Thus, these designs integrate quality engineering with statistical methods and emphasize that design rather than inspection is important for quality control. The conventional quality control relies on the inspection of the products (such an inspection is actually for manufacturing processes). Taguchi brought design (before manufacturing) to the center of quality control process.


Genichi Taguchi

The basic premise of the Taguchi design is a loss function based on deviation from the target value. This means that if the target of a treatment process is to provide relief to at least 62% of cancer patients, but if the actual relief rate is 61%, there is a loss. In the conventional quality control group, 61% may be within the tolerance range and ignored for any corrective steps. In the Taguchi method, there is no defined tolerance range, and any deviation from the target contributes to the loss. It is concerned with the design of the product such that the variance around the target value is minimized.

Actions taken after the manufacturing process in the conventional setup increase the cost, whereas in the design, the cost is much less, because it is constructed before the process.

Historically, the best way of controlling variation was by the use of specifications: for example, $\tau \pm \Delta x$, where $\tau$ is the target and $\Delta x$ is the tolerance limit. Within the specification, the product was deemed satisfactory, and outside the specification, the product was deemed to be unsatisfactory. Say, we had $\tau = 100$ and $\Delta x = 10$: the specification was (90–110). With a stream of units with values 108, 102, 96, 92, and 90, the units passed and the production process were deemed satisfactory because all of them are within limits. However, suppose the sixth unit had a value of 89 and everyone suddenly panicked! Inspectors inspected incoming raw materials, and project teams brainstormed the process. All of this greatly increases the cost associated with the production of this product. Most importantly, note that the difference between the fifth unit (90) and the sixth unit (89) was less than any of the differences between earlier successive units, yet the fifth unit was deemed to be satisfactory and the sixth unit was deemed to be unsatisfactory in the conventional method! There was a real need for a better way of controlling variation in the manufacturing process, and that is where Taguchi designs came in.

There are instances of useful applications of the Taguchi approach in health and medicine. Emami et al. [2] discussed formulation and optimization of solid lipid nanoparticle formulation for pulmonary delivery of budesonide using the Taguchi design. They studied the impact of various process variables such as surfactant type and concentration, lipid content and organic and aqueous volume, and sonication time on particle size, zeta potential, mean dissolution time, and so on. Azadeh and Sheikhalishahi [3] discussed how this approach can be useful in performance optimization of health, safety, environment, and ergonomics in generation companies based on sensitivity analysis and maximum correlation.

The study by Cogdill and Drennen [4] is a useful resource in this respect. For further details of Taguchi designs, see Ryan [5].


1. Taguchi G. *System of Experimental Design*, Vols 1 and 2. UNIPUB/Kraus, 1987.
2. Emami J, Mohiti H, Hamishehkar H, Varshosaz J. Formulation and optimization of solid lipid nanoparticle formulation for pulmonary delivery of budesonide using Taguchi and Box-Behnken design. *Res Pharm Sci* 2015 Jan–Feb;10(1):17–33. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4578209/
3. Azadeh A, Sheikhalishahi M. An efficient Taguchi approach for the performance optimization of health, safety, environment and ergonomics in generation companies. *Saf Health Work* 2015 Jun;6(2):77–84. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4476202/
4. Cogdill RP, Drennen JK. Risk-based quality by design (QbD): A Taguchi perspective on the assessment of product quality, and the quantitative linkage of drug product parameters and clinical performance. *J Pharma Innovation* 2008 Mar;3:23–9. http://link.springer.com/article/10.1007/s12247-008-9025-3?no-access=true
5. Ryan TP. *Modern Experimental Design*. Wiley, 2007.


## Tanimoto dichotomy coefficient, *see* association between dichotomous characteristics (degree of)

## tapping

Health measurement scales (particularly the **Likert scale**) are based on the concept of "tapping" [1]. Likert developed the concept of measuring participants' attitudes to a topic by asking them to respond to statements about that topic (particularly as to how much they agreed with these statements), thereby tapping into the cognitive components of these attitudes. Tapping here relates to the measurement of the underlying phenomenon under investigation. This affects the design of a **questionnaire** by helping to decide what items in the questionnaire to use. This is important since no amount of statistical maneuvering after completion of the questionnaire can compensate for poorly chosen questions.

The motivation for developing a new tool (such as a Likert scale/questionnaire) comes about because the researcher believes that previously used methods are inadequate for one reason or another or do not completely cover the construct under study. The researchers can start with new items that may come from four different sources: clinical observation, theory, research, or expert opinion, although the lines between these sources are not firm. New items are chosen based on how well they are expected to tap into the underlying construct. Once the items have been chosen from one or more of these four sources, the researcher would ideally have more items than will actually end up on the scale. Conversely, we can argue whether the scale has enough items to tap with for adequately covering the domain under investigation. The technical term for this is **content validity**.

These concepts arose from achievement testing in schools where students are assessed to determine if they have learned the material in a specific content area, such as in the passing of medical school exams. Each item on the test should tap into one of the course objectives so that the items have content relevance. Conversely, each part of the syllabus should be represented by one or more questions so that the content is adequately covered. The criterion for selecting items is to eliminate those that are ambiguous or incomprehensible and to include items on those aspects that are left uncovered.

See Streiner and Kottner [2] for a good working example on this topic.


1. Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use*, Third Edition. Oxford, 2013.
2. Streiner DL, Kottner J. Recommendations for reporting the results of studies of instrument and scale development and testing. *J Advanced Nurs* September 2014;70(9):1970–9. http://onlinelibrary.wiley.com/doi/10.1111/jan.12402/full


## target population, *see* population (the concept of)

## Tarone–Ware test, *see also* log-rank test (Mantel–Cox test)

This test compares two survival curves in a manner similar to what a **log-rank test** does, but provides differential weights to different time points, weight being dependent on the number of subjects available at various time points.

The log-rank test compares two survival curves without reference to any particular point of time but is valid when all time points are equally important. Sometimes, time points with a higher number of subjects at risk are given more weight. Naturally, later time points have lesser number of subjects because of accumulation of deaths (or failures) as the time passes. In that case, another method such as the **Breslow test**, which gives weight proportional to the number of subjects at risk at different time points, could be used. However, this sometimes becomes overdone as the initial time points where the number of subjects at risk is higher tend to decide the statistical significance. As a middle path between the log-rank test and the Breslow test, some prefer the Tarone–Ware test [1], which gives

weight proportional to $\sqrt{n_t}$, where $n_t$ is the number of subjects at risk at time point $t$. All these are **nonparametric tests** since they do not require any specific pattern of survival duration, and they belong to the same family of tests [2].

Consider the equation to calculate the $\chi^2$ value in the log-rank test where group 1 contributes $(\Sigma d_{1t} - \Sigma e_{1t})^2$, where $d_{1t}$ is the number of deaths at time point $t$ in group 1 and $e_{1t}$ is the expected number of deaths at this time under the null hypothesis of equality of survival curves. This gives equal weight to each failure time when combining (observed – expected) failures. This does not have to be the case and a weight $w_j$ can be given to each failure time and use $[\Sigma w_t (d_{1t} - e_{1t})]^2$. The most common weighting is $w_j = n_j$, where $n_j$ is the total number at risk in the two groups, referred to earlier as the Breslow test. This has been observed to be more powerful in detecting specified differences when the survival distribution is **log-normal** but may have less power in the case of heavy **censoring**. On the one hand, the log-rank test is more powerful for **Weibull** survival distributions and equal censoring at different time points. The Tarone–Ware test, being the middle path with $w_j = \sqrt{n_j}$, can have reasonable power across a wide range of survival functions, although it may not be as good in some specific situations as listed.

Sanna et al. [3] calculated *P*-values for log-rank, Breslow, and Tarone–Ware tests for their data on survival using quartiles of NT-pro-BNP levels in outpatients undergoing elective cardioversion of persistent atrial fibrillations and reported nearly similar results by the three tests. Statistical software packages generally have options to perform all the three tests and you can choose the one most appropriate for the data in hand.


1. Tarone RE, Ware J. On the distribution free tests for equality of survival distributions. *Biometrika* 1977;64:156–60. http://www.jstor.org/stable/2335790
2. Marubini E, Valsecchi MG. *Analysing Survival Data from Clinical Trials and Observational Studies*. Wiley, 1995:pp. 99–100.
3. Sanna T, Sonaglioni A, Pieroni M, Dello Russo A, Pelargonio G, Casella M, Zichichi E, La Torre G, Narducci ML, Bellocci F. Baseline NT-Pro-BNP levels and arrhythmia recurrence in outpatients undergoing elective cardioversion of persistent atrial fibrillation: A survival analysis. *Indian Pacing Electrophysiol J* 2009;9(1):15–24. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2615058/


## telephone sampling

Telephone sampling is an effective instrument in developed countries as almost every household has a phone, and the listing is available separately for homes and commercial establishments (white and yellow pages). Any of the methods of probability or nonprobability sampling can be used to select telephone numbers. However, keep in mind four rather severe limitations: (i) Many households have two or more phone numbers. (ii) Some households without telephones can never appear in the telephone sampling. This is a serious limitation for areas where many households are without phones. These households are likely to have a relatively low socioeconomic status and a bias can creep in if the condition under investigation is associated with the socioeconomic status. (iii) Some households suppress their telephone numbers and their numbers are not listed. (iv) Attrition (nonresponse) rate is generally very high in telephone surveys compared with personal interviews, and this can increase the bias. However, in telephone sampling, a replacement is easy to make for nonresponse. This kind of sampling enables a much cheaper and faster data collection from geographically scattered samples compared to field interviews and may also be better than postal surveys because of the personal voice contact.

According to an evaluation of telephone sampling design carried out in 1978 by Landon and Banks [1], the increasing cost of personal interviewing and the higher crime rate in cities have made personal interviewing less practical. Telephone sampling is feasible in areas where telephone penetration is nearly 100%, but the basic problem is in eliciting responses in the first place and getting the right response in the second. Since the nonresponse could be substantially high, the resulting bias can render results doubtful. The latest telephone directory, particularly the printed one, if available, lends itself to a good **sampling frame**, but **stratified random sampling**, say, by age or sex, or even socioeconomic status, is nearly impossible. Only a **simple random sampling** can be done with all its advantages and disadvantages. For refusals or not-at-home responses, callbacks can be arranged, but the sample finally available with responses may still be biased.

Thomas and Purdon [2] discuss the pros and cons of telephone sampling.


1. Landon EL, Banks SK. An evaluation of sampling design. *Advances in Consumer Research* 1978;5;103–8. http://www.acrwebsite.org/search/view-conference-proceedings.aspx?Id=9408
2. Thomas R, Purdon S. Telephone methods for social surveys. *Social Research Update*. University of Surrey. Winter 1994; Issue 8. http://sru.soc.surrey.ac.uk/SRU8.html


## tertiles, *see* quantiles

## test criterion

A statistical test criterion is a validly established rule by which certain sample summaries are considered relatively probable and others relatively improbable for a particular hypothesis [1]. Such a criterion plays a crucial role in empirical sciences where the decisions are based on the available values. These criteria are also called tests of statistical significance or just **test statistics**. The distributional form of these criteria has been studied and is mostly known. This depends on (i) the nature of the data (qualitative or quantitative); (ii) the form of the distribution such as Gaussian or non-Gaussian when the data are quantitative; (iii) the number of groups to be compared (one, two, or more than two); (iv) the parameter under consideration (it can be the mean, median, correlation coefficient, etc., in case of quantitative data; it is always a proportion $\pi$ or a ratio of proportions in the case of qualitative data), or difference; (v) the size of the sample (small or large); and (vi) the number of variables considered together (one, two, or more). All this is analogous to saying that the criterion for assessing the health of a person depends on the age, gender, purpose, general health or organ-focused health, and so on.

Besides the considerations mentioned in the preceding paragraph, the criterion is constructed in a manner that it involves the parameter of interest, and the distribution of the criterion, especially under the null hypothesis, is known. Then, only the probable and improbable values can be identified. The decision depends on what is called the ***P*-value**. The procedure to obtain a *P*-value corresponding to a null hypothesis requires identifying a criterion (such as **Student *t***, **chi-square**, and ***F***) that is suitable for the scenario in hand. For example, the criterion $F = \dfrac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ involves the parameters $\sigma_1^2$ and $\sigma_2^2$ whose distribution under $H_0$: $\sigma_1^2 = \sigma_2^2$ is $F$ with $(n_1 - 1)$ and $(n_2 - 1)$ df, where $n_1$ and $n_2$ are the respective sample sizes. It is only when such distribution under the null is known that *P*-values can be obtained.

As another example, consider $z = \dfrac{\left(\frac{1}{2}\ln\frac{1-r}{1+r}\right) - \left(\frac{1}{2}\ln\frac{1-\rho}{1+\rho}\right)}{1/\sqrt{n-3}}$, where $r$ is the sample correlation coefficient and $\rho$ is the population correlation coefficient, which becomes $z = \dfrac{\left(\frac{1}{2}\ln\frac{1-r}{1+r}\right)}{1/\sqrt{n-3}}$ under $H_0$: $\rho = 0$ and we know that it follows a **Gaussian** (Normal) **distribution** for large $n$.

The conventional statistical criteria generally require that the sample values come from a specified distribution. For example, Student $t$ requires that the values are from a Gaussian distribution. Certain very general conditions such as independent observations and large $n$ may also be required for some criteria. Nonparametric criteria are relatively free from such encumbrances but still want independence of values.

In addition to the test criteria, there are other statistical criteria for purposes other than tests of hypothesis. Examples are the **Akaike information criterion**, which is used to determine the adequacy of a model without "test of hypothesis" overtones, and **coefficient of determination**, which is also used as a criterion for the same purpose.

Statistical test criteria have a long history—for an interesting account, see Ekstrom [1].

1. Ekstrom J. On statistical criteria: Theory history and applications. *Department of Statistics Papers, UCLA*. http://escholarship.org/uc/item/87t603ns

## tests of hypothesis (philosophy of)

Statistical tests of hypotheses are commonly used but are also abused because of lack of understanding of their philosophy. This is amply illustrated by the following examples. Ptolemy in the second century AD propounded that the sun revolves around the earth. This remained the "truth" for 14 centuries until Galileo came up with evidence against it in the 16th century and established a new truth, which says that the earth revolves around the sun. Newton's laws of mechanics were accepted until Einstein discussed circumstances in which some of them did not hold. Peptic ulcers were believed to be caused by acidity until sometime ago when *Helicobacter pylori* was found to be a culprit in some cases. As of today, coronary heart disease is not considered to be caused by any infection, but who knows that a rebuttal will come soon. The moral is that the "truth" changes as new convincing evidence emerges. Tests of hypothesis in statistics also follow the same system and look for evidence against a given hypothesis.

The philosophy is best understood with the help of an example of a court decision in a crime case. Consider the possibilities mentioned in Table T.4a. When a case is presented before a court of law by the prosecution, in most societies, the judge is supposed to start with the presumption of innocence. It is up to the prosecution to present evidence against the innocence of the person. The evidence should be enough to change the initial opinion of the judge. Guilt should be proven beyond reasonable doubt. If the evidence is not sufficient, the person is acquitted whether he/she actually committed the crime or not. Sometimes, the circumstantial evidence is strong and an innocent person is wrongly pronounced guilty. This is considered a very serious error. Special caution is exercised to guard against this type of error even at the cost of acquitting some criminals.

An empirical strategy in science is to find evidence against a hypothesis that is stated in **null** form. If this evidence is sufficient, the null hypothesis is rejected; otherwise, it continues to be considered "truth" by default. By analogy, in the development of a new therapeutic regimen, it seems reasonable to demand evidence against the hypothesis that there is no effect at all. Similar is the case with diagnosis (Table T.4b) where enough indication must be present that the disease is present. Statistical testing also follows the same process (Table T.4c). Decisions in all the three setups involve the same kind of errors. Statistical **Type I** and **Type II** errors that occur in testing of hypothesis setup are explained separately.

This process is difficult to implement without setting up a null hypothesis. As testing proceeds, philosophy involves not accepting any null (statistically, this is different from rejecting any null) but only rejecting whenever sufficient evidence is present. The statistical hypothesis testing parallels the court judgment just illustrated. This process is also called the **test of significance** because the effect seen in samples is checked whether it is significant enough to be considered most likely real. Martin explores the analogy further [1].

Since samples by their very nature are uncertain, the conclusion depends on what sort of data are obtained from the subjects. Statistical tests of significance are precisely meant to deal with uncertainties arising from **sampling fluctuations**. They provide the answer to the question: What is the likelihood of sample values given that null hypothesis is true? If this likelihood is exceedingly small, say, less than 5%, the null is considered implausible and rejected. At the same time, it would be ridiculous to accept a hypothesis whose likelihood of giving the obtained sample is only 15%. Thus, the only conclusion drawn in this case is that the sample fails to provide sufficient evidence to reject the null hypothesis. It is not accepted and the situation reverts to what it was before that study.

Many would consider it idiosyncratic that the sample values are searched for evidence against a null without recourse to finding what they are for—what they support. Perhaps another example would be more convincing. Consider the claim of a pharmaceutical company

**TABLE T.4**
**Errors in Court Judgment, Diagnosis, and Statistical Decision**

| (a) Court Setting | | | (b) Diagnosis | | | (c) Statistical Decision | | |
|---|---|---|---|---|---|---|---|---|
| | Assumption of Innocence | | | Disease Actually Present | | Statistical Decision | Null Hypothesis | |
| Judgment | True | False | Diagnosis | No | Yes | | True | False |
| Pronounced guilty | Serious error | ✓ | Disease present | Misdiagnosis | ✓ | Rejected | Type I error | ✓ |
| Pronounced not guilty | ✓ | Error | Disease absent | ✓ | Missed diagnosis | Not rejected | ✓ | Type II error |

that their drug is superior to the existing angiotensin-converting enzyme inhibitors in improving insulin sensitivity in diabetic hypertensives. In a trial of matched cases, improvement was seen in 7 out of 10 patients who were given the new drug compared with 6 out of 10 who were given the existing drug. It is evident that the sample size $n = 10$ in each group is small and the difference is too small to provide confidence to pronounce that the new drug is better. The difference could have arisen as a result of sampling fluctuation. If so, the claim of superiority is not tenable. The manufacturer needs to withdraw the claim forever or until such time that more evidence is available for scrutiny. The medical fraternity is expected to continue with their existing practice and not take cognizance of the claim until the claim is adequately substantiated.

If a null hypothesis, denoted by $H_0$, turns out to be false, what alternative is there? This is what is named the **alternative hypothesis**, denoted by $H_1$. This must be true when $H_0$ is found not to be so. For example: if the null ($H_0$) is that a drug is not effective, and this is shown not to be true, then the alternative ($H_1$) has to be that it is effective. In the preceding example, the claim is that of the superiority of the new drug. This is the alternative hypothesis. This is a one-sided alternative if inferiority is ruled out.

The following example may help one to further understand the steps and the philosophy behind the statistical tests. Suppose an argument erupts concerning the percentage of births in a community that are premature. Based on everyday experience, a practitioner asserts that exactly 10% of births are premature, neither less nor more. To test this assertion, a random sample of 60 births is systematically observed and 8 of them are found to be premature. This is 13.3%. Can we conclude that the percentage of premature births in the population is not 10%? Assertion of 10% premature births in this case is the null hypothesis. This is what is to be tested (and possibly to be refuted). Thus, $H_0$: $\pi = 0.10$. Since there is no assertion about a particular direction of difference, it could be negative or positive, and the alternative hypothesis is $H_1$: $\pi \neq 0.10$. Let the **level of significance** be fixed at $\alpha = 0.05$. That is, the chance of error of Type I should be less than 5%. Quantity of consequence in this case naturally is the proportion, $p$, actually observed in the sample. For this sample of 60 births, $p = 8/60 = 0.1333$. Since $n$ is 60 and $np \geq 8$, invoke the **central limit theorem** and assume that $np$ will approximately have a Gaussian distribution. Since $H_1$ is two sided, it is easily seen that $P = 0.39$. This is certainly very high in comparison with the conventional threshold 0.05. If $H_0$ is rejected, the chances of its being wrongly rejected are as much as 0.39 (or 39%). Since this is too high an error, we cannot reject the null.

The sample in this example can provide sufficient evidence to reject some other values of $\pi$. For the purpose of illustration, change $H_0$ to $\pi = 0.25$. Under this value of $H_0$, the usual **$z$-test** reveals $P = 0.038$. Since this is less than 0.05, this $H_0$ is not plausible and is rejected. It is concluded that the percentage of premature births is not 25. The percentage (13.33%) observed in the sample is sufficiently different from 25 in a statistical sense but not sufficiently different from 10. Thus, the $H_0$ that the premature births are 25% can be rejected but not the $H_0$ that they are 10%. The example illustrates the philosophy of statistical tests, explaining how sample values are considered consistent with some hypothesized values of the parameters and not consistent with other values.

The above example also illustrates one more dilemma that some of us face: if we decide to accept, which value should we accept? If we start accepting in place of not being able to reject, there would be multiple values for acceptance. In this example, we are not able to reject the null that the percentage is 10%. If the procedure is repeated for $H_0$: $\pi = 0.11$, this will also not be rejected, and we will not be able to reject $H_0$: $\pi = 0.12$ either. If we accept, which value do we accept? Thus, the only conclusion reached is that the null is not rejected. It only means that the value propounded under the null cannot be denied.

1. Martin MA. "It's like... you know": The use of analogies and heuristics in teaching introductory statistical methods. *J Stat Educ* 2003;11, http://www.amstat.org/publications/jse/v11n2/martin.html

## tests of significance, *see also* tests of hypothesis (philosophy of)

*Tests of significance* is the other name for **tests of hypotheses** and follows the same philosophy. We have discussed a large number of tests of significance in this book under different topics. Table T.5a

### TABLE T.5a
### Statistical Tests of Significance for Qualitative Data (Proportions)

| Parameter of Interest and Setup | Conditions | Main Criterion for Test of Significance |
|---|---|---|
| **Proportions in All of the Following Small-Sized Tables** | | |
| One dichotomous variable | Independent trials | |
| | Any $n$ | Binomial |
| | Large $n$ | Gaussian $Z$ |
| One polytomous variable | Independent trials | |
| | Large $n$ | Goodness-of-fit chi-square |
| | Small $n$ | Multinomial |
| Two dichotomous variables ($2 \times 2$) | Two independent samples | |
| | Large $n$ | Chi-square or Gaussian $Z$ |
| | Small $n$ | Fisher exact |
| | Detecting a medically important difference—large $n$ | Gaussian $Z$ |
| | Equivalence test | TOSTs |
| | Matched pairs | |
| | Large $n$ | McNemar |
| | Small $n$ | Binomial |
| | Crossover design | |
| | Large $n$ | Chi-square |
| | Small $n$ | Fisher exact |
| **Bigger Tables, No Matching** | | |
| Association | $2 \times C$ tables—large $n$ | Chi-square |
| Trend in proportions | $2 \times C$ tables—large $n$ | Chi-square for trend |
| Dichotomy in repeated measures | Many related $2 \times 2$ tables | Cochran $Q$ |
| Association | $R \times C$ tables | |
| Association | Three-way tables | Chi-square |
| | Test of full independence | Chi-square |
| | Test of other types of independence (log-linear models) | $G^2$ |
| $I \times I$ table | Matched pairs | McNemar–Bowker |
| Association in stratified tables | Stratified into many $2 \times 2$ tables | Mantel–Haenszel chi-square |

**TABLE T.5b**
**Statistical Tests of Significance for Relative Risk (RR) and Odds Ratio (OR)**

| Parameter of Interest and Setup | Conditions | Main Criterion for Test of Significance |
|---|---|---|
| Relative risk, odds ratio, and attributable risks | Large $n$ required | |
| ln(RR) or ln(OR) | One group | Wald |
| ln(RR) or ln)OR) | Two independent samples | Gaussian $Z$ or chi-square |
| RR or OR | Matched pairs | Gaussian $Z$ or McNemar |
| | Stratified | Mantel–Haenszel chi-square |
| AR | Two independent samples | Chi-square or Gaussian $Z$ |
| | Matched pairs | McNemar |
| Homogeneity of RRs or ORs across strata | Large sample | Breslow–Day |
| | Small sample | Zelen test |

summarizes when to use a particular test for qualitative data (proportions), Table T.5b summarizes when to use a particular test for odds ratio and relative risk, and Table T.5c summarizes when to use a particular test for quantitative data including (survival) durations and distributions. This is still not an exhaustive list but covers those that are relatively more commonly used. Consider this list as just a guideline as the conditions are given in brief—for details, see the topic on the concerned test. Choose the test from this list and go to the topic that describes that test to find the actual conditions where the test is appropriate.

## test–retest reliability

**Reliability** is the ability of a tool to perform and maintain functions in routine circumstances, as well as under unexpected circumstances that do not vary much from the routine. In a slightly different form, reliability is **repeatability** that is concerned with stability of the responses with repeated use of the instrument. The measurement of repeatability essentially involves administering the test twice or more to the same subjects under the same condition within a short period. This is therefore known as test–retest reliability. It can be validly calculated for an instrument that does not provide any learning to the respondent, and the second-time responses are not affected by the first-time responses. The scores or the responses obtained on the two occasions are checked for agreement. These can be checked itemwise, but generally the total score is compared. The extent of agreement is measured by **intraclass correlation** (ICC) if the responses are quantitative and by **Cohen kappa** if they are qualitative. If the value of these measures is more than 0.7, the test is considered reliable, and a value less than 0.5 is considered unacceptable.

Suppose one wants to find out whether laboratory A is more reliable than laboratory B in measuring plasma glucose levels. One strategy for this could be to split each blood sample into six parts. Send three parts to laboratory A and three to laboratory B. Do this for, say, 25 subjects and keep the laboratories blind regarding sending aliquots of the same sample. If the laboratory is reliable, the measurement on the different parts of the same sample will give nearly the same values, yielding a high value of the ICC coefficient. The laboratory with significantly higher ICC would be more reliable.

**TABLE T.5c**
**Statistical Tests of Significance for Quantitative Data (Means, Variances, Correlations, Survival)**

| Parameter of Interest and Setup | Conditions | Main Criterion for Test of Significance |
|---|---|---|
| | **Mean or Central Value** | |
| One group | Comparison with prespecified—Gaussian | |
| | σ known | Gaussian $Z$ |
| | σ not known | Student $t$ |
| Comparison of two independent groups | Paired—Gaussian (σ not known) | Student $t$ |
| | Paired—non-Gaussian | |
| | Any $n$ | Sign test |
| | $5 \le n \le 19$ | Wilcoxon signed-ranks $W_S$ |
| | $20 \le n \le 29$ | Standardized $W_S$ referred to Gaussian Z |
| | $n \ge 30$ | Student $t$ |
| | Unpaired—Gaussian | |
| | Equal variances | Student $t$ |
| | Unequal variances | Welch |
| | Unpaired—non-Gaussian | |
| | $n_1, n_2$ between (4, 9) | Wilcoxon rank-sum $W_R$ |
| | $n_1, n_2$ between (10, 29) | Standardized $W_R$ referred to Gaussian Z |
| | $n_1, n_2 \ge 30$ | Student $t$ |
| | Crossover design—Gaussian | Student $t$ |
| | Detecting medically important difference | Student $t$ |
| | Equivalence tests | Student $t$ (TOSTs) |
| Comparison of three or more independent groups, equal variances | One-way, two-way, or multi-way layout—Gaussian | ANOVA $F$ |
| | One-way nonparametric | |
| | $n \le 5$ | Kruskal–Wallis $H$ |
| | $n \ge 6$ | $H$ referred to chi-square |
| | Two-way layout—Gaussian | ANOVA $F$ |
| | Two-way nonparametric (one observation per cell) | |
| | $J \le 13$ and $K = 3$ | Friedman $S$ |
| | $J \le 8$ and $K = 4$ | Friedman $S$ |
| | $J \le 5$ and $K = 5$ | Friedman $S$ |
| | Larger $J$, $K$ | $S$ referred to chi-square |
| | Multiple comparisons—Gaussian | |
| | All pairwise | Tukey $D$ |
| | With control group | Dunnett |
| | Few comparisons | Bonferroni |

(*Continued*)

**TABLE T.5c (CONTINUED)**
**Statistical Tests of Significance for Quantitative Data (Means, Variances, Correlations, Survival)**

| Parameter of Interest and Setup | Conditions | Main Criterion for Test of Significance |
|---|---|---|
| Repeated measures | Gaussian | ANOVA $F$ with Huynh–Feldt correction for sphericity |
| Comparison of three or more groups, unequal variances | Gaussian | |
| | Large samples | Welch |
| | Small samples | Brown–Forsythe |
| | **Variance** | |
| One group | Comparison with prespecified—Gaussian | Variance ratio $F$ |
| Two independent groups | Gaussian | Variance ratio $F$ |
| | Non-Gaussian (Mild) | Levene |
| More than two independent groups | Gaussian | Bartlett |
| | Non-Gaussian (Mild) | Levene |
| Homogeneity of covariance matrices | Gaussian | Box $M$ |
| Outliers | Gaussian | Grubbs |
| | **Correlation** | |
| One sample | Gaussian | $z$-test after Fisher $z$ transformation |
| Comparison of two independent groups | Gaussian | $z$-test after Fisher $z$ transformation |
| In repeated measures | Gaussian | Mauchly |
| Autocorrelation | Gaussian | Durbin–Watson |
| | **Survival Curve** | |
| Comparison of two independent groups | Nonparametric—large samples | |
| | Same weight to all time points | Log-rank |
| | Weight proportional to $n_t$ | Breslow |
| | Weight proportional to $\sqrt{n_t}$ | Tarone–Ware |
| | **Distribution** | |
| One sample | Nonparametric—large sample | Kolmogorov–Smirnov |
| | Gaussian or non-Gaussian—small sample | Anderson–Darling |
| | Gaussian—moderate sample | Shapiro–Wilk |
| Two samples | Nonparametric—large samples | Kolmogorov–Smirnov |

Note that the reliability may be higher even when there is a constant bias in the measurement since correlation coefficient, including the ICC, ignores constant bias. It is possible that a laboratory has an ICC of 0.90 but consistently reports values lower or higher than the actual value. There must be an external check for this kind of bias.

Russo et al. [1] tested a 36-item health survey form (SF-36) on 36 outpatients with chronic schizophrenia (incidentally, in this example, the sample size $n = 36$ and the number of items = 36 are equal). The form was completed in writing as well as orally by each patient. The test–retest reliability showed that SF-36 is an appropriate outcome measure for schizophrenic outpatients. Mathewson et al. [2] investigated the short-term test–retest reliability of resting regional power and asymmetry in a sample of 38 active older adults in Canada and observed ICC in excess of 0.90 for regional power at different sites (closed eyes), indicating that individual differences before and after the task were clearly preserved.


1. Russo J, Trujillo CA, Wingerson D, Decker K, Ries R, Wetzler H, Roy-Byrne P. The MOS 36-item Short Form health survey: Reliability, validity and preliminary findings in schizophrenic outpatients. *Med Care* 1998;36:752–6. http://www.ncbi.nlm.nih.gov/pubmed/9596066
2. Mathewson KJ, Hashemi A, Sheng B, Sekuler AB, Bennett PJ, Schmidt LA. Regional electroencephalogram (EEG) alpha power and asymmetry in older adults: A study of short-term test-retest reliability. *Front Aging Neurosci* 2015 Sep 16;7:177. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4584992/


## test statistic, *see* test criterion

## tetrachoric correlation

This is the correlation between two quantitative variables each with a **Gaussian** (normal) **distribution** but observed in dichotomous categories. Suppose you are dealing with HDL cholesterol (HDL-C) and LDL cholesterol (LDL-C) levels in healthy subjects—both of which have Gaussian distribution but the observations are made only regarding how many subjects have an HDL-C level of 60 mg/dL or less and how many have an LDL-C level of 100 mg/dL or more. The correlation coefficient between such dichotomous categories would be obtained by tetrachoric correlation. Tetra is derived from Greek for "four," which comes from a $2 \times 2$ classification. If the number of categories is 3 or more, this is called a *polychoric correlation*.

Although other measures of association such as **odds ratio** and **contingency coefficient** are available for double dichotomous data, those do not use Gaussian pattern of values even if known. Also, the term *correlation* is more familiar and seems to have a wider appeal. Tetrachoric correlation ranges from −1 to +1 just as the ordinary correlation coefficient does, but different cutoffs of the underlying values can provide different values of the tetrachoric correlation. Software packages easily calculate tetrachoric correlation by considering various combinations that yield a **2 × 2 table** closely corresponding to the one actually observed.

To understand this correlation, consider that the distributions of the two random variables are put at right angles to each other (Figure T.1),

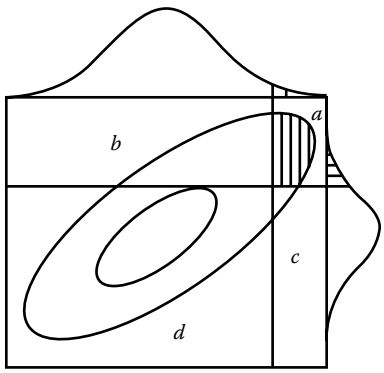



**FIGURE T.1** A graphic of a correlation ellipse divided into four regions based on $x$ and $y$ cutpoints on two Gaussian distributions.

as suggested by R Graphic Manual [1]. The tails of each distribution are marked out using two cutpoints ($x_0$, $y_0$), say, by two raters: $x_0$ on the $x$ axis and $y_0$ on the $y$ axis. Draw two ellipses, one inside another. Within the frame, the shaded region has $x > x_0$ and $y > y_0$. To the left of this, $x < x_0$ and $y > y_0$; underneath that, $x_0 < x$ and $y_0 < y$; and at the bottom right-hand corner, $x > x_0$ and $y < y_0$. The proportions $a$, $b$, $d$, and $c$, respectively, denote the proportion of cases that fall in these regions defined by the two thresholds. For example, $d$ is the proportion below both raters' thresholds (bottom left-hand corner) and is therefore considered "negative" by both. A software package is used to find the values of $a$, $b$, $c$, and $d$ for the given data and to calculate the value of the tetrachoric correlation. An approximation is as follows:

$$\text{Tetrachoric correlation: } r = \frac{\theta - 1}{\theta + 1},$$

where $\theta = \left(\frac{ad}{bc}\right)^{\pi/4}$ ($\pi = 3.14159$) if $a, b, c, d > 0$. If $a = 0$ or $d = 0$, then $r = -1$, and if $b = 0$ or $c = 0$, then $r = +1$.

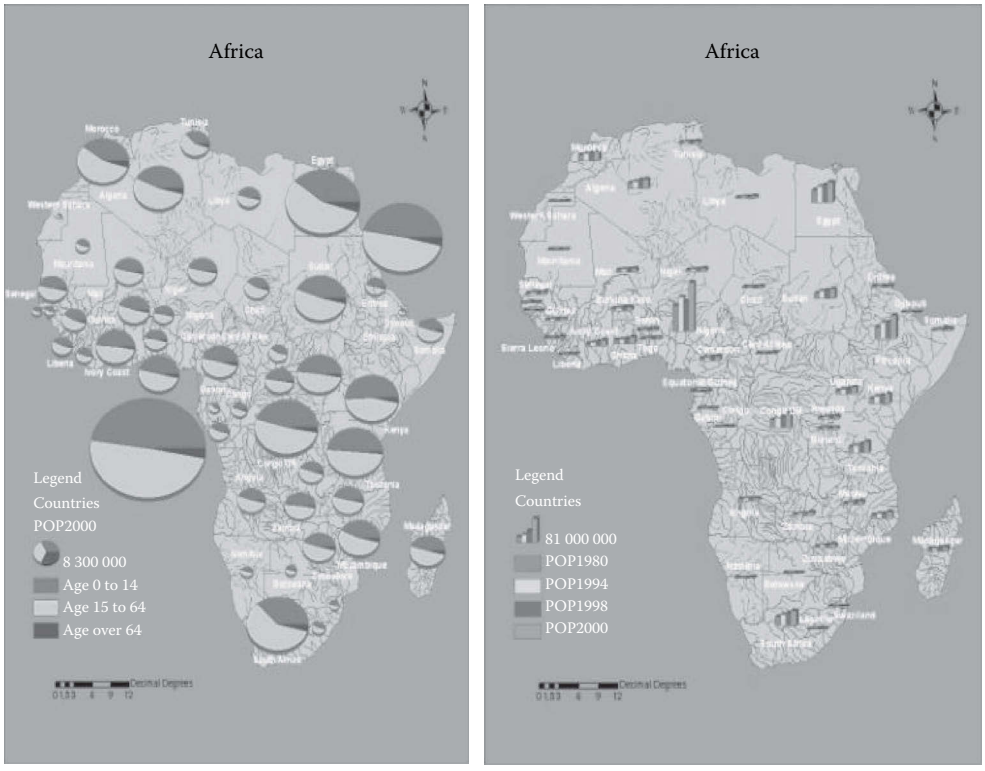
Brunault et al. [2] used tetrachoric correlation for testing the factor structure in a factor analysis of dichotomized food addiction scale values and eating behavior scale values. Even though this kind of correlation is meant to be used when both the variables are intrinsically quantitative but dichotomized, there are instances when this condition was not strictly followed. For example, Behrsin et al. [3] used this for correlation between the diagnostic accuracy of adenosine deaminase (P-ADA) with histopathology in pleural tuberculosis cases in Brazil where P-ADA was divided with an optimum cutoff value of 40 IU/l as determined by the ROC but histopathological findings were not quantitative. The tetrachoric correlation coefficient was 0.563 but was termed high correlation.

See Uebersax [4] for a good exposition of the subject.


