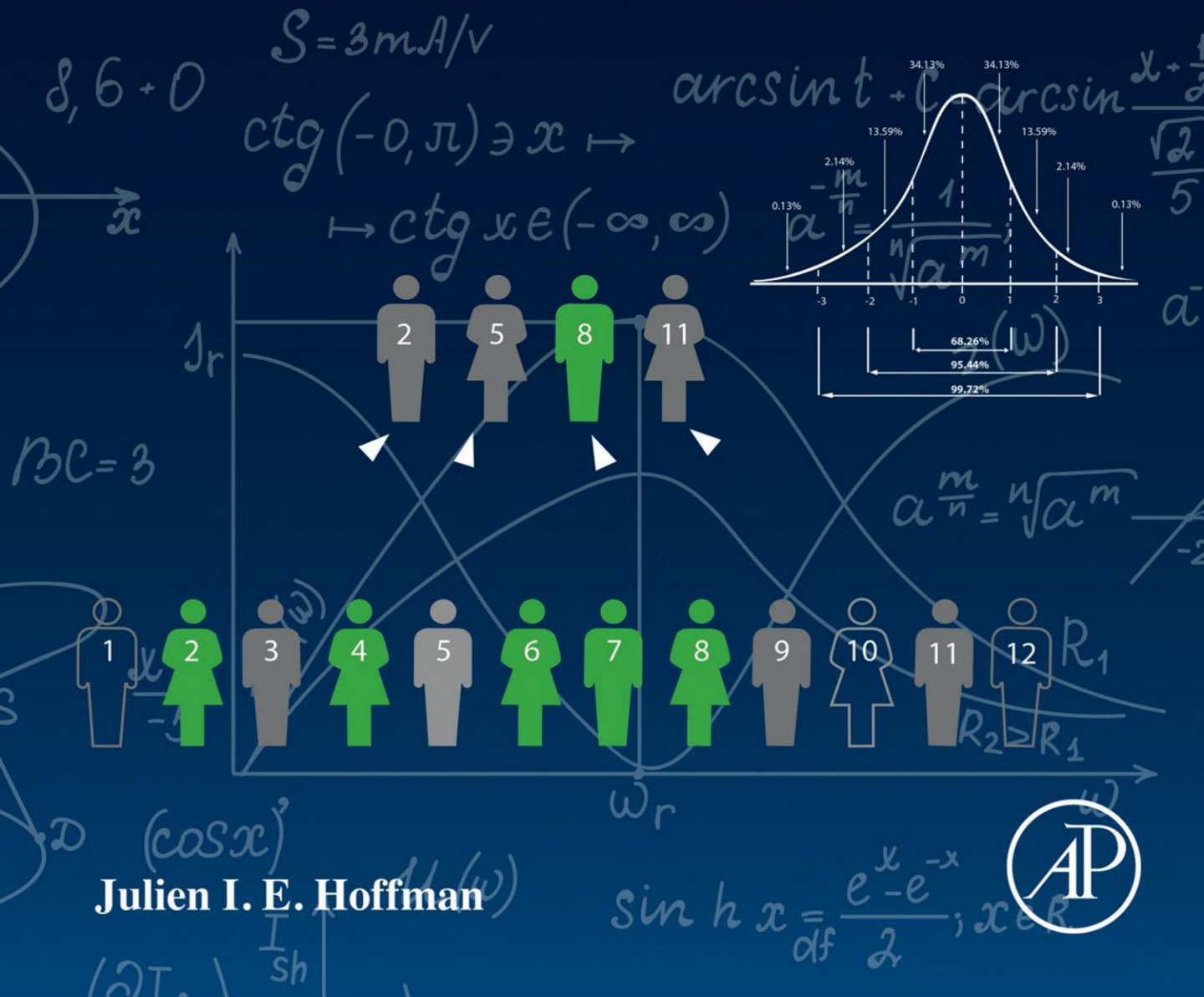


Basic Biostatistics *for Medical and Biomedical Practitioners*



BASIC BIOSTATISTICS FOR MEDICAL AND BIOMEDICAL PRACTITIONERS

Second Edition

JULIEN I.E. HOFFMAN



ACADEMIC PRESS

An imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
525 B Street, Suite 1650, San Diego, CA 92101, United States
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

© 2019 Elsevier Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

ISBN 978-0-12-817084-7

For information on all Academic Press publications
visit our website at <https://www.elsevier.com/books-and-journals>