1. R Graphical Manual. Draw a correlation ellipse and two normal curves to demonstrate tetrachoric correlation. http://rgm3.lab.nig.ac.jp/RGM/R_rdfile?f=psych/man/draw.tetra.Rd&d=R_CC
2. Brunault P, Ballon N, Gaillard P, Réveillère C, Courtois R. Validation of the French version of the Yale food addiction scale: An examination of its factor structure, reliability, and construct validity in a nonclinical sample. *Can J Psychiatry* 2014 May;59(5):276–84. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4079141/
3. Behrsin RF, Junior CT, Cardoso GP, Barillo JL, de Souza JB, de Araújo EG. Combined evaluation of adenosine deaminase level and histopathological findings from pleural biopsy with Cope's needle for the diagnosis of tuberculous pleurisy. *Int J Clin Exp Pathol* 2015 Jun 1;8(6):7239–46. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4525955/
4. Uebersax JS. *Introduction to the Tetrachoric and Polychoric Correlation Coefficients*. 8 Sep 2015. http://john-uebersax.com/stat/tetra.htm


## thematic map, *see also* choroplethic map

This is a type of statistical mapping for values of an indicator in different areas—thus, it displays spatial pattern of values in contrast to the usual maps that show locations, boundaries, land features, and so on. Using different shades or colors is one type of display that makes it what is called a **choroplethic map**, but other methods such as dots and proportional symbols can also be used in a thematic map. Generally, only one indicator (theme), such as density of doctors, is shown; for two or more indicators, a combined index




(a) (b)

**FIGURE T.2** Thematic maps of African countries with regard to the age structure (a) by pie and (b) by bar diagram. (From Andrea P. Building geodatabase 6, Thematic maps, simple analyses methods. *Digitalis Tankonyvtar.* http://www.tankonyvtar.hu/en/tartalom/tamop425/0027_BGD6/ch01s02.html.)

T

may have to be devised such as the **human development index** for the combination of education, income, and health. Nonetheless, it is possible to produce maps to show multivariate data by imposing a pie diagram, a bar diagram, and so on, on the areas, although that may defy the defining principle of keeping a map simple for the user to be able to visualize the theme. Marks and symbols should be easy to understand. Figure T.2 shows a thematic map of age structure (0–14, 15–64, and 65+ years) in African countries where it is shown by pie diagrams on the left side and by bar diagrams on the right side. The size of the pies and bars is according to the total population. Thus, they depict not only the percentage of population in the different age-groups but also the differences in the total population among the countries [1]. See this reference for the method to draw such maps.

Quite often, the categories in thematic maps are arbitrarily chosen of equal width and the number of categories also remains arbitrary. For example, for excess relative risk of heart diseases, Haining et al. [2] used categories 0.3–1.0, 1.0–1.5, and 1.5–2.0. This introduces considerable subjectivity in the cognition and perception obtained from the maps. A discussion on this aspect is given by Indrayan and Kumar [3] who advocate that the categories should be natural, as dictated by the data, in place of arbitrary choices. These natural categories can be identified by consensus in the groups obtained by various methods of clustering and the picture thus obtained can be substantially different from the one based on equal-width categories.

Thematic maps can be used for exploratory purposes, spatial data analysis such as identifying patterns, generating or even confirming a hypothesis, and of course data presentation.


1. Andrea P. Building geodatabase 6, Thematic maps, simple analyses methods. *Digitalis Tankonyvtar.* http://www.tankonyvtar.hu/en/tartalom/tamop425/0027_BGD6/ch01s02.html
2. Haining R, Law J, Maheswaran R, Pearson T, Brindley P. Bayesian modelling of environmental risk: Example using a small area ecological study of coronary heart disease mortality in relation to modelled outdoor nitrogen oxide levels. Department of Geography, University of Cambridge. http://www.geog.cam.ac.uk/research/projects/camgis/bayesenvironment/, last accessed December 12, 2015.
3. Indrayan A, Kumar R. Statistical choropleth cartography in epidemiology. *Int J Epidemiol* 1996;25181–9. http://ije.oxfordjournals.org/content/25/1/181.short

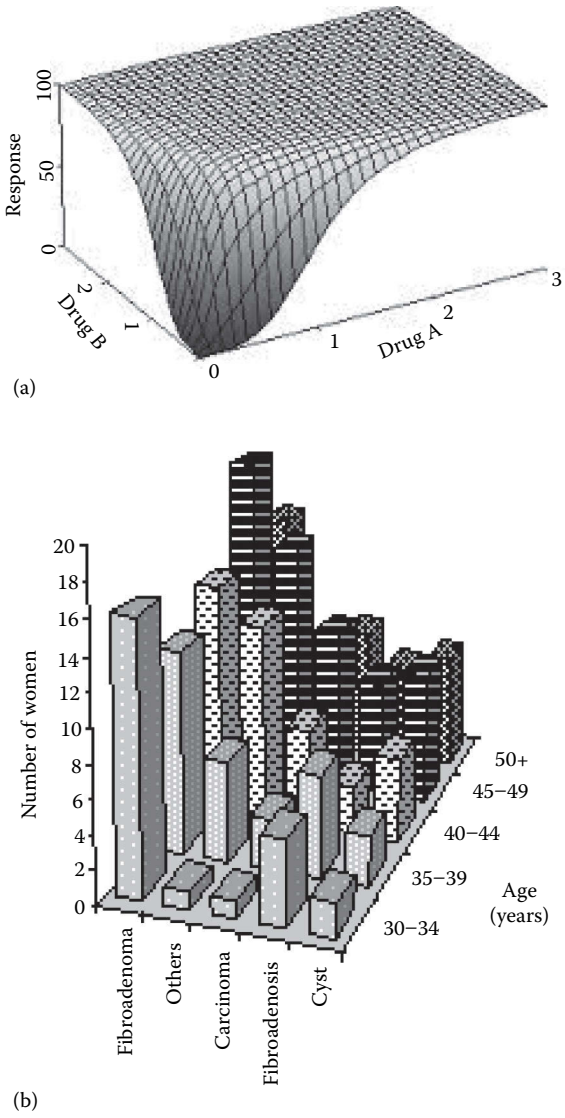




**FIGURE T.3** (a) Response surface for interaction of two drugs. (From Minto C. *Response Surface Modelling of Drug Interactions.* http://www.eurosiva.org/Archive/Vienna/abstracts/speakers/minto.htm.) (b) Distribution of palpable breast cases in different age-groups by histopathological diagnosis.

## therapeutic trials, *see* clinical trials

## 30 × 7 sampling, *see* cluster sampling

## three-dimensional diagrams, *see also* response surface

These are the diagrams that show height also in addition to length and breath. Thus, a three dimensional (3-D) diagram can depict three variables simultaneously.

Many computer packages readily draw 3-D diagrams. The additional variable is shown on a third axis (hence, here we have *x*, *y*, and *z*). A 3-D representation of a given relationship is a **response surface** of the type shown in Figure T.3a. This shows how the response flattens when two drugs have a high dose [1].

Another 3-D diagram is in Figure T.3b, which shows the distribution of palpable breast cases in different age-groups by their histopathological diagnosis. This is a bar diagram in 3-D with three axes. You can also have a 3-D diagram with just two axes, where, for example, bars are shown in 3-D but there is no third variable.


1. Minto C. *Response Surface Modelling of Drug Interactions.* http://www.eurosiva.org/Archive/Vienna/abstracts/speakers/minto.htm


## three-way tables, *see* chi-square test for three-way contingency tables

## three-sigma limits, *see also* six-sigma methodology

Among several properties of a **Gaussian** (Normal) **distribution**, one is that the limits from (mean – 3σ) to (mean + 3σ) cover the measurements of exactly 99.7% of subjects, where σ is for the standard

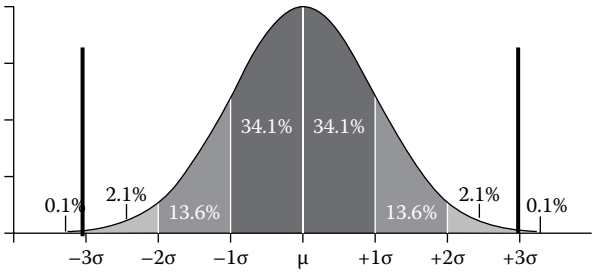



**FIGURE T.4** Three-sigma limits in a Gaussian distribution.

deviation (SD) (Figure T.4). These are referred to as ±3SD limits or sometimes as three-sigma limits.

Three-sigma limits are primarily used in **quality control** charts to set the upper and lower control limits. If the measurement of a product in a line reaches beyond these limits, the product is considered defective as its chance of being "within control" is less than 1%. However, this works when the inherent variation, measured by the standard deviation (σ), is small. Another use of these limits is in ***Z*-scores** of growth parameters of children such as

$$Z\text{-score} = (\text{Weight} - \text{Mean})/\text{SD}$$

for weight. Since Z-score already has SD in the denominator, the limits for this are from −3 to +3. Any weight with Z-score beyond these limits is considered unusually high or unusually low.

## ties in values, *see* ranks

## time-dependent covariate, *see* Cox regression

## time series, *see also* periodogram, autoregressive moving average (ARMA) models

Recording the values of a variable at regular intervals over a long period gives rise to what is known as a time series. The observed movement and fluctuations within such a time series are generally made up of four components: (i) "secular trend" (the underlying smooth movement of a time series), (ii) seasonal variation, (iii) cyclical variation, and (iv) irregular variation called random. For example, the incidence of an infectious disease recorded monthly over several years in a population is a time series, where all these components may operate. In some time series, such as on incidence of some noncommunicable diseases, one or more of these components may be absent.

The primary purpose of studying time series is to explore and understand the trend over time and sometimes to forecast values for the near future. Time series analysis comprises decomposing the series into seasonal, cyclical, and secular trend—the remaining being random variation as just mentioned. Secular trend provides the long-term trend such as declining incidence of infectious diseases over the past 50 years. This trend is generally assumed linear, which implies a constant rate of either growth in case of an increase or decay in case of a decrease. The systematic variation could be cyclic, such as a 5-year cycle for dengue in India similar to the one depicted in Figure T.5. Parainfluenza is seen to have a 2-year cycle [1] whereas dengue hemorrhagic fever is seen to have a 5-year cycle in Latin America [2]. Seasonal trend is mostly within a year, such as that of swine flu occurring at the start of winter and that of dengue with a peak around the month of October in India every year, as is clear from Figure T.5. Such cyclic trend is visibly seen each day as circadian rhythm in blood pressure, heart rate, and so on, or when a drug is taken such as three times a day. It is necessary that the hourly/daily/weekly/monthly values are available to study these components, and any time series has to be necessarily long for any adequate analysis, that is, for, say, at least values for 50 time points.

Time series models are very useful for short-range forecasting problems. They assume that whatever forces have influenced the variables in question in the recent past will continue to operate in the near future. Because of the complexity owing to seasonal, cyclical, and secular trend, forecasting in a time series setup has to follow certain steps. As a first step, the series is smoothened so that the random variation is removed as much as possible. This involves some form of local averaging, such as **moving average**, which tends to cancel out the negative and positive irregular (random) component. If real outliers are present, perhaps the local median of a few successive values can be tried in place of the arithmetic mean generally used in moving "average." The appropriate lag for this is identified by **periodogram** analysis. Other smoothing methods such as **cubic splines** can also be used. Being local, these methods tend to highlight seasonal pattern but not cyclical pattern. The cyclical pattern is explored mostly with trigonometric functions such as sine waves. For secular trend, regular regression methods can be tried—a simple linear method can be used if the trend is anticipated to follow a linear trend, whereas a curvilinear method can be used if the long-term trend seems to follow a curve.



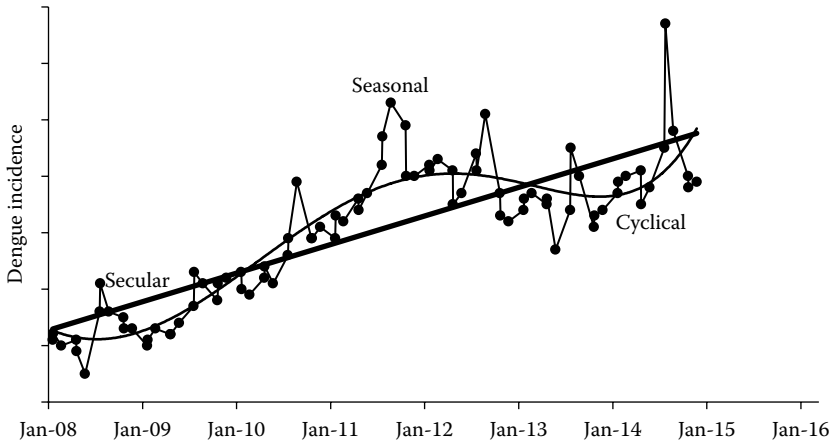



**FIGURE T.5** Seasonal, cyclic, and secular trend in a time series.

In analyzing time series data, it is necessary to take account of any serial correlation (also known as **autocorrelation**), since the values depend serially on the value on previous occasions. Such serial correlation can occur because the time effect by itself may not be able to take care of other factors that may be simultaneously changing, such as population size.

Further analysis can be carried out using the **autoregressive moving averages (ARMA)** method. Other methods can also be used. For example, Been et al. [3] studied monthly incidence of asthma cases and respiratory tract infections in four UK counties over the period 1997 to 2013. They used generalized additive mixed models since their objective was to see the effect of smoke-free legislation introduced at that time. However, this also accounted for autocorrelation between data points. Ngo et al. [4] used spectral analysis to study electroencephalogram time series in the United States.

For more information, see Montgomery and Jenning [5].


1. Came PE, Caliguiri LA (Eds.). *Chemotherapy of Viral Infections*. Springer, 2013:p. 41.
2. Torres JR, Casto J. The health and economic impact of dengue in Latin America. *Cademos de Saude Publica* 2007;23 (Suppl 1):S23–31. http://www.scielosp.org/pdf/csp/v23s1/04.pdf
3. Been JV, Mackay DF, Millett C, Pell JP, van Schayck OCP, Sheikh A. Impact of smoke-free legislation on perinatal and infant mortality: A national quasi-experimental study. *Scientific Rep* 2015;5:13020. http://www.nature.com/articles/srep13020
4. Ngo D, Sun Y, Genton MG, Wu J, Srinivasan R, Cramer SC, Ombao H. An exploratory data analysis of electroencephalograms using the functional boxplots approach. *Frontiers in Neuroscience*. 2015;9:282. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4541028/
5. Montgomery DC, Jenning CL. *Introduction to Time Series Analysis and Forecasting*. Wiley, 2015.


## titration study

A titration study is where a patient is sequentially administered graded doses of a drug with predetermined rules to determine at what dose the response is optimal. Optimality could be defined in terms of maximum response without substantial side effects or in any other appropriate manner. Typically, the study starts with a low dose that is gradually increased until the dose can no longer be tolerated or until the desired response is obtained. This kind of study can obviously be done only when the drug has no adverse response, where "response" is usually some objective measurement: for example, the reduction of blood pressure below a certain level. Thus, a titration study is actually a dose-finding study, mostly oriented to an individual.

The work by Blonde et al. [1] provides an example of titration in practice, although this was done on a large group of patients. The aim of their study was to compare the efficacy and safety of two fasting plasma glucose (FPG) titration targets using a patient-directed algorithm for once-daily basal insulin in insulin-naïve subjects with type 2 diabetes suboptimally treated with an oral antidiabetes drug. In this 20-week randomized controlled study, 244 insulin-naïve subjects with type 2 diabetes were equally randomized to one of two treatment arms using 80–110 or 90–110 mg/dL FPG as titration targets. Once-daily insulin detemir (a long-acting human insulin analog for maintaining the basal level of insulin) doses were adjusted using algorithm-guided, patient-directed titration to achieve the target FPG values. Another such example is a continuous positive airway pressure titration study where the objective is to find the right amount of air pressure with sleep disorder [2].

Chuang [3] used the life table method and a dose–response model to illustrate the analysis of a titration study that investigated the efficacy of an anti-hypertensive compound. Care is required since treatment effect may be confounded with time, and the titration process might be correlated with safety issues as mentioned earlier. For further details of titration studies, see Chow [4].


1. Blonde L, Merilainen M, Karwe V, Raskin P, TITRATE Study Group. Patient-directed titration for achieving glycaemic goals using a once-daily basal insulin analogue: An assessment of two different fasting plasma glucose targets—The TITRATE study. *Diabetes Obes Metab* 2009 Jun;11(6):623–31. http://www.ncbi.nlm.nih.gov/pubmed/19515182
2. Law M, Naughton M, Ho S, Roebuck T, Dabscheck E. Depression may reduce adherence during CPAP titration trial. *J Clinical Sleep Med* 2014;10(2):163–9. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3899318/
3. Chuang C. The analysis of a titration study. *Stat Med* 1987;6(5):583–90. http://www.ncbi.nlm.nih.gov/pubmed/3659668
4. Chow S-C. *Design and Analysis of Clinical Trials: Concepts and Methodologies*, Third Edition. Wiley, 2013.


## *T*-max, *see* pharmacokinetic parameters ($C_{max}$, $T_{max}$) and pharmacokinetic studies

## tolerance interval

A tolerance interval covers a stated proportion of a population distribution with a given **confidence level**. It is used to determine whether or not a process is capable in the sense that the process does not exceed a certain failure rate as determined using present specifications. The basic difference with confidence interval (CI) is that CI is for a population parameter whereas tolerance interval is for a specified proportion of the population. For example, the interest may be in tolerance interval within which at least 90% ($\pi = 0.90$) of blood samples have values with probability $(1 - \alpha) = 0.95$ after a treatment. A CI, such as $\bar{x} \pm 1.96s/\sqrt{n}$ for mean, does not consider the sampling fluctuation in the estimate $\bar{x}$ and $s$, whereas tolerance interval considers this fluctuation. Consequently, it becomes mathematically complex. The CI can be greatly reduced by increasing the sample size, but not the tolerance interval because it also considers variance in addition to sampling fluctuations in the estimates. A large tolerance interval implies too much variation in the values.

Figure T.6 presents the concept of a tolerance interval graphically. The solid line in this figure represents the actual population distribution. The dotted line distributions result from the uncertainty of knowing the true location of the population mean. The difference in location of the mean as a result of this uncertainty is defined by the CI. Because of this, the tolerance interval extends beyond the tail areas of the actual population distribution. For tolerance interval, it is necessary to consider the uncertainty of knowing just where the mean of the population distribution lies.

A two-sided tolerance interval can be calculated as follows:

$$\bar{x} \pm k_2 s,$$

where $\bar{x}$ is the sample mean, $s$ is the sample standard deviation (SD), and $k_2$ is a factor for a two-sided tolerance interval defining the number of sample SDs required to cover the desired proportion of the population. This is also called the tolerance factor and is different if 50% of the population values are to be covered than, say, 75% of the values of the population.

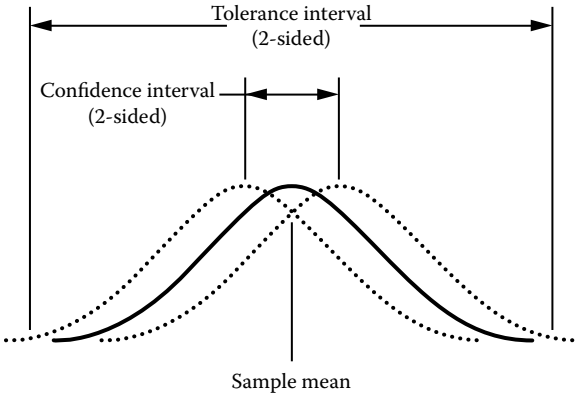



**FIGURE T.6** Graphical depiction of a tolerance interval. (From Propharma Group. *Understanding Statistical Intervals Part III. News and Blog*, 2015. http://www.propharmagroup.com/blog/understanding-statistical-intervals-part-iii-tolerance-intervals.)

Exact values of $k_2$ are tabularized in ISO 16269-6 [2]. These values of $k_2$ were calculated iteratively using a numerical integration process. However, in practice, a reasonable approximation of $k_2$ can be obtained using the formula below:

$$k_2 = \sqrt{\frac{(n-1)\left(1+\frac{1}{n}\right)z^2_{(1-\pi)/2}}{\chi^2_{1-\alpha,n-1}}},$$

where

$n$ is the sample size

$z_{(1-\pi)/2}$ is the standard Gaussian (Normal) variate corresponding to $(1 - \pi)/2$, where $\pi$ is the proportion of the population to be covered

$\chi^2_{(1-\alpha),n-1}$ is the critical value of the chi-square distribution with $n - 1$ degrees of freedom surpassed with probability $(1 - \alpha)$, the statistical confidence

The formula and tables for one-sided tolerance intervals are provided by Natrella [3].

Consider an example of prostatic specific antigen (PSA) measured for a random sample of $n = 40$ healthy males with mean = 3.5 ng/mL and SD = 0.7 ng/mL. What are the tolerance limits for $\pi = 90\%$ of the subjects with confidence $(1 - \alpha) = 0.95$? For $(1 - \alpha) = 0.95$ and df = 40 − 1 = 39, the critical value of chi-square $\chi^2_{0.95,39} = 54.572$, and $z_{(1-\pi)/2} = z_{0.05} = -1.645$. These give $k_2 = \sqrt{\frac{(40-1)\left(1+\frac{1}{40}\right)(-1.645)^2}{54.572}} = \sqrt{\frac{39.975 \times 2.7060}{54.572}} = \sqrt{1.9822} = 1.41$.

Thus, the 95% tolerance range for 90% population is $3.5 \pm 1.41 \times 0.7$, or 2.11 to 4.89. This means that in this population, there is a 95% chance that 90% of the subjects have PSA between 2.11 and 4.89 ng/mL. Note that this is much larger than the 90% CI.

The validity of a tolerance interval is highly dependent on the distribution of the underlying data. In practice, data are assumed to be distributed in a **Gaussian** manner; this assumption can be verified by using the **Anderson–Darling test**.

There is another statistical use of the term *tolerance*. This is to measure the effect of **multicollinearity** on the variance of the model's parameter estimates. A formal measure of collinearity between covariate $x_k$ and the other covariates is given by the proportion of the variance in $x_k$ explained by the other covariates when $x_k$ is regressed on the other covariates. In fact, this is the square of the **multiple correlation coefficient**, denoted here by $R_k$. It can be shown that the variance of the regression coefficient $b_k$ is proportional to $1/\left(1-R_k^2\right)$; thus, large values of $R_k^2$ are associated with imprecise estimates of $\beta_k$. The **variance inflation factor (VIF)** for a covariate $x_k$ is defined to be $1/\left(1-R_k^2\right)$, and the tolerance is the reciprocal of this: $1 - R_k^2$. A small value of tolerance, for instance, less than 0.10, indicates that collinearity is a problem and needs to be tackled.

1. Propharma Group. *Understanding Statistical Intervals Part III. News and Blog*, 2015. http://www.propharmagroup.com/blog/understanding-statistical-intervals-part-iii-tolerance-intervals
2. ISO 16269-6:2014. Statistical interpretation of data—Part 6: Determination of statistical tolerance intervals. http://www.iso.org/iso/catalogue_detail.htm?csnumber=57191
3. Natrella MG. *Experimental Statistics*, NBS Handbook 91. US Department of Commerce, 1963. http://water.usgs.gov/osw/bulletin17b/1963_Natrella.pdf

## total fertility rate, *see* fertility indicators

## tracking

Tracking is a term often used for follow-up observations from earlier values as sometimes used for examining **longitudinal data**. At face value, tracking would imply that the subjects with the largest values of the response variable at the start of the study will tend to continue to have relatively larger values throughout the study. This is measured by a quantity called *tracking coefficient* that could be simply correlation coefficient or any other such measure. If a correlation coefficient is used for tracking, it is highly unlikely to be negative in this setup, and the effective range is 0 to 1. If the effect of some factors is to be removed, the adjusted correlation coefficient can be computed. A value of at least 0.6 is considered a good correlation for tracking. Note that a tracking coefficient is just a pure measure of longitudinal changeability and does not tell the direction of individual change.

An example illustrating this is given by Kagura et al. [1]. They studied hypertension in South Africa to examine if hypertension in adulthood can be traced back to childhood. Since there is a scarcity of longitudinal data on pediatric blood pressure (BP) in African populations, the authors set out to assess the prevalence of hypertension in children and to evaluate BP tracking between childhood and late adolescence through a cohort of South African black children. A Birth to Twenty cohort of children born in Soweto, Johannesburg, in 1990 was available ($n = 3273$, 78.5% black). Data on BP were collected at six follow-up periods between ages 5 and 18 years (14 years). Pearson correlation coefficients and relative risk (RR) were used to describe tracking of BP between childhood and late adolescence. Tracking coefficients ranged from 0.20 to 0.57 for systolic BP and 0.17 to 0.51 for diastolic BP in both sexes over the period of 14 years. RR of having elevated BP ranged from 1.60 at 5 years to 2.71 at 18 years of age. The authors concluded that this study reported high prevalence of elevated BP, which tracks from early childhood into late adolescence. They emphasized the importance of early identification of children at risk of developing elevated BP and related risk factors plus timely intervention to prevent hypertension in adulthood.

Ulmer et al. [2] have explained tracking well in the context of cardiovascular risk factors.


1. Kagura J, Adair LS, Musa MG, Pettifor JM, Norris AN. Blood pressure tracking in urban black South African children: Birth to twenty cohort. *BMC Pediatrics* 2015;15:78. http://www.biomedcentral.com/1471-2431/15/78
2. Ulmer H, Kelleher C, Diem G, Concin H. Long-term tracking of cardiovascular risk factors among men and women in a large population-based health system. *Eur Heart J* 2013;24(11):1004–13. http://eurheartj.oxfordjournals.org/content/24/11/1004


## transformations

Four types of transformations are popular in statistics: (i) linearizing, (ii) normalizing, (iii) variance stabilizing, and (iv) distributional Gaussianizing. For the last, see **Box-Cox power transformation**. This includes **logarithmic transformation** for highly skewed data. The others can be explained as follows.

### *Linearizing Transformation*

Linear relation means that the rate of change in one variable with respect to another is uniform, and it can be depicted by a straight line. Linear relations are easy to understand and easy to model and use. In mathematical form, this is expressed as $y = a + bx$. Linearizing transformations are those that convert a nonlinear or curvilinear relation to linear form—thus making it easy to interpret. The most simple of linearizing transformations is the logarithmic, which is used for multiplicative values such as titers and radiation doses. This arises from the fact that $\log(m*p) = \log m + \log p$. This kind of transformation works very well for the exponential relationship $z = ce^{bx}$ since then $\ln z = \ln c + bx$, which is the same as $y = a + bx$ with $y = \ln z$ and $a = \ln c$. In many setups, the linearizing transformation is obtained by running the **regression**. Suppose the regression of total glomerular filtration rate (GFR) on plasma creatinine level (in milligrams per deciliter) in patients with chronic renal failure is GFR = 2.7 + 150(1/creatinine). This implies that the reciprocal of creatinine linearizes the relationship. After this transformation, the relationship can be represented by a straight line. Without the transformation, it is a curve.

### *Normalizing Transformation*

Normalizing is transforming the data within the range [0, 1], keeping the order preserved. This is achieved by using the transformation

$$(x - \text{Min})/(\text{Max} - \text{Min}).$$

This is different from standardization: $(x - \text{Mean})/\text{SD}$ that leads to ***Z*-scores**. For further details, see the topic **normalization**.

Recognizing the association of *Z*-scores with the so-called normal distribution, some believe that the *Z*-score transformation is "normalizing" and that this can make their data normally distributed. This actually is not so as these are linear transformations that do not alter the basic shape of the distribution.

### *Variance Stabilizing Transformations*

These transformations are used to achieve **homoscedasticity** so that the variances across groups are nearly equal as is required for some statistical methods such as analysis of variance and regression. These transformations are separately discussed under the topic **variance stabilizing transformations**. One of these, the logarithmic, can be explained as follows.

Sometimes, the residuals from regressing the data are large for larger values of the dependent variable. Such a trend often occurs when the change in the outcome variable is often not an absolute value but a percentage of the value. For the same percent error, a larger value of the variable means a larger absolute error, leading to the residuals also becoming larger. In this case, taking logarithm helps in obtaining uniform variance.

## trend

Trend is the smoothed pattern of values over a period, or over values of the other variables. For example, the trend of systolic level with age is upward linear for males and a slight curvilinear for females with quick upward move after menopause. A trend is called linear when the change in the value of one variable ($y$) is the same per unit change in the value of another variable ($x$), at least on average, and is called curvilinear when the rate of change in $y$ changes with $x$, such as fast rise for initial values of $x$, slow for middling values of $x$, and fast decline for large values $x$ (this is the trend lung functions follow as age increases). For details, see **linear regression** and **curvilinear regression**. We can also have **nonlinear** trend as discussed in this topic. All regressions are for the mean of $y$ for given $x_s$. Regression is one method of obtaining smooth trend. For others, see **smoothing methods**.

The term *trend* is also used for seasonal, cyclic, and secular as used for **time series** for intra-year and long-term pattern. This term is also used for the pattern of frequencies or of proportions in ordinal categories such as mortality in none, mild, moderate, serious, and critical conditions. This is assessed by the **Cochran test for linearity of trend**. The trend will be increasing mortality with increasing severity of disease, but the increase is sharp for some diseases and shallow for other diseases.

Graphs and diagrams, particularly the line diagram, are very helpful in trend assessment, as they can give at least a gross idea of what kind of trend, if any, exists in the data. Precise statistical study is done by fitting regressions, time series analysis, trend in proportion, and so on, as just mentioned. Studying trends helps us evaluate the result of our actions, such as for control of disease, and to project the future, and become prepared accordingly such as devising a strategy for changing the trend. This is mostly done by studying the effect of various covariates on the outcome of interest. For example, the study reported by Rahman et al. [1] has shown that satellite-based vegetation health indices are highly applicable for the malaria assessment trend in Bangladesh. Blaxill [2] observed that comparison of autism rates in the United States and the United Kingdom by year of birth for specific geographies provides the strongest basis for trend assessment. For world health situation and trend assessment from 1948 to 1988, see the article by Uemura [3].


1. Rahman A, Roytman L, Goldberg M, Kogan F. Comparative analysis on applicability of satellite and meteorological data for prediction of malaria in endemic area in Bangladesh. *J Trop Med* 2010;2010:914094. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3025368/
2. Blaxill MF. What's going on? The question of time trends in autism. *Public Health Rep* 2004 Nov–Dec;119(6):536–51. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1497666/
3. Uemura K. World health situation and trend assessment from 1948 to 1988. *Bull World Health Organ* 1988;66(6):679–87. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2491154/

## trials, *see* clinical trials

## trimmed mean

This is the mean obtained after discarding some extreme values from the data. In some analyses, the question of rejection or correction of data values may not be an issue, but it is well known that occasional extreme values do occur in many medical setups. They may or may not be **outliers**. It is natural to look for a method so that the outliers do not significantly affect the results and might seek an estimator that is less influenced than the mean (for example) by such extreme values. The sample median is the most widely used method in such situations, but another method is the trimmed mean, which is calculated after deleting generally a fixed proportion of values in each tail of the distribution of values. This is also called the *truncated mean* and is more **robust** than the usual mean.

Note that removing transparently erroneous data is not trimming. Trimming occurs when you discard data values that are legitimate (or cannot be identified as illegitimate) but small or large with respect to the sample as a whole. But isn't throwing away data a bad idea? In truth, the mean with small trimming is nearly as good as $\bar{x}$ for approximately normal data and a much safer bet than $\bar{x}$ for heavy-tailed data.

There are two types of trimmed mean that deserve mention: α-trimmed mean and α-**Winsorized mean**, where "α" is not as it is usually used in statistics; that is, it is not the probability of a Type I error. Alpha (α)-trimmed mean requires deleting a proportion α of the observations from both ends of the sample and calculating the mean of the remainder. If $x_{[1]}, x_{[2]}, \ldots, x_{[n]}$ represent the ordered sample values from the minimum to the maximum, then

$$\alpha_{\text{trimmed mean}} = [1/(n-2K)]\sum_{i=K+1}^{n-K} x_{[i]},$$

where $K$ is the smallest integer greater than or equal to $\alpha^* n$. For simplicity, we have considered the same trimming in both the tails but they can be unequal.

For example, a 5% trimmed mean is obtained by deleting 5% of the highest values and 5% of the lowest values, leaving 90% of the original values upon which a mean is then calculated. If we have the following 20 data points:

3.52, 6.60, 2.95, 2.18, 7.60, 5.21, 6.45, 5.87, 3.65, 6.99, 0.83, 1.21, 7.38, 4.74, 8.53, 5.08, 3.04, 7.24, 17.54, 2.43,

proceed by first putting the data in ascending order:

0.83, 1.28, 2.18, 2.43, 2.95, 3.04, 3.52, 3.65, 4.74, 5.08, 5.21, 5.87, 6.45, 6.60, 6.99, 7.24, 7.38, 7.60, 8.53, 17.54,

then 5% trimming removes 1 lowest out of 20 and 1 out of 20; that is, 0.83 and 17.54 are deleted. The 5% trimmed mean is based on the 20 − 2 = 18 remaining values, and in this example, this is 5.04. The untrimmed (usual) mean of all 20 values is 5.46. The difference is primarily due to the value 17.54, which is too high relative to the other values.

Incidentally, the median of a distribution is nearly a 50% trimmed mean, that is, you remove 50% of the lowest data values and 50% of the highest data values and are left with one number as an estimate of the center of the distribution. You can think of 25% trimmed mean that would be based on the middle 50% values. The trimmed mean sits between the mean and the median in terms of its statistical properties, with advantages of both and possibly not so many disadvantages. The only difficulty is with regard to its mathematics—its statistical distribution, standard error, confidence interval, test of hypothesis, and so on. For this reason, it fails to find favor with statisticians. Discarding some valid values is also not considered a good statistical strategy.

The alpha (α)-Winsorized mean is another method of estimating the mean of a population in a way that is less affected by the presence of extreme values or outliers than is the usual mean. Essentially, the $K$ smallest and $K$ largest observations, where $K$ is the smallest integer greater than or equal to $\alpha^* n$, are replaced in value to the next remaining observation and counted as though they had these values. Specifically,

$$\alpha_{\text{Winsorized mean}} = 1/n^* \left[ (K+1)(x_{[K+1]}) + (K+1)(x_{[n-K]}) + \sum_{i=K+2}^{n-K-1} x_{[i]} \right],$$

where $x_{[1]}, x_{[2]}, \ldots, x_{[n]}$ are the ordered sample values. Winsorizing differs from trimming since a certain proportion of lowest and highest values are replaced by the lowest and highest values in the remaining data set, and not deleted.

Returning to the previous example, if these data are Winsorized at 10%, then the smallest and largest values after the two lowest and two highest (10% of 20 = 2) values are 2.18 and 7.60, respectively. The Winsorized data set becomes

2.18, 2.18, 2.18, 2.43, 2.95, 3.04, 3.52, 3.65, 4.74,
5.08, 5.21, 5.87, 6.45, 6.60, 6.99, 7.24, 7.38, 7.60, 7.60, 7.60.

The 10% Winsorized mean is the average of these 20 values with the repeats, which is 5.02, as compared to the usual sample mean of 5.46. Similarly, Winsorized standard deviation can also be calculated.

It is not right to use the standard method of constructing confidence intervals for mean when using trimmed means (for one, the remaining values in the trimmed data set are no longer independent or identically distributed since the trimming is not at random). The correct standard error for an α-trimmed mean is $s_w/[(1-\alpha)\sqrt{n}]$, where $s_w$ is the standard deviation of the α-Winsorized sample. The correct standard error for the 10% trimmed mean in our example is 0.58, and a 95% confidence interval for the 10% trimmed mean is $5.02 \pm 2.13(0.58) = (3.78, 6.26)$, where 2.13 is the appropriate critical value from the Student $t$ distribution with 15 (the number of values left after deleting two smallest and two largest − 1) = 15 degrees of freedom (i.e., $t_{0.975,\,15} = 2.13$). It is interesting to note that this CI in our example omits the lowest eight and the highest eight values and just keeps the middle four.

## triple blinding, *see* blinding

## trivariate distribution

This is the distribution of subjects by three characteristics. An example is in Table T.6 where 279 users of tobacco are classified by (i) age, (ii) sex, and (iii) type of tobacco use. In this distribution, the values of a continuous variable would be grouped such as for age in Table T.6.

Ordinal and nominal groups may appear as such without any numeric assigned as are sex and types of tobacco use in Table T.6, and continuous variables also remain as such in a distribution although they may have to be categorized for tabular presentation. For a discrete variable, the values may appear as such, for example, 0, 1, 2, 3, 4, and 5+ for parity. The last category in this case is also a group.

A one-way **contingency table** describes a univariate distribution, a two-way table describes a **bivariate distribution**, and a three-way table describes a **trivariate distribution**.

**TABLE T.6**
**Distribution of 279 Tobacco Users by the Type of Tobacco**

| Age-Group (Years) | Chewing Only | | | Smoking Only | | | Dual Use | | | Total Tobacco Users | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | M | F | P | M | F | P | M | F | P | M | F | P |
| <19 | 1 | 0 | 1 | 7 | 2 | 9 | 2 | 0 | 2 | **10** | **2** | **12** |
| 20–29 | 16 | 11 | 27 | 19 | 13 | 32 | 13 | 6 | 19 | **48** | **30** | **78** |
| 30–39 | 5 | 1 | 6 | 53 | 42 | 95 | 2 | 7 | 19 | **60** | **50** | **110** |
| 40–49 | 10 | 1 | 11 | 14 | 4 | 18 | 2 | 7 | 9 | **26** | **12** | **38** |
| 50+ | 3 | 4 | 7 | 9 | 14 | 23 | 6 | 5 | 11 | **18** | **23** | **41** |
| **Total** | **35** | **17** | **52** | **102** | **75** | **177** | **25** | **25** | **50** | **162** | **117** | **279** |

*Note:* F, female; M, male; P, person.

It is worth mentioning **marginal distributions** and *conditional distributions* at this stage. The former refers to the probability distribution of a single variable (or indeed a number of variables) when the other variables are collapsed. In Table T.6, the bottom row is the marginal (bivariate) distribution of the tobacco users by sex and type of tobacco use. The age has been collapsed. Similarly, the last column is the marginal distribution of the tobacco users by age disregarding sex and type of tobacco.

Conditional refers to the probability distribution of a random variable (or maybe the joint distribution of a number of variables) when one or more other random variables are held fixed. In Table T.6, the first three columns contain conditional distribution of tobacco chewers by age and sex. This is conditioned on type of tobacco use = chewing only. Similarly, the distribution in the next three columns is the conditional distribution for tobacco use = smoking only. The marginal distributions help in deriving conclusions for specific variables and conditional for specific groups.

## truncated values/distributions

Truncation in health and medicine often occurs owing to limitations of the tools such as truncation of CT projection by limited field of view relative to large patient anatomies. They do provide truncated data and can introduce serious artifacts. However, the concern in this section is as follows.

Truncated data are values where larger or smaller values than a given fixed value are either not recorded or not observed. Where the ignored values are larger, this leads to right truncation, and where the ignored values are smaller, this leads to left truncation. Truncation is a variant of **censoring** that occurs when the incomplete nature of the observation is due to the study design. Censoring occurs when an observation is not complete because of some random cause, whereas truncation is deliberate and can also be done after the data are available. Figure T.7 has one full Gaussian distribution (broken lines) and two truncated distributions: one (thick solid line) left truncated at −0.25 and another (thin solid line) truncated at +1. It should be evident that the mean and standard deviation (SD) of the truncated distribution will not be the same as of a full distribution.

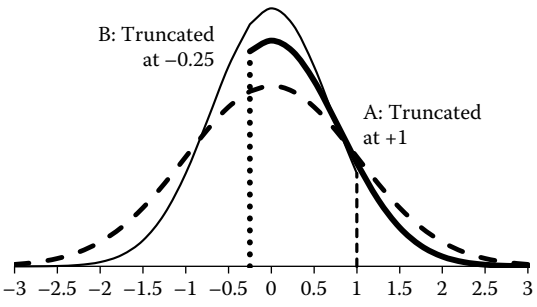



**FIGURE T.7** Truncated Gaussian distributions.

Truncation might be an operational necessity. For example, a study on cataract may restrict itself to only people of age 60 years or more. In this case, the age distribution will be left truncated at age = 60 years. The mean age of onset in these people cannot be termed as mean age of onset of cataract in the general population because people less than 60 years are not included. Age at onset of these who are not included will not be known. If people more than 90 years are excluded for some reason, the age distribution will be right truncated at age = 90 years. We can have truncation on both sides such as when including people of age 60 to 90 years. Note that this is truncation with respect to age. If the outcome of interest is the visual acuity, there is no truncation with respect to this outcome. This truncation can happen if, for example, you decide to exclude people with normal vision after obtaining the data on all.

Values with a truncated distribution cannot be analyzed in the usual way because the underlying requirement of many parametric methods may be violated. This, however, applies to only the stochastic variable such as the dependent variable in regressions. Such truncation does not affect the regression if it is for independent variables. In our example, if age is one of the regressors, its truncation to between 60 and 90 years will not affect the logistic or ordinary regression where the outcome is visual acuity except that the regression is valid for people with age 60 to 90 years. However, if age is the dependent variable, the regression will have restricted utility.

The problem compounds when truncation is applied to some values and others are fully observed. Piccorelli and Schluchter [1] discussed modeling for serial measurement of pulmonary function ($FEV_1$% predicted) and survival in cystic fibrosis patients using registry data where some patients enter late and not followed from birth, whereas for others, full data from birth is available. In this case, the truncation is with respect to age and not $FEV_1$%.

Truncation is frequently used in survival studies. This is in addition to censoring as mentioned elsewhere. For analysis of such data, see Kline and Moeschberger [2].

1. Piccorelli AV, Schluchter MD. Jointly modeling the relationship between longitudinal and survival data subject to left truncation with applications to cystic fibrosis. *Stat Med* 2012 Dec 20;31(29):3931–45. http://www.ncbi.nlm.nih.gov/pubmed/22786556
2. Kline JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data*: Springer, 2010.

## *T*-score, *see also* *Z*-scores

Generally speaking, a *T*-score of a value is how many sample standard deviation (SD) away is this value from sample mean. That is $t = \frac{x - \bar{x}}{s}$, where $\bar{x}$ and $s$ are the sample mean and sample SD, respectively. The difference with ***Z*-score** is that $Z$ uses the population mean and population SD; that is, $Z = \frac{x - \mu}{\sigma}$. Since these parameters are rarely known, they are replaced by their sample counterparts to give a *T*-score.

However in medical applications, particularly for bone mineral density (BMD), *T*-score is used to convey how far a person's BMD is from the ideal. That is, in this case, the mean is replaced by the ideal value, which, in the case of BMD, is generally considered to be the average BMD of 30-year-old healthy adults when BMD is usually is at its peak. This is also referred to as the expected BMD. The denominator should be the population SD $\sigma$ for this *T*-score that also may be known from population studies. Thus, in this case,

$$T\text{-score} = \frac{x - \mu_0}{\sigma_0},$$

where $\mu_0$ is the peak average BMD and $\sigma_0$ is the SD of BMD of 30-year-old healthy adults. A value of $T \geq -1.0$ is considered normal, $-2.5 \leq T < -1.0$ indicates osteopenia, and $T < -2.5$ suggests osteoporosis. For easy understanding, *T*-scores are simultaneously converted to have mean = 50 and SD = 10. Thus, you can say, for example, that the BMD of a person is 20 points above the expected when the actual *T*-score is +2.0. In the case of BMD, *Z*-score is calculated for each age and sex using mean and SD for that exact age and sex, whereas *T*-score always uses the ideal BMD.

Gourlay et al. [1] have made an ingenious use of *T*-scores to determine the time required for women 67 years and older with normal BMD or mild osteopenia ($T \leq -1.50$) to progress to osteoporosis ($T \leq -2.5$), and estimated it to be as much as 15 years. However, Doshi et al. [2] argue that the testing interval for BMD should be guided by an assessment of risk factors and not *T*-scores.


1. Gourlay ML, Fine JP, Preisser JS, May RC, Li C, Lui LY, Ransohoff DF, Cauley JA, Ensrud KE; Study of Osteoporotic Fractures Research Group. Bone-density testing interval and transition to osteoporosis in older women. *N Engl J Med* 2012 Jan 19;366(3):225–33. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3285114/
2. Doshi KB, Khan LZ, Williams SE, Licata AA. Bone mineral density testing: Is a T score enough to determine the screening interval? *Cleve Clin J Med* 2013 Apr;80(4):234–9. http://www.ccjm.org/view-pdf.html?file=uploads/media/media_dadd345_234




## Tschuprow coefficient, *see* association between polytomous characteristic (degree of)

## *t*-test, *see* Student *t*-tests

## Tukey test for additivity

Tukey test for additivity is used to test the statistical significance of **interaction** between two factors in a **two-way ANOVA** setup when there is only one observation per cell. When the interaction is absent, the factor effects are additive for their combined effect, which is the reason that it is called test for additivity. The usual method of testing interaction between factors does not apply when the value for only one subject is available for each combination of levels of factors because then the error sum of squares becomes zero. Unfortunately, sometimes it is not feasible to make more than one observation under the same conditions. For instance, it is not possible to ask the same participant the same question twice, when only a limited number of subjects are available. Tukey test for additivity is a great way for testing the significance of the interaction in a two-way analysis of variance with one observation per cell, although there are certain restrictions.

Suppose there are 3 drugs under trial and each has 4 doses. Thus, there are a total of $3 \times 4 = 12$ combinations. If only one subject gets each of these combinations, there will be 12 subjects in this trial. This is no replication. In this case, the interaction between drug and dose cannot be tested by the usual methods because of just one observation per cell but can be tested by the Tukey test for additivity. The test was introduced by John Tukey in 1949 [1].

Consider the scenario of two factors in which the response is modeled with $\alpha_j$ as the main effect of the $j$th level of the first factor and $\beta_k$ as the main effect of the $k$th level of the second factor. Let the interaction between the $j$th level of factor 1 and the $k$th level of factor 2 be denoted by $\gamma_{jk}$. For simplicity, we consider the case when both the factors are fixed and assume $\gamma_{jk} = D\alpha_j\beta_k$. This is the restriction that we mentioned earlier and says that the interaction is multiplicative of the individual effects. Under these conditions, the test is given by

$$\text{Tukey test for additivity: } T = \frac{SS_{AB}}{MSE},$$

where $SS_{AB} = \frac{[\Sigma_{jk} y_{jk}(\bar{y}_{j\cdot} - \bar{y}_{\cdot\cdot})(\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot})]^2}{\Sigma_j(\bar{y}_{j\cdot} - \bar{y}_{\cdot\cdot})^2 \Sigma_k(\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot})^2}$ and $\text{MSE} = \frac{SSE - SS_{AB}}{JK - (J + K)}$.

This follows an $F$ distribution with $(1, \nu)$ df, where $\nu = JK - (J + K)$. The basic difference is the way $SS_{AB}$ is computed in this case.

The critical assumption in this case is $\gamma_{jk} = D\alpha_j\beta_k$, where $D$ is some constant. This is based on the consideration that if $\gamma_{jk}$ is any second-degree polynomial function of $\alpha_j$ and $\beta_k$, then it must be of this form [2].


1. Tukey JW. One degree of freedom for nonadditivity. *Biometrics* 1949;5:232–42. http://www.jstor.org/stable/3001938
2. Neter J, Wasserman W. *Applied Linear Statistical Models*. Irwin, 1974:p. 610.


## Tukey test for multiple comparisons, *see also* multiple comparisons

The Tukey test is one of the several methods of **multiple comparisons** of group means after the analysis of variance (ANOVA) results indicate that there is a significant difference among the groups somewhere. In the case of pairwise comparisons, for example, if there are 4 groups, the comparisons are group 1 with group 2, group 1 with group 3, group 1 with group 4, group 2 with group 3, group 2 with group 4, and group 3 with group 4. There are a total of six pairwise comparisons. In addition, one can test the difference between, say, group 1 with the average of groups 2, 3, and 4 combined, and so many others of this type. This is what we mean by multiple comparisons. Means of two groups are generally compared by the Student *t*-test, but repeated application of this test at, say, 5% level of significance on the same data blows up the total probability of **Type I error** to an unacceptable level. If there are 15 tests on the same data, each done at the 5% level, then the overall (experiment-wise) Type I error could be as high as $1 - (1 - 0.05)^{15} = 0.54$. Compare this with the desired 0.05.

Type I error allowed for each individual comparison is called *comparison-wise error rate*, whereas the total error for all the comparisons together is called *experiment-wise error rate*. To keep the probability of (experiment-wise) Type I error within a specified limit such as 0.05, many procedures for multiple comparisons are available. Each of these is generally known by the name of the scientist who first proposed it. Among them are **Bonferroni, Tukey, least significant difference, Duncan,** and **Dunnett**. The Bonferroni and Tukey procedures are commonly used in medical and health literature and are also the most suitable ones.

The Tukey test is the most suitable when the interest is in *all* pairwise comparisons. The procedure works as follows. Corresponding to the df($J$,$v$), where $J$ is the number of groups to be compared and $v$ is the df associated with **mean square error** (MSE) in the ANOVA table, a value $Q$ is obtained from what is called the **Studentized range distribution** for a specified **level of significance** α such as 0.05. This distribution is based on the maximum difference in means, which is the difference between the largest mean and the smallest mean. Critical values of $Q$ are available in tables for different numbers of groups and different error df when α is 0.05 [1]. Use this $Q$ value and calculate

$$\text{Tukey test: } D = Q\sqrt{\frac{\text{MSE}}{n/J}},$$

where $n$ is the total number of subjects in all the groups together and MSE is the mean square error in the ANOVA table for those data. Any pairwise difference exceeding $D$ in group means is considered statistically significant. This keeps the experiment-wise probability of Type I error limited to the specified α level. This test was developed by John Tukey in 1949 [2] and is also called *honestly significance difference* test.

John Tukey

The requirements of the Tukey test are essentially the same as those of Student $t$-test for independent groups, namely, the homogeneity of variance, independent observations, and Gaussian (Normal) distribution or large sample, with independence being the most important and Gaussian distribution being the least important especially if the group sizes are equal. The Tukey test need not be preceded by the ANOVA $F$-test, but the MSE is required anyway.

Consider an observational study on the effect of maternal smoking (duration and quantity—both observed in three categories each) on child birth weight. Suppose that the effect of duration of smoking was statistically significant and the effect of quantity was not significant. To find which category of duration or categories were making a significant impact, researchers compared mean birth weights for different categories of duration. Hence, all pairwise comparisons were needed, which makes a Tukey test appropriate. The number of levels of this factor is $J = 3$ in this example. Also, the total number of subjects in this study is $n = 75$. The ANOVA table obtained from a computer package revealed MSE = 0.022 and the error df $v = 66$ since there were three categories of quantity of smoking as well, and the interaction term was also required. From Studentized range tables, the value of $Q$ at $\alpha = 0.05$ for (3, 66) df is approximately 3.39. Thus,

$$D = 3.39\sqrt{\frac{0.022}{75/3}} = 0.10.$$

This is the critical difference that should be present between the means for it to be statistically significant at $\alpha = 0.05$. The mean birth weights (in kilograms) in the three duration-of-smoking categories were 3.44, 3.39, and 3.25, respectively. These categories were <18 weeks, 18–31 weeks, and 32+ weeks, respectively. The difference in mean birth weight between <18 weeks and 18–31 weeks does not exceed the critical value but exceeds between other durations. Thus, the real culprit in this example is smoking for 32+ weeks, which significantly lowered the mean birth weight.

Multiple comparison tests, particularly the Tukey test, may sometimes give results at variance with the results of the $F$-test. It is possible for the $F$-test to be significant without any of the pairwise differences being significant. Conversely, the $F$-test may not show significance, but comparison for one or more specific pairs may still be significant. The reason for this is that both the $F$-test and the Tukey test require a Gaussian pattern, but the underlying distribution may not be exactly Gaussian. The $F$-test and the Tukey test behave differently in this case. The problem may arise more frequently for small $n$ and then for large $n$ because large $n$ is an insulation against violation of a Gaussian pattern in most cases.

It may be instructive to compare the Tukey test with Student $t$-test. In our example, the critical difference between two means for Student $t$-test is $d = 1.98\sqrt{\frac{0.022}{25/2}} = 0.083$, where 1.98 is the value of Student $t$ at 25 + 25 − 2 = 48 df (the sample size in each group being 25 in this example). Note how small this $d$ is compared with Tukey critical difference $D = 0.10$. This tells that the critical difference should be larger in pairwise comparisons for keeping the Type I error at the same level.

1. Neter J, Kutner M, Wsserman W, Nachtsheim Ch. *Applied Linear Statistical Models*. McGraw-Hill/Irwin, 1996.
2. Tukey J. Comparing individual means in the analysis of variance. *Biometrics* 1949;5 (2):99–114. http://www.jstor.org/stable/3001913

## twin studies

In 1875, Francis Galton wrote "Twins have a special claim upon our attention; it is, that their history affords means of distinguishing between the effects of tendencies received at birth, and those that were imposed by the special circumstances of their after lives" [1] ("after life" referring to life after birth).

The classical twin study design involves studying twins raised in the same family circumstances. Monozygotic (identical) twins share all of their genes, while dizygotic (not identical) twins share only approximately 50% of them. In addition, they share exactly the same environment in utero. Thus, when researching a particular characteristic, if a researcher compares the similarity between one set of identical twins and the similarity between sets of nonidentical twins, then any excess likeness between the identical twins should be due to the genes rather than the environment.

Identical twins have the same genome as they develop from a single fertilized egg. Hence, any differences between identical twins arise from their environments, not from genetics. The insight gained

from studying twins improves our understanding of nature versus nurture and how the two work together. For a very long time now, researchers have compared traits in twins trying to determine the extent to which certain traits are inherited, such as eye color, and which traits are acquired from the environment, such as diet.

When only one of the identical twins gets a disease, the environment is suspected to be the culprit, but if both get the same disease almost at the same time, the suspicion is on the genetic component. This theory is strengthened when several twins are seen to follow the same pattern. For example, in schizophrenia, it has been found that in approximately 50% of identical twins, both have the disease, while this is true in only approximately 10%–15% of nonidentical twins [2]. Hence, one can argue that there is a strong genetic component in susceptibility to schizophrenia. If both identical twins in a pair do not develop the disease 100% of the time while others do, this indicates that other environmental factors are also involved.

Analysis of quantitative data from twin studies is usually done by **variance components** analysis by considering the observed value $y = g + e + r$, where $g$ is the genetic component (called heritability), $e$ is the effect of shared environment, and $r$ is the residual. The extent of agreement between monozygotic and dizygotic twins can be measured by **intraclass correlation**, and ***F*-test** can be used to check statistical significance when the conditions of this test are fulfilled. For more information on analysis of twin data, see Balding et al. [3].


1. Galton F. Inquiries into human faculty and its development, in: *History of Twins*. J M Dent & Sons, 1907:pp. 155–73. http://dx.doi.org/10.1037/10913-025
2. University of Utah, Health Sciences. *Insights from Identical Twins*. http://learn.genetics.utah.edu/content/epigenetics/twins/
3. Balding DJ, Bishop M, Canning C (Eds.). *Handbook of Statistical Genetics*, Third Edition. Wiley, 2007.


## two-by-two tables

A two-by-two contingency table is a table with two rows and two columns formed by cross-classifying a group of subjects by two binary characteristics. One could be like characteristic present or absent, and the other could be group I and group II such as with and without disease, with disease A and with disease B, treatment and control, male and female, young and old, or any other such groups. These groups are independent and the scenario is essentially bivariate. Since both the variables have two categories, this is also known as a *fourfold table*. It is a key table when dealing with most proportions at the analysis stage.

The general form of such a table is given in Table T.7, where $O_{rc}$ is the observed count in the $(r, c)$th cell and $\pi_{rc}$ is the corresponding probability ($r = 1, 2$; $c = 1, 2$); dots denote the sum for the corresponding row or column.

**TABLE T.7**
**General Structure of a 2 × 2 Contingency Table**

| | Variable 1 (Antecedent) | | |
|---|---|---|---|
| **Variable 2 (Outcome)** | **Present** | **Absent** | **Total** |
| Present | $O_{11}$ ($\pi_{11}$) | $O_{12}$ ($\pi_{12}$) | $\boldsymbol{O_{1\bullet}(\pi_{1\bullet})}$ |
| Absent | $O_{21}$ ($\pi_{21}$) | $O_{22}$ ($\pi_{22}$) | $\boldsymbol{O_{2\bullet}(\pi_{2\bullet})}$ |
| **Total** | $\boldsymbol{O_{\bullet 1}(\pi_{\bullet 1})}$ | $\boldsymbol{O_{\bullet 2}(\pi_{\bullet 2})}$ | $\boldsymbol{n}$ |

### *Structure of a 2 × 2 Table in Different Types of Study*

Table T.7 is stated in the classical antecedent–outcome format. The following three scenarios are possible in this setup.

*Structure in a **Prospective Study***: Because the investigation is from antecedent to outcome in a prospective study, the column totals $O_{\bullet 1}$ and $O_{\bullet 2}$ are fixed in advance. They can also be denoted by $n_1$ and $n_2$, respectively. These are the numbers of exposed and nonexposed subjects followed up for appearance of outcome. The row totals $O_{1\bullet}$ and $O_{2\bullet}$ become known only after the investigation is over and are random variables. The relevant **null hypothesis** in this case is $H_0$: $\pi_{11} = \pi_{12}$. This states that the incidence rate of outcome in the two groups is the same. In this case, $\pi_{11} + \pi_{21} = 1$ and $\pi_{12} + \pi_{22} = 1$. Thus, $H_0$ is equivalent to $\pi_{21} = \pi_{22}$. The analysis of the data in this setup is mostly done in terms of **relative risk**, although **chi-square** can be used to test association.

*Structure in a **Retrospective Study***: The direction of the investigation in a retrospective study is from outcome to antecedent. Thus, the row totals $O_{1\bullet}$ and $O_{2\bullet}$, say, with and without disease, are fixed in advance and the column totals $O_{\bullet 1}$ and $O_{\bullet 2}$ are obtained during the study. The fixed row totals can also be denoted by $n_1$ and $n_2$. The null hypothesis now is that the rate of presence of antecedent in those with a positive outcome is the same as in those with a negative outcome; that is, $H_0$: $\pi_{11} = \pi_{21}$. In this case, $\pi_{11} + \pi_{12} = 1$ and $\pi_{21} + \pi_{22} = 1$. $H_0$ implies $\pi_{12} = \pi_{22}$ as well. Analysis of data in this setup is mostly done in terms of **odds ratio**, although chi-square can also be used to test association in this case.

*Structure in **Cross-Sectional Study***: In this case, $n$ subjects are simultaneously cross-classified by the antecedent and outcome. Neither the column totals nor the row totals are fixed in advance and both become known only after study of the participants is over. In this case, $\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22} = 1$. According to the law of multiplication of probabilities, the antecedent and outcome are independent if and only if $H_0$: $\pi_{rc} = \pi^*_{j\bullet}\,\pi_{\bullet c}$ ($r, c = 1, 2$) holds, where $\pi_{r\bullet} = \pi_{r1} + \pi_{r2}$ and $\pi_{\bullet c} = \pi_{1c} + \pi_{2c}$. Analysis of data in this setup is mostly done by chi-square.

$H_0$ in the prospective and retrospective scenarios is called the hypothesis of homogeneity (column homogeneity and row homogeneity, respectively), and $H_0$ in the cross-sectional scenario is called the hypothesis of independence. All these situations can be viewed as participants divided by two qualitative characteristics with the objective of investigating whether one characteristic has any association with the other—whether one is occurring more commonly with the other than expected by chance. Luckily, the statistical test for all the three scenarios is the same in terms of chi-square for large samples and of the **Fisher exact test** if any cell count is less than 5.

## two one-sided tests (TOSTs), *see* equivalence, superiority, and inferiority tests

## two-phase sampling

This is a sampling method that involves two distinct phases. In the first phase, information about a particular auxiliary variable of

interest is collected on every member of a relatively large sample. This may already be available in records, and personal contact may not be needed. In the second phase, a subsample of the individuals in the original sample is taken and additional information about these is collected. This is also called **double sampling** and is used for **ratio estimate** of the parameter of interest or for **stratification**. Do not confuse it with two-stage sampling, where the sampling unit for the two stages is different such as cities in the first stage and hospitals in the second stage. In a two-phase sampling, some of the same units are generally sampled again for collecting additional information, although another set of units can also be selected. Also, some researchers use the term *double sampling* whenever a population is sampled twice for some reason. This is a nontechnical use of the term and not actually double sampling in the sense that we describe here.

An example of where this method of sampling might be useful is when estimating prevalence based on results provided by a potentially fallible, but inexpensive and easy-to-use diagnosis of the true disease state of all the sampled individuals. The diagnosis might then be confirmed for a subsample through the use of a more accurate diagnostic test.

This methodology can be extended in several ways. (i) The second sample may not be out of the first sample but the second sample should also collect the auxiliary information. (ii) You can use this method to estimate mean instead of proportion. For mean, there is another method called regression estimate. (iii) Double sampling is also used for **stratification** where the first sample is used to estimate stratum sizes. These sizes are required for estimating the mean or proportion based on **stratified random sampling**. Some details are given later in this section, and more details of all these are available in the classic book by Cochran [1]. (iv) Two-phase sampling is also used in **acceptance sampling** where the objective is to determine whether the lot (of, say, medical devices) has defectives less than a prefixed threshold. In this situation, accept the lot if the proportion of defectives in a sample is less than, say, $p_1$, reject the lot if the proportion of defectives is more than $p_2$, and take another sample if the proportion of defectives is between $p_1$ and $p_2$. In this setup, the second sample is not out of the first.

To formalize, the conventional method of two-phase sampling can be described in the following manner:

Step 1: From the available population, select a first-phase sample of large size $n_1$ and observe $x$, where $x$ is some auxiliary information that can be easily obtained.

Step 2: Treat the first-phase sample $n_1$ as the population and select a second-phase sample of size $n_2$, and observe another variable $y$. Information on $x$ for this sample is already available from the first phase.

Here, the probability of being selected as a second-phase sample is often determined by the value of $x$ obtained from the first-phase sample. Because of this, the probability for inclusion of a subject in the second-phase sample is a random variable: its value changes as the first-phase sample changes.

There are other variations of this method. Yiannoutsos et al. [2] used a double-sampling methodology for improving the estimate of HIV mortality in patients on combination antiretroviral therapy. The two samples were patient outreach and vital registries. Ekelund et al. [3] used the term *double sampling* for two blood samples in pregnancies affected by trisomy 21 and control—one at the early stages of pregnancy and the other at a late stage to assess their screening performance.

### *Two-Phase Sampling for Stratification*

Stratification helps in obtaining reliable estimates for each stratum and can improve the reliability of the overall estimate as well. In some situations, the required information on the stratifying characteristic is not available and has to be collected. This is done in phase I of double sampling by selecting, say, $n_1$ units. These units are stratified, and suppose we find that there are $n_{1h}$ units in stratum $h$. The first phase helps estimate the sample proportion in each stratum by $w_h = \frac{n_{1h}}{n_1}$. This is a random variable in this setup in contrast to stratified random sampling where the stratum size and hence proportion is fixed and known in advance. In this case, $w_h$'*s* also give an idea of how the subjects are distributed. More importantly, $w_h$ serves as weight for estimating the means after phase II as per the following procedure.

Phase II sampling is done using these strata with $n_{2h}$ units from the $h$th stratum, $\Sigma_h n_{2h} = n_2$. Let the value of the $i$th unit of the $h$th stratum be denoted by $y_{ih}$. The mean of these would be $\bar{y}_h = \frac{\Sigma_i y_{ih}}{n_{2h}}$. Then, the mean of all the units in the second sample can be obtained as

$$\text{Mean in two-phase sampling for stratification: } \bar{y}_{st} = \Sigma_h w_h \bar{y}_h.$$

The weights $w_h$ are based on the first-phase sampling results as already stated. This would be an **unbiased estimator** of the population mean when the sampling in both phases is simple random.

For example, consider a study on estimation of prevalence of diabetic neuropathy in diabetic subjects. This is suspected to be related to sex. Take a random sample of, say, $n$ = 1000 diabetic subjects from something like a diabetes registry and find, for example, that 580 of them are females and 420 are males. This gives the ratio of female diabetic subjects to male diabetic subjects. Now, take a phase II sample of $n_2$ = 200 subjects from these 1000, 100 for each sex after stratification. If the prevalence of diabetic neuropathy is 5% in females and 8% in males, what is the estimate of diabetic neuropathy in those appearing in the registry? This is an example of proportions and not means, but the method is the same. The estimate for prevalence of diabetic neuropathy is $\bar{y}_{st} = \frac{580}{1000} \times 0.05 + \frac{420}{1000} \times 0.08 = 0.0626$, or 6.26% of diabetics in this registry have diabetic neuropathy. Note that this estimate, on the basis of only the phase II sample of $n_2$ = 200, would have been (5 + 8)/2 = 6.5%. These two estimates are close but not the same. In some situations, they can be very different.

### *Two-Phase Sampling for Ratio Estimate*

In case there is a strong relationship between the variable $x$ on which information is collected in the first phase and the variable $y$ on which information is collected for a subsample (information on $x$ is already available for this sample), the reliability of the estimate can be improved. For details, see the topic **ratio estimator**.

Further information about two-phase sampling can be found in a book by Ahmad et al. [4]


1. Cochran WG. *Sampling Techniques*, Third Edition. Wiley, 1977.
2. Yiannoutsos CT, Johnson LF, Boulle A, Musick BS, Gsponer T, Balestre E, Law M, Shepherd BE, Egger M; International Epidemiologic Databases to Evaluate AIDS (IeDEA) Collaboration. Estimated mortality of adult HIV-infected patients starting treatment with combination antiretroviral therapy. *Sex Transm Infect* 2012 Dec;88 Suppl 2:i33–43. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3512431/

3. Ekelund C, Wright D, Ball S, Kirkegaard I, Nørgaard P, Sørensen S, Friis-Hansen L et al. Prospective study evaluating performance of first-trimester combined screening for trisomy 21 using repeat sampling of maternal serum markers PAPP-A and free β-hCG. *Ultrasound Obstet Gynecol* 2012 Sep;40(3):276–81. http://onlinelibrary.wiley.com/doi/10.1002/uog.12266/full, last accessed December 14, 2014.
4. Ahmad Z, Shahbaz MQ, Hanif M. *Two Phase Sampling*. Cambridge Scholars, 2013.

## two-sided alternative, *see* one- and two-tailed alternatives/tests

## two-way ANOVA, *see also* analysis of variance (ANOVA)

Refer to the topic **analysis of variance (ANOVA)** for the fundamentals of ANOVA. Consider a **clinical trial** in which three doses (including a placebo) of a drug are given to a group of anemic male and female subjects to assess the rise in hematocrit (Hct) level. Thus, there are two factors in this trial, namely, dose and sex, and the analysis of such data would require a two-way ANOVA. If it is suspected that the response for different doses may be different for males than for females, conclusion about the effect of factors would require that this **interaction** is also investigated. This makes it imperative to adopt a two-way design. The details of interaction are given a little later in this section. Consider the design aspect first.

### *Two-Factor Design with Fixed Effects*

There are two factors in the trial just mentioned: the dose of the drug and the sex of the subjects. These factors are fixed in the sense that the interest is limited to these specific doses and sex. The response of interest in this experiment is a quantitative variable, namely, the rise in Hct level. The objective of the trial is to find the effect of dose and sex and their interaction on the response. Such a setup with two factors is called a two-way ANOVA situation. Note that there are three dose groups of male subjects in this trial and another three dose groups of female subjects. The researcher may wish to have $n = 10$ subjects in each of these six groups, making a total of 60 subjects. To minimize the role of other factors causing variation, these 60 subjects should be as homogeneous as possible with respect to all the other characteristics that might influence the response; for example, they may be of the same age-group, they may be of normal build (say, BMI between 20 and 25), and they may not be suffering from any ailment that may alter the response. Once 30 male and 30 female eligible subjects meeting the inclusion and exclusion criteria are identified, they need to be randomly allocated 10 each to the three dose levels. Such allocation then increases the confidence in asserting that any difference that occurs is mostly, if not exclusively, attributed to the factors under study, namely, the dose of the drug and the sex of the subjects in this example. A post hoc analysis can be done to check that other influencing factors, such as age and BMI, are indeed almost equally distributed in the six groups under study. Informed consent and other requirements of **ethics**, in any case, must be met. Blinding may be required to eliminate possible bias of the subjects, of the observers, and even of the data analyst.

The following notation would be clear if you can imagine another trial in which a drug is administered in $I = 3$ doses, a dietary supplement of $K = 4$ types (e.g., vitamin A, vitamin D, zinc, no supplement) is given to a group of subjects, and bone mineral density (BMD) gain ($y_{ijk}$) from baseline is measured after 90 days for the $i$th subject in the $j$th dose group and the $k$th diet group. The gain could be negative also in some cases, which means that the BMD declined. Under this notation, for example, $y_{423}$ is the BMD gain of the fourth person ($i = 4$) getting a second ($j = 2$) dose and a third ($k = 3$) supplement. Response in such a two-factor design with $J$ levels of factor 1 and $K$ levels of factor 2 with $n$ subjects at each treatment combination (**balanced design**) can be modeled as

$$y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{jk};\ i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, J;\ k = 1, 2, \ldots, K,$$

where $y_{ijk}$ is the response of the $i$th subject at the $j$th level of factor 1 and the $k$th level of factor 2, $\mu$ is the overall mean response (mean BMD gain of all persons combined in our example), $\alpha_j$ is the main effect of the $j$th level of factor 1 (such as the effect of dose 1 on BMD gain), $\beta_k$ is the main effect of the $k$th level of factor 2 (such as the effect of vitamin D on BMD gain), $(\alpha\beta)_{jk}$ is the effect of interaction between the $j$th level of factor 1 and the $k$th level of factor 2, and $\varepsilon_{ijk}$ is the error, which is the remainder after all these effects are considered.

In another two-factor design, say, with three doses of drug A and two doses of drug B, you can randomly allocate subjects to any of the six combinations of the doses. However, if factor 2 is something like sex, there is no way that you can randomly allocate subjects to these six combinations—you will have to allocate males (randomly) to one of the three doses and females to one of the three doses as well. If factor 1 is also something like obesity (thin, normal, or obese), the subject's category is fixed—there is no random allocation. The best option in this situation could be random selection of subjects from male obese and female obese subjects in the population.

### *The Hypotheses and Their Test in a Two-Way ANOVA*

The null hypotheses that can be tested in a two-way ANOVA situation are as follows:

- Levels of factor 1 have no effect on the mean response; that is, each level of factor 1 has the same response on average. This translates into

$$H_{0a}: \mu_{1\bullet} = \mu_{2\bullet} = \ldots = \mu_{J\bullet},$$

  where $J$ is the number of levels of factor 1.
- Levels of factor 2 have no effect on the mean response. That is,

$$H_{0b}: \mu_{\bullet 1} = \mu_{\bullet 2} = \ldots = \mu_{\bullet K},$$

  where $K$ is the number of levels of factor 2.
- There is no interaction between factor 1 and factor 2.

See the example that follows shortly.

The total sum of squares in a two-way ANOVA is broken into the **sum of squares** due to factor 1, due to factor 2, due to interaction, and the residual (also called error). These are divided by the respective degrees of freedom to get **mean squares**. Each of these mean squares is an independent estimate of the same variance $\sigma^2$ of the response $y$ when the corresponding $H_0$ is true. Mean square due to error (MSE) is an estimate of $\sigma^2$ even when $H_0$ is false. Other mean squares are compared with MSE, and the criterion $\boldsymbol{F}$ is calculated separately for each of the two factors and for their interaction. The

*P*-value is obtained as usual corresponding to the calculated value of *F*. The pair of df for different *F*'s are (i) for factor 1: $(J-1)$, $JK(n-1)$; (ii) for factor 2: $(K-1)$, $JK(n-1)$; and (iii) for interaction between factor 1 and factor 2: $(J-1)^*(K-1)$, $JK(n-1)$, where $n$ (≥2) is the number of subjects with the $j$th level of factor 1 and the $k$th level of factor 2 ($j = 1, 2, \ldots, J$; $k = 1, 2, \ldots, K$). This assumes that the trial has the same number of subjects for each combination. Separate decisions for factor 1, factor 2, and the interaction are made regarding their statistical significance. When the interaction is not significant, the factors are called **additive**. If the interaction is significant, the inference for any factor cannot be drawn in isolation, and it has to be in conjunction with the level of the other factor.

A two-factor trial can also be done with only $n_{ij} = 1$ subject in each group. In this case, however, interaction cannot be easily evaluated (see **Tukey test for additivity**). For simplicity, the preceding example had the same number of subjects in each group, and it is more than 1. If the situation so demands, you can plan a trial or an experiment with unequal $n$ in different groups. This is called an **unbalanced design** and is illustrated in Table T.8 for birth weight of children born to women with different amount of smoking and duration of smoking as explained in a short while.

The analysis of an unbalanced design such as in Table T.8 is slightly more complex, although the concepts remain the same. For example, the total df would be $\Sigma(n_{jk} - 1)$, where $n_{jk}$ (≥2) is the number of subjects for the $j$th level of factor 1 and the $k$th level of factor 2. When using a statistical software package, be careful in selecting an appropriate routine if the data are unbalanced. Our advice, though, is to plan and conduct a balanced design (equal $n$ for each group) wherever feasible so that such complications are avoided. Incidentally, equal $n_{jk}$'s also maximize the power for any given total sample size.

As an example, consider a study done by Wang et al. [1] on the effect of maternal smoking on birth weight in the United States. For our purpose, vary this study and assume that some women, who are otherwise habitual smokers, give up smoking off and on but not completely during pregnancy. Nonsmokers are excluded in this example. Suppose that such women are asked at the time of their last antenatal visit just before birth about the duration of smoking (factor 1) and the amount of smoking (factor 2). At the end of the pregnancy, the former is categorized as <18 weeks, 18–31 weeks, and ≥32 weeks. These are the three levels of factor 1. The amount of smoking is categorized as mild (1–9 cigarettes per day), moderate (10–19 cigarettes per day), and heavy (20+ cigarettes per day). The days in this calculation are only those when at least some smoking was done and the amount of smoking is the average per smoking day. These are the three levels of factor 2. The outcome of interest (response variable) is the birth weight of the children born to these women.

Some possible confounders in this case are race, age, nutrition, education, and parity since all these can affect birth weight. Suppose these factors could be satisfactorily matched in different groups such that the effect of such extraneous factors is minimal. Let the mean birth weight in different groups be as given in Table T.8. There are, for instance, 15 women who smoked an average of 1–9 cigarettes per day for a total of less than 18 weeks during the entire pregnancy. The average birth weight of their babies was 3.45 kg. There are 8 women who smoked an average of 1–9 cigarettes per day for a total of 18–31 weeks, and the average birth weight of their babies was 3.38 kg, and so on. Because of unequal $n$, the design is unbalanced as mentioned earlier.

The question to be answered is whether the difference in mean birth weight in different groups is really present in this population of antenatal women or whether it is just a chance occurrence in this sample. This can be examined in the following three ways:

- Differences in mean birth weight in the last row of Table T.8 for different amounts of smoking are significant or not.
- Differences in mean birth weight in the last column of Table T.8 for different durations of smoking are significant or not.
- Difference in mean birth weight, for instance, in those mildly smoking for 32+ weeks and those heavily smoking for 32+ weeks compared with the difference in those with similar smoking for <18 weeks are significant or not. This is the interaction.

The calculations are done on the original values of birth weight in $\Sigma n_{jk} = 75$ children. Suppose a software package reveals $P > 0.05$ for $F$ when calculated for amount of smoking (factor 2) and $P < 0.01$ for $F$ when calculated for duration of smoking (factor 1). The first is not statistically significant but the second is. The conclusion then is that this sample of women does not provide sufficient evidence to conclude that the amount of smoking makes a difference in birth weight on average, but the duration of smoking in the pregnancy does make a difference. Let $P < 0.05$ for $F$ for interaction. Thus, an interaction between amount of smoking and duration of smoking is present and implies that the effect of duration of smoking on birth weight is not uniform in the three categories of amount of smoking. This requires additional care in interpreting the significance of the effect of amount of smoking as that requires separate conclusion regarding this effect of smoking. ANOVA considers all categories nominal. In the above example, both factor 1 and factor 2 are ordinal, but this feature is overlooked in this analysis. Also, this analysis is valid only when the birth weights follow a Gaussian distribution



**TABLE T.8**
**Average Birth Weight of Children Born to Women with Different Amount and Duration of Smoking**

| Duration of Smoking in Pregnancy | Amount of Smoking | | | |
|---|---|---|---|---|
| | Mild | Moderate | Heavy | All |
| <18 weeks | 3.45 ($n$ = 15) | 3.42 ($n$ = 12) | 3.43 ($n$ = 7) | **3.44 ($n$ = 34)** |
| 18–31 weeks | 3.38 ($n$ = 8) | 3.40 ($n$ = 10) | 3.39 ($n$ = 6) | **3.39 ($n$ = 24)** |
| 32+ weeks | 3.35 ($n$ = 5) | 3.30 ($n$ = 3) | 3.18 ($n$ = 9) | **3.25 ($n$ = 17)** |
| All | **3.41 ($n$ = 28)** | **3.40 ($n$ = 25)** | **3.32 ($n$ = 22)** | **3.38 ($n$ = 75)** |

*Note:* Entries are average birth weight in kilograms.

and have the same variance across groups. Women in this setup are likely to be independent in the sense that the smoking status of one woman will not affect the smoking status of the others and thus the birth weights are independent as well.