Working together
to grow libraries in
developing countries

www.elsevier.com • www.bookaid.org

Publisher: Stacy Masucci

Acquisition Editor: Rafael E. Teixeira

Editorial Project Manager: Sandra Harron

Production Project Manager: Maria Bernard

Designer: Victoria Pearson

Typeset by SPi Global, India

ABOUT THE AUTHOR

Julien Hoffman was born and educated in Salisbury (now named Harare) in Southern Rhodesia (now named Zimbabwe) in 1925. He entered the Faculty of Medicine at the University of the Witwatersrand in Johannesburg, South Africa, and obtained a B.Sc. Hons in Anatomy and Physiology in 1945 and a medical degree (M.B., B.Ch.) in 1949. After working for almost 3 years at the Central Middlesex Hospital in London, England as a house officer and registrar in Medicine he returned to Johannesburg as a senior registrar in the Department of Medicine for 18 months, and then returned to England to work for the Medical Research Council. In 1957 he went to Boston Children's Hospital to study congenital heart disease, and this was followed by 15 months as a Fellow in the Cardiovascular Research Institute at the University of California in San Francisco.

In 1962 he joined the faculty at the Albert Einstein College of Medicine in New York City as an Assistant Professor of Pediatrics and Internal Medicine, and in 1966 returned to the University of California in San Francisco as Associate Professor of Pediatrics and member of the Cardiovascular Research Institute. He was a clinical pediatric cardiologist, taking care of children with heart disease, but spent about 50% time running a research laboratory to study the pathophysiology of the coronary circulation.

His interest in statistics began while taking his Science degree. After going to England he took short courses in Biostatistics from Bradford Hill and his colleagues. On returning to South Africa, as the only department member to know anything about statistics, he was assigned to perform statistical analyses for other members of the department. This was a period of learning by trial and error, helped by Dr. J. Kerrich, head of the University's Division of Statistics.

When he went to San Francisco as a Fellow, he was assigned to give statistics lectures to the Fellows and Residents, and when he returned to San Francisco in 1966 he gave an officially sanctioned course in Biostatistics to research Fellows and Residents. These lectures were given annually for over 30 years. He was a member of the Biostatistics group, a semiformal group that supervised statistical teaching and consultation. He also was a statistical consultant to the journal *Circulation Research*, and was often assigned manuscripts by other journals, mostly from the American Heart Association, to check on the statistical procedures used.

PREFACE

In my 40-year experience of helping medical research workers to analyze their studies, certain problems arose frequently. Many investigators need to compare two Poisson distributions, yet some introductory books on Biostatistics give little attention to the Poisson distribution, even though it answers a frequent question about how often an uncommon event (such as a rare form of cancer) occurs. Few people can navigate the minefield of multiple comparisons, involved when several different groups are compared, often done incorrectly by performing multiple t -tests, yet most elementary texts do not deal with this problem adequately. Problems of repeated measures analysis in which several measurements are made in each member of the group and thus are not independent occur frequently in medical research but are not often discussed. Prediction and tolerance tests are often needed to set ranges of normal values so that a single measurement can be assessed as normal or abnormal (such as a single fasting blood glucose concentration). Because most basic books do not discuss this problem, most people incorrectly set confidence limits that should apply only to mean values. An incentive for this book was the lack of books in introductory Biostatistics that could be understood relatively easily, but nevertheless were advanced enough that most readers would not need to consult several additional books or hunt through unfamiliar journals for appropriate tests.

The book is intended to help physicians and biologists who had a short course on Statistics several years ago but have forgotten all but a few of the terms and concepts and have not used their knowledge of statistics for reading the literature critically or for designing experiments. Most of them have neither time nor interest in taking formal courses in Statistics. The general aim is to extend their knowledge of statistics, to indicate when various tests are applicable, what their requirements are, and what can happen when they are used inappropriately.

This book has four components.

1. It covers the standard statistical approaches for making descriptions and inferences—for example, mean and standard deviation, confidence limits, hypothesis testing, t -tests, chi-square tests, binomial, Poisson and normal distributions, analysis of variance, linear regression and correlation, logistic regression, and survival analysis—to help readers understand what hypotheses are being tested, how the tests are constructed, how to look for and avoid using inappropriate tests, and how to interpret the results. Examples of injudicious use of these tests are given. Although some basic formulas are presented, these are not essential for understanding what the tests do and how they should be interpreted. However, following simple algebraic equations leads to better understanding of the technique. If you

can remember that $(a - b)^2 = a^2 - 2ab + b^2$ you will be able to follow almost all the formulas provided.