There is no random allocation in this example, and the results of ANOVA are valid only if the women are randomly chosen from their respective groups. This is apparently missing in the example. The stipulation is that the women included in the study adequately represent the "population" of such antenatal women with respect to the influence of amount and duration of smoking on birth weight.

### *Main Effect and Interaction (Effect)*

ANOVA is primarily designed to determine the statistical significance of the differences between group means, but it can also be used to estimate the average effect of various levels of the factors, called main effects, and the magnitude of different interaction effects. In our birth weight example, the main effect of smoking for 32+ weeks is estimated by the difference of the mean in this category from the overall mean. This is 3.25 − 3.38 = −0.13 kg. That is, smoking for 32+ weeks in pregnancy reduced birth weight on average by 130 g compared with the overall average. (The average birth weight of 3.38 kg is determined from those who smoked since the study is based on smokers only.) Similar main effects can be estimated for the other categories of duration of smoking. They can be calculated for different categories of the amount of smoking as well, but that is inadvisable in this case because these differences are not statistically significant.

Main effects and interactions are easy to understand with the help of the following notations:

Estimated main effect of the $j$th level of factor 1:

$$\alpha_j = (\bar{y}_{\bullet j \bullet} - \bar{y}_{\bullet\bullet\bullet}); \; j = 1, 2, \ldots, J$$

Estimated main effect of the $k$th level of factor 2:

$$\beta_k = (\bar{y}_{\bullet\bullet k} - \bar{y}_{\bullet\bullet\bullet}); \; k = 1, 2, \ldots, K.$$

In our birth weight and smoking example, the main effect of smoking for less than 18 weeks is 3.44 − 3.38 = +0.06 kg and that of moderate smoking is 3.40 − 3.38 = +0.02 kg. The positive effect is not surprising since these are the effects on the birth weight *compared with the overall mean* of all categories that include heavy and long-duration smokers. It is easy to show by some algebra that under these notations, $\Sigma_j \alpha_j = 0$ and $\Sigma_k \beta_k = 0$. Consequently, only $(J - 1)$ α's and $(K - 1)$ β's are independently determined and each one is automatically determined by these conditions.

The interaction is obtained separately for each combination of the levels of the two factors. This is the excess mean after adjustment for the main effects of the concerned level of factor 1 and factor 2. Thus, the estimated interaction effect between the $j$th level of factor l and the $k$th level of factor 2 is

$$\begin{aligned}(\alpha\beta)_{jk} &= (\bar{y}_{\bullet jk} - \bar{y}_{\bullet\bullet\bullet}) - (\bar{y}_{\bullet j \bullet} - \bar{y}_{\bullet\bullet\bullet}) - (\bar{y}_{\bullet\bullet k} - \bar{y}_{\bullet\bullet\bullet}) \\ &= (\bar{y}_{\bullet jk} - \bar{y}_{\bullet j \bullet} - \bar{y}_{\bullet\bullet k} + \bar{y}_{\bullet\bullet\bullet}),\end{aligned}$$

where $\bar{y}_{\bullet jk}$ is the mean of $n$ (now assumed equal for easy notations) subjects in the $(j, k)$th group. Mathematically, this implies $\Sigma_j(\alpha\beta)_{jk} = 0$ and $\Sigma_k(\alpha\beta)_{jk} = 0$. For example, the estimate of the interaction effect between moderate smoking and smoking for 32+ weeks in Table T.8 is 3.30 − 3.25 − 3.40 + 3.38 = +0.03 kg. Thus, this combination of duration and amount of smoking increases birth weight by 30 g on average in this sample *relative to the means in respective categories*. Note, again, that the overall mean is based on all women including those who are heavy and long-duration smokers.

This example also illustrates that if the interaction is significant, one ought to condition the conclusion about one factor's effect on the level of the other factor. When the interaction is not significant, the focus is on the main effects. In that case, factor 1 levels should not be compared within factor 2 levels because factor 1 effects are not significantly different. Only the average is adequate.

The above equations define the main effects and the interactions relative to the overall mean, which is the most commonly used definition. However, if so desired, you can define them relative to a base value. In our example, the base could be the category with least smoking where the mean birth weight is 3.45 kg. In that case, for example, the estimated main effect of heavy smoking is 3.32 − 3.45 = −0.13 kg. That is, heavy smoking reduces birth weight by 0.13 kg on average compared with the mild smokers for less than 18 weeks. A similar interpretation can be given for the other main effects and interactions when measured from a base value.

Note the following for the two-way ANOVA procedure:

- As for a one-way ANOVA, criterion $F$ in a two-way ANOVA also provides an overall test about difference being present *somewhere*. **Multiple comparisons** procedures as described separately can be used after ANOVA to find the group or groups that are different from others. These can be done for factor 1 effects separately from factor 2 effects.
- Extension of ANOVA to three or more factors is straightforward with main effects and interactions similar to those in a two-way ANOVA. However, there would now be several two-factor interactions, three-factor interactions, and so on. The details are beyond the scope of this book. Interested readers may consult Doncaster and Davey [2].
- It is desirable to test higher-order interactions before lower-order ones because it is difficult to attach meaning to the lower-order interactions when higher-order interactions are present.
- As always, lack of statistical significance of interaction does not mean interaction is absent. It only means that this could not be detected from the available data. This might be more so in the case of ANOVA since the sample size is generally planned to detect main effects whereas a higher sample size is required for detecting interactions.

For two-way ANOVA with random effects, see **mixed effects models** when one factor is random and **random effects ANOVA** when both the factors are random.

1. Wang X, Tager IB, van Vunakis H, Speizer FE, Hanrahan JP. Maternal smoking during pregnancy, urine cotinine concentrations, and birth outcomes: A prospective cohort study. *Int J Epidemiol* 1997;26:978–88. http://www.ncbi.nlm.nih.gov/pubmed/9363518
2. Doncaster CP, Davey AJH. *Analysis of Variance and Covariance: How to Choose and Construct Models for the Life Sciences*. Cambridge, 2007.

## two-way designs, *see also* one-way designs

Take a study on hormone replacement therapy (HRT) in mice with continuous HRT, cyclic HRT, and no HRT, and introduce one additional factor, that is, the mice are normal estrous with ovary or ovariectomized. This factor has these two levels. With three levels of treatment and two levels of estrous status, this is a 3 × 2 experiment,

**TABLE T.9**
**A Two-Way Design (3 × 2) with 10 Mice in Each Group (Mice Number in Each Group Shown after Random Allocation) to HRT Categories**

| | Factor 2 | |
|---|---|---|
| **Factor 1** | **Normal Estrous** | **Ovariectomized** |
| Continuous HRT | 3, 7, 8, 14, 18, 30, 31, 37, 48, 52 | 1, 5, 10, 11, 15, 20, 27, 42, 43, 49 |
| Cyclic HRT | 4, 9, 13, 21, 32, 39, 44, 47, 50, 57 | 6, 17, 26, 33, 34, 40, 51, 56, 58, 59 |
| No HRT (control) | 12, 22, 23, 25, 36, 41, 45, 53, 55, 60 | 2, 16, 19, 24, 28, 29, 35, 38, 46, 54 |

necessitating a two-way design. With a total of 6 subgroups and 10 mice in each, this may look like as shown in Table T.9.

Each subgroup has the same number of mice in this example, but that is not a requirement for any *K*-way design. Only equal numbers make the analysis and interpretation a lot easier. A design with equal numbers is called **balanced**; otherwise, it is an unbalanced design.

The response in this experiment is the mitotic index. The two-way design allows study of response variation between levels of factor 1 as much as between levels of factor 2. It is possible in this case that the difference in mitotic index between continuous and cyclic HRT is not the same in estrous mice as in the ovariectomized mice. This is called **interaction**. In view of the importance of interaction in medical experiments, this is explained in detail under a separate topic. For analysis of the data from two-way designs, see the topic **two-way ANOVA**.

## two-way tables, *see* contingency tables

## Type I error, *see* level of significance

## Type II error, *see also* power (statistical) and power analysis

Data-based statistical decisions regarding rejecting or not rejecting a **null hypothesis** ($H_0$) can have primarily two types of statistical errors. See **level of significance** for details of the Type I error that occurs when a true null hypothesis is rejected. The Type II statistical error fails to reject $H_0$ when it is false. This occurs when a study fails to detect a real difference because the available data do not contain sufficient evidence. This corresponds to a missed diagnosis in a clinical setup as well as to pronouncing a criminal not guilty in a court of law. In a clinical trial scenario, this is equivalent to declaring a drug ineffective when it is actually effective. When this occurs, the society will continue to be without the drug just as it was before the trial and a drug that could possibly provide better relief to scores of patients is denied entry into the market. If the manufacturer believes that the drug is really effective, it will carry out further trials and collect further evidence. Thus, the effect of Type II error in this case is that the introduction of the drug is delayed but not denied, which is not so much of an error. Type I error is considered more serious, and the tolerance threshold for the chance of Type I error is always fixed in advance with a special name, that is, the level of significance.

The probability of Type II error is denoted by $\beta$ and is calculated for a specific value under the **alternative hypothesis** ($H_1$) after fixing the level of significance $\alpha$. For example, if the null is $\mu = 15$, the Type II error would be different for $H_1$: $\mu = 25$ and for $H_1$: $\mu = 20$.

Although the statistical procedures are based on chance of Type I and Type II errors, once a decision has been made, the researcher could have made only one error—wrongly reject the null or wrongly not reject the null—or have made no error. He/she would not know that this was the case unless he/she is divine to know the truth. Also, note that there is no Type I or Type II error if there is no test of hypothesis. For example, these errors do not arise while estimating the mean or proportion from the sample values. For more details, see the topic **power**, which is the complement of the probability of Type II error.

### *Balancing Type I and Type II Errors*

There is a trade-off between Type I and Type II errors. In terms of an analogy to the judiciary system, the courts' desire to convict a guilty person has a direct bearing on an innocent person being convicted. The latter is the Type I error. Type II error is preferred over Type I error in court to uphold civil rights. If the evidence is not sufficiently convincing, the person is set free because of doubt on whether the crime was committed or not. Similar is the case with statistical testing of hypothesis. It is better not to reject a false null than to reject a true null hypothesis.

Type II error and the level of significance have a direct relationship. The lower the level of significance, the higher the Type II error. Using 1% level of significance instead of the conventional 5% makes it difficult to reject a false null, and the Type II error increases. A relationship between significance level $\alpha$ and power $(1 - \beta)$ can be expressed by a receiver operating characteristic (ROC) curve similar to the **ROC curve** between sensitivity and (1 − specificity). Just how useful this is is debatable since $\alpha$ is mostly fixed in advance.

Also, the higher the **variance** of the data values, the higher the Type II error. When values are highly variable from participant to participant, the difference between two or more groups will not be easy to detect. Moreover, irrespective of the values of $\alpha$ and of variance, a smaller difference is more difficult to detect such that the corresponding Type II error is high. The larger the minimum **medically important effect**, the lower the Type II error (higher power to detect). Sample size is an all-important factor in this scenario. An adequately large sample can override any limitations brought about by low level of significance, high variance, and small, medically relevant effect for detection.

For further details of these two errors, see Rogers [1].

1. Rogers T. *Type I and Type II Errors—Making Mistakes in the Justice System. Amazing Applications of Probability and Statistics.* http://intuitor.com/statistics/T1T2Errors.html, last accessed December 15, 2015.

## Type III error

Type III error [1] occurs when the researcher chooses to use the wrong direction of the effect. For example, if the population value of the parameter under test is actually more than the null value, the sample might give a value so much below the null value that the researcher rejects the null and concludes that the population value is also below the null value. Similarly, the error may occur in the other direction as well. This can occur when the **alternative hypothesis** is two tailed.

Leventhal and Huynh [2] wrote an interesting article on what they call the "directional two-tailed test of statistical significance." They argue that the traditional two-tailed test does not allow directional decisions since the conclusion on rejecting the

null is that the tested parameter does not have a value equal to the value specified in the null hypothesis, and nothing can be stated in which of the two directions the actual value of the parameter differs from the null value. Where there is a direction of difference as in one-tailed $H_1$, we commonly conclude that the population follows the sample. Another possible procedure is the directional two-tailed test, also called a *three-choice test*. The three hypotheses entertained are parameter = null value, parameter < null value, and parameter > null value. With the three-choice test, one may make **Type I errors**, **Type II errors**, or Type III errors. Procedures for this are not well defined yet, but an indication of what possibly can be done is discussed in the next paragraph.

To reiterate, a Type III error concerns with the direction of the alternative hypothesis. When the null for a parameter is $H_0$: $\theta = \theta_0$, the alternative could be (i) $H_1$: $\theta > \theta_0$, (ii) $H_1$: $\theta < \theta_0$, or (iii) $H_1$: $\theta \neq \theta_0$. The last one is called nondirectional and is the one that can give rise to Type III error. When testing the usual nondirectional hypotheses, one can correctly reject the null yet make a Type III error. In this case, the probability of making a Type III error is included in a Type II error. In view of this, Leventhal and Huynh [2] suggest a revised definition of power: "the conditional probability of correctly rejecting the null hypothesis and *correctly identifying the true direction of difference* between the population value of the tested parameter and the null value." This includes direction that the conventional definition does not do. They also demonstrate that when conducting three-choice tests, the revised definition of power as above results in power being somewhat less than when using the traditional definition. They also show how to adjust traditional power and sample size calculations to take into account the possibility of a Type III error. Many researchers say that the probability of a Type III error in most setups will be so small that it is not really necessary to make any adjustment when computing power or sample sizes.

There is another interpretation of a Type III error. Kimball [3] wrote about "errors of the third kind in statistical consulting." The error he was referring to was that of giving the right answer to the wrong problem. He surmised that this error originated in poor communication between the statistical consultant and his client, and suggested that statistical consultants need to be taught communication skills or "people involving" skills. It is also defined as correctly rejecting the null hypothesis for a wrong reason. An example of this kind of Type III error is assessing the causes of interindividual variation within a population when the research question of interest is about causes of difference between populations or between periods. Schwartz and Carpenter [4] illustrate the consequences of this kind of a Type III error in investigating obesity.

1. Kaiser H F Directional statistical decisions. *Psychol Rev* 1960;67:160–7. http://psycnet.apa.org/psycinfo/1961-01476-001
2. Leventhal L, Huynh C-L, Directional decisions for two-tailed tests: Power, error rates, and sample size. *Psycho Methods* 1996;1:278–92. http://www.researchgate.net/publication/232605183_Directional_decisions_for_two-tailed_tests_Power_error_rates_and_sample_size
3. Kimball A W. Errors of the third kind in statistical consulting. *J Am Stat Assoc* 1957;57:133–42. http://www.jstor.org/stable/2280840?seq=1#page_scan_tab_contents\
4. Schwartz S, Carpenter KM. The right answer for the wrong question: Consequences of type III error for public health research. *Am J Public Health* 1999 Aug;89(8):1175–80. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1508697/

# U

## unbalanced design, *see* balanced and unbalanced design

## unbiased estimator

Before the unbiased estimator, we need to explain a couple of other terms in biostatistics that would make it easy to understand this term. First, technically, a *statistic* is something we calculate on the basis of sample values. Note that this is singular, the plural of which is *statistics*. Sample mean, sample median, sample standard deviation (SD), etc. are all statistics in this sense. The values of the sample statistics vary from sample to sample, i.e., they are also variables. Second, the sample statistic formula used to estimate the value of the population **parameter** is called its *estimator*. For example, population mean $\mu$ is estimated by sample mean $\bar{x}$—thus, $\bar{x}$ is the estimator of $\mu$. Generally, the corresponding sample analogue is the estimator of any population parameter. Sample SD is the estimator of the population SD, and sample median is the estimator of the population median. Whether or not they are good at providing an adequate estimate is another question.

The validity of statistical estimators is generally assessed in terms of its bias [see **bias (statistical)**] and is different from **precision**. Unbiasedness is the ability of an estimator to reach the correct value of the parameter when the average of the values of the statistic over all possible samples of that kind is the same as the value of the population parameter. Statistically, this average of the statistic over all possible samples is called the *expected value*. A statistic is called an unbiased estimator of the parameter when it is established that the statistic has a tendency to reach the true value of the parameter when averaged over repeated samples. An unbiased estimator of parameter $\tau$ is the statistic whose average over different possible samples is the same as the correct value of $\tau$. Technically, this is stated as that a statistic is an unbiased estimator of a parameter when the expected value of the estimator is the same as the value of that parameter. Unbiasedness is one of the important criteria to judge whether an estimator is good or not. For most distributions, an unbiased estimator can be theoretically established without worrying about drawing so many samples.

An unbiased estimator easily can be illustrated with the help of a hypothetical example. Suppose you know for sure from a complete enumeration of a population that 9% of all women of age 30–39 years in an area have raised blood sugar (BS) level. Now, you take a sample of 200 women of that age from that population and find that 15 (7.5%) of them have raised BS level. You take another sample of 200, some of which may be the same women as in the first sample (called **sampling with replacement**), and find that the percentage now is 10.0. The third sample gives 8.5%, the fourth 11.5%, the fifth 9.0%, etc. The big question is what the average of such percentages will be when all possible samples are considered. If the average of all possible samples is 9%, which is the same as the true percentage in the population, the sample percentage is an unbiased estimator of the population percentage; otherwise, it is not. In brief, the sample mean is an unbiased estimator of the population mean in the case of random samples, but the sample median is generally not an unbiased estimator of the population median except for symmetric **distributions**. The sample variance is an unbiased estimator of the population variance only when the denominator is $(n - 1)$ instead of $n$. This is the reason that $(n - 1)$ is used in place of $n$ in this case.

When sampling is with replacement, as in this example, note that an infinite number of samples can be drawn. It is not practically feasible to study all possible samples. Thus, the help of theory is taken to find out whether an estimator is unbiased or not. Theoretical derivations have established that sample mean is an unbiased estimator of the population mean provided that the sample is drawn randomly. This does not hold for nonrandom samples. This explains why random samples are important for statistical methods. Sample variance is not an unbiased estimator of the population variance unless the divisor is $(n - 1)$ instead of $n$. Strange as it may sound, sample SD is not the unbiased estimator of the population SD even when the denominator is $(n - 1)$. This "aberration" occurs due to the square root sign in the SD, which does not occur in the variance. Sample proportion (or percentage) is an unbiased estimator of the population proportion. Sample median is an unbiased estimator of population median when the distribution is Gaussian but not necessarily so for other distributions. A qualifier in all these statements is that the samples are random.

## unbiased sample, *see* biased sample

## uncertainties, *see* aleatory uncertainties, epistemic uncertainties, medical uncertainties, parameter uncertainty

## uncertainty analysis, *see also* sensitivity analysis

Uncertainty analysis is the process to measure the impact on the result of changing values of one or more key inputs about which there is uncertainty. The study of the effect of varying the value of parameters included in the analysis is called **uncertainty analysis**. The need for this mostly arises due to the use of sample estimates for the parameters of the model. For uncertainty analysis, the inputs are varied over a reasonable range that can occur in practice. Uncertainty analysis can also be viewed as a method for checking the robustness of the results that would be established if the result does not change much when input values are changed within reasonable limits.

Models, particularly statistical models, are not unique. Different sets of parameters may reasonably reproduce the same sample values. Thus, agreement between a model and the observed data does not imply that the model assumptions accurately describe the underlying process, only that it is one of the plausible explanations and empirically adequate. Thus, the model needs to be checked under varying conditions. While **sensitivity analysis** refers to the study of the effect of changes in the basic premise, such as individual and

societal preferences or the assumptions made at the time of model development, uncertainty analysis is for the study of the effect of varying the value of estimates of the parameters included in the analysis. In essence, sensitivity analysis is primarily for **epistemic uncertainties**, whereas uncertainty analysis is mostly for **aleatory uncertainties**.

As is evident, uncertainty analysis pertains to **parameter uncertainty**. For example, in calculation of **disability-adjusted life years (DALYs)**, such parameters are incidence, prevalence, and duration of disease and mortality rates for various health conditions. The value of DALYs will naturally change if any of these are changed. Correct estimation of these parameters is vital to the validity of the DALYs obtained. The uncertainty surrounding these parameters can be reduced through more accurate measurement and by adopting a scientifically sound methodology of estimation. This is not possible for the preferences as required for sensitivity analysis.

In medicine, uncertainty analysis is generally done for a model that relates an outcome with its antecedents. For example, urinary excretion of creatinine can be reasonably predicted by age, dietary constituents, and lean body mass. If age is reported approximately in multiples of 5, such as 45 years instead of the exact 43, will the model still be able to predict nearly the correct value of urinary creatinine? If this example is not convincing, consider the effect of changes in diet from day to day on the urinary excretion of creatinine. If the outcome prediction remains more or less the same despite this change, the model is considered robust.

Uncertainty analysis incorporates three types of aleatory variations in the inputs: (i) First are random measurement errors, such as approximate age in the creatinine example. (ii) Second, natural variation that can occur in the input parameters such as dietary constituents can change from one day to another. If the model is based on the average diet over a month, the model may or may not be able to reflect the effect of daily changes. If it is based on diet of the previous day, what happens if the diet changes the next day? (iii) Third is variation in the multipliers used in the prediction. If ln(creatinine in mmol/day) is predicted as (0.012 × height in cm − 0.68) in children, what happens if the multiplier of height is 0.013 or 0.011? The multiplier 0.012 is an estimate, which is subject to sampling fluctuation, and this multiplier can change depending upon the subjects that happen to be in the sample. For further illustration, consider the following example.

Similar to the DALYs example cited earlier, consider the example of calculation of **health-adjusted life expectancy (HALE)** for any population. This requires three inputs: (i) life expectancy at each age, (ii) estimates of prevalence of various nonhealthy conditions at each age, and (iii) a method of valuing nonhealthy period in comparison to full health. All three are estimates and have built-in sampling and other variation. When the life expectancy of females at birth in Brazil is estimated as 72 years in 2015, it could actually be 71 years or 73 years. Life expectancy is relatively robust, but such variation is more prominent in prevalence rates of various diseases. Many of these prevalences are not known, and they are estimated by indirect methods. A statistical distribution can be imagined around all such estimates. Salomon et al. [1] have described a method to calculate the uncertainty interval around HALE for each member country of the World Health Organization based on such a premise. They used computer simulations for generating statistical distributions around input values and thus propagating uncertainty intervals.

An uncertainty interval is very different from the confidence interval (CI). In our HALE example, the CI would be based exclusively on the standard error (SE) of HALE without considering the variation in input values, and the SE can be reduced at will by increasing the sample size. The uncertainty interval would be much larger as it incorporates the effect of variation in each of the inputs and will not depend on the sample size. This variation depends on the repeatability of instruments, number of measurements, and other sources of variation that can contribute to disagreement between the predicted and the actual result. For a model reported by Xiridou et al. [2] that described the transmission of HIV and chlamydia among men who have sex with men (MSM) in the Netherlands, uncertainty analysis was carried out for model parameters without establishing the validity of the model.

Uncertainty analysis is strongly advised if the proposed model has serious consequences. This is also recommended when it is necessary to disclose the potential bias associated with models that use a single value of the parameters, particularly when your calculations indicate the need for further investigation before taking any action. Our discussion is in the context of health outcomes, but the major application of uncertainty analysis has been in the context of costing. Health care cost can substantially vary depending on the quantity and quality of various inputs, and uncertainty analysis helps to delineate the range of costs for varying inputs.

1. Salomon JA, Mathers CD, Murray CJL, Ferguson B. Methods for life expectancy and healthy life expectancy uncertainty analysis. *Global Programme on Evidence for Health Policy Working Paper No. 10*. World Health Organization, 2001. http://www.who.int/healthinfo/paper10.pdf
2. Xiridou M, Vriend HJ, Lugner AK, Wallinga J, Fennema JS, Prins JM, Geerlings SE et al. Modelling the impact of chlamydia screening on the transmission of HIV among men who have sex with men. *BMC Infect Dis* 2013 Sep 18;13:436. http://www.biomedcentral.com/content/pdf/1471-2334-13-436.pdf

## uncertainty principle, *see also* equipoises

In the context of clinical trials, the uncertainty principle espouses that the researcher is genuinely uncertain about the outcome of the trial for the type of patients proposed to be enrolled. This is considered a moral prerequisite for proceeding with a trial. This principle also applies to the individual patients and says that the outcome must be uncertain in the patient chosen for the trial. If the researcher already knows which regimen is better, rightly or wrongly, there is no point in conducting the trial. Thus, the uncertainty principle provides a valid basis not just for random allocation of the patients but also for conducting the trial.

The term *uncertainty principle* actually belongs to quantum mechanics in the context of the position and momentum of a particle, where if one is known with high precision, the other will not. This can be directly applied to tissue cells, such as done by Barbieri et al. [1], who adopted the uncertainty principle associated with the task of detection of position and orientation as the main tool to provide quantitative bounds on the family of simple cells concretely implemented in the primary visual cortex. In clinical trials, uncertainty about the outcome is helpful in producing valid results without worrying much about other aspects. Using the uncertainty principle should allow the process of providing information and gaining consent to become much closer to what is appropriate in normal medical practice [2]. Although the principle has been adopted in some trials, it is many times impossible to gauge whether this is genuine or just a pretense by a particular trialist, since this could also be based on wrong information or wrong assessment. That the concept is trialist centered is among the severest criticisms of this principle.

The uncertainty principle is intimately related to the concept of **equipoise**. This could be personal equipoise similar to the uncertainty

just discussed, clinical equipoise for collective uncertainty of the medical community, or patient equipoise. For a debate on the uncertainty principle and clinical equipoise, see Weijer et al. [3].


1. Barbieri D, Citti G, Sarti A. How uncertainty bounds the shape index of simple cells. *J Math Neurosci*. 2014 Apr 17;4(1):5. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3991892/
2. Peto R, Baigent C. Trials: The next 50 years: Large scale randomised evidence of moderate benefits. *BMJ*. 1998 Oct 31;317(7167):1170–1. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1114150/
3. Weijer C, Shapiro SH, Glass KC. Clinical equipoise and not the uncertainty principle is the moral underpinning of the randomised controlled trial. *BMJ* 2000 September 23;321:756–8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1127868/


## under-5 mortality rate, *see* mortality rates

## uniform distribution

A variable $x$ is said to have a uniform distribution when all its values within a specified range have the same chance of occurrence. The simplest illustrations are tossing a coin and throwing a die, where each face has the same chance. You can say that sex at birth has a uniform distribution, with equal probability for male and female, and possibly, incidence of cancer is uniformly distributed over various months of a year. The latter in effect means that season does not increase or decrease the chance of onset of cancer. Uniform distribution can be used to check whether or not the event of interest is equally likely to occur within a range. For example, one may want to investigate whether myocardial infarctions (MIs) are uniformly distributed over the 24 h period in a day, whether there is preponderance of any particular time such as early morning [1], or whether a particular material such as EndA is uniformly distributed in the membrane of cells and not localized [2]. It can also be used to check **digit preference**.

Graphically, a uniform distribution for a continuous variable looks like a line over its range $(c, d)$ (Figure U.1a), and has an obvious mean $(c + d)/2$ and not-so-obvious variance $(d - c)^2/12$. Compare this with, for example, a **Gaussian distribution**, which has a peak in the middle and tapering off on both the sides. Here, there is no peak and no tapering off, but it, too, is symmetric. If the variable is discrete, it will be depicted by bars (or spikes) (Figure U.1b) with as many bars as its possible values. If the number of possible values is $K$, the mean is $(K + 1)/2$, and the variance is $(K - 1)(K + 1)/12$. If the 24 h period is divided into 8 intervals of 3 h each for investigation of preponderance of MI in any time interval, there would be $K = 8$ such bars, one for each interval, and the number (or the proportion) of MIs in different time intervals would be on the vertical axis. If the distribution is uniform, the mean time of MI would be $(K + 1)/2 = 9/2$, or the 4.5th time interval, i.e., 12 noon when the time is measured from 00:00 hours. If the mean time of MI is substantially different from this, you know that the distribution of time of MI is not uniform over the 24 h period. This example also illustrates how the mean could be an absurd quantity in some setups.

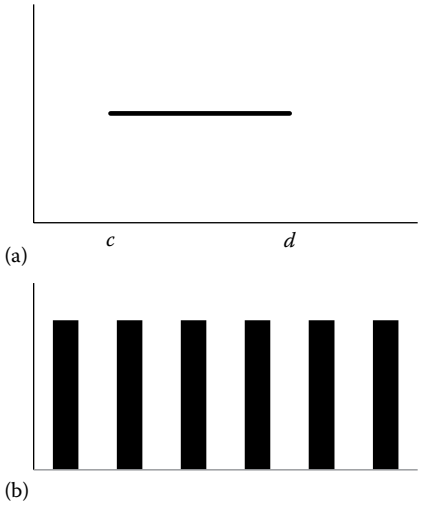



**FIGURE U.1** Uniform distribution: (a) continuous; (b) discrete.


1. Kanth R, Ittaman S, Rezkalla S. Circadian patterns of ST elevation myocardial infarction in the new millennium. *Clin Med Res* 2013 Jun;11(2):66–72. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3692391/
2. Bergé MJ, Kamgoué A, Martin B, Polard P, Campo N, Claverys JP. Midcell recruitment of the DNA uptake and virulence nuclease, EndA, for pneumococcal transformation. *PLoS Pathog*. 2013;9(9):e1003596. http://www.plospathogens.org/article/info%3Adoi%2F10.1371%2Fjournal.ppat.1003596


## union and intersection of events, *see* Venn diagrams

## unit (statistical)

Unit is an element that is complete in itself in some sense, although this can be a part of a system or a process. In the context of medicine, we can think of a unit of investigation, e.g., blood or urine, a person or a patient for an inquiry, an organ such as an eye or bone for examination, an area or a family such as for sampling, etc. We also have a unit of measurement, such as milligram per deciliter for cholesterol level, and a unit of analysis, such as a test group and a control group in a clinical trial. Important among these for biostatistics are the unit of inquiry, unit of sampling, and unit of analysis.

### *Unit of Inquiry*

The **unit of inquiry** is the entity on which the information is obtained. In an investigation on household pollution, a household will be the unit of inquiry; for infant mortality rate (IMR), a city or a state can be the unit of inquiry; for growth of children, a child is the unit of inquiry; for cancer survival, a patient would be the unit of inquiry. The database will contain information on each unit with no further breakdown for subunits. For example, if the unit of inquiry for IMR is a state, no city-wise breakdown would be available. But city can also be the unit of inquiry for IMR. Thus, the unit of inquiry should be carefully decided depending on the objectives of the study.

### *Sampling Unit*

A **sampling unit** is that which is used for selection of the subjects. In a community survey on protein energy malnutrition, the sampling unit could be a family, but the unit of inquiry could be a child younger than 5 years. One sampling unit can have multiple units or no unit to inquire on. Sometimes, the sampling is done in stages, such as selection of some hospitals in the first stage, wards or departments within the selected hospitals in the second stage, and then patients in the selected wards in the third stage. These are the sampling units at

various stages. The unit of inquiry could be a patient, but sampling units are multiple in this kind of sampling.

In the case of multistage sampling, the first-stage sampling unit is generally termed **primary sampling unit (PSU)**. For example, in the US National Health and Nutrition Examination Surveys, the PSU is typically a county [1]. The size of the PSU and the **intraclass correlation** among subunits within the PSUs can substantially alter the variance of the estimates.

### *Unit of Analysis*

This is the basic unit for which the results would be available. For example, for an average, the unit is a group—there must be individual measurements within each group. The term *group*, in a generic sense, incorporates any aggregation; for example, it could be the average of one person over time when repeated measurements are taken, or a school whose children have been studied for, say, vision problems. A single value cannot be the unit of analysis, as this cannot be "analyzed," although a single value could be the unit of observation, such as an outlier, which may be an important observation by itself. An outlier by itself does not involve any statistical analysis.

The unit of analysis intimately depends on the objectives of the study. In a clinical trial, the treatment and control groups are the conventional units of analysis in the sense that all estimates are obtained for these two groups, but it can also be different time points in the case of repeated measures, such as to find the time with maximum response. Other examples of unit of analysis are age group, sex, disease severity group, cases responding early to a treatment, and vaginal swabs from women aged at least 50 years. In a multistage sampling, such as a state, county, and family, the unit of analysis may be any or all of these for, say, computing averages or generating estimates: this setup may require **hierarchical models** for analysis.

1. Johnson CL, Dohrmann SM, Burt VL, Mohadjer LK. National Health and Nutrition Examination Survey: Sample design, 2011–2014. *Vital and Health Statistics* March 2014;2(162):4. http://www.cdc.gov/nchs/data/series/sr_02/sr02_162.pdf

## univariate analysis

This analysis considers one variable at a time. Contrary to general perception, this refers to the **stochastic variable**: "fixed variables" can be many in a univariate analysis. In a classical regression setup, all the regressors are considered fixed, and only the dependent variable is considered stochastic. Thus, even if you have $K$ regressors, the analysis is still univariate. For $K \geq 2$, this is called multiple regression and not multivariate regression. The same is true with logistic regression, although you may occasionally find this wrongly referred to as multivariate logistic regression in the literature, particularly the old literature, when the number of regressors is more than one. The following example illustrates a univariate analysis.

If the presence or absence of kidney disease is being studied as dependent on creatinine level, glomerular filtration rate (GFR), age, and sex, this would be a univariate logistic regression. Similarly, if creatinine level, GFR, and blood urea nitrogen are studied separately as dependent on age, sex, and weight, each of these is a univariate analysis. It becomes multivariate when all three kidney functions are studied together in a holistic manner. Examples of multivariate analysis methods are cluster analysis, factor analysis, multivariate analysis of variance (MANOVA), and multivariate multiple regression. Chi-square can be used for both multivariate and univariate analyses, but Student $t$-test, analysis of variance, and ordinary and logistic regression are univariate analysis methods unless specifically mentioned as multivariate.

Results obtained by a combination of univariate analyses could be very different from the results of multivariate analysis because the multivariate methods also consider the correlation between the variables, which univariate analysis does not do. However, univariate analyses are relatively very simple to do, are easy to interpret, and in many cases provide focused conclusions. For example, if three kidney functions are studied together in a multivariate setup and the dependence is statistically significant, it would be difficult to say which of the three kidney functions is affected and which not. Univariate results would answer this kind of question but fail to account for the correlations between different kidney functions. Thus, in some situations, both analyses may be required to give a comprehensive account of what is going on.

## universe (statistical), *see* population (the concept of)

## UPGMA method of clustering, *see* average linkage method of clustering

## unweighted mean, *see* weighted mean

## up-and-down trials

An up-and-down trial is used for calibrating the dose where a series of patients is involved in the rule of stepping up the dose by a prefixed amount following a negative response and decreasing the dose following a positive response. The prerequisite is that the higher dose of the regimen is more likely to give a positive response. This can also be used for dose escalation studies as in a phase I cancer trial. The trial can be carried out on one patient also, provided that the validity conditions of ***n*-of-1 trials** are met. The first test should be performed at a dose close to your guess of the median effective dose. The method is also known as *sensitivity experiment*, and also as the Dixon and Massey method, although this is believed to have been introduced by Dixon and Mood [1] in 1948. The up-and-down methodology is rarely used for the usual drug trials because of the risk of an unacceptable level of toxicity. For details, see Le Tourneau et al. [2].

Consider a trial starting with 7 mg of intrathecal hyperbaric bupivacaine for an anesthetic effect in lower limb surgeries that failed to produce a satisfactory response in a particular patient. If the increment fixed in advance is 1.5 mg, the next patient will get 8.5 mg. If that also fails, the next will get 10 mg. If that succeeds, the next will now get a lower dose, 8.5 mg. If that also succeeds, then next will get 7 mg. Success will decrease the dose, and failure will increase the dose each time by the fixed margin (Figure U.2). Some clinicians may find the up-and-down method very convincing for identifying a minimum critical dose that is effective. Although the dose in this method is increased or decreased by a fixed margin, the up-and-down method helps estimate the exact mean dose, which could be in between those tried. In most situations, the dose quickly settles near the mean or median. The method generally requires fewer tests than a method with groups on preassigned doses.

As just mentioned, for this kind of trial, fix a dose step that you think would be appropriate escalation or reduction at successive steps, and also guess a median dose using your clinical acumen or past experience. For stepping up or down, generally, 1 standard

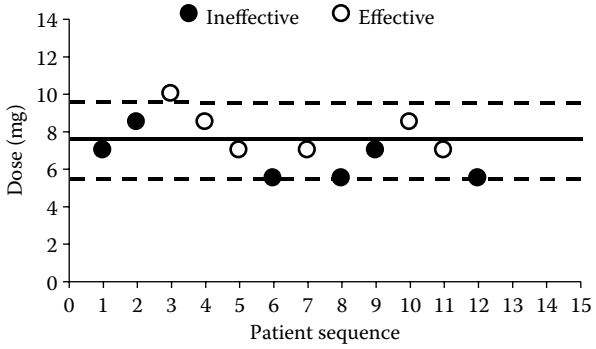



**FIGURE U.2** A trial with up-and-down method.

deviation (SD) works fine, particularly for a Gaussian distribution, after taking the logarithm of dose. Give the anticipated median dose to an eligible patient and assess if it is effective. If effective, step down the dose by the prior fixed step; if not effective, step up the dose for the second patient; and so on. An essential prerequisite for this kind of trial is that such variation in dose should not be unethical.