2. Some chapters include a section on advanced methods that should be ignored on a first reading but provide information when needed, and others have an appendix where some simple algebraic proofs are given. As this is not intended to be a mathematically rigorous book most mathematical proofs are omitted, but a few are important teaching tools in their own right and should be studied. However, knowledge of mathematics (and differential calculus in particular) beyond elementary algebra is not required to use the material provided in this book.
3. Scattered throughout the chapters are variations on tests that are often needed but not frequently found in basic texts. These sections are often labeled “Alternative Methods,” and they should be read and understood because they often provide simpler and more effective ways of approaching statistical inference. These include:
 - a. Robust statistics for dealing with grossly abnormal distributions, both univariate and bivariate.
 - b. Equivalence or noninferiority testing, to determine if a new drug or vaccine is equivalent to or not inferior to those in standard use.
 - c. Finding the break point between two regression lines. For example, if the lactate: pyruvate ratio remains unchanged when systemic oxygen delivery is reduced below normal until some critical point is reached when the ratio starts to rise, how do we determine the critical oxygen delivery value?
 - d. Competing risks analysis used when following the survival of a group of patients after some treatment, say replacement of the mitral valve, and allowing for deaths from noncardiac causes.
 - e. Tolerance and prediction testing to determine if a single new measurement is compatible with a normative group.
 - f. Cross-over tests, in which for a group of subjects each person receives two treatments, thus acting as his or her own control.
 - g. Use of weighted kappa statistics for evaluating how much two observers agree on a diagnosis.
 - h. Bowker’s test for evaluating more than two sets of paired data.
4. Some chapters describe more complex inferences and their associated tests. The average reader will not be able to use these tests without consulting a statistician but does need to know that these techniques exist and, if even vaguely, what to look for and how to interpret the results of these tests when they appear in publications. These subjects include:
 - a. Poisson regression ([Chapter 34](#)), in which a predicted count, for example, the number of carious teeth, is determined by how many subjects have 0, 1, 2, and so on carious teeth.

- b. Resampling methods ([Chapter 37](#)), in which computer-intensive calculations allow the determination of the distributions and confidence limits for mean, median, standard deviations, correlation coefficients, and many other parameters without needing to assume a particular distribution.
- c. The negative binomial distribution ([Chapter 19](#)) that allows investigation of distributions that are not random but in which the data are aggregated.
- d. Metaanalysis ([Chapter 36](#)), in which the results of several small studies are aggregated to provide a larger sample, for example, combining several small studies of the effects of beta-adrenergic blockers on the incidence of a second myocardial infarction, is often used. The pitfalls of doing such an analysis are seldom made clear in basic statistics texts.
- e. Every reader should be aware of multiple and nonlinear regression techniques ([Chapter 30](#)) because they may be important in planning experiments. They are also used frequently in publications, but usually without mentioning their drawbacks.

With the availability of personal computers and statistical software, it is no longer necessary to detail computations that should be done by computer programs that save time and prevent arithmetic errors. There are many simple free online programs that calculate most of the commonly used statistical descriptions as well as commonly used inferential tests along with their associated graphics, and hyperlinks are provided for these. More complex tests require commercial programs.

Problems are given in appropriate chapters. They are placed after a procedure is described so that the reader can immediately practice what has been studied to make sure that the message is understood. Although the problems are simple and could be done by hand, it is better to use one of the recommended online calculators. This frees up time for the reader to consider what the results mean.

The simpler arithmetic techniques, however, are still described in this book because they lead to better understanding of statistical methods and show the reader where various components of the calculation come from, and how the components are used and interpreted. In place of tedious instructions for doing the more complex arithmetic procedures there is a greater concentration on the prerequisites for doing each test and for interpreting the results. It is easier than ever for the student to think about what the statistical tests are doing and how they contribute to solving the problem. On the other hand, we need to resist the temptation to give a cookbook approach to solving problems without giving some understanding of their bases, even though this may involve some elementary algebra. As [Good and Hardin \(2009\)](#) wrote: “Don’t be too quick to turn on the computer. Bypassing the brain to compute by reflex is a sure recipe for disaster.”