The trials can be conducted on as many patients as decided in advance, but generally, a trial on 10–15 patients is considered adequate. In most cases, the dose will converge to a definite dose that is the minimum to achieve a positive response. In any case, once the data are available for successive doses and success–failure information, they can be analyzed to estimate least mean effective dose, and confidence interval (CI) can also be obtained using the method mentioned in a short while. This type of trial cannot easily estimate other locations, such as $ED_{95}$—the dose effective in 95% cases. But one drug can be compared with another for the least mean effective dose; for example, hyperbaric bupivacaine can be compared with isobaric bupivacaine, and the one with a lower mean dose can be recommended.

The following method can be used to analyze data from an up-and-down trial when the response is of the yes/no type. The analysis is in terms of estimating the least effective dose and obtaining a CI. Generally, the least effective dose in this case is defined as the same as the median effective dose. The procedure for obtaining the median effective dose is as follows:

Step 1. Calculate the logarithm for all the doses you tried from minimum to maximum. Find the mean of the differences of successive log-doses. Denote this mean difference by $\bar{d}$.

Step 2. Separately list doses that were effective in different patients after excluding the ineffective ones. Calculate the mean of the logarithm of these doses and SD of these logarithms. Denote them by $\bar{y}$ and $s_y$, respectively.

Step 3. Median effective dose = $\exp(\bar{y} - 1/2 * \bar{d})$.

Calculation of the CI requires a constant *G*, which is obtained from a figure given by Dixon and Massey [3]. See this reference for the method to obtain the CI. The following example illustrates the method.

There is considerable uncertainty regarding the minimum dose requirement of local anesthetics administered intraspinally for surgery. Sell et al. [4] conducted a study to estimate the minimum effective dose of levobupivacaine and ropivacaine in hip replacement surgery. Forty-one patients were randomly allocated to one of the two local anesthetic groups in a double-blind manner. The authors used the up-and-down strategy, determining the initial dose based on previous experience. The minimum effective dose was defined as the median effective dose. The up-and-down method found the minimum effective dose of levobupivacaine to be 11.7 mg (95% CI, 11.1–12.4) and that of ropivacaine 12.8 mg (95% CI, 12.2–13.4). The authors concluded that these doses are smaller than those reported earlier for single-shot anesthesia. The overlapping CIs indicate that the doses of the two anesthetic agents under this trial are not significantly different.

Dixon [5] has presented a modified method that should work for small samples under which median = (final test level) − *k*(step-up dose), where *k* is the value from the table he provided in his paper.

1. Dixon WJ, Mood AM. A method for obtaining and analyzing sensitivity data. *J Amer Stat Assoc* 1948;43:109–26. http://www.tandfonline.com/doi/abs/10.1080/01621459.1948.10483254?journalCode=uasa20
2. Le Tourneau C, Lee JJ, Siu LL. Dose escalation methods in phase I cancer clinical trials. *JNCI Journal of the National Cancer Institute*. 2009;101(10):708–20. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2684552/
3. Dixon WJ, Massey FJ Jr. *Introduction to Statistical Analysis*, Second Edition. McGraw Hill, 1957: p. 324.
4. Sell A, Olkkola KT, Jalonen J, Aantaa R. Minimum effective local anaesthetic dose of isobaric levobupivacaine and ropivacaine administered via a spinal catheter for hip replacement surgery. *Br J Anaesth* 2005;94:239–42. http://bja.oxfordjournals.org/content/94/2/239.long
5. Dixon WJ. The up-and-down method for small samples. *J Am Stat Assoc* December 1965;60(312):967–78. http://www.meduniwien.ac.at/typo3/ismed/fileadmin/hirnforschung/teaching/Testfolder/Testdokument2.pdf, last accessed December 20, 2015.

## U-shaped curve/distribution, *see also* bathtub curve/distribution

This refers to the curve obtained when the response (*y*) is high for low values of the antecedent (*x*), levels off to a low level for middling values of *x*, and is high again for high values of *x* (Figure U.3a). This kind of relationship has been observed by Merghani et al. [1] between exercise and cardiac morbidity, where morbidity is extremely high when exercise is absent or very low and also when exercise is very strenuous. This is similar to the **J-shaped curve**, but now, the low tip of *J* on the left rises to make a *U* shape.

Figure U.3b is an inverted U-shaped curve between age and lung function, showing that lung function increases with age to a limit, plateaus between ages 30 and 40 years, and then declines with ageing. This curve is not exactly inverted U-shaped and may apparently look Gaussian, but it is not so, because of the flat peak and the sharp rise and decline on either side. Moreover, the term *Gaussian* is used for a statistical **distribution**, while here, we are discussing just a relationship between any two variables. It becomes a distribution when the vertical axis is frequency or proportion, which is not the case with either of the two curves we have shown. However, Leboeuf-Yde et al. [2] have reported a U-shaped distribution of "percentage of weeks in one year with bothersome low back pain in all 7 days" by patients with chronic low-back pain in Denmark. The distribution of patients reporting a low percentage of weeks was very high, and that of patients reporting a high percentage was also very high, while that of patients reporting pain in 30–70% of weeks (15–35 weeks in a year) was low. This, in effect, means that the patients with chronic low-back pain mostly had either a few weeks of persistent pain in a year or had many weeks of persistent pain.

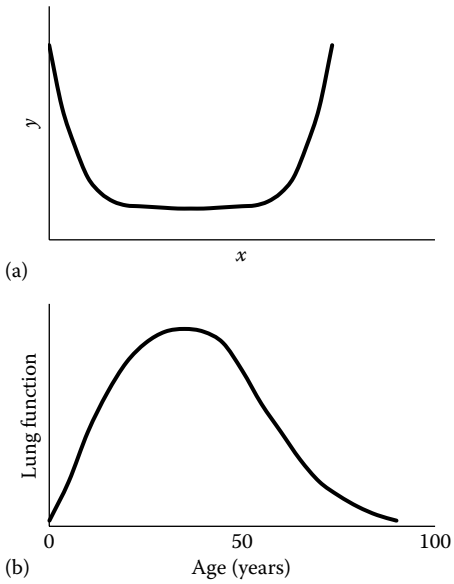



**FIGURE U.3** (a) A U-shaped curve between exercise and cardiac mortality and (b) an inverted U-shaped curve between age and lung function.

Davis and Svendsgaard [3] studied risk assessment implications of a U-shaped dose–response curve and concluded that this kind of curve raises important issues in the identification of no-observed-effect levels and trade-offs between potential risks and benefits of the intervention, where *no observed effect* pertains to the low levels in the U-shaped curve.

1. Merghani A, Malhotra A, Sharma S. The U-shaped relationship between exercise and cardiac morbidity. *Trends Cardiovasc Med* 2015 Jun 18. pii: S1050-1738(15)00171-1. http://www.ncbi.nlm.nih.gov/pubmed/26187713
2. Leboeuf-Yde L, Jensen RK, Wedderkopp N. Persistence of pain in patients with chronic low back pain reported via weekly automated text messages over one year. *BMC Musculoskeletal Disorders* 2015; 16:299. http://www.biomedcentral.com/1471-2474/16/299
3. Davis JM, Svendsgaard DJ. U-shaped dose–response curves: Their occurrence and implications for risk assessment. *J Toxicol Environ Health* 1990 Jun;30(2):71–83. http://www.ncbi.nlm.nih.gov/pubmed/2192070

## U-test, *see* Wilcoxon rank-sum test

# V

## vaccine trials, *see* clinical trials

## validation sample/study, *see also* validity, robustness

In the context of biostatistics, validation is the process of confirming the adequacy of a medical tool or of the results of a medical study for the specified application. The tool can be a **questionnaire**, a diagnostic test or criterion, a **scoring system**, a treatment strategy, or any such device proposed to be used in a particular setup. Study results also are validated for their applicability to a particular setup.

As is evident, validation is done not just for newly discovered tools and procedures but also for established tools when they are proposed to be applied to a new setup. The new setup, where a particular tool is proposed to be used, may have some peculiarities, and you may not have sufficiently convincing previous evidence of applicability of these tools to this new setup. If so, it is desirable, if not essential, that the tool be checked for its **validity** to perform well in this particular setup.

Among the procedures available for validation, the simplest is *peer validation*. This may be of two types: (i) find out from other potential users about their experience in a setting similar to the new one where the tool is proposed to be utilized, and (ii) look at a full-scale or pilot study of other researchers who may have carried out validity checks of similar tools in a similar setup. Often, the workers who develop the tool also do some kind of validation. They may do *internal validation* by splitting the available sample of subjects in two equal or unequal parts—use one for developing the tool and the other for validation, sometimes called the training set and the validation set, respectively. The tool developed on the first part of the sample is tried on the second part of the sample to check if it is working satisfactorily. Each part must be reasonably large and adequate representative to generate confidence. Since internal validation is on part of the same sample, the tool under test may give good results even when it is not really as good for applications elsewhere. Thus, *external validation* is preferred where an independent validation study is done on another sample drawn from the same milieu—sometimes even from a different milieu—to check if the tool works in altered settings or not. In this case, this other sample is the validation sample. If your own setup does not reasonably match with the setup of the sample on which the tool was developed or validated, it is desirable to carry out your own validation study on a random sample of subjects for whom this tool is intended. For example, the McGill pain score has been validated to work in different languages and different setups.

See **validity** for further details.

## validity, *see also* validity (types of)

The validity of a tool is its ability to assess correctly or its appropriateness for use in a particular setup. In medicine, the term is used for devices, factors, data, designs, estimates, evidence, tests, results, scoring systems, studies, and a host of other tools. Do not confuse this with accuracy. Accuracy is generally for a single measurement, whereas validity is the ability to hit the target on average.

There is some confusion in the literature about the terms *validity* and **reliability**. The difference between the two is aptly illustrated in Figure V.1. Reliability is hitting the same point in repeated attempts, although this may be far away from the target (see third panel). Validity is hitting around the target (second panel). If hits are close together and near the target every time, both reliability and validity are high (fourth panel). Lack of validity can be due to systematic bias or for any other reason. See **bias in medical studies and their minimization** for further details on the types of bias that can affect validity. If the bias is small, we say that the tool has high validity, and if the bias is large, we say that the validity is low.

What signs and symptoms or hematological parameters (such as hematocrit) can correctly identify cases of dengue fever in early stages (because serological evidence takes time)? Is BP > 140/90 mmHg the right definition for hypertension, or is BP > 160/95 more correct? Both definitions are used. (See Refs. [1,2]. These, incidentally, are nearly consecutive articles in the same journal but with varying definitions.) This means that those with BP = 150/92 are considered hypertensive by one definition but not by the other. How can one distinguish, without error, a hypothyroid condition from euthyroid goiter? How can one correctly ascertain the cause of death in an inaccessible rural area where a medical doctor is not available? How can age be correctly assessed when a dependable record is not available? Can the diet content of an individual be adequately assessed by taking a 3-day history? How can physical activity be validly measured? Many other examples can be cited where validity is an important concern.

Validity is context specific; for example, high validity of a tool in hospital A does not necessarily mean that this will be so in hospital B as well. A highly valid tool in better-nourished children may not

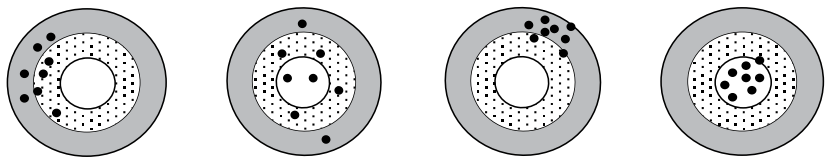



**FIGURE V.1** Darts illustrating the difference between validity and reliability.

work well in undernourished children. It intimately depends on the case mix. Thus, validity is qualified for the type of subjects evaluated.

### *Validity of Diagnostic Tests*

In practice, no medical test is 100% perfect. A computed tomography (CT) scan can give a false-negative or false-positive result, and a negative histologic result for a specimen is no guarantee that proliferation is absent, although in this case, positive **predictivity** is nearly 100%. The values of measurements such as creatinine level, platelet count, and total lung capacity are indicative of different medical conditions rather than absolute; i.e., they mostly estimate the *likelihood* of a disease. Signs and symptoms seldom provide infallible evidence and have to be supplemented with other clinical information to come to any conclusion worth implementing.

Because all these tools are imperfect, decisions based on them are also necessarily probabilistic rather than definitive. Their validity in a particular setup is generally measured in terms of indices such as **sensitivity and specificity** and positive and negative **predictivities** (see these topics for details). These terms are applied not just to conventional laboratory and radiological tests but also to other clinical evaluation tools, such as signs and symptoms, and attitudes and behavior. These indices measure the validity of these tools on a scale of 0 to 1, with 0 indicating that they are not helpful at all and 1 indicating that they are perfect in reaching a right conclusion. The details under the respective topics will tell you that sometimes a highly sensitive test and sometimes a highly specific test is more valid for a particular application. Also, sensitivity and specificity are evaluated on subjects who are already known to have or not have a disease or condition, whereas predictivities provide a valid measurement of the diagnostic utility of tools in a local context. You may also want to review the topic **receiver operating characteristic (ROC) curves**; these provide information on overall performance of quantitative tests on the basis of sensitivity–specificity of the quantitative tests at different cutoffs.

### *Validity of Medical Information*

Besides diagnostic tools, other "tools" used in health and medicine include information on a patient or a person obtained in a variety of ways—previous records, interview, examination, and investigations. None provides perfect information that would be right in all cases all the time. Which method of eliciting information provides more valid information regarding the health of a person?

Laboratory and radiological investigations are generally believed more than others as they are objective and have the least human element. However, validity of laboratory results, for example, depends on whether or not the apparatus, reagents, and procedures used are fully standardized. Even then, as mentioned earlier, they do provide **false-negative and false-positive** results in some cases. Moreover, some clinicians believe that nothing surpasses the feeling they get of the condition of the person by clinical examination. For them, investigations are just for support or confirmation and do not have much value by themselves. This is because, first, laboratory results can be in error; second, some people have so-called abnormal levels even when they are completely healthy; and third, health is more of a perception of the affected individual than his/her organic state. The same is generally true for radiological investigations. In most cases, the holistic picture obtained by the combination of clinical information and investigation results provides much more valid information about the condition of the patients.

The preceding discussion may have given you an idea of how difficult assessment of validity is. Each tool has multifarious aspects, and its performance in providing correct information depends on a host of features. Consider, for example, forms for eliciting and recording information, as commonly used in working up patients in clinics and most research studies, including surveys. A form could be a **questionnaire**, **schedule**, **or proforma**. See this topic for their relative merits and demerits. The validity of the information obtained by these tools depends on, among others, their contents. For example, the burden of smoking is assessed generally by the duration and number of cigarettes smoked. Sometimes, information on the type of smoking (cigarette, cigar, pipe, water pipe, bidi, etc.) is also obtained. In some rare cases, the age at initiation and time elapsed since quitting by ex-smokers is also obtained as done for the Global Adult Tobacco Surveys in different countries [3]. But how to validly combine all these aspects of smoking into a single index is a question addressed by the **Indrayan smoking index**.

Validity of response also depends on interview technique, patient cooperation, biases and prejudices of both the interviewer and the interviewee, rapport between them, and the like. It also depends on the sequence of questions, the length of the questions and of the questionnaire, their wording, etc., besides the statistical **scale** (metric, ordinal, nominal polytomous, or nominal dichotomous) used to elicit and record the response. For example, a **Likert scale** is commonly used to elicit opinion or satisfaction with gradation from most adverse to most favorable. The response on this scale may differ if you use scores from 0 to 6 compared to scores from –3 to +3 despite both being 7-point scales. This illustrates how the concept of validity becomes intricate as we go deeper.

In the context of a questionnaire or a schedule, validity is assessed in terms of, for example, its ability to distinguish between a sick and a healthy subject or a very sick and a not-so-sick subject. This means that the scores received or the responses obtained from one group of subjects should be sufficiently different from those of another group when the two groups are known to be different with respect to the characteristic under assessment. Consider a questionnaire containing 20 items measuring quality of life. If the response on one particular item, say on physical independence, from healthy subjects is nearly the same as from mentally retarded subjects, this item is not valid for measuring quality of life in those types of subjects, and the item can be deleted. A shorter questionnaire is always better than a longer questionnaire—thus, deletion always helps. But the number of items finally left should be adequate to assess all the domains of the construct under evaluation.

### *Validity of the Design, Variables under Study, and Assumptions*

Medical researchers commonly face the problem of validity of their design, variables to be studied, and assumptions. When any of them lacks validity, the results will be affected.

A valid design is one that would rightly answer the question the study is intended to answer. For example, a **one-way design** cannot provide the right answer on the effect of two **nominal** factors on a **quantitative** outcome. Kaestner [4] found that the preintervention versus postintervention design with a **control group** for studying the effect of state Medicaid expansion on mortality in low-income adults is invalid because prior trends in the treatment and control states differed significantly. A **case–control** design, for that matter, any **observational study**, is generally not valid for reaching a **cause–effect relationship** as this type of study provides associational results except when a set of other strict criteria is met.

Similarly, one has to choose the right variables or factors to get valid results from a study. For example, for studying correlates of an outcome such as duration of hospitalization, severity of condition

at admission by itself may not be valid, and you may have to consider other factors also such as age, treatment strategies, nursing care, and overall health of the person for response to the treatment. Operational definitions of each of these variables should also be valid. For example, nursing care could be assessed in terms of responding to the specific needs of the patient throughout the stay, sticking to the prescribed regimen, proper entries in bedhead tickets, etc. If something is missing or not properly recorded, validity will suffer in the sense that the right answers to the questions of what correlates affect the duration of hospitalization and by how much will not be obtained. For statistical methods such as regression, the choice of regressors is a big consideration in getting valid results. For example, Creemers et al. [5] selected a core set of valid variables for ankylosing spondylitis activity out of a large number of possibilities for developing a disease activity score.

Assumptions for a particular method are requirements under which that method gives valid results. This applies to statistical methods as much as to most other methods. For example, an important requirement for analysis of variance (ANOVA) is that values have the same variance in different groups. For running a regression, linearity of the relationship and Gaussian distribution of the residuals is a frequent assumption. Even nonparametric methods require independent values. Different statistical methods have different assumptions, and checking the validity of these assumptions is an important part of the statistical analysis. We have presented methods for checking their validity.

### *Validity of Data, Measurements, Scoring Systems, Estimates, Statistical Methods, and Models*

Data are considered valid when they reflect the true status. If a person's age is 36 years and is inadvertently recorded as 63 years, if a male person is shown to be pregnant, or a woman of age 27 years married 4 years ago is shown to have a daughter of age 7 years from the same marriage, the error is obvious. In a hospital setup, this can occur when, for example, the date of surgery is recorded as before the date of admission, liver abscess is wrongly shown as biliary disease, diagnosis and type of surgery are entered in switched columns, etc. If a person's sex in wrongly entered, the error may not be detectable at all at the time of analysis—thus, due care is required at the time of data entry. **Missing data** and intentional wrong reporting as in some medicolegal cases also reduce the validity of the data.

For validity of a measurement, realize that many characteristics can be measured in a variety of ways. Choosing the right index that measures a characteristic correctly can be difficult in some setups. Consider measuring obesity in the context of its impact on cardiovascular diseases (CVDs). This can be measured by body mass index (BMI), waist–hip ratio, waist–height ratio, Broca index, skinfold thickness, etc. Which of them is more valid in the context of CVDs is a moot question. In addition, how valid is it to divide BMI into thin, normal, obese, and morbid categories, or should one always use the actual values without any categorization? **Categories** have their own advantages but also involve risks, as discussed under that topic. Categorization does affect the validity of the measurement, and different categorizations can give different results. Some characteristics defy measurement: intense debate is ongoing regarding how to measure stress and strains in life, psychological well-being, and personality traits. In the absence of a direct measure, **surrogates** are used. The validity of these surrogates for measuring the actual characteristic remains a concern.

A large number of **scoring systems** are used in the practice of medicine and research. They are primarily used for gradation of severity (e.g., **APACHE score**) and secondarily for diagnosis (e.g., risk of malignancy index [6]). The validity of severity scores is assessed in terms of correct prediction of prognosis, and of diagnostic scores in terms of correct classification of patients into diseased and nondiseased. For example, the authors of APACHE II found that prediction of survival or hospital death by this score is correct in 85% of cases for cases admitted to critical care in US hospitals [7]. This is the index of validity of this score. The risk of malignancy index has a diagnostic sensitivity of 85% and specificity of 97% for ovarian cancer [6]. These two are the indices of validity in this case. If you are using a scoring system in your practice or research, examine its validity for that particular use. Using sensitivity and specificity as indicators of diagnostic efficacy is an example of their invalid use.

There are questions about the validity of statistical measures also, such as odds ratio in a prospective study setup. Relative risk is a valid measure only when the two groups under comparison have identical features except that the factor for which relative risk is being computed is present in one group and absent in the other. A large number of examples of such statistical measures can be cited whose validity is confined to a narrow sense. Validity of statistical estimates is generally assessed in terms of their bias [see the topic **bias (statistical)**]. This is different from **precision** since precision is for reliability and not validity. In brief, sample mean is an unbiased estimate of the population mean, but sample median generally is not an unbiased estimate of the population median except for symmetric distributions. The sample variance is an unbiased estimate of the population variance only when the denominator is $(n - 1)$, which explains why $(n - 1)$ is used in place of $n$ in this case. An unbiased estimate of **parameter** $\tau$ is one whose average over all different possible samples is the same as the value of $\tau$. For most distributions, unbiasedness can be theoretically established without worrying about drawing so many samples.

Validity of a statistical method is its ability to provide correct inference as evidenced by the data. This is mostly assessed by the validity of the underlying assumptions, as already discussed. However, in isolated cases, there might be gross deviations, for example, some use relative risk for odds ratio, and vice versa. The term *validity* is also used for **statistical analysis** where, for example, it becomes invalid when a wrong computer command is used while running a statistical package. An invalid design command for computation while using a statistical package will result in an error message in some cases, but in many others, the package will give you an output that is not what you want. This can easily happen for hierarchical design or designs with random effects where computer commands require sufficient expertise. These computer commands require some training and some understanding of the implications of different commands.

The extent of validity of statistical models is evaluated on two criteria: (i) their biological plausibility, which says that a possible mechanism should be available that links the outcome with the inputs as stipulated in the model, and (ii) their ability to correctly explain or predict the underlying phenomenon. Since models, by their very nature, represent simplified versions of the actual process, they are not expected to be 100% valid. In case a model is for predicting a **dichotomous** or **polytomous** outcome, percent correctly classified is an accepted measure of validity. For example, the Dimodent equation, which uses measurements of certain teeth, was able to correctly identify sex in 78.7% of cases in a Lebanese population [8]. This quantifies the validity of this equation in that population, but the validity could be different in some other population. For ordinal outcome, the ***C*-statistic** can used to quantify the validity. For diagnosis of ovarian tumors as benign, borderline, primary invasive, or

metastatic on the basis of a risk prediction model, the value of the *C*-statistic was found to be between 0.57 and 0.64 in external validation for Belgium as reported by van Calster et al. [9]. On a scale of 0 to 1, this value is not particularly high to inspire confidence about his model. For continuous outcome, a predictive model is valid when the mean of residuals (predicted − observed) is close to 0. This is easy to achieve, for example, in regression, by suitably choosing the intercept. In this case, validity can also be assessed by percent correctly predicted, although the primary consideration in such models is the reliability measured by the variance of the residuals instead of the validity. The validity of many outcome models is assessed in terms of biological plausibility, as stated earlier, and their ability to serve their intended purpose.

### *Validity of Results and Conclusions*

Validity of results and conclusions can be divided into two broad categories. The first is *internal validity*, that pertains to the validity of the information available after the study, and second is *external validity*, which pertains to their generalizability to the target population. Internal validity depends on the use of proper methods on one hand and consistency of different results on the other. If alcohol intake is positively associated with liver disease in a segment of the population, and it is also observed that that mine workers take more alcohol, you should get the result that liver diseases are more common in mine workers. If this is not so in a sample, and no reason can be identified for this inconsistency, the results are not internally valid, and suspicion arises about the method of collection of the data or their analysis. Maybe the subjects have given wrong information, or confounding was not properly accounted for in the analysis.

External validity depends mostly on the adequacy of the subjects for a particular generalization. A representative sample is a definite help for validity of results of **descriptive studies**, but that is not a prerequisite for **analytical studies**. Some results based on subjects from a hundred years ago remain valid for the present population; e.g., "smoking causes lung cancer" is the conclusion derived from a study of people decades ago but is valid for the present population also. A cause–effect type of relationship generally transcends time and population. A second factor affecting the external validity is correctness of the information obtained from the study subjects (see **bias**). Remember that wrong information cannot lead to right conclusions howsoever immaculate analysis might be. Third, it also depends on the proper choice of variables and indicators for differentially assessing the antecedents and the outcomes.


1. Chambless LE, Shahar E, Sharrett AR, Heiss G, Wijnberg L, Paton CC, Sorlie P, Toole JF. Association of transient ischemic attack/stroke symptoms assessed by standardized questionnaire and algorithm with cerebrovascular risk factors and carotid artery wall thickness. The ARIC study, 1987–1989. *Am J Epidemiol* 1996;144:857–66. http://aje.oxfordjournals.org/content/144/9/857.full.pdf
2. Dwyer JH, Li L, Dwyer KM, Curtin LR, Feinleib M. Dietary calcium, alcohol and incidence of treated hypertension in the NHANES I epidemiologic follow-up study. *Am J Epidemiol* 1996;144:828–38. http://aje.oxfordjournals.org/content/144/9/828.full.pdf
3. World Health Organization. *Tobacco Free Initiative—Global Adult Tobacco Survey*. http://www.who.int/tobacco/surveillance/gats/en/, last accessed December 15, 2013.
4. Kaestner R. Mortality and access to care after Medicaid expansions. *N Engl J Med* 20 December 2012;367:2453–4. http://www.nejm.org/doi/full/10.1056/NEJMc1212920
5. Creemers MC, van 't Hof MA, Franssen MJ, van de Putte LB, Gribnau FW, van Riel PL. Disease activity in ankylosing spondylitis: Selection of a core set of variables and a first set in the development of a disease activity score. *Br J Rheumatol* 1996 Sep;35(9):867–73. http://rheumatology.oxfordjournals.org/content/35/9/867.long
6. Jacobs I, Oram D, Fairbanks J, Turner J, Frost C, Grudzinskas JG. A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. *Br J Obstet Gynaecol* 1990 Oct;97(10):922–9.
7. Knaus WA, Draper EA, Wagnwer DP, Zimmerman JE. APACHE II: A severity of disease classification system. *Crit Care Med* 1985;13:818–29. http://www.ncbi.nlm.nih.gov/pubmed/3928249
8. Ayoub F, Cassia A, Chartouni S, Atiyeh F, Rizk A, Yehya M, Majzoub Z, Abi-Farah A. Applicability of the Dimodent equation of sex prediction in a Lebanese population sample. *J Forensic Odontostomatol* 2007 Dec;25(2):36–9. http://www.iofos.eu/Journals/JFOS%20Dec07/ayoub%20article.pdf
9. Van Calster B, Valentin L, Van Holsbeke C, Testa AC, Bourne T, Van Huffel S, Timmerman D. Polytomous diagnosis of ovarian tumors as benign, borderline, primary invasive or metastatic: Development and validation of standard and kernel-based risk prediction models. *BMC Med Res Methodol* 2010 Oct 20;10:96. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2988009/


## validity (types of)

Important types of validity for the purpose of medical applications are face, criterion, concurrent, content, and construct validity.

**Face validity** is an apparent correspondence between what is intended and what is actually obtained. Grossly speaking, ovarian cancer cannot occur in males, and prostate cancer cannot occur in females. Cleft lip, deafness, and dental caries cannot be causes of death. If any data show such inconsistency, these are not face valid. Face validity is also violated when a patient is shown as discharged from a hospital and cause of death within the hospital is also recorded. Current age cannot be less than the age at first childbirth. Thus, face validity is achieved when the observations look just about right. If a man states his age as 20 years but looks like a 40-year old, the response is not face valid. If a family reports consuming food amounting to 2800 calories per unit per day but the children are grossly undernourished, then again, the response or the assessment is not face valid.

An instrument or a method is called **criterion valid** if it gives nearly the same information that a criterion with established validity would. Carter et al. [1] report on the criterion validity of the Duke Activity Status Index with respect to standard pathologic work capacity indices in patients with chronic obstructive pulmonary disease. Whenever a new index, score, or any other tool is developed, it is customary that its criterion validity be established by comparing its performance with a standard. Our review of literature suggests that **correlation** is the statistical method generally used for this purpose, whereas the right assessment of criterion validity should be obtained by evaluating predictivity. If negative and positive **predictivities** are high, the tool can be considered criterion valid.

It is not uncommon in medical setups that no validated standard is available. For example, no valid measure is available to assess physical balance in people with vestibular dysfunction. The Berg Balance Scale is often used for this purpose but is not considered a fully valid measure. A relatively new tool is the Dynamic Gait Index, and the two can be investigated for **agreement**. This is different from correlation. If the agreement is good, the two methods can

be called **concurrently valid**. This means that both are equally good (or equally bad!). This, however, does not establish superiority of one over the other. Superiority can be inferred only when two methods under evaluation are compared with a known **gold standard**.

The fourth type is **content validity**. This is based on the domain of the content of the measurement or the device. Would you consider kidney function tests such as urea clearance and diodrast clearance content valid for assessing the overall health of kidneys? Such tests are restricted to specific aspects and do not provide a complete picture—thus, they are not fully valid for assessing the health of kidneys. Content validity is often established through qualitative expert reviews. Wynd and Schaefer [2] explain how the content validity of an osteoporosis risk assessment tool was established through a panel of experts. This is the most favored method for assessing the content validity of any tool.

The last is **construct validity**, which seeks agreement of a device with its theoretical concept. Body mass index is construct valid for overall obesity but not for central obesity, whereas waist–hip ratio is construct valid for central obesity and not for overall obesity. Construct validity is sometimes assessed by the statistical method of **factor analysis**, which reveals "constructs." If these constructs correspond well with the ones that are otherwise theoretically expected, the tool is considered construct valid.


1. Carter R, Holiday DB, Grothues C, Nwasuruba C, Stocks J, Tiep B. Criterion validity of the Duke Activity Status Index for assessing functional capacity in patients with chronic obstructive pulmonary disease. *J Cardiopulm Rehabil* 2002;22(4):298–308. http://journals.lww.com/jcrjournal/Fulltext/2002/07000/Criterion_Validity_of_the_Duke_Activity_Status.14.aspx
2. Wynd CA, Schaefer MA. The osteoporosis risk assessment tool: Establishing content validity through a panel of experts. *Appl Nurs Res* 2002;15(3):184–8. http://www.ncbi.nlm.nih.gov/pubmed/12173169


## variability, *see* variation (measures of)

## variables

A variable is a characteristic that tends to vary from person to person, time to time, unit to unit, etc. All biological characteristics—age, sex, birth order, body temperature, duration of survival, disease severity, etc.—are variables. If the characteristic is a quality such as blood group or site of lesion, it is called a **qualitative** variable, and if the characteristic is numerical such as age, blood pressure (BP), or parity, it is called a **quantitative** variable. The latter is also called a **metric** variable.

Statisticians find it easier to work with notations, as they help in making generalized statements, denoting a characteristic by variable $x$. If many characteristics are considered, notations such as $x_1$, $x_2$, $x_3$, etc., or $x$, $y$, $z$, etc., can be used.

It is easier to work with variables when they are represented by numbers. Quantitative variables are already observed as numbers, and you can say that $x = 132$ mmHg for a particular person where $x$ denotes systolic BP. Each category of a qualitative variable is also assigned a number, called a **code**, such as 1, 2, 3, and 4 for blood groups O, A, B, and AB. If a person has blood group B, you can say that $x = 2$ for him/her where $x$ now is the notation for blood group. This coding, however, does not make the variable quantitative. In the case of qualitative variables, the main interest generally is in the number of subjects in different categories—how many of blood group A, how many of blood group B, etc., or the proportion or percentage; obviously, a quantitative summary such as mean is not applicable to these variables.

### *Dummy and Indicator Variables*

A **dichotomous** characteristic such as sex can be represented by $x = 0$ for females and $x = 1$ for males. This is called a **dummy variable** or an **indicator variable**. For blood groups with four categories, the dummy variables could be defined as follows:

$x_1 = 1, x_2 = 0, x_3 = 0$ for blood group A
$x_1 = 0, x_2 = 1, x_3 = 0$ for blood group B
$x_1 = 0, x_2 = 0, x_3 = 1$ for blood group AB
$x_1 = 0, x_2 = 0, x_3 = 0$ for blood group O

Three dummy variables are required to represent four **categories**, and these together uniquely identify the category. In general, $(K - 1)$ dummy variables are needed to identify $K$ categories.

Dummy variables are extensively used in statistics. We give one simple example that will illustrate their use. Consider the following simple linear regression of systolic blood pressure (sysBP) on age in apparently healthy adults:

$$\text{sysBP} = 120 + ½(\text{age}) - ⅛(\text{age})*x, \qquad \text{(V.1)}$$

where age is in years and $x = 0$ for females and $x = 1$ for males. This is the dummy variable. Note that this regression is

$$\text{sysBP} = 120 + ½(\text{age}) \text{ for females } (x = 0)$$

and

$$\begin{aligned}\text{sysBP} &= 120 + \tfrac{1}{2}(\text{age}) - \tfrac{1}{8}(\text{age}) \text{ for males } (x = 1)\\ &= 120 + \tfrac{3}{8}(\text{age}).\end{aligned}$$

Equation V.1 is now able to provide two regressions with the help of a dummy variable—one for males and the other for females. We have not used notations $y$ and $x$ for the variables sysBP and age in these equations, so that it remains easy to understand, but those notations can also be used.

### *Deterministic and Stochastic (Random) Variables*

A variable whose value is already known is called a **deterministic variable**. If you know that a person is male, $x = 1$ for him is deterministic. If a person with a particular disease is selected, that disease is deterministic. On the contrary are **stochastic variables**, whose values are not known and cannot be predicted with certainty, such as outcomes of the treatment. These are also called **random variables**. When a person comes to a clinic with certain complaints, his/her signs and symptoms are elicited as these can vary even when the disease is known. Stochastic means probabilistic—something that depends at least partly on chance. When a patient is selected for a clinical trial on the basis of presence or absence of, say, liver cirrhosis, it is not possible to correctly predict his/her age or diet pattern as these would vary from patient to patient. Thus, age and diet are stochastic (random) variables, particularly when diet is also identified by a numerical code, age already being numerical. Core statistical methods for inference are for random variables including,

but not restricted to, the effect of deterministic variables on random variables.

Strange as it may sound, **independent variables** in a **regression** setup are considered deterministic. In Equation V.1, sex represented by $x$ and age in years as is both are deterministic. This equation tells us what sysBP is expected in that population for adults of a specific age and sex. The equation works only when age and sex are *known* and plugged into the right side of the equation. Details given under the topic **regression** tell you that the regression gives the *average* of the dependent $y$ for fixed values of the independent $x$'s—in our example, average sysBP for fixed age and sex.

### Discrete and Continuous Variables

Some variables can take only a small number of values. Gender has only two possible values, and blood group has four possible values. Some others can take any of a large number of values, such as cholesterol level, ranging from 100 to 200 mg/dL, which can even be 132.7 mg/dL if an instrument giving such accuracy is available. On the other hand, the parity of a woman can be 1 or 2 but never 1.6. A variable that can take only a finite, generally small, number of values in a range is called **discrete**. Number of deaths in a hospital in a day, site of lesion, and diagnosis are other examples of discrete variables. Almost all qualitative variables are discrete, whereas quantitative variables can be discrete or continuous. Parity and family size are examples of quantitative but discrete variables.

It is common in medicine that discrete variables take only nonnegative *integer* values, but the definition does not require it to be so. Shoe size can be 7, $7\frac{1}{2}$, 8, $8\frac{1}{2}$, etc. but is still discrete. It can take only these four values between 7 and $8\frac{1}{2}$. Compare this with a variable such as age that can be measured accurately as 7.2613 years when needed. Although recording of age in terms of completed years is often considered adequate, particularly for adults, theoretically, age can take an infinite number of values between 7 and $8\frac{1}{2}$. Thus, this is not a discrete variable.

A variable that can take an infinite number of values in a range is called continuous. Age, cholesterol level, body temperature, enzyme level, and weight are examples of continuous variables, although in practice, it may be redundant to measure them to several decimal places. For example, it does not help to say that the hemoglobin (Hb) level is 14.038 g/dL; just 14.0 is enough. Weight is generally measured to the nearest kilogram, BP to the nearest millimeter of mercury, and Hb level to one-tenth of a gram per deciliter. It can be argued that such approximation makes the variable discrete, but the continuous character is not lost in practice as long as the measurement is sufficiently accurate. Inversely, variables such as heart rate, platelet count, and respiration rate are in fact discrete yet are considered continuous because of the large number of possible values. *Only variables that can take a small number of values, say less than 10, are generally considered discrete for practical applications.* Others can be treated as continuous for most practical purposes even when they are theoretically discrete. It would be instructive to note that variables such as heart rate and respiration rate per minute can be in decimals but noted as integers. In 1 min, you may find that the heart rate is 67, but if this is observed for 10 min, you may find 673 beats, giving the average of 67.3 per minute. All this can become intricate when better **accuracy** is required.

Classification of a variable into discrete or continuous is important because statistical methods for the two types of variables are different. For example, methods based on **binomial**, **multinomial**, and **Poisson distributions** are used for exact analysis of discrete variables, whereas methods based on **Gaussian distribution** are mostly used for continuous variables.

### Categorical Variables

A variable is categorical when its values identify a **category** instead of a single value. Statistical methods distinguish three kinds of categorical variables depending on the kind of category they identify.

First are nominal categories, which are just names with no order, such as disease names (migraine, dementia, cancer, etc.) and signs (pallor, distension, palpable spleen, etc.). Either codes or mostly dummy variables are used for analysis of data with these kinds of categories. The variable for two categories (yes/no or male/female, etc.) is called **dichotomous**, and the variable for more than two categories is called **polytomous**. The statistical method of choice in this case is **chi-square** for large samples, and binomial or multinomial for small samples.

Second are ordinal categories, such as disease severity categorized into none, mild, moderate, serious, and critical. These categories have an underlying continuum but are observed as categories generally because no satisfactory metric is available to measure the continuum. The concerned variable is called an **ordinal variable**. Sometimes, **scores** are used to measure this continuum. If so, ordinal categories become **metric categories**, which is the third kind of category. Metric categories are often used for **quantitative data** also. Age categories of child, adolescent, adult, and old are intrinsically metric since each category can be defined numerically in terms of years of age. Similarly, for example, parity can be divided into 0, 1–2, 3–4, and 5+ categories.

Statistical methods for analysis of ordinal variables are not fully developed yet. Mostly methods based on **ranks** are used, such as **nonparameric methods**. Sometimes, ranks are considered as the actual values, and the usual methods for quantitative variables are used on these ranks. This, however, has severe limitations because the difference, for example, between mild and moderate (ranks 1 and 2) is considered the same as between moderate and serious (ranks 2 and 3). In addition, the analysis that uses ranks as the actual values also will consider, for example, that 3*mild = 1*serious because both have value 3. In a study reported by Zagouri et al. [1] on heat shock protein 90 (Hsp90) in lobular neoplasia of the breast, the intensity of Hsp90 stain was an ordinal variable (0 = negative, 1 = low, 2 = moderate, 3 = high), but mean and standard deviation (SD) were calculated as though they were usual quantities, although they used a nonparametric **Wilcoxon test**. Further, the categorization may not be sharp and objective, since a patient may look mild to one physician and moderate to another. Ranks do need extra care in analysis and interpretation.

**Statistical analysis** of categorical variables for metric categories also has limitations. First, the actual values are lost. If the category for, say, pain score is 0–3, score 1 is treated the same as score 2 or score 3. Thus, the advantage of metric numbers is not available, and the results may not be the same as expected from exact values. Second, such categories are mostly arbitrary. You may form categories 0–2, 3–5, etc., and I may form categories 0–3, 4–7, etc. Results based on two different categorizations do not lend themselves well for valid comparison and can give different results for the same data. Third, a variable for metric categories can be analyzed as metric by using the midpoint as the actual value, which assumes all the subjects in that category have the same value as the midpoint. This obviously is not true but works reasonably well in cases where the sample size is large and categories are small.

### Dependent and Independent Variables

These are discussed in detail under the topic **dependent and independent variables**. This topic also includes other related terms

such as outcome, response, and target variables on one hand, and regressor, predictor, explanatory, concomitant, confounding variables, etc. on the other. All these variables can be either qualitative or quantitative.

### *Instrumental Variables*

The ordinary regression setup, which in its most simple form is $y = \alpha + \beta x + \varepsilon$, requires that the relationship between $y$ and $x$ is direct and the errors $\varepsilon$ are uncorrelated with $x$. However, in some situations, the error and $x$ values are correlated (e.g., the higher the value of $x$, the higher the error), which would imply that changes in $y$ are associated not just with changing $x$ but also with changing $\varepsilon$. This aberration produces a biased estimate of the regression coefficient $\beta$ and can happen when (i) $x$ cannot be directly measured but can be measured with some error that contributes to the error $\varepsilon$—thus, $x$ and $\varepsilon$ are correlated; (ii) $x$ causes $y$ and $y$ causes $x$; or (iii) $x$ is an endogenous variable. A more valid solution in such cases can be obtained by including another variable, called an instrumental variable, which is not of interest here but can be exactly measured and has a direct relationship with $x$ and not with the errors $\varepsilon$. In case there are many regressors, as in multiple regression, the instrumental variable/s should be uncorrelated with all the $x$'s.

Let the instrumental variable be denoted by $z$. Find the relationship between $x$ and $z$. Since they are directly related, the regression of $z$ on $x$ will be uncontaminated and believable. This information can be used to improve prediction of $y$ using the two-stage procedure of regression with instrumental variables, as discussed by Greenland [2].

Consider the example of dependence of the length of hospital stay of critical patients on the severity of condition at admission. Severity cannot be directly measured, and the error in measurement of severity will affect the duration of hospital stay. Suppose a measure such as an APACHE score is available that is closely related to severity at admission but is not related to the errors. If so, the APACHE score can be used as an instrumental variable for getting the relationship between severity and length of hospital stay. Nead et al. [3] used body mass index as the instrumental variable to assess whether insulinemia and type 2 diabetes are causally associated with endometrial cancer.


1. Zagouri F, Nonni A, Sergentanis TN, Papadimitriou CA, Michalopoulos NV, Lazaris AC, Patsouris E, Zografos GC. Heat shock protein 90 in lobular neoplasia of the breast. *BMC Cancer* 2008 Oct 28;8:312. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2588461/pdf/1471-2407-8-312.pdf
2. Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol* 2000;29(4):722–9. http://ije.oxfordjournals.org/content/29/4/722.full
3. Nead KT, Sharp SJ, Thompson DJ, Painter JN, Savage DB, Semple RK, Barker A et al. Evidence of a causal association between insulinemia and endometrial cancer: A Mendelian randomization analysis. *J Natl Cancer Inst* 2015 Jul 1;107(9). pii: djv178. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4572886/


## variable selection, *see* stepwise methods, regressors (choice of), best subset method of choosing predictors

## variance, *see* standard deviation, confidence interval (CI) for variance

## variance components analysis

As the name suggests, variance component analysis is the partitioning of total variance into its components. **Analysis of variance (ANOVA)** also does the same with partitions, such as between factors and within factors, but the difference is that in conventional ANOVA, the effect of various factors is considered fixed, whereas variance component analysis is for random effects. You may want to review the topic **fixed and random effects** to understand the basics of these effects. The concept of random effects can be briefly explained as follows.

Consider a study in 4 hospitals and 10 patients undergoing a particular cancer surgery from each of these hospitals for studying their duration of stay in the hospital after the surgery. Duration of stay will surely vary from patient to patient, but the average can also vary because of the difference in nursing care from hospital to hospital, the hospital policies regarding when to discharge, etc. Patients in any case are considered a random sample, but if the inference is restricted specifically to these 4 hospitals, the hospital effect is fixed. If these 4 hospitals are a random sample from, say, a **population** of 50 hospitals and a generalized conclusion for these 50 hospitals is desired, the hospital effect becomes random. When the effect is random, the variance components method is used. Note that including all 50 hospitals in the study is generally not feasible, and if one has the resources to do this, the number of parameters for estimation in the statistical analysis, as illustrated later, becomes 49. This is too large and would require an extremely large sample. In a random-effect situation, these reduce to just one parameter, as explained later within this section. The fixed-effect model works well only for a small number of levels of the factors—in this example, the number of levels is 4, which is the number of hospitals. This is small, and a fixed effect surely can be considered if the interest is restricted to these 4 hospitals only.

The random-effect factor should be **categorical** for it to be treated by the variance components method. You can have more than one random-effect factor, and you can, at the same time, also have other factors with fixed effects—in which case this becomes a **mixed effects model**.

In the example just cited, there is only one factor, namely, the hospital. It has four levels in this example; consider these fixed for the time being. The outcome is the duration of hospital stay, which is quantitative. This is a **one-way ANOVA** setup where the **main effect** of hospital $j$ ($j = 1, 2, 3, 4$), if fixed, as in a conventional setup, can be denoted by $\alpha_j$. The corresponding mathematical formulation for conventional ANOVA is

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij},$$

where, in the context of our example, $y_{ij}$ is the hospital stay of the $i$th patient in the $j$th hospital, $\mu$ is the overall mean duration of stay of all such patients (estimated by the mean of all 40 patients in the study), $\alpha_j$ is the effect of the $j$th hospital (measured from overall mean $\mu$), and $\varepsilon_{ij}$ is whatever remains, called error. In this model, $\varepsilon_{ij}$ is considered to be independent and have a **Gaussian distribution** with mean 0 and variance $\sigma_\varepsilon^2$. The error $\varepsilon_{ij}$ is the only random quantity in this fixed-effect model and arises from the variation in hospital stay of individual patients. There are four hospital effect parameters, one for each hospital, namely, $\alpha_1$, $\alpha_2$, $\alpha_3$, and $\alpha_4$. When the natural condition, such as $\Sigma\alpha_j = 0$, is imposed, the effective number of these parameters is three.

When a hospital has a random effect, the interest shifts from the fixed effect of each hospital to the variance of the hospital effect—variance because now, the variation between hospitals is random.

The effect of the $j$th hospital can be denoted by $a_j$, and this will be considered to have, say, a Gaussian (normal) distribution with mean 0 and variance $\sigma_H^2$. This is written as $a_j \sim N(0, \sigma_H^2)$, where $N$ is for normal distribution. We prefer to call this distribution Gaussian since *normal* has a very different meaning in medicine. This says that hospital effects have a distribution—some hospitals will have a negative effect (when measured from the mean), some will have a positive effect on the duration of hospital stay, and the pattern is Gaussian. The interest is not in the effect of any particular hospital—example, the difference between hospital 2 and hospital 3 is not a consideration—but in how much the difference is between hospitals, that is, in the variance of the effects. The lesser the variance $\sigma_H^2$, the less is the effect of hospitals on duration of stay; that is, the hospitals then are more homogenous with respect to the duration of stay after that particular surgery. The **null hypothesis** in this case would be $H_0$: $\sigma_H^2 = 0$, which implies that the effect of all the hospitals is the same. If this cannot be rejected, you can say that the effect of hospitals on the duration of stay is not significant. Note that now, hospital effect is represented by just one parameter, namely, the variance $\sigma_H^2$.

The statistical problem now is to decompose the total variance across all 40 patients in our example into the variance among 4 hospitals and between 10 patients within hospitals. Incidentally, this decomposition is the same as done in the fixed-effect case, but the interpretation is different. See the topic **analysis of variance** for details of the decomposition in the fixed-effect case. In a one-factor case, as in our example, and when the number of units within each level of the factor is the same (10 patients from each hospital in our example), the mean sum of squares due to hospitals (MSB) theoretically can be shown to be an estimate of $\sigma_H^2 + n\sigma_H^2$, where $n$ is the number of patients within *each* hospital. In our example, $J = 4$ and $n = 10$. In any case, the estimate of $\sigma_\varepsilon^2$ is the mean square error (MSE). These terms are explained under the topic **one-way ANOVA**. Thus,

$$\text{estimate of } \sigma_H^2 = \frac{\text{MSB} - \text{MSE}}{n}.$$

Total variance = $\sigma_H^2 + \sigma_\varepsilon^2$, and the variance contributed by the hospitals is

$$\text{variance partitioning coefficient} = \frac{\sigma_H^2}{\sigma_H^2 + \sigma_\varepsilon^2}.$$

The validity requirement of this is that both kinds of random effects $a_j$ and $e_{ij}$ are uncorrelated with one another within and between groups. In the context of our example, this implies that the effect of any hospital is unrelated to the effect of any other hospital, the effect of the hospital is unrelated to the effect of the patient, and the effect of any patient is unrelated to the effect of patients in a different hospital. However, the effect of one patient is correlated with the effect of patients in the same hospital since they share exposure to the same nursing care, the same surgeons, the same policies, etc. This correlation among patients of the same hospital is measured by

$$\text{intraclass correlation coefficient (ICC): } \rho_I = \frac{\sigma_H^2}{\sigma_H^2 + \sigma_\varepsilon^2},$$

which is the same as the variance partitioning coefficient. Thus, the **intraclass correlation coefficient** is the fraction of the total variance contributed by the factor.

Variance components have an especially useful application to the analysis of data from hierarchical designs where one factor is nested within another. For example, in **multistage random sampling**, say, districts are randomly selected, then enumeration blocks within the selected districts, and families within the selected blocks. The response may be affected by the district, by the enumeration block, and by the family. All three are random effects in this case. When two or more factors are present, **interaction** between them may also be important. If districts are not randomly selected but all districts of interest are included in the study (this is easy when the number of districts of interest is small, say, ≤5), the district effect becomes fixed. In that case, the mixed effects model will be applicable since the effect of enumeration blocks and of families continues to be random. In either case, the value of $F$ is not necessarily obtained by dividing the concerned mean sum of squares by MSE. This is where the procedure is different from conventional fixed-effects ANOVA. Depending on whether the factor is random or fixed and depending on the hierarchical rank, sometimes, the divisor may be MSB or MSAB (mean square due to interaction) or some other value. For details, see Cox and Solomon [1].

1. Cox DR, Solomon PJ. *Components of Variance*. Chapman & Hall/CRC Press, 2002.

## variance-covariance matrix, *see* dispersion matrix

## variance inflation factor

In a multiple-variable situation, variance inflation factor (VIF) is defined as

$$\text{variance inflation factor: VIF} = \frac{1}{1 - R^2},$$

where $R^2$ is the **multiple correlation** coefficient. Depending on where used, $R^2$ may be of **dependent** $y$ on the **independent** $x$'s or of one $x$ on other $x$'s. This is for linear regressions, but if nonlinear regression is considered, $R^2$ is replaced by the **coefficient of determination** $\eta^2$. It is called the variance inflation factor because it estimates how much the variance of a coefficient is "inflated" because of linear dependence on other predictors.

The most pronounced application of VIF is in assessing **multicollinearity** in a **multiple regression** situation. In this setup, the standard error (SE), and hence variance, of the estimates of regression coefficients increases when the regressors have multicollinearity. In a multiple regression equation that looks like $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + + \beta_k x_k + \ldots + \beta_K x_K$, VIF for the regressor $x_k$ is the ratio by which the variance of $b_k$ increases due to multicollinearity, where $b_k$ is the estimate of $\beta_k$. A VIF of 1.8 tells us that the variance (the square of the SE) of a particular coefficient is 80% larger than it would be if that predictor were completely uncorrelated with all the other predictors. If VIF = 4, the SE is $\sqrt{4} = 2$ times what it would be without multicollinearity. The VIF has a lower bound of 1 but no upper bound. VIF > 2.5 or $R^2 > 0.6$ is considered high for multicollinearity, and VIF > 5 (i.e., $R^2 > 0.8$) surely worth taking action.

VIF for regression coefficient $b_k$ is given by

$$\text{variance inflation factor for } b_k\text{: VIF}_k = \frac{1}{1 - R_k^2},$$

where $R_k^2$ is the multiple correlation coefficient of the variable $x_k$ when regressed on the remaining regressors in the model. This may have to be obtained for each regressor so that you will have $(K - 1)$ VIFs. Obviously, VIF = 1 means no multicollinearity of $x_k$, and a value more than 4 is generally considered large enough to be on guard. A value more than 10 for a particular variable is a definite indication that this variable can be removed from the list of regressors without any substantial effect on the efficacy of the regression equation. O'Brien [1] cautions against this approach as well as other approaches such as **ridge regression** and against combining two or more variables into a single index, and advises instead that VIF needs to be interpreted in the context of other factors that influence the stability of the estimates.

Sometimes, the term *tolerance* is used in a specific context; it is the inverse of VIF, i.e., tolerance = $1 - R_k^2$.

1. O'Brien RM. A caution regarding rules of thumb for variance inflation factors. *Qual Quant* 2007; 41:673–90. http://www.nkd-group.com/ghdash/mba555/PDF/VIF%20article.pdf, last accessed December 25, 2013.

## variance ratio, *see also* *F*-test

This is the ratio of two variances $\sigma_1^2/\sigma_2^2$ that measures how different variation in one population is from another population on a ratio scale. The term was first used by Fisher way back in the 1920s. If you measure fasting blood glucose levels of healthy and diabetic subjects, there will be a huge difference not just in their mean levels but also in their variances. Blood glucose level is much more volatile across diabetic subjects than in healthy subjects. This probably is true for many measurements in healthy and sick subjects, underscoring the importance of the variance ratio. Statistically, this is important since a variance ratio different from 1 invalidates procedures such as ANOVA and regression. In their conventional format, these procedures examine difference in means provided that variances are the same between groups or for different values of the regressors, and this premise is violated when the variance ratio is very different from 1.

Statistically, the operational quantity is the sample analogue $s_1^2/s_2^2$. This is related to

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}.$$

Under the null hypothesis $H_0$: $\sigma_1^2 = \sigma_2^2$, this reduces to $F = s_1^2/s_2^2$. When the two groups are independent and the values follow a Gaussian (normal) distribution, this ratio is $F$-test with $(n_1 - 1, n_2 - 1)$ degrees of freedom (df's), where $n_1$ and $n_2$ are the number of subjects in the two groups, respectively. Thus, this sample variance ratio can be used for testing equality of population variances when the just-stated conditions of its validity are fulfilled. This can be a two-tailed test for testing the alternative $\sigma_1^2 \neq \sigma_2^2$, but it is customary to call the one with larger variance group 1 so that we need to worry only about $F > 1$. In case needed, the property that $1/F$ is also $F$ with reversed df's and $(1 - P)$ probability can be used for the other alternative. The most common application of the variance ratio is in one-way ANOVA, where $s_1^2$ is the mean sum of squares between groups and $s_2^2$ is the mean sum of squares within groups.

Among other applications, an interesting one is illustrated in assessment of 24 h intake of vitamin K in older adults in Canada [1]. The study found intraindividual variation from day to day to be significantly higher than interindividual variation. The authors conclude that such pronounced intraindividual variation indicates that vitamin K intake assessment should be done at least six times for each individual to get a fair picture of the intake in older adults. They also worked out confidence interval (CI) for the variance ratio in their population. In general, for Gaussian distribution, this CI is given by

$$(100 - \alpha)\%\ \text{CI for variance ratio: } \frac{s_1^2/s_2^2}{F_{1-\alpha/2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2/s_2^2}{F_{\alpha/2}},$$

where $F_{1-\alpha/2}$ and $F_{\alpha/2}$ are the values of $F$ distribution for probabilities $(1 - \alpha/2)$ and $\alpha/2$, respectively, at $(n_1 - 1, n_2 - 1)$ df's.

1. Presse N, Payette H, Shatenstein B, Greenwood CE, Kergoat MJ, Ferland G. A minimum of six days of diet recording is needed to assess usual vitamin K intake among older adults. *J Nutr* 2011 Feb;141(2):341–6. http://jn.nutrition.org/content/141/2/341.full.pdf+html

## variance-stabilizing transformation

In case variance across groups is substantially different, variance-stabilizing transformation is used to achieve near equality. Equality of variances, technically called **homoscedasticity**, may be required for valid analysis of data, such as by analysis of variance (ANOVA).

Let the groups be denoted by **deterministic variable** $x$, that is, $x = 1$ for group 1, $x = 2$ for group 2, etc. These could be, for example, thin, normal, overweight, and obese groups with respect to body mass index (BMI). Let the outcome be denoted by **stochastic variable** $y$ such as, say, random blood glucose (RBG) level in this example. Restrict the subjects to those who are not under any treatment. As $x$ increases, it is possible that the mean of $y$ increases, but at the same time, the variance of $y$ may also show an increase. In our example, RBG level may not only be higher but may have much more variability from person to person in the obese group than in the group with normal BMI. Among a large number of possibilities, one simple possibility is that variance for any particular $x$, i.e., var($y|x$), is equal to $a^*\mu_x$, where $\mu_x$ is the mean of $y$ for a particular value of $x$. This says that variance in different groups is proportional to the mean in that group, and not uniform—thus violating homoscedasticity.

In our example, suppose the data on RBG are as in Table V.1, where you can see that variance $\approx 0.64^*\mu_x$. It can be theoretically established that square-root transformation stabilizes the variance when variance is proportional to the mean, provided that all values are positive (if not positive, add a constant to each value and adjust this after the results of the analysis are obtained). In this example,

**TABLE V.1**
**Mean, Standard Deviation (SD), and Variance of Random Blood Glucose Level in Various BMI Groups**

| | Random Blood Glucose Level in mg/dL ($y$) | | |
|---|---|---|---|
| **BMI Group** | **Mean ($\mu_x$)** | **SD** | **Variance (in sq. units)** |
| Thin ($x = 1$) | 132 | 9.2 | 84.64 |
| Normal ($x = 2$) | 141 | 9.5 | 90.17 |
| Overweight ($x = 3$) | 155 | 9.9 | 98.01 |
| Obese ($x = 4$) | 172 | 10.5 | 110.41 |

if you consider the square root of the RBG level instead of the RBG level itself, the variance would be nearly the same in each BMI group, and the usual analysis can be done after this transformation. However, this would require that the original units be retrieved by squaring the values after the results are obtained so that proper interpretation can be done.

If variance is proportional to the square of the mean, then log transformation will stabilize the variance. Log transformation may be appropriate also when **residuals** are a certain percentage of the $y$ values; for example, residuals are 7% of the original values rather than, say, around 7 mg/dL in absolute value. If the variance of $y$ increases very substantially with $x$, say in proportion of $\mu_x^4$, then inverse transformation may be appropriate. This can happen when most of the values of $y$ are close to 0 (but not 0) when $x$ is small and some values show a huge increase as $x$ increases.

For further details, see Weisberg [1].

1. Weisberg S. *Applied Linear Regression*, Fourth Edition. Wiley, 2013.

## variation (measures of), *see also* coefficient of variation

Variation in statistics is the extent to which values differ from each other. This is also called **dispersion** and can also be understood as the spread of values or their scatter. In health and medicine, variation occurs due to a large number of sources, such as biologic, environmental, behavioral, and instrumental differences in the subjects and measurements. Some of the variation is due to unknown sources, called chance. Some characteristics show small variation, whereas others show large variation. Body temperature in healthy subjects is fairly stable from person to person, whereas total cholesterol shows huge variation. Opinions and behavioral variables generally show large variation not only across individuals but also within a person from time to time.

The following example may make it clear why a measure of variation is needed. Consider the systolic blood pressures (BPs) measured for two groups of five persons each:

Group 1: (in mmHg) 134, 132, 124, 132, 128
Group 2: (in mmHg) 110, 140, 118, 150, 132

Both groups have the same mean (130 mmHg) and the same median (132 mmHg). Despite this equality in groups with regard to central values, the groups are very different. The BP values in group 1 vary from 124 to 134 and in group 2 from 110 to 150—thus, their scatter or dispersion is different. Therefore, representation of data merely by a central value is incomplete, and it needs to be supplemented by a measure of dispersion as well. Figure V.2 shows two **Gaussian** (normal) **distributions** with the same mean but different variation.

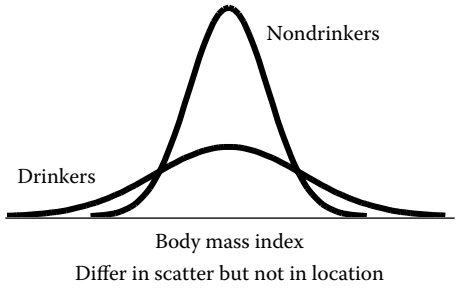



**FIGURE V.2** Body mass index in drinkers and nondrinkers illustrating Gaussian distributions with same mean but different variation.

When variation is large, obviously any, conclusion is liable to a wide margin of error. In this situation, the **reliability**, also called precision in statistical parlance, suffers, and our confidence level in the conclusion is low. A measure of variation gives us an idea of the reliability of the measurements. The most commonly used measures are described next.

### *Range*

In the case of quantitative values, one measure of variation could be from minimum to maximum. This is called the **range**. You can say that normal body temperature varies from 97.8°F to 99.1°F, and normal LDL cholesterol ranges between 50 and 120 mg/dL. That certainly gives an idea of the variation among healthy people, although it does not rule out a lower or higher value than these limits. In the case of actual data, range is easily affected by extreme values. For survival duration of a sample of 20 cancer patients, the range could be from 1 to 10 years, but if one patient happens to survive for 20 years, the range shoots to 1–20 years. This example also highlights another limitation of range. Most values may be close to 3 and 4 years, and if only a few values are very different, the range can alter very substantially. Range is based entirely on the minimum and the maximum, and we would like to have a measure that considers all the values and not just the values at the two ends.

### *Mean Deviation, Variance, and Standard Deviation*

Since variation is the difference of values with one another, one can think of finding all these differences and calculating, say, their average. But this could require enormous calculations—10 values will have $(10 \times 9)/2 = 45$ differences, and 100 values will have $(100 \times 99)/2 = 4950$ differences. Let us look for something ingenious that requires fewer calculations and gives a more meaningful assessment. It has been theoretically established that the best measure of variation in most situations is obtained when differences are obtained from one central value rather than from one another. In most cases (though not all), the mean is considered most suitable central value for this purpose. In these cases, the difference of values from the mean, called deviation, is obtained. In notations, if $x$ is the value of the variable, deviation = $(x - \bar{x})$. Since this is from the mean, some deviations will always be negative and some positive, and their mean will always be 0. One alternative is to ignore the sign, consider their absolute value, and calculate the average as follows:

$$\text{Mean deviation} = \frac{\Sigma|x - \bar{x}|}{n},$$

where $n$ is the number of values. More fully, this is the **mean deviation** from the mean since we can also have mean deviation from the median. Sometimes, this is called mean absolute deviation, and the expression we have mentioned is the sample estimate. Mean deviation can be an adequate measure of variation, but the primary difficulty with this is that absolute values are not easy to handle mathematically. Negative and positive values, when occurring together, are easily handled by squaring. This gives

$$\text{variance} = \frac{\Sigma(x - \bar{x})^2}{n}.$$

Thus, variance is the average of the squared deviations from the mean. This is the most acceptable measure of variation but is in squared units. Units are restored by taking the square root, which becomes the popular standard deviation (SD):

$$\text{standard deviation, SD} = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}\ .$$

Note that all these three measures are 0 when all values are equal since then, there is no variation. See **variance** for further details of SD and variance, and see **coefficient of variation** for another measure of variation based on mean and SD.

### *Semi-Interquartile Range*

Since the mean is not a good representative central value when extreme values are present or, in other words, when the distribution is highly skewed, SD, since it is based on the mean, is not a good measure of dispersion in this situation. A generally accepted alternative is the **interquartile range** (IQR) for this setup. This range is from the first **quartile** to the third quartile. Briefly, the first quartile ($Q_1$) is the value below (or at) which 25% of the values lie, and the third quartile ($Q_3$) is the value below (or at) which 75% of values lie. They are also called lower and upper quartiles, respectively. Thus, IQR encompasses the middle 50% of values and is considered an improvement over range in the sense that the extreme values (top 25% and bottom 25% of values) do not affect this measure. Often, half of IQR is used a measure of variation possibly to get a value comparable with SD. This gives

$$\text{semi-interquartile range} = \frac{Q_3 - Q_1}{2}\ .$$

To illustrate the calculations, consider the following strange-looking data on satisfaction with nursing services in a hospital on a scale from 0 to 6 from complete dissatisfaction to full satisfaction from 10 recently discharged patients after a complicated surgery:

6, 5, 0, 0, 5, 5, 0, 6, 1, 5

Three patients provided a score of 0, one patient gave a score of 1, four patients gave a score of 5, and two patients gave 6 out of 6. What do you think of the variation in these scores?

Range = 6 − 0 = 6
Mean = 3.3; median = 5
Deviations from mean: +2.7, +1.7, −3.3, −3.3, +1.7, +1.7, −3.3, +2.7, −2.3, +1.7
(You can check that their mean is 0.)
Mean deviation = (2.7 + 1.7 + 3.3 + 3.3 + 1.7 + 1.7 + 3.3 + 2.7 + 2.3 + 1.7)/10 = 2.4

$$\text{SD} = \sqrt{\frac{2.7^2 + 1.7^2 + \ldots + 1.7^2}{10}} = 2.5$$

$Q_1$ = 2.5th value in order = (0 + 0)/2 = 0; $Q_3$ = 7.5th value in order = (5 + 5)/2 = 5
Semi-interquartile range = (0 + 5)/2 = 2.5

### *Measure of Variation in Qualitative Data*

Perhaps the concept of variation is not applicable to qualitative data, and no good measure of variation is available for such data. Yet, one can think of variation = 1 (highest) when all categories have equal frequencies and no variation when all values are concentrated in just one category. If there are $n$ values divided into $K$ categories, equal frequency would mean $n/K$ in each category. If maximum frequency in any category is $f_m$, the excess is $[f_m - (n/K)]$, which can be, at most, $(n - n/K)$ when all $n$ values are concentrated in one category and there is no variation. Thus, the following can be considered as an index of variation on a scale of 0 to 1:

Index of variation for qualitative data:

$$\textit{variation ratio} = 1 - \frac{f_m - (n/K)}{n - (n/K)}\ .$$

Say the distribution of patients coming to a clinic is no illness, 10%; mild illness, 40%; moderate illness, 30%; serious illness, 15%; and critical illness, 5%. Variation ratio = 1 − (40 − 100/5)/(100 − 100/5) = 0.75. The higher the concentration in any one category, the less is the variation. In our example, if all patients coming to the clinic have mild illness, we can say that they do not have any variation with regard to the severity of illness. A drawback of this index is that it ignores the scatter of frequencies in the categories other than the maximum, although they also contribute to variation.

## varimax rotation

Varimax rotation is a special kind of rotation of axes in graphical form, generally used for multidimensional phenomena, which helps to present a clearer picture of what is happening. This strategy is commonly used for some statistical procedures such as **factor analysis** and **principal component** analysis.

See Figure V.3a, where many features of the bar diagram are obscure but become clear when the figure is rotated (Figure V.3b). In this case, we have rotated the $x$- and $y$-axes by 20° each. This figure is in three dimensions, but the same can be conceived for multidimensions. In Figures V.3, the axes continue to be perpendicular with each other (called *orthogonal*), although they may not appear so in the figure. However, methods are available that would do oblique rotation so that the axes are no longer orthogonal.

Rotation is just change of coordinates that define location. In the usual two dimensions, if a regression line has a slope of 30° and you rotate the ($x$, $y$) axes by 30°, the line will look parallel to the $x$-axis. This may be a big convenience in interpreting the implication of the line just as rotation in Figure V.3b is for clearly seeing various bars. Algorithms are available to devise rotations for increasing or decreasing the scatter.

*Varimax* stands for "maximizing the variance," where the variance is chosen depending on the application. For example, in the case of factor analysis, where this rotation is most commonly used, the variance to be maximized is the sum of the $K$-factor sample variances of the standardized **loadings**. The factors are supposed to be on different axes in this setup, and axes are orthogonal so that the factors are independent of each other. The rotation minimizes the complexity of the factors by making the large loadings larger and the small loadings smaller within each factor—thus maximizing the variance. This tends to minimize the number of variables with large loadings in each factor and helps in achieving clarity regarding the variables appearing in each factor. After this rotation, the variables with small loadings can be ignored, and only the factors with high loadings are considered for interpreting the factors. See **factor analysis** for more details.

There are other rotational methods. *Quartimax rotation* works in a somewhat inverse manner as it makes large loadings larger and

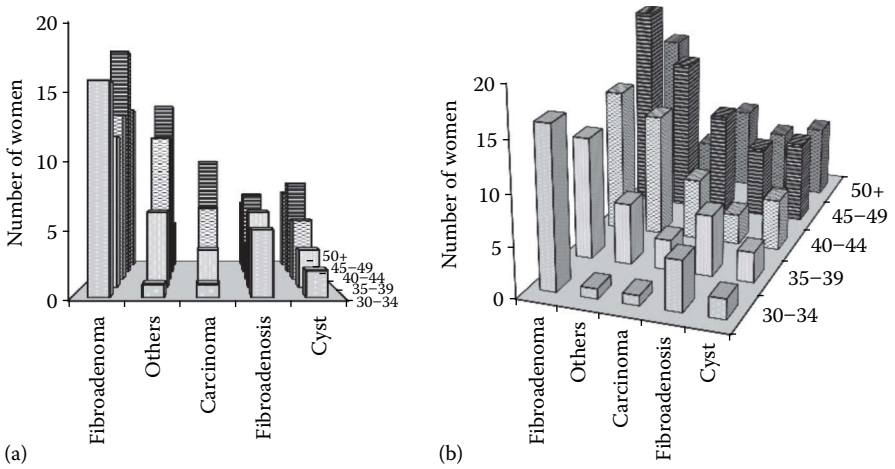



**FIGURE V.3** A three-dimensional bar diagram (a) without rotation of axes and (b) with rotation of axes.

small loadings smaller for factors within each variable. This works well when variables are sought to be explained by the factors. Note that factors in factor analysis consist of variables, whereas here variables, are considered to contain factors. *Equamax rotation* is a compromise that attempts to simplify both factors and variables. These are all orthogonal rotations. Oblique rotations are used when the factors are not orthogonal, i.e., correlated.

## velocity of growth in children

Velocity is the speed of movement in the aimed direction and is calculated as the distance traveled in the desired direction divided by the time taken. Although the term is used elsewhere in medicine, such as blood velocity and pulse wave velocity, we restrict it to its most common usage in assessing growth in children.

Growth velocity is the rate of growth per unit of time. Among various parameters of growth in children, height and weight are the most commonly used. For height, velocity is higher at the beginning of life and tapers off as age increases with a slight upswing, called midgrowth, around 6 or 7 years of age in some cases. A definite spurt is seen in adolescence. Velocity is indicated by the steepness of the curve and represents the incremental growth per unit of time. For body mass index (BMI) in children in the United States, the velocity is negative (BMI decreases with age) from 2 to 5 years and then high till the age of 15 years before slowing down (see Figure V.4 for girls). Velocity is also used to

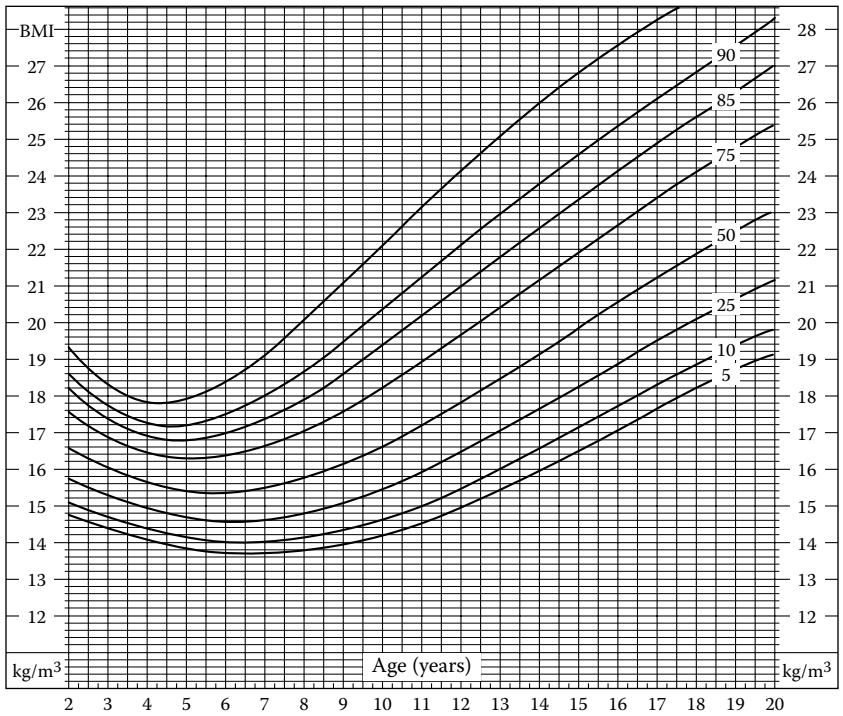



**FIGURE V.4** Age–BMI Centers for Disease Control and Prevention (CDC) chart for girls of age 2–20 years in the United States. (From CDC. *Body Mass Index-for-Age Percentiles: Boys*. http://www.cdc.gov/growthcharts/data/set1clinical/cj41c023.pdf.)

monitor growth and requires longitudinal measurements. A velocity less than normal for a child of a particular age indicates failure to thrive. Different percentiles may have different velocities, as in Figure V.4. In this figure, the fifth percentile has slow velocity and 95th percentile relatively high velocity, as indicated by the differential steepness of the curves. This really means that children with already-high BMI increase their BMI faster than those with lower BMI.

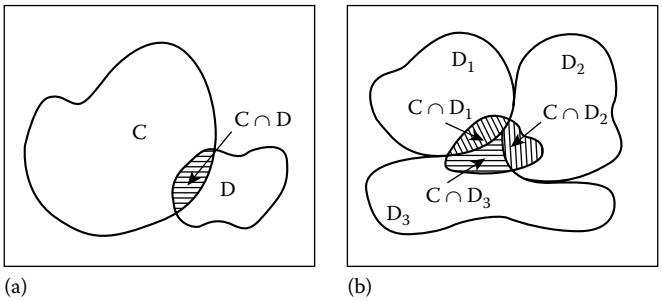
Conventional velocity charts involve two charts and hence are difficult to adopt in practice. A 3-in-1 weight-monitoring chart has been devised [2] for infants. It consists of conventional weight centiles complemented with extra lines called thrive lines, where the slope defines a cutoff for failure to thrive. The weight must be measured at 4-week intervals for this chart to be useful. This chart needs to be field-tested in different populations.

Weight velocity and height velocity can be used as indicators of growth in the immediate past and thus helps in detecting acute malnutrition. A sudden decline in weight velocity (or in weight gain) in a particular child may provide better insight into the existence of a health problem than a weight-for-age measurement. Weight for age is relatively slow to react and slow to show that a problem exists. Height for age and weight for age are also affected by hereditary factors, particularly the size of the parents, while velocity, when compared with the previous velocity of the same child, is relatively independent of heredity.


1. CDC. *Body Mass Index-for-Age Percentiles: Boys*. http://www.cdc.gov/growthcharts/data/set1clinical/cj41c023.pdf
2. Cole TJ. 3-in-1 weight-monitoring chart. *Lancet* 1997 January 11;349:102–3. http://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736(05)60886-0.pdf


## Venn diagrams

A Venn diagram depicts combinations and commonalities among two or more collections of units of interest, called sets. In medicine, the group of patients with a particular complaint, for example, is one set, and the group of people with another complaint is another set. These two sets may be mutually exclusive in the sense that no unit is common (no patient has both complaints) or may overlap (some patients with complaint 1 also have complaint 2). In Figure V.5a, one set is denoted by C, and the other set is D. The shaded area is the area that is common to both. This is called the intersection and is denoted by $C \cap D$. The combination of C and D is called the union and denoted by $C \cup D$. This, however, contains $C \cap D$ only once, although this is part of C as well as D. We have used irregular shapes to depict sets—many use circles of equal radius, as shown in Figure V.6, even when the sets have unequal size.




(a) (b)

**FIGURE V.5** Venn diagrams showing unions and intersections of sets (a) two sets, (b) three sets.

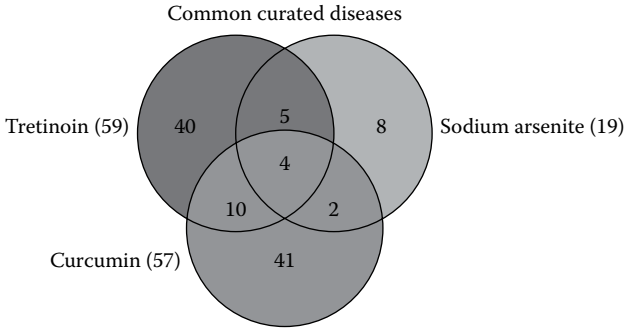



**FIGURE V.6** Venn diagram showing overlaps of different common curated diseases. (From Davis AP, King BL, Mockus S et al. *Nucleic Acids Res* 2011 Jan;39(Database issue):D1067–72. http://www.ncbi.nlm.nih.gov/pubmed/20864448, with permission.)

Figure V.5b illustrates another Venn diagram. This has three **mutually exclusive** sets denoted by $D_1$, $D_2$, and $D_3$ with no overlap. Set C is in the middle and intersects all the three sets. That is, union of $D_1$, $D_2$, and $D_3$ contains all of C. Thus, set C is the combination of $(C \cap D_1)$, $(C \cap D_2)$, and $(C \cap D_3)$, and these three also are mutually exclusive subsets in this case. Thus $C = (C \cap D_1) \cup (C \cap D_2) \cup (C \cap D_3)$. This illustrates how a Venn diagram makes it easy to understand the unions and intersections of various groups of people or events.

Application of Venn diagrams in statistics is in understanding and computing probabilities of two or more events that may or may not occur simultaneously, as in our example of complaints. As another example, suppose 12% of adults have hypertension and 7% have diabetes in a population. How many have both, and how many have only one of these two ailments? Such questions can be easily understood well with the help of a Venn diagram. Let C be for hypertension and D be for diabetes—thus, the intersection $C \cap D$ represents those who have both the diseases. If 3% have both the ailments, 9% will have hypertension alone (no diabetes), and 4% will have diabetes alone (no hypertension). A total of $12 + 7 - 3 = 16\%$ have either hypertension or diabetes or both.

Davis et al. [1] have discussed the use of Venn diagrams for discovering overlaps and unique attributes of any set of chemicals, genes, or diseases in the context of a comparative toxicogenomics database. One such diagram is in Figure V.6. For example, in this figure, there are 10 cases common to tretinoin and curcumin that are not in sodium arsenite. Note how different overlaps are clearly visible. This facility is rarely available in other kinds of depictions.


1. Davis AP, King BL, Mockus S, Murphy CG, Saraceni-Richards C, Rosenstein M, Wiegers T, Mattingly CJ. The comparative toxicogenomics database: Update 2011. *Nucleic Acids Res* 2011 Jan;39(Database issue):D1067–72. http://www.ncbi.nlm.nih.gov/pubmed/20864448


## visual display (of data), *see* graphs and diagrams

## vital statistics

Colloquially, vital statistics refer to bust, waist, and hip measurements, particularly of female adults, but they have a different meaning in the context of health and medicine. *Vita* means "relating to life," and vital statistics are the data on the most important events of life. These include birth, death, and marriage (and divorces), sometimes including migration also. An accurate recording of all births, deaths, and migration helps in keeping an updated register of the

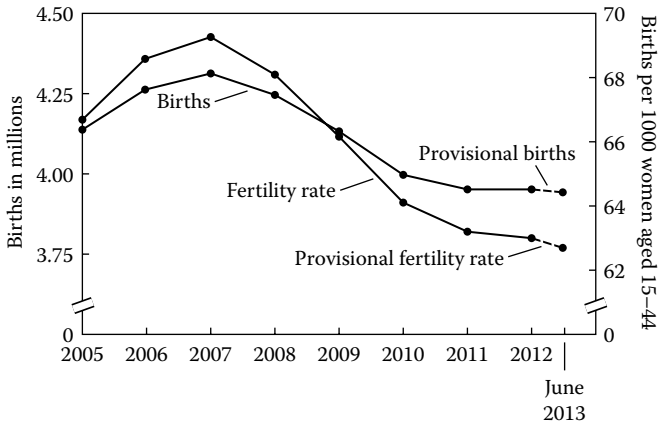



**FIGURE V.7** Trend of birth rate and fertility rate in the United States, 2005–2013. (From Hamilton BE, Sutton PD. *NCHS Health E-Stat*. http://www.cdc.gov/nchs/data/hestat/births_fertility_june_2013/births_june_2013.htm, last accessed January 4, 2014.)

population. When properly done, this provides a ready source of the count of people in an area.

When accurate vital statistics are available over a period of time, the trends can provide useful information regarding not just the increase or decrease but the rate also. For example, see Figure V.7 reported by Hamilton and Sutton [1], which shows that births and fertility rate increased in the United States between 2005 and 2007, and declined thereafter, with fertility rate exhibiting a more prominent trend than births.

There are finer points regarding some of the vital statistics. Births are sometimes defined to include live births and fetal deaths, and for this to be complete, a count of abortions, miscarriages, and stillbirths should also available. For comparison over time and across areas, birth rate, abortion rate, stillbirth rate, etc. can be calculated, which can tell a lot about the health of women and their health needs. Similarly, deaths generally include the cause and age at death, which help in assessing the gravity of various causes in different segments of the population. These details are important ingredients for assessing the health needs of people from the vital statistics. Trend over time can also provide clues on how effective health programs have been in controlling the mortality by different diseases, and projections can help in planning services. Age at death is used to construct life tables and to estimate **life expectancy**.

The best source of vital statistics is birth and death registration under what is called the civil registration system. This operates in almost all countries around the world but is nearly complete only in developed nations. In most developing countries, this registration is incomplete, and one has to rely on indirect estimation using other sources such as decennial census, sample surveys from time to time, and statistical models.

1. Hamilton BE, Sutton PD. Recent trends in births and fertility rates through June 2013. *NCHS Health E-Stat*. http://www.cdc.gov/nchs/data/hestat/births_fertility_june_2013/births_june_2013.htm, last accessed January 4, 2014.

## volunteer-based studies

Medical research by its very nature is invasive. Laboratory investigations require samples of blood, urine, etc., and radiological investigations require the invasion of x-rays and ultrasounds. Medical interviews look benign but many times seek intimate information, and in any case, the patient's time and privacy is invaded. For this reason, the ethics of medical research require that informed consent must be taken from the subjects of research. A step forward is volunteer-based studies. Volunteers are those who not only provide consent but are willing, sometimes keen, to participate in a research endeavor either to further the cause of science or because they find a ray of hope for themselves or others in that research. By definition, volunteers do not expect anything in return for themselves, but indirect benefits accrue in terms of the satisfaction of doing something useful for society at large, a natural wider social acceptance, building up capacity, and the confidence to face challenges. Recent research has shown that volunteers get indirect health benefits in addition to social benefits [1]. Many institutions and organizations seek volunteers openly through their websites or otherwise for research studies after ethical clearance, and they do get some volunteers. This is relatively easy when the volunteers required are healthy subjects but may be difficult if the study is on subjects with specific diseases. However, in some situations, defining healthy subjects as generally needed to serve as controls in many studies may be difficult—they could be just those with an absence of any overt manifestation of disease, those who have a healthy hemoglobin level, or those able to cope with adverse conditions. Gierthmühlen et al. [2] discussed this aspect in the context of quality sensory testing (QST)-based studies, which illustrates the kinds of problems one can face.

Volunteer-based medical studies bring forth statistical challenges. Early **phases of clinical trials** are often done on volunteers who select themselves as the subjects. Volunteers tend to be very different from the general class of subjects as many of them are either hopeless terminal cases or are people with exceptional courage. Both affect the response, and the results are not applicable to the general class of patients. Notwithstanding this limitation, volunteer studies have a definite place in medicine as they do provide important clues on the toxicity of the regimen under test, the dose level that can be tolerated, the potential for further testing of the modality, etc.

1. Grimm, Jr. R, Spring K, Dietz N. *The Health benefits of Volunteering: A Review of Recent Research*. Corporation for National and Community Service, 2007. http://www.nationalservice.gov/pdf/07_0506_hbr.pdf
2. Gierthmühlen J, Enax-Krumova EK, Attal N, Bouhassira D, Cruccu G, Finnerup NB, Haanpää M et al. Who is healthy? Aspects to consider when including healthy volunteers in QST-based studies—A consensus statement by the EUROPAIN and NEUROPAIN consortia. *Pain* 2015 Nov;156(11):2203–11. http://www.ncbi.nlm.nih.gov/pubmed/26075963

# W

## waist–hip ratio, *see also* obesity (measures of)

Evenly distributed fat is probably not as harmful as accumulation around the waist. In place of just waist measurement, its comparison with hip measurement has been observed to be a more effective indicator of truncal obesity. In adults, this is measured by

$$\text{waist} - \text{hip ratio} = \frac{\text{waist circumference}}{\text{hip circumference}}.$$

The unit of measurement does not matter, but the numerator and the denominator must be measured in the same unit—either inches, centimeters, or any other. This ratio measures central **obesity** or abdominal obesity (or truncal obesity). The normal ranges are 0.8–1.0 for men and 0.70–0.85 for women. Note that a unit increase in this ratio is enormous, if not unlikely—thus, its coefficient in a regression (logistic or ordinary least squares) is to be interpreted with full caution.

A large waist–hip ratio (WHR) is found to be nearly as much associated as body mass index with the risk of some serious conditions such as ischemic stroke [1] regardless of height and build. It is also associated with diabetes and heart diseases. Siddiqui et al. [2] reported for Indian adults that a high WHR is associated with high concentrations of malondialdehyde and a low concentration of antioxidant enzyme, resulting in increased oxidative stress, which is a major causative factor of acute myocardial infarction. However, Borruel et al. [3] observed that waist circumference and body mass index are more accurate surrogate markers of visceral adiposity in young adults of Spain than WHR.

While waist size seems easy to measure, accurate measurements rely on the method of measurement and body posture. Waist circumference is best measured at its narrowest point when the person stands with the feet 25–30 cm apart with a relaxed abdomen. Measurement should be taken by positioning the tape horizontally placing a graduated, flexible but inelastic tape snugly against the skin without compressing the underlying soft tissue. The circumference is generally measured to the nearest 0.1 cm after the subject breathes out. The hip is similarly measured at its widest portion of the buttocks.


1. Wang A, Wu J, Zhou Y, Guo X, Luo Y, Wu S, Zhao X. Measures of adiposity and risk of stroke in China: A result from the Kailuan study. *PLoS One* 2013 Apr 17;8(4):e61665. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3629147/
2. Siddiqui AH, Gulati R, Tauheed N, Pervez A. Correlation of waist-to-hip ratio (WHR) and oxidative stress in patients of acute myocardial infarction (AMI). *J Clin Diagn Res* 2014;8(1):4–7. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3939583/
3. Borruel S, Moltó JF, Alpañés M et al. Surrogate markers of visceral adiposity in young adults: Waist circumference and body mass index are more accurate than waist hip ratio, model of adipose distribution and visceral adiposity index. *PLoS One* 2014;9(12):e114112. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4257592/


## Wald test

Among several versions, the most common form of the criterion of the Wald test for the **null hypothesis** $H_0$: $\theta = \theta_0$ is

$$\text{Wald criterion: } W = \left( \frac{\hat{\theta} - \theta_0}{\text{estimated SE}(\hat{\theta})} \right)^2,$$

where $\hat{\theta}$ is the estimate of the **parameter** $\theta$ and SE is its **standard error**, and estimated SE is the SE when $\theta$ is replaced by $\hat{\theta}$. The criterion $W$ follows a **chi-square** distribution for large $n$. Thus, the ***P*-value** for judging **statistical significance** can be easily obtained. The **degrees of freedom (df's)** of chi-square vary from situation to situation. This test is very versatile and can be used in a wide variety of situations for large samples, and it is attributed to Abraham Wald, from his paper published in 1943 [1].

Abraham Wald
(Courtesy of Konrad Jacobs. Archives of the Mathematisches Forschungsinstitut Oberwolfach.)

Apparently, the Wald criterion looks like square of the **Student *t***, but that really is not so. First, the Student $t$ requires that the underlying distribution is **Gaussian**. This is not a strict requirement for the Wald test. Second, the Student $t$ in the case of the mean, for example, uses the sample standard deviation with denominator $(n - 1)$, whereas Wald uses $n$. In this respect, this test is more like the ***z*-test**, although the $z$-test also requires Gaussian distribution and the SE itself in the denominator instead of its estimate.

The Wald test, as just mentioned, is for **quantitative variables**. If the variable is **qualitative**, the test can still be used, although in that setup, the estimated SE is difficult to compute. In fact, the most common use of the Wald test is in testing the statistical significance of coefficients in **logistic regression**, where the variable can be qualitative. For testing the significance of any one coefficient, df = 1, but the test can be used for testing the significance of two or more coefficients simultaneously, for which the df's will accordingly change. This test can be used as a replacement for the classical **likelihood ratio test** for finding whether or not the contribution of one or more regressors in a logistic regression is statistically significant. Asymptotically (i.e., for very large $n$), both are equivalent. In its general form, the Wald test can be used for a **multivariate** setup also.


1. Wald A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans Am Math Soc* 1943;54:426–82. http://www.jstor.org/stable/1990256

## Ward method of clustering, *see also* cluster analysis

**Clustering** is the process of putting together quantitatively similar units in one group and dissimilar ones in different groups. Generally, the number of **clusters** is not predetermined, and as many clusters are formed as needed for internal homogeneity and external isolation of the units. It is hoped that these will be natural clusters as dictated by the data. Among popular algorithms to do this are **hierarchical** agglomerative and hierarchical divisive. The agglomerative algorithm is a bottom-up approach where units are successively merged into entities containing similar units one after the other starting from $n$ units, whereas the divisive algorithm is a top-down approach where all units are considered to form one big entity to begin with, and then this entity is sequentially split into smaller ones. Each of these algorithms has several methods. The Ward method is one of the methods of hierarchical agglomerative clustering.

While most other methods use the inverse of distance measures such as **Euclidean distance** for clustering, the Ward method, proposed in 1963 by Joe Ward Jr. [1], considers the within-entity variance as the criterion to decide which entities should be merged in an agglomerative algorithm. If there are four entities at a hierarchical stage containing, say, $n_1$, $n_2$, $n_3$, $n_4$, units, within-entity variance is calculated for entities obtained by merging $n_1$ and $n_2$ units, by merging $n_1$ and $n_3$ units, by merging $n_1$ and $n_4$ and then by merging $n_2$ and $n_3$ units, and by merging $n_2$ and $n_4$ units and then by merging $n_3$ and $n_4$ units. Whichever merging gives the least variance is done. Others are left as such. This process is sequentially followed till such time that the entities continue to have internal homogeneity in terms of small variance. If at some hierarchical stage, the within-entity variance suddenly inflates due to merging, the process can be stopped since this indicates that two very different entities that need to be in different groups are being merged. Entities obtained at such conclusion of the process are the required clusters obtained by the Ward method.

Sakagami et al. [2] used this method to identify groups of asthma patients with greater decline in lung functions and similar clinical features (a total of right variables). The method divided 86 patients into three clusters. Cluster 1 comprised women with late onset of asthma; cluster 2 had men and women with early onset, atopy, and long duration of disease; and cluster 3 mostly had older men with late onset. These were considered relevant groups to follow clinical guidelines. Gerlinger et al. [3] found the Ward method to be better than single, average, and complete linkage in discovering distinct clusters of records of bleeding diaries of women who were on various sex hormones containing drugs. They found four desirable and two undesirable bleeding patterns. Statistically, the Ward method prefers to join entities of nearly the same size rather than entities of widely different sizes. It is also very sensitive to outliers because they tend to inflate the variance.


1. Ward JH Jr. Hierarchical grouping to optimize an objective function. *J Amer Stat Assoc* 1963;58:236–44. http://www.jstor.org/stable/2282967
2. Sakagami T, Hasegawa T, Koya T, Furukawa T, Kawakami H, Kimura Y, Hoshino Y et al. Cluster analysis identifies characteristic phenotypes of asthma with accelerated lung function decline. *J Asthma* 2014 Mar;51(2):113–8. http://www.ncbi.nlm.nih.gov/pubmed/24102534
3. Gerlinger C, Wessel J, Kallischnigg G, Endrikat J. Pattern recognition in menstrual bleeding diaries by statistical cluster analysis. *BMC Womens Health* 2009 Jul 16;9:21. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2717079/

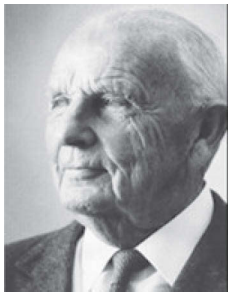

## washout period, *see* crossover designs

## Weibull distribution

Mathematically, a variable $x$ is considered to have a Weibull distribution if its **frequency curve** can be expressed as

$$\text{Weibull distribution: } f(x) = \alpha\beta x^{\beta-1} e^{-\alpha x^{\beta}},\ x \geq 0,\ \alpha > 0\ \beta > 0,$$

where α is called the scale parameter and β is called the shape parameter. These are its parameters, just as mean μ and standard deviation σ are for a **Gaussian distribution**. Note in Figure W.1a and b how the shape changes when the value of β changes but the shape basically remains same (peak toward left—Figure W.1b and c) when β remains the same but α changes. A large variety of patterns can be depicted by Weibull distribution. This is named after Waloddi Weibull, who described it in 1939 [1].


Waloddi Weibull

As explained for the term **distribution**, this is the pattern values follow in a population. For example, a **Gaussian distribution** has its peak (highest frequency) in the middle, and the pattern of decline on both the sides follows a specific and symmetric pattern. On the other hand, *skewed distributions* have their peak toward one side. A

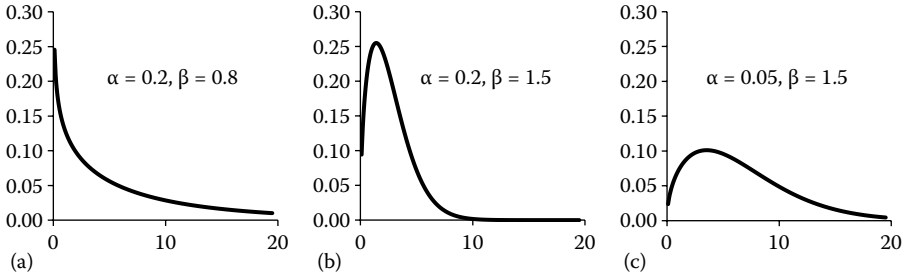



**FIGURE W.1** Different shapes of Weibull distribution.

Weibull distribution has a specific right-skewed pattern, with several important applications in medical sciences.

A Weibull pattern is usually followed by time-to-event (durations) type of variables. For example, duration of surgery or duration of hospital stay may follow a Weibull pattern since many patients have low duration, some have long duration, and few have very long duration. Kanters et al. [2] used a Weibull pattern to model duration of survival of HIV-positive men and women on antiretroviral therapy in Uganda and observed that part of the difference between sexes was due to late initiation of treatment among men.

Weibull distributions have an especially useful application in the study of hazards. For an explanation of this term, see **hazard** in this volume. Duration of survival in real life has a feature that the hazard of death increases as age increases in both healthy and sick subjects. Increasing hazard with higher values of the variable $x$ can be modeled to follow a Weibull distribution with $\beta > 1$. For $\beta < 1$, it represents decreasing hazards with the passing of time, called hardening. This may give rise to the pattern seen in Figure W.1a. An example is decreased "hazard" of birth with increasing age of women beyond 40 years. A constant hazard that remains the same for every value of $x$ gives rise to an **exponential distribution**. The hazard of occurrence of dengue fever can be the same for each age.

For further details, see Abernethy [3].


1. Weibull W. A statistical theory of the strength of materials, *Ingeniörsvetenskapsakademiens HandlingarNr* 151, 1939, Generalstabens Litografiska Anstalts Förlag, Stockholm. http://www.barringer1.com/wa_files/Weibull-1939-Strength-of-Materials.pdf
2. Kanters S, Nansubuga M, Mwehire D, Odiit M, Kasirye M, Musoke W, Druyts E et al. Increased mortality among HIV-positive men on antiretroviral therapy: Survival differences between sexes explained by late initiation in Uganda. *HIV AIDS (Auckl)*. 2013 May 29;5:111–9. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3677809/pdf/hiv-5-111.pdf
3. Abernethy RB. *The New Weibull Handbook*, Fifth Edition. Abernethy, 2004. http://www.barringer1.com/tnwhb.htm


## weighted least squares

In its simplest form, weighted least squares is a **least squares method** of estimation that gives due weight to the variance of **residuals** (as in **regression**) and is used when the variance is not constant but varies for different values of the **independent variables**. If residual variance for a given value of $x_i$ is $\sigma_i^2$, then weight $w_i = 1/\sigma_i^2$. This gives more weight to the values with smaller variance of residuals and less weight to the values with larger variance. That sounds intuitively reasonable too as it rightly gives more importance to the values that give more precise information. Weighting overcomes the problem of non-**homogeneity of variances** since that is one of the requirements of the least squares method.

In Figure W.2, the variance of residuals is increasing with higher values of the independent $x$. This can happen, for example, in a study of the effect of age on systolic blood pressure (BP) in females in the general population since not only average BP increases with age, but the variation is also seen to increase as age increases. This pattern will remain the same when residuals are obtained after adjusting for the effect of age. If an increase in variance with age is statistically significant, the ordinary least squares method will not give valid results. Weighted least squares will provide valid estimates and their valid standard error.

In a study on the effect of yoga experience (lifetime hours) on positive attitude on a 7-point scale in women over the age of 45 years [1], attitude scores had high variation when yoga experience was short.

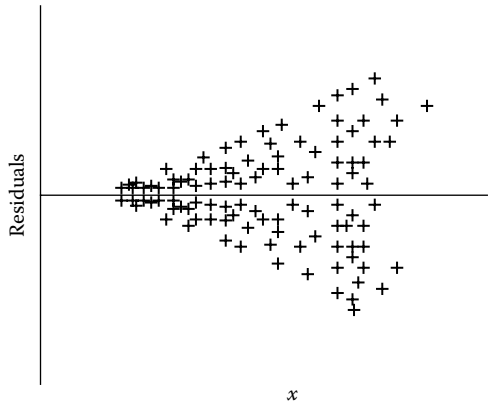



**FIGURE W.2** Fanning of residuals—increasing variance with higher value of $x$.

The scores stabilized as yoga experience increased. Thus, weighted least squares was the right method for regression as used in this study.

For running weighted least squares, check the options in your software package. Most packages have this facility under general linear models.


1. Moliver N, Mika E, Chartrand M, Haussmann R, Khalsa S. Yoga experience as a predictor of psychological wellness in women over 45 years. *Int J Yoga* 2013 Jan;6(1):11–9. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3573537/


## weighted mean

The weighted mean is the mean after adjusting for the differential importance of different values. If the weight of $x_i$ is $w_i$,

$$\text{weighted mean: } \bar{x}_w = \frac{\Sigma w_i x_i}{\Sigma w_i}.$$

As an extreme example, consider the number of children of 100 couples who have been married for 5 years. The number of children would be 0, 1, 2, or 3 when more children in 5 years of marriage are ruled out. It would be foolish to say that the average number of children in these couples is $(0 + 1 + 2 + 3)/4 = 1.5$. The average would depend on how many couples have no child, how many 1 child, how many 2 children, and how many 3 children. If these numbers out of 100 are 30, 50, 19, and 1, respectively, the average number of children is $(30 \times 0 + 50 \times 1 + 19 \times 2 + 1 \times 3)/100 = 0.91$. The numerator is the total number of children to these couples, and the denominator is the total number of couples. In this case, the values of $x$ are 0, 1, 2, and 3, and the weights are the number of couples with different values of $x$. The statistical name for these numbers is **frequency**, denoted by $f$. Then weighted mean = $\Sigma f_i x_i / \Sigma f_i$.

Besides the situation of differing frequencies as described in the preceding paragraph, weighted mean has other applications also. Particularly in large-scale surveys, such as the **Demographic** and **Health Surveys**, each person in the sample represents a specified number of people of his/her type. Consider data in Table W.1 on smokers and the number of cigarettes smoked by male and female adults. In this population, there are 400,000 males and 500,000 females. But the sample happens to have 1000 males and 2000 females. Then, each male in the

**TABLE W.1**
**Weighted Mean When Samples Have Different Representations**

| Particulars | Males | Females | Combined | |
|---|---|---|---|---|
| Number in the population | 400,000 | 500,000 | 900,000 | |
| Number in the sample | 1000 | 2000 | 3000 | |
| Number of current smokers in the sample | 200 | 100 | 300 | |
| Prevalence of current smokers in the sample | 20% | 5% | Unweighted (20 + 5)/2 = 12.5% | Weighted by Sample Size (20 × 1000 + 5 × 2000)/3000 = 10% |
| Average cigarettes smoked per day by current smokers in the sample | 12 | 5 | Unweighted (12 + 5)/2 = 8.5 | Weighted by Sample Size (200 × 12 + 100 × 5)/300 = 9.67 |
| | **Weighted by Population Size** | | | |
| Estimated number of current smokers in the population | 20% of 400,000 = 80,000 | 5% of 500,000 = 25,000 | 80,000 + 25,000 = 105,000 | |
| Prevalence of current smokers in the population | 20% | 5% | (20 × 400,000 + 5 × 500,000)/900,000 = 11.67% | |
| Estimate of average cigarettes smoked per day by current smokers in the population | 12 | 5 | (80,000 × 12 + 25,000 × 5)/105,000 = 10.33 | |

sample is representing 400 persons, and each female in the sample is representing 250 persons.

Two kinds of weighting are shown in Table W.1: first with respect to sample size and second with respect to the population numbers. In the sample, 20% of males and 5% of females are current smokers. What is the prevalence when males and females are combined? The **unweighted mean** of these two is (20 + 5)/2 = 12.5%. But this is incorrect since their numbers in the sample are not equal. When you consider that the sample has 1000 males and 2000 females, the weighted mean is 10%. This is the prevalence in the sample of a total of 3000. However, in this case, the sample sizes of males and females are not in proportion to their respective numbers in the population. When this is considered, the estimate of the prevalence of current smoking in the population becomes 11.67%. This is the correct estimate. Similarly, the average number of cigarettes smoked by current smokers is 9.67 when weighted for sample numbers and 10.33 when weighted for population numbers. The latter is correct. Most studies will erroneously report the sample weighted mean and percentages, and conveniently forget about population weighting.

## Welch test

Developed by Bernard Welch in 1938 [1], this test is used for assessing equality of sample means in two or more groups in situations where the variances in different groups are different. The usual test for equality of means is the ***F*-test**, whose special case is the **Student *t*-test** for two groups, but these tests are valid only when the variances across groups are the same—a condition popularly known as **homoscedasticity**. This can be relaxed to some extent for these two tests when the group sizes are not much different. When homoscedasticity is violated, particularly if group sizes also are markedly unequal, the Welch test is used. The Welch test also requires that the values follow a **Gaussian** (normal) **distribution**, just as is the requirement for the *F*-test and Student *t*-test. This requirement can be waived when the sample sizes in each group are large. However, the Welch test also fails when the samples are small and very unequal, particularly if the variances differ in the opposite direction.

The steps generally followed for testing a difference in means is as follows. First, ensure that the values are independent of one another. This will come from the design for collecting the data. Second, check that they follow an approximate Gaussian pattern (see **Gaussianity, how to check**). This can be relaxed for large *n*. Third, check for homoscedasticity. For this, generally, the **Levene test** is used. If this test reveals that the variances are significantly different, use the Welch test. Thus, this test effectively makes the Levene test for equality of variances redundant, as this can be directly used. For two samples, the criterion for the Welch test also follows Student *t* distribution but with different **degrees of freedom** (df's):

$$\text{Welch test (two-sample): } t_v = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

where the df

$$v = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}.$$

For multiple samples, the formula becomes complicated, but in this case, the ***F*-test** remains the appropriate test (just as in equal variances situation), although the df changes, as for the two-sample situation just described. The df of the Welch test is not a fixed number but depends on the sample values. One study with $n_1 = 10$ and $n_2 = 15$ can give different df than in another study of the same sizes. In the case of Student *t*, the df will remain the same so long as the sample sizes are the same. The Welch test is an approximate solution to the **Beherens–Fisher problem**, Beherens–Fisher also used an approximation.

1. Welch BL. The significance of the difference between two means when the population variances are unequal. *Biometrika* Feb 1938;29(3/4);350–62. http://www.stat.cmu.edu/~fienberg/Statistics36-756/Welch-Biometrika-1937.pdf

## WHO growth charts

The World Health Organization (WHO) growth charts provide **percentile** trends over age of growth parameters such as height and weight of a cross-section of healthy children around the world. These are available in a WHO document [1]. One such chart is shown in Figure W.3 for body mass index (BMI) for boys of age 0–24 months. To construct these charts, data on healthy children were obtained from Brazil, Ghana, India, Norway, Oman, and the United States. Growth parameters covered in this document are length/height for age, weight for age, weight for length, weight for height, and BMI for age. These charts serve as important tools to assess whether a child or a group of children in an area are on the expected track of good health or not, and the extent of shortfall if any, or whether there is any excessive growth. Interventions and policies can be accordingly formulated.

As just mentioned, growth charts are drawn on the basis of measurements seen in a healthy segment of children. The WHO document [1] emphasizes the distinction between standard and reference charts. Standards are those values that *should* be attained by all children, whereas reference provides values that *are* attained. WHO growth charts are for standards and not for reference in this sense. The title of this document indicates that these efforts started for developing reference but ended up developing standards.

These WHO charts are based on a combination of **longitudinal** measurements of children during early life (<2 years) and mostly **cross-sectional** measurements thereafter. The total sample size was 8440 children. Data were cleaned for **outliers** and **missing values**. In addition, to discount the likelihood of extreme values, children with measurements beyond ±3 standard deviation (SD) for their age and sex were excluded. Since the distributions generally were highly **skewed** to the right, measurements beyond +2 SD were also excluded from the cross-sectional sample.

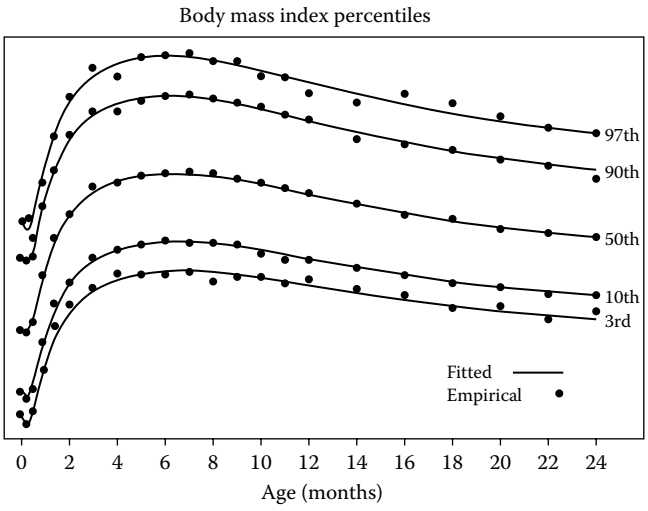



**FIGURE W.3** WHO chart for BMI for boys of age 0–24 months. (From WHO Multicentre Growth Reference Study Group. *WHO Child Growth Standards: Length/Height-for-Age, Weight-for-Age, Weight-for-Length, Weight-for-Height and Body Mass Index-for-Age: Methods and Development*. p. 239. World Health Organization, 2006. http://www.who.int/childgrowth/standards/technical_report/en/index.html.)

You can see that a large sample is imperative for constructing any such chart. The statistical method used to obtain the WHO growth charts is the **Box–Cox power exponential (BCPE)** since this considers **kurtosis** also in addition to skewness of the distribution of growth measurements. The method of **cubic splines** was used for smoothing of the percentile curves after age transformation. For testing goodness of fit, the ***Q*-test** was used. A nonmathematical explanation of these steps is given by Indrayan [2]. Figure W.4 shows that the obtained percentile curves after such rigorous methods are very close to the observed means at different ages. For further details, see the WHO document [1].

1. WHO Multicentre Growth Reference Study Group. *WHO Child Growth Standards: Length/Height-for-Age, Weight-for-Age, Weight-for-Length, Weight-for-Height and Body Mass Index-for-Age: Methods and Development*. World Health Organization, 2006. http://www.who.int/childgrowth/standards/technical_report/en/index.html
2. Indrayan A. Demystifying LMS/BCPE methods of centile estimation for growth and other health parameters. *Indian Pediatr* 2014;51(1):37–43. http://www.indianpediatrics.net/jan2014/jan-37-43.htm

## Wilcoxon rank-sum test

The Wilcoxon rank-sum test [1] is a **nonparametric test** for assessing whether or not two independent samples have come from a population with the same **location**. The corresponding parametric test for this purpose is the **Student *t***, where the **parameters** under test are means. In the case of nonparametric tests, the mean is not the target parameter; rather, it is just the location (Figure W.4). The distributions can have any shape but should have same shape. Nonparametric tests compare the distribution as a whole for location and not any particular parameter. But they can be considered to compare medians, and some books describe them as **median tests**. Since the Student *t*-test has more **efficiency**, it can be used for large samples, and we can use the Wilcoxon test for small samples when the distribution is far from **Gaussian**. This is equivalent to the Mann–Whitney ***U*-test** [2]; both came at about the same time and were independently worked out. Thus, this is also called the **Mann–Whitney–Wilcoxon test**.

Suppose there are $n_1$ observations in the first sample and $n_2$ in the second sample. For the Wilcoxon rank-sum test, the two samples are combined, and these $n$ ($= n_1 + n_2$) observations are assigned **ranks** from lowest to highest. Ties, if any, are given average ranks. For convenience, assume that the labels are such that $n_1 \le n_2$. Now,

Wilcoxon rank-sum test: $W_R$ = sum of the ranks assigned to the $n_1$ observations in the first sample.

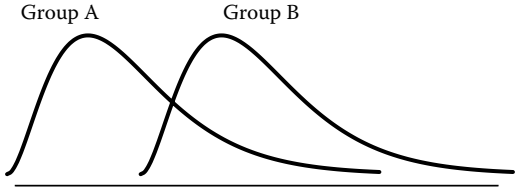



**FIGURE W.4** Two distributions with the same shape but differing in location.

Critical values of $W_R$ for $\alpha = 0.05$ under the **null hypothesis** $H_0$ of no difference between the two groups are given in Table W.2 for one-sided and two-sided **alternative hypothesis** $H_1$.

Frank Wilcoxon

Reject $H_0$ if the calculated value of $W_R$ is equal to or beyond the critical value for your $n_1$ and $n_2$. The table is only for values of ($n_1$, $n_2$) between 4 and 9. The Wilcoxon rank-sum test for two independent samples will not give any statistical significance at the 5% level if both $n_1$ and $n_2$ are less than or equal to 3 (also see the note below Table W.2). For the right-sided alternative, use the upper value, and for the left-sided alternative, use the lower value. For a two-sided alternative, use both values, as shown in Table W.1. An $\alpha = 0.025$ for a **one-tail alternative** is equivalent to $\alpha = 0.05$ for a **two-tail alternative** because the Wilcoxon criterion is symmetric.

When any of these sample sizes is 10 or more, use the following Gaussian approximation.

$$z = \frac{W_R - \mu_{W_R}}{\sigma_{W_R}},$$

where

$$\mu_{W_R} = \frac{n_1(n+1)}{2}, \text{ and } \sigma_{W_R} = \sqrt{\frac{n_1 n_2 (n+1)}{12}}, \; n_1 \le n_2 \text{ and } n = n_1 + n_2.$$

This $z$ follows a **standard normal distribution**, and the same critical values can be used for testing equality of locations. If each $n_1$ and $n_2$ is 30 or more, there is no need to use this criterion. A Student $t$-test for means can be safely used for such large sample sizes in practically every situation.

Note that the Wilcoxon rank-sum test, for that matter, any test based on ranks, is based on the order of the values and not the actual values. If one value is 10.1 and the next higher is 18.2, this receives the same rank as when the second value is 10.2. Thus, quantities lose half their relevance. For this reason, the Wilcoxon test and all other rank-based tests are sometimes not considered adequate. Many nonparametric enthusiasts consider this a strength of nonparametric tests. In any case, they are the best procedures available so far for small samples when the underlying distribution is far from Gaussian.

This test is valid only when all the rearrangements of the data are equally likely under the null. This condition is met if and only if the groups differ in location only and nothing else. For example, if variances differ, the test ceases to be a test for difference in location. It mixes the variance differences also.


1. Wilcoxon F. Individual comparisons by ranking methods. *Biometr Bull* 1945;1(6):80–3. http://www.jstor.org/discover/10.2307/3001968
2. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat* 1945;18(1):50–60. http://projecteuclid.org/download/pdf_1/euclid.aoms/1177730491
3. Wilcoxon F, Wilcox RA. *Some Rapid Approximate Statistical Procedures*. Lederle Laboratories, 1964.
4. Hollander M, Wolfe, DA. *Nonparametric Statistical Methods*, Second Edition. Wiley, 1999.


## Wilcoxon signed-ranks test

The null hypothesis under test in the Wilcoxon signed-ranks test [1] is that two **paired samples** have the same location. Pairing could be before–after measurements or one-to-one matching. For the meaning of location difference, see Figure W.4 under **Wilcoxon rank-sum test**. The Wilcoxon signed-ranks test is

**TABLE W.2**
**Lower and Upper Critical Values of Wilcoxon Rank-Sum Test $W_R$ for ($n_1$, $n_2$) between 4 and 9**

| | α | | $n_1$ | | | | | |
|---|---|---|---|---|---|---|---|---|
| $n_2$ | One-Tailed | Two-Tailed | 4 | 5 | 6 | 7 | 8 | 9 |
| 4 | 0.05 | 0.10 | 11, 25 | | | | | |
| | 0.025 | 0.05 | 10, 26 | | | | | |
| 5 | 0.05 | 0.10 | 12, 28 | 19, 36 | | | | |
| | 0.025 | 0.05 | 11, 29 | 17, 38 | | | | |
| 6 | 0.05 | 0.10 | 13, 31 | 20, 40 | 28, 50 | | | |
| | 0.025 | 0.05 | 12, 32 | 18, 42 | 26, 52 | | | |
| 7 | 0.05 | 0.10 | 14, 34 | 21, 44 | 29, 55 | 39, 66 | | |
| | 0.025 | 0.05 | 13, 35 | 20, 45 | 27, 57 | 36, 69 | | |
| 8 | 0.05 | 0.10 | 15, 37 | 23, 47 | 31, 59 | 41, 71 | 51, 85 | |
| | 0.025 | 0.05 | 14, 38 | 21, 49 | 29, 61 | 38, 74 | 49, 87 | |
| 9 | 0.05 | 0.10 | 16, 40 | 24, 51 | 33, 63 | 43, 76 | 54, 90 | 66, 105 |
| | 0.025 | 0.05 | 14, 42 | 22, 53 | 31, 65 | 40, 79 | 51, 93 | 62, 109 |

*Source:* Wilcoxon F, Wilcox RA. *Some Rapid Approximate Statistical Procedures*. Lederle Laboratories, 1964.

*Note:* No difference can be statistically significant by this test at $\alpha = 0.05$ (two-tailed) or $\alpha = 0.025$ (one-tailed) when both $n_1 \le 3$ and $n_2 \le 3$. However, significance at this level can be achieved if $n_1 = 2$ and $n_2 \ge 8$, or $n_1 = 3$ and $n_2 \ge 5$. Critical values for these situations are not given in this table. If needed, see Hollander and Wolfe [4].

used for small samples with non-**Gaussian distribution** because for large samples and for Gaussian distribution, the **paired *t*-test** performs better. The method for the Wilcoxon signed-ranks test is as follows:

Step 1. Let there be $n$ pairs and the observed value for the $i$th pair $(x_i, y_i)$, $i = 1, 2, \ldots, n$. Calculate the difference $d_i = (x_i - y_i)$ in the variable values for each pair. For the purpose of illustration, consider $x_i$ as the value before the treatment and $y_i$ the value after the treatment.

Step 2. Ignore the + or − sign of $d_i$ and assign rank starting from rank 1 for the smallest $|d_i|$ to rank $n'$ for the largest $|d_i|$, where $n'$ is the number of pairs with nonzero difference. The pairs with zero difference are omitted. Thus, $n' \le n$. If two or more $|d_i|$ are equal (ties), they are each assigned the average rank of the ranks they would have received individually if ties in the data had not occurred.

Step 3. Reaffix the + or − sign of the difference to the respective ranks. That is, indicate which ranks arose from negative $d_i$s and which from positive $d_i$s.

Step 4. Calculate the Wilcoxon test criterion as the sum of the positive ranks. That is,

Wilcoxon signed-ranks test:
$W_S$ = sum of the ranks with positive sign.

The minimum possible value of $W_S$ is 0 when no rank has a positive sign. The maximum occurs when all ranks are positive. If there were no difference between the groups, the criterion $W_S$ would take on a value close to its mean $n'(n' + 1)/4$. A big difference from this mean would be evidence against the **null hypothesis** ($H_0$) of no difference between the *before* and *after* measurements.

Step 5A. Find the ***P*-value** corresponding to the calculated value of $W_S$. Reject $H_0$ if $P$ is less than the predetermined **level of significance**. Described here is a procedure based on critical values of $W_S$ corresponding to $\alpha = 0.05$. For this $\alpha$ and $n' \le 19$, reject $H_0$ if

a. $W_S \ge w_r$ if $H_1$ is that *before* measurements are *higher* than *after* measurements (right-sided $H_1$)
b. $W_S \le w_l$ if $H_1$ is that *before* measurements are *lower* than *after* measurements (left-sided $H_1$)
c. Either $W_S \le w_1$ or $W_S \ge w_2$ if $H_1$ is that *before* measurements are *different* from *after* measurements (two-sided $H_1$)

The critical values $w_r$, $w_l$, or $w_1$ and $w_2$ for $\alpha = 0.05$ for $n'$ from 5 to 19 are given in Table W.3. The one-sided critical value for $n' = 5$ is 0 or 15. These are the minimum and maximum possible values with $n' = 5$. Thus, the decision to reject $H_0$ in favor of one-sided $H_1$ can be reached in this case only if all five differences are positive (or all negative). When $n' \le 4$, no difference can be statistically significant at the 5% level when $H_1$ is one-sided. If $H_1$ is two-sided, no difference can be statistically significant by this test at the 5% level for $n' \le 5$.

Step 5B. For $n' \ge 30$, it is safe to use the Student $t$-test for means. If $n'$ is between 20 and 29, use the Gaussian approximation to the Wilcoxon test. This is given by

$$z = \frac{W_S - \mu_{W_S}}{\sigma_{W_S}},$$

where $\mu_{W_S}$ is the mean value of $W_S$, given by $\mu_{W_S} = n'(n'+1)/4$, and $\sigma_{W_S}$ is the standard deviation (SD) of $W_S$ given by $\sigma_{W_S} = \sqrt{n'(n'+1)(2n'+1)/24}$. A one-tailed or two-tailed test can be carried out as usual with this $Z$.

**TABLE W.3**
**Critical Values of Wilcoxon Signed-Ranks Test $W_S$ for Matched Pairs ($\alpha = 0.05$)**

| | Right-Sided $H_1$ | Left-Sided $H_1$ | Two-Sided $H_1$ | |
|---|---|---|---|---|
| $n'$ | $w_r$ | $w_l$ | $w_1$ | $w_2$ |
| ≤4 | [a] | [a] | [b] | [b] |
| 5 | 15 | 0 | [b] | [b] |
| 6 | 19 | 2 | 0 | 21 |
| 7 | 25 | 3 | 2 | 26 |
| 8 | 31 | 5 | 3 | 33 |
| 9 | 37 | 8 | 5 | 40 |
| 10 | 45 | 10 | 8 | 47 |
| 11 | 53 | 13 | 10 | 56 |
| 12 | 61 | 17 | 13 | 65 |
| 13 | 70 | 21 | 17 | 74 |
| 14 | 80 | 25 | 21 | 84 |
| 15 | 90 | 30 | 25 | 95 |
| 16 | 101 | 35 | 29 | 107 |
| 17 | 112 | 41 | 34 | 119 |
| 18 | 124 | 47 | 40 | 131 |
| 19 | 137 | 53 | 46 | 144 |

*Source:* Wilcoxon F, Wilcox RA. *Some Rapid Approximate Statistical Procedures*. Lederle Laboratories, 1964.

[a] No difference can be statistically significant by this test at $\alpha = 0.05$ (one-tailed) when $n' \le 4$.

[b] No difference can be statistically significant by this test at $\alpha = 0.05$ (two-tailed) when $n' \le 5$.

A good statistical package will do all these steps for you. It will give the $P$-value, which you can interpret to draw conclusions. The following comments may be helpful:

1. The Wilcoxon test is less powerful than the Student $t$-test if the underlying distribution is Gaussian. Thus, it should not be used when the data follow a Gaussian pattern. Also, the **sign test** is less powerful than the Wilcoxon test because it ignores the magnitude of differences.
2. The Wilcoxon signed-ranks test, for that matter, almost all the nonparametric tests for quantitative data, is based on ranks. These tests are discrete and not continuous. For this reason, it is rarely possible to specify a critical value for exact $\alpha = 0.05$. The critical values given in Table W.3 actually have $\alpha \le 0.05$. In some cases, this could be substantially less. But that is the nearest one could reach if $\alpha$ is to be 0.05.

As an example, suppose the role of obesity in prolonged labor is examined in pregnant women of age more than 35 years with persistent occipitoposterior presentation. Seven obese (say, body mass index [BMI] ≥ 30) and seven nonobese women with such presentation, one-to-one matched for age, parity, hemoglobin (Hb) level, etc.,

**TABLE W.4**
**Duration of Labor in Obese and Nonobese Women**

| | Duration of Labor (Hours) | | | | | | |
|---|---|---|---|---|---|---|---|
| Obese | 18 | 15 | 17 | 20 | 14 | 12 | 18 |
| Nonobese | 17 | 15 | 18 | 18 | 11 | 10 | 14 |

**TABLE W.5**
**Calculation for Wilcoxon Signed-Ranks Test for Data in Table W.4**

| $d_i$ | 1 | 0 | −1 | 2 | 3 | 2 | 4 |
|---|---|---|---|---|---|---|---|
| Rank of $\|d_i\|$ | 1.5 | – | 1.5 | 3.5 | 5 | 3.5 | 6 |
| Signed ranks | +1.5 | – | −1.5 | +3.5 | +5 | +3.5 | +6 |

are included in the study. The data obtained are in Table W.4. Is the evidence sufficient to conclude that obese women have longer labor?

The calculations are shown in Table W.5. For positive ranks,

$$W_S = 1.5 + 3.5 + 5 + 3.5 + 6$$
$$= 19.5.$$

Since the pair with $d_i = 0$ is ignored, $n' = 6$ in this example. Note how the ranks have been assigned to the ties. Two pairs have the same absolute difference, namely, 1 h. They would have received ranks 1 and 2 but now receive rank 1.5 each because of the tie. The case with ranks 3 and 4, where the difference is 2 h, is also similar.

The alternative hypothesis here is one-sided ($H_1$: obese women have a longer duration of labor than nonobese women) because there is no stipulation for obese women to have a shorter duration of labor. Since the calculated value of $W_S$ is more than the cutoff 19 in Table W.3 for $n' = 6$, reject $H_0$ of no difference at $\alpha = 0.05$. Conclude that obese women do indeed have a longer duration of labor.


1. Wilcoxon F. Individual comparisons by ranking methods. *Biometr Bull* 1945;1(6):80–3. http://www.jstor.org/discover/10.2307/3001968
2. Wilcoxon F, Wilcox RA. *Some Rapid Approximate Statistical Procedures*. Lederle Laboratories, 1964.


## Wilks lambda (Λ)

Wilks lambda (Λ) is the multivariate generalization of the ***F*-test** for assessing equality of a set of means in three or more groups. Samuel Wilks first proposed this in 1938 [1]. This has become the basic test in one-way **multivariate analysis of variance (MANOVA)**.


Samuel Wilks

Suppose you measure three lung functions (say, forced vital capacity, forced expiratory volume in 1 s, and total lung capacity) for male adults of age 40–44 years working in tire factories, paint factories, and coal mines, and as traffic police for at least 10 years. These are some of the workplaces where lung functions can be compromised because of prolonged exposure to dust, fumes, and chemicals. The interest is in finding whether or not average lung functions in these four groups are the same. Since there is a set of three measurements on each person and there are four groups, MANOVA is the statistical method of choice to find if the means are different. The interest is not in individual lung functions but to consider the set of measurements together. This is what makes it a multivariate setup. A statistical package will be needed for calculations, and that will give the value of Wilks Λ. The software will also convert this Λ to an *F*-test with appropriate degrees of freedom and will provide the ***P*-value** so that you can decide whether to reject or not reject the **null hypothesis** of equality of the set of means in these four groups.

Wilks lambda can be interpreted as the proportion of the variance in the outcomes that is not explained by the group differences. As for *F*, Wilks Λ also requires **independence** of observations, **homoscedasticity**, and **multivariate Gaussian**-ity of the residuals. Homoscedasticity in this case means equality of **dispersion matrices**. For further details, see **multivariate analysis of variance (MANOVA)**.


1. Wilks SS. Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika* 1938;3:23–40. http://link.springer.com/article/10.1007%2FBF02287917#page-1


## Wilson interval

The standard error (SE) of the sample proportion $p$ is $\sqrt{[\pi(1-\pi)/n]}$, but the conventional method is to replace the population proportion $\pi$ with the sample proportion $p$ for calculating this SE. Recollect that when $\sigma$ is replaced by the sample SD $s$ while working out the SE of the mean for, say, calculating the **confidence interval (CI)**, the distribution becomes Student $t$ in place of Gaussian (normal). No such adjustment is done while calculating the CI for $\pi$, and Gaussian $\pm z_{(1-\alpha/2)}$*SE($p$) interval is used even after replacing $\pi$ with $p$. This introduces approximation that may not work even under **Gaussian conditions**.

Edwin Wilson, in his paper published in 1927 [1], refined the CI for a population proportion $\pi$ for a large sample using the algebraic quadratic equation to solve $(p-\pi)/\sqrt{\pi(1-\pi)/n} = \pm z_{1-\alpha/2}$. This is given by

$$\text{Wilson interval for } \pi\text{: } \frac{1}{1+z^2_{\alpha/2}/n}\left[p+\frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}+\frac{z^2_{\alpha/2}}{4n^2}}\right],$$

where $z_{1-\alpha/2}$ is the **Gaussian deviate** corresponding to the two-sided $100(1-\alpha)\%$ **confidence level** and $p$ is the proportion observed in the sample. For confidence level 95%, $z_{1-\alpha/2} = 1.96$. Compare this with the crude 95% CI given by $p \pm 1.96$*SE($p$), where SE($p$) is the **standard error** of $p$ estimated by $\sqrt{p(1-p)/n}$. Since the Wilson interval is based on the quantity $(p-\pi)/\sqrt{\pi(1-\pi)/n}$, which is also called score, this interval is sometimes also known as the Wilson *score* interval. This interval has good properties even for small $n$ and very small or very large values of $p$.

**TABLE W.6**
**Comparison of Gaussian Interval and Wilson Interval for Some Values of *n* and *p***

| | | 95% Gaussian Interval | | 95% Wilson Interval | |
|---|---|---|---|---|---|
| *p* | *n* | Lower Limit | Upper Limit | Lower Limit | Upper Limit |
| 0.1 | 10 | 0 (–0.086) | 0.286 | 0.018 | 0.404 |
| 0.1 | 30 | 0 (–0.007) | 0.207 | 0.035 | 0.256 |
| 0.1 | 100 | 0.041 | 0.159 | 0.055 | 0.174 |
| 0.4 | 10 | 0.096 | 0.704 | 0.168 | 0.687 |
| 0.4 | 30 | 0.225 | 0.575 | 0.246 | 0.577 |
| 0.4 | 100 | 0.304 | 0.496 | 0.309 | 0.498 |


Edwin Wilson

Presented in Table W.6 are the Gaussian CIs and Wilson CIs for two values of $p$, namely $p = 0.1$ and $p = 0.4$, and $n = 10$, 30, and 100. The Gaussian interval is far too deficient as it gives a negative lower limit of the CI when the observed sample proportion is 0.1 (even when $n$ is 30), which is obviously impossible for any probability (or population proportion), and we have to consider it to be 0. Only for $n = 100$, the limits corresponds well, particularly when the observed $p$ is 0.40.

The Wilson interval is a better approximation than the Gaussian and can be used for relatively small $n$, but this would still be different from the **exact confidence interval** for binomial $\pi$. The exact CI is valid for small samples too as it is based on **binomial distribution**.

1. Wilson EB. Probable inference, the law of succession, and statistical inference. *J Amer Stat Assoc* 1927; 22:209–12. http://www.jstor.org/discover/10.2307/2276774

## Winsorized mean, *see* trimmed mean

## Wishart distribution

The Wishart distribution is the multivariate analogue of **chi-square** distribution, devised by John Wishart in 1928 [1]. The sample **dispersion matrix** follows this distribution when a set of variables observed together has a **multivariate Gaussian distribution**.


John Wishart

A dispersion matrix is the row–column arrangement of variances and covariances of the variables in the set. Thus, this can be used to test the **null hypothesis** that the sample has come from a multivariate Gaussian distribution with a diagonal dispersion matrix. A diagonal dispersion matrix in this case implies that variables in the set are independent. If your set contains variables measuring functions of different organs, for example, bilirubin level for liver function, creatinine level for kidney function, systolic blood pressure for cardiac function, and vital capacity for lung function, you can find out whether any of these affects the others. It is known that various kidney functions affect one another, but it probably is not known whether or not measurements on different organs are related. This might be useful when tested for, say, people who have a low hemoglobin (Hb) level and for people who have a normal Hb level. This might generate a hypothesis regarding functions of which organs, if any, are simultaneously compromised by a low Hb level. However, the Wishart distribution is applicable only if the variables jointly have multivariate Gaussian distribution.

1. Wishart J. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika* 1928;20A (1–2):32–52. http://www.jstor.org/stable/2331939

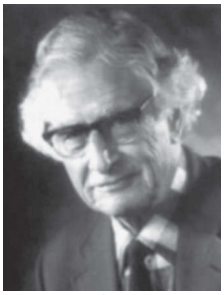
## worm plots, *see* quantile-by-quantile plot.

# Y

## Yates correction for continuity

This is the correction done to the **chi-square** criterion for testing the presence of association between two attributes in contingency tables for its better approximation to a continuous distribution. The correction is advised only for 2 × 2 tables and not for bigger ones. A contingency table contains the number of subjects in different cells—thus, it yields discrete values of the criterion. However, chi-square itself is a continuous distribution. When the Yates correction is applied, the value of the criterion tends to become continuous. This requires that 0.5 be subtracted from the difference between the observed and the expected frequencies in the numerator of the chi-square. The corrected chi-square criterion is as follows:

$$\text{Yates correction for continuity: } \chi^2 = \sum \frac{[(O-E)-0.5]^2}{E}.$$



Frank Yates

The correction was first suggested by Frank Yates in 1934 [1]. Any such correction becomes infructuous when the sample size is large because the result either way is nearly the same. Thus, this correction is advocated when any **cell frequency** in the contingency table is small, say, less than 5. The effect of this correction is that the value of the chi-square criterion is reduced because of this subtraction of 0.5 from each difference in the numerator that increases the *P*-value and makes it difficult to reject the null hypothesis of no association. Although this correction was accepted for a long time, it has been found to be too severe and to cause overcorrection in many situations, resulting in loss of statistical power to detect an association. The value of chi-square without correction has been found a better approximation than the one with the correction [2]. In any case, for small cell frequencies, we have the **Fisher exact test**, which does not need any such correction.


1. Yates F. Contingency table involving small numbers and the $\chi^2$ test. *J Royal Stat Soc* (Suppl) 1934;1(2):217–35. http://www.jstor.org/stable/2983604
2. Feinstein AR. *Principles of Medical Statistics*. Chapman & Hall/CRC, 2002: p. 251.


## years of life lost, *see* potential years of life lost (PYLL)

## Youden index, *see also* receiver operating characteristic (ROC) curve

The Youden index is used to assess the performance of quantitative diagnostic tests in terms of their **sensitivity and specificity** and is defined as

Youden index: $J$ = maximum value of (sensitivity + specificity − 1).

Note that this is maximum where sensitivity + specificity is maximum. Thus, the Youden index is a measure of the best performance of the test in terms of a combination of **sensitivity and specificity**. This was suggested by William Youden in 1950 [1].

For finding the Youden index, sensitivity and specificity are calculated for each value of the quantitative test. This is easily done when **receiver operating characteristic (ROC) curve** is drawn. In that case, the Youden index maximizes the vertical distance from the line of equality (diagonal line) to the point ($x$, $y$) on the ROC curve, as shown in Figure Y.1. The $x$-axis represents (1 − specificity), and the $y$-axis represents sensitivity. In other words, the Youden index $J$ is the largest distance of the ROC curve from the line of equality. The main aim of the Youden index is to maximize the difference between the true-positive rate ($s_n$) and the false-positive rate ($1 - s_p$), and a little algebra yields $J = \max(s_n + s_p - 1)$. For example, the Youden index of serum interleukin-18 (IL-18) at commencement of renal therapy of critically ill patients with acute kidney injury for hospital mortality was found to be 0.65 [2]. This is the maximum correct classification of the negative and positive cases by this test.

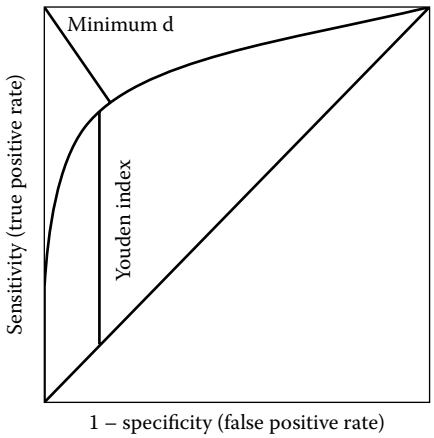



**FIGURE Y.1** ROC curve and Youden index.

Another purpose of the Youden index is to help locate the cutoff point of the quantitative test where its performance is best. For IL-18 in the example just cited, the best cutoff was 1786 pg/mL.

For further details and more applications of the Youden index, see Schisterman et al. [3].

1. Youden, WJ. Index for rating diagnostic tests. *Cancer* 1950;3:32–5. http://www.ncbi.nlm.nih.gov/pubmed/15405679
2. Lin CY, Chang CH, Fan PC, Tian YC, Chang MY, Jenq CC, Hung CC, Fang JT, Yang CW, Chen YC. Serum interleukin-18 at commencement of renal replacement therapy predicts short-term prognosis in critically ill patients with acute kidney injury. *PLoS One* 2013 May 31; 8(5):e66028. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3669263/
3. Schisterman EF, Perkins NJ, Liu A, Bondell H. Optimal cut-point and its corresponding Youden index to discriminate individuals using pooled blood samples. *Epidemiology* 2005;16(1):73–81. http://www4.stat.ncsu.edu/~hdbondel/OptimCut.pdf

## Yule Q, *see* association between dichotomous characteristics (degree of)

# Z

## Zelen design

Also called **randomized consent design**, this design was originally introduced to deal with some of the ethical problems facing clinicians as they enter patients into randomized clinical trials. For example, the patients in the control arm who receive standard treatment are not given the chance to be treated with the new treatment even if the new treatment is potentially better than the standard treatment, unless of course it is committed in the study protocol that the participants in the control arm will be given the chance to take the new treatment after the trial is over. Also, those allocated to the new treatment may want to opt for the standard treatment.

Once the patient's eligibility for entering the trial has been established, the patient is randomized to one of the treatments: A or B. Under the randomized consent design, patients randomized to one particular group are approached for consent: are they willing to receive the allocated treatment for their illness? The researcher and the patient openly discuss all potential risks, benefits, and treatment options. If the patient does consent, he/she receives the allocated treatment. On the other hand, after the discussion has taken place, if the patient does not give consent, he/she is assigned to receive the other treatment. This could be the standard treatment that the patient would get in any case if he/she were not part of the trial. This, in any case, is part of the process of **informed consent** that is generally followed before assigning treatment, but under a randomized consent design, this is done after the allocation. This kind of design was suggested by Marvin Zelen in 1979 [1].


Marvin Zelen

The informed consent process of the conventional randomized controlled trial (RCT) design may be ethical on some occasions but is not always ethical from the patients' perspective. In some cases, obtaining prior consent is difficult if not impossible because of the age or the condition of the subjects or because of the difficulty in explaining the scientific issues. Many patients show reluctance in participating in randomized clinical trials, as they are suspicious of the outcome. Sometimes, the detailed explanation process exhausts patients, even brings a *nocebo effect* (the opposite effects of placebo), resulting in refusal to participate in the trial, which is truly a waste of energy for both the patients and the investigators [2]. The Zelen design seeks to overcome this difficulty.

This design has some distinct advantages over the conventional RCT. Foremost is that it is perfectly ethical; it is also able to include all eligible subjects, and the problem arising from seeking informed consent is avoided. A major problem, however, is that this design does not allow double-blinding, as the allocation is known—and this can create serious bias in the responses. Also, the analysis of data from this kind of trial is difficult because the number of participants in each arm becomes a random number instead of a fixed number. The problem is compounded if a large number of eligible subjects refuse to remain in their allocated group.

Nevertheless, the design has been used with some success. Hatcher et al. [3] describe a protocol of a study on a treatment package including problem-solving therapy compared to the standard treatment in people who present to hospital after self-harm in New Zealand where a Zelen design is proposed. Nio et al. [4] used this type of design for a trial in Japan on neoadjuvant chemotherapy with tegafur plus uracil for gastric cancer.


1. Zelen M. A new design for randomized clinical trials. *N Engl J Med* 1979;300:1242–5. http://www.nejm.org/doi/full/10.1056/NEJM197905313002203
2. Hamajima N, Yuasa H, Nakamura M, Tajima K, Tominaga S. Nested consent design for clinical trials. *Jpn J Clin Oncol* 1998;28(5):329–32. http://jjco.oxfordjournals.org/content/28/5/329.full
3. Hatcher S, Sharon C, House A, Collings S, Parag V, Collins N. The ACCESS study a Zelen randomised controlled trial of a treatment package including problem solving therapy compared to treatment as usual in people who present to hospital after self-harm: Study protocol for a randomised controlled trial. *Trials* 2011 May 26;12:135. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3117717/
4. Nio Y, Koike M, Omori H, Hashimoto K, Itakura M, Yano S, Higami T, Maruyama R. A randomized consent design trial of neoadjuvant chemotherapy with tegafur plus uracil (UFT) for gastric cancer—A single institute study. *Anticancer Res* 2004 May–Jun;24(3b):1879–87. http://ar.iiarjournals.org/content/24/3B/1879.long


## Zelen test

This is an exact test for comparing the **odds ratio (OR)** or **relative risk (RR)** in several $2 \times 2$ tables that can arise due to stratification. If there are $K$ strata, the data would be expressed in a $K \times 2 \times 2$ table. The usual procedure for homogeneity of ORs across strata is the **Breslow–Day test**, and then Zelen test is an exact counterpart of this test. The test was devised by Marvin Zelen in 1971 [1].

The Zelen test can be regarded as an extension of the **Fisher exact test** as this also uses the probability of all possible $K \times 2 \times 2$ tables in favor of the alternative hypothesis with the same row, column, and stratum totals as the observed table and with the same sum of cell frequencies. The $P$-value is the probability of the observed frequencies conditioned on the fixed margins. Significance would mean that the OR is different in different strata, which also means, in this case, that the three-way interaction is significant. Patil [2] generalized this procedure for $I \times J \times K$ tables. However, Halperin et al. [3] cited examples where this test contradicts the data. Hirji et al. [4] have presented alternative approaches that are too intricate for our audience.

1. Zelen M. The analysis of several 2 × 2 tables. *Biometrika* 1971;58(1); 129–37. http://biomet.oxfordjournals.org/content/58/1/129.abstract
2. Patil DK. Interaction test for three-dimensional contingency tables. *J Amer Stat Assoc* 1974;69:164–8. http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1974.10480145
3. Halperin M, Ware JH, Byar DP, Mantel N, Brown CC, Koziol J, Gail M, Green SB. Testing for interaction in an $I \times J \times K$ contingency table. *Biometrika* 1977;64(2):271–5. http://www.jstor.org/stable/2335693
4. Hirji KF, Vollset SE, Reis IM, Afifi AA. Exact tests for interaction in several 2 × 2 tables. *J Comp Graphical Stat* 1996;5(3):209–24. http://www.jstor.org/stable/1390800

## zero-inflated models

Zero-inflated models are those that account for an unusually high number of 0's in the data.

In many medical situations, zero value of the variable of interest is quite common. This can happen with number of cigarettes smoked per day, where it is 0 for nonsmokers; number of extramarital sex encounters in 1 month by married persons, number of days with sickness in a month; number of angina attacks in a month in patients with cardiovascular diseases; etc. These examples are for counts, but 0 can commonly occur with continuous measurements also, such as with differences between values before and after a procedure. One such example is the deviation of organs with a tumor when irradiated for treatment. Sometimes, the deviations do not occur at all in this procedure, and sometimes, they are so small that even the most sophisticated computed tomography (CT) scan fails to detect them. Thus, the value would be 0o for many subjects. Just for context, if the deviation is large, radiation could damage the adjacent organs—thus, close track of deviations is kept.

Analysis of data with many 0's is problematic because such data do not conform to any of the established statistical distributions. Counts otherwise generally follow a **Poisson distribution**, but if mean $\mu = 2$, $P(x = 0) = e^{-2}\lambda^0/0! = 0.135$, i.e., only 13.5% of values are expected to be 0 under this model. If they are substantially more, say, 20% or more, the Poisson model will not be adequate to represent this variable. Similarly, if the measurement is continuous with mean = 1 mm and standard deviation (SD) = 0.5 mm for deviations of organs in our tumor radiation example (note in this case that the deviations in either direction are measured as positive—there are no negative values), and if any deviation < 0.25 mm is undetectable, the probability of zero value is $P(0 < x < 0.25|\ \mu = 1, \sigma = 0.5) = 0.044$ for Gaussian distribution. If zero or undetectable deviation is actually observed in half of the subjects against the expected 4.4%, obviously, the Gaussian distribution is not appropriate. In addition, in this case, Gaussian distribution will not be able to rule out negative values.

Among the many approaches suggested for analyzing data with many 0's, one is the **negative binomial distribution**. If this does not fit well, consider zero-inflated Poisson distribution for counts, which is stated as

$$P(x) = \pi + (1 - \pi)e^{-\lambda} \text{ for } x = 0 \text{ and}$$

$$P(x) = (1 - \pi)\ e^{-\lambda}\lambda^x/x! \text{ for } x = 1, 2, 3, \ldots,$$

where $\pi$ is the expected proportion of extra 0's and can be negative in a zero-deflated model, where the number of 0's is much less than expected with a Poisson distribution, and $\lambda$ is the usual Poisson parameter. This adjustment has been discussed by Rideout et al. [1]. The other approach suggested by Fletcher et al. [2] for continuous variables is combining ordinary (quantitative) and logistic regressions. Briefly, this requires splitting the data in two parts—the first with $x = 0$ and the second with $x > 0$. Two separate analyses are done: first logistic regression for the $x = 0$ part, and second using ordinary regression for the $x > 0$ part. These analyses are combined to get a unified conclusion.

**TABLE Z.1**
**Distribution of Multiple Births (Twins, Triplets, and Quadruplets) by Sex**

| Number of Female Children | Number of Male Children | | | | | |
|---|---|---|---|---|---|---|
| | **0** | **1** | **2** | **3** | **4** | **Total** |
| 0 | × | × | 17 | 5 | 0 | **22** |
| 1 | × | 25 | 6 | 2 | × | **33** |
| 2 | 12 | 3 | 0 | × | × | **15** |
| 3 | 4 | 0 | × | × | × | **4** |
| 4 | 1 | × | × | × | × | **1** |
| **Total** | **17** | **28** | **23** | **7** | **0** | **75** |

Our discussion is for observed 0's, that is, a value other than 0 was possible but happened to be 0 for some or many subjects in our sample. The other possibility is that zero value is bound to occur, called *structural zero*. Consider a study on multiple births—twins, triplets, and quadruplets. The births are classified by sex. The distribution may be as displayed in Table Z.1. There are some observed 0's in this table. In addition, there is a "×" sign in some cells where no frequency is possible. Such tables are called **incomplete tables**. Special methods are required to analyze such tables. For a summary of such methods, see Agresti [3].

1. Rideout M, Demetrio CGB, Hinde J. Models for count data with many zeros. *Int Bioletric Conf*, Cape Town, 1998. https://www.kent.ac.uk/smsas/personal/msr/webfiles/zip/ibc_fin.pdf, last accessed December 21, 2015.
2. Fletcher D, Darryl MacKenzie D, Villouta E. Modelling skewed data with many zeros: A simple approach combining ordinary and logistic regression. *Environ Ecol Stat* 2005;12(1):45–54. http://link.springer.com/article/10.1007/s10651-005-6817-1, last accessed September 5, 2015.
3. Agresti A. *Categorical Data Analysis*. Wiley, 1990: pp. 244–50.

## *Z*-scores

The *Z*-score of a quantitative value of measurement is an expression of how far this value is from its mean in standard deviation (SD) units. This is defined as

$$Z\text{-score: } Z = \frac{x - \mu}{\sigma},$$

where $x$ is the value of the measurement, $\mu$ its mean, and $\sigma$ its SD. Suppose the body temperature in a healthy person is 97.8°F and it is known that the population mean $\mu$ of healthy people is 98.6°F and population SD $\sigma$ is 0.4°F; then the *Z*-score of the person's body temperature is (97.8 – 98.6)/0.4 = –2. In case the distribution of body temperature has a **Gaussian** (normal) pattern, this can be translated to, say, that the chance of this kind of temperature or more extreme (less than or equal to 97.8°F) is only 2.5%, where this chance comes from the probabilities of the Gaussian variate.

When the distribution is not Gaussian, it is difficult to assign a probability to any measurement, but *Z*-scores are still used because of their extremely useful property, namely, that mean = 0 and SD = 1. This comes directly from the way *Z*-scores are defined. Because mean = 0 and SD = 1, this is also called **standardized deviate** or **standardized variate**. The process of subtracting the mean and dividing by the SD is called **standardization of values**.

Whereas the *Z*-score is defined in terms of population mean and population SD, the standardized deviate or standardization of values can be done using the sample mean and sample SD also. If there are 15 subjects in a sample, calculate the mean and SD of these values, subtract the sample mean $\bar{x}$ from each value, divide by the sample SD $s$, and get standardized values = $\frac{x-\bar{x}}{s}$. This is demonstrated in Table Z.2 for serum urea levels observed in 15 subjects in the sample. The sample mean in these data is 34 mg/dL, and the sample SD is 14.21267 mg/dL, but the mean and SD of the standardized values are 0 and 1, respectively. Extreme values do not matter in this case. In our example, one person had a serum urea level of 78 mg/dL, whereas all others were between 15 and 45 mg/dL. This has not affected the mean or SD of the standardized values; neither is a Gaussian pattern required for this process.

A big advantage of standardized values is that they can be compared for entirely different measurements. For example, if one person's standardized value of serum urea level is −0.1407 and his/her standardized value for intelligence quotient (IQ) is +1.7384, it can be easily concluded that the serum urea level of this person is slightly less than the average, while IQ is much above the average. When comparable measurements, such as serum urea and serum creatinine, are available, this type of comparison of standardized values can provide useful insight to the clinician as to what measurement could be the culprit. Precisely for this reason, standardized values are sometimes used in regression to obtain the standardized **regression coefficients** because then, these can be compared across the regressors to find which regressor is contributing more to the variation in response than the others.

**TABLE Z.2:**
**Mean and SD of *Z*-scores for Sample Values**

| | Serum Urea (mg/dL) $x$ | Standardized Values $\frac{x-\bar{x}}{s}$ |
|---|---|---|
| | 32 | −0.1407 |
| | 45 | 0.7740 |
| | 23 | −0.7740 |
| | 27 | −0.4925 |
| | 36 | 0.1407 |
| | 31 | −0.2111 |
| | 29 | −0.3518 |
| | 36 | 0.1407 |
| | 42 | 0.5629 |
| | 78 | 3.0958 |
| | 26 | −0.5629 |
| | 15 | −1.3368 |
| | 27 | −0.4925 |
| | 30 | −0.2814 |
| | 33 | −0.0704 |
| **Sum** | **510** | **0.0000** |
| **Mean ($\bar{x}$)** | **34** | **0.0000** |
| **Sample SD ($s$)** | **14.21267** | **1.0000** |

Sometimes, particularly for highly skewed distributions, such as of weight in children, you may occasionally find that the median is used for calculating *Z*-score in place of the mean. If that is done, the property of mean = 0 and SD = 1 may not hold, and the transformed values cannot be called "standardized." Use of the median is done for pragmatic reasons since the mean in the case of a highly skewed distribution is not a representative value.

## *z*-test

This is the most basic of all statistical tests and is based on the ***Z*-score** of the sample summary. The *Z*-score for individual values is $\frac{x-\mu}{\sigma}$, where $x$ is the value of the measurement, $\mu$ its mean, and $\sigma$ its standard deviation (SD), and the mean of $Z$ is always 0 and variance always 1. Replace $x$ with any sample summary, $\mu$ with its mean, and $\sigma$ with its SD, now called standard error (SE), and get the criterion for the $z$-test for random samples. That is, in its most generalized form,

$$z\text{-test: } z = \frac{\text{sample summary} - \text{its mean}}{\text{its SE}}.$$

The sample summary could be the sample mean, sample median, sample proportion, difference between these two groups, etc. When the underlying distribution of $x$ is **Gaussian** (normal), the distribution of most sample summaries is also Gaussian, and in that case, this $z$ has a Gaussian distribution with mean = 0 and SD = 1. For testing a hypothesis, the value of this $z$ is calculated under the null hypothesis. A large value of $|z|$, say, more than 2, would indicate that the value of the sample summary is improbable (***P*-value** is small) under the null, and thus, the null is rejected. That is, the value of $z$ obtained from the data under the null is checked against a Gaussian table for finding the one-tailed or two-tailed *P*-value depending on whether the alternative is **one-tailed or two-tailed**, and is rejected if the *P*-value is less than the prefixed **level of significance**. This is popularly known as the $z$-test.

If the underlying distribution is different from Gaussian, the **central limit theorem** comes to the rescues for large samples, particularly for mean and proportion, and the test then can still be used under **Gaussian conditions**. Just to remove any confusion, note that the same criterion becomes **Student *t*** under Gaussian conditions when the SE in the denominator is replaced by the *estimated* SE. The $z$-test requires that the "population" SE be used in the denominator, but for many applications, estimated SE is used in place of the population SE in the realization that Student $t$ also becomes Gaussian for a sufficiently large sample. In summary, the $z$-test, say, for a sample mean, is exact for small samples when the underlying distribution is Gaussian and the SE is known and not estimated. In all other cases, it is approximate and requires a large sample. The requirement of a large sample is twice as important for this because of two approximations: (i) for approximating the Student $t$ with Gaussian $z$ and (ii) for using the result of the central limit theorem for non-Gaussian distribution of the sample summary under test. For the first, generally, a sample of size at least 30 is considered adequate, but for the second, a very large sample may be required for some sample summaries, such as odds ratio.

The $z$-test takes a different form depending upon what sample summary is being used for the test. This may require independence of observations, and independence of samples in the case of comparison of two groups, but the equality of variances is not required except for simplifying the calculations. For $z$-tests for different

situations, see the topics **comparison of two or more correlation coefficients**, **comparison of two or more odds ratios**, **comparison of two or more proportions**, **comparison of two or more relative risks**, and **comparison of one-sample mean and proportion with a specified value**. In the last case, a *z*-test is used for proportion and not for the mean (Student *t* is used for the mean), and a *z*-test in all these situations is used for one and two groups but not for more than two groups. For all *t*-tests, such as for comparison of one sample mean with a specified value, comparison of intercepts with a specified value or in two simple linear regressions, comparison of two means, and comparison of a regression coefficient with a specified value or of two regression coefficients, it becomes a *z*-test in the rare case where the population SD is known and not estimated from the sample.