Each chapter has a bibliography for those who want to learn more about Statistics. Some references are to books that contain tables or formulas that may occasionally be needed. Some references give the origins of the data sets discussed. Some discuss tests

not described in detail in this book. Finally, some are basically nonmathematical discussions of the subject; reading these will expand your knowledge.

Many people devise and carry out their own experiments, and for them a good knowledge of statistics is essential. There are many statistical consultants, but not enough of them to advise every investigator who is designing an experiment. An investigator should be able to develop efficient experiments in most fields and should reserve consultation only for the more complex of these. But more numerous and important are those who do not intend to do their own research. Even if the person is not a research worker, he or she is still responsible for assessing the merit of the articles that they are reading. It is no longer good enough for such a reader to know technical information only about the drugs used, the techniques for measuring pressure and flow, or to understand the physiologic and biochemical changes that take place in the disease in question. It is as essential for the reader to learn to read critically and to know how to evaluate the statistical tests used. Who has not been impressed by the argument that because Fisher's exact test showed a probability of 0.003, the two groups were different, or that because the probability was 0.08 the two groups were not different? Who has heard of Fisher's exact test? Of McNemar's test? Of the Kolmogorov-Smirnov two-group test? Of Poisson regression? And if the reader has not heard of them, how can he or she know if they should have been used, or have been used and interpreted correctly? With the pace at which new knowledge is appearing and being incorporated into medical practice, everyone engaged in medical research or practice needs to be well grounded in Statistics. **Statistical thinking is not an addendum to the scientific method but is an integral component of it.**

What can you expect to achieve after studying this book?

1. How to take account of variation in assessing the importance of measured values or the difference between these values in two or more groups.
2. A better understanding of the role of chance in producing unexpected results, and how to make decisions about what the next steps should be.
3. An appreciation of Bayes' theorem, that is, how to improve estimates of probability by adding in potential explanatory variables, and the importance of the population prevalence in making clinical predictions.
4. An ability to perform simple statistical tests (e.g., *t*-tests, chi-square, linear regression and correlation, McNemar's test, simple analysis of variance, calculate odds ratios and their confidence limits, and calculate simple survival tables), as well as understanding their limitations.
5. Realization that the usual attitude that $P < .05$ has ruled out the null hypothesis is open to question and should not be used by itself to reach a final conclusion.
6. Appreciate that there are many statistical techniques that can be used to address specific problems, such as metaanalysis, bioassay methods, nonlinear and multiple regression, negative binomial distributions, Poisson regression, and time series analysis.

It is unlikely that after studying this book you will be able to perform these analyses on your own, but you should know that such methods exist and that a statistical consultant can help you to choose the right analytic method.

If an investigator is not well acquainted with principles of experimental design, for example, random selection, sample size, the expenditure of time and money will be wasted. This is unethical and is even more important in clinical research because subjects may be submitted to an unreliable experiment.

Statistical procedures are technologies to help investigators interpret their data and are not ends in themselves. Like most technologies, Statistics is a good servant but a bad master. Francis Galton wrote in 1889: “Some people hate the very name of statistics, but I find them full of beauty and interest. Whenever they are not brutalized but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated phenomena is extraordinary (Galton, 1889).”

REFERENCES

- Galton, F.I., 1889. *Natural Inheritance*. Macmillan and Company, London, p. 62.
Good, P.I., Hardin, J.W., 2009. *Common Errors in Statistics (and How to Avoid Them)*. John Wiley & Sons, Hoboken, NJ, p. 274.

ACKNOWLEDGMENTS

Many people have helped me eradicate errors and clumsy phrasing. I owe a debt of gratitude to Joseph P. Archie, Jr., Stanton Glantz, Gus Vlahakes, Calvin Zippin and particularly to Raul Domenech. My special thanks to my wife for help with editing.

SECTION I

Basic Aspects of Statistics

CHAPTER 1

Basic Concepts

INTRODUCTION

Statistics is a crucial part of scientific method for testing theories derived from empirical studies. In most scientific studies the investigator starts with an idea, examines the literature to determine what is and is not known about the subject, formulates a hypothesis, decides what and how measurements should be made, collects and analyzes the results, and draws conclusions from them. Excellent introductions to these processes are provided by [Moses \(1985\)](#), [Altman \(1992\)](#), [Hulley et al. \(2007\)](#), and [Easterling \(2015\)](#).

One of the main aims of scientific study is to elucidate causal relationships between a stimulus and a response: if A, then B. Because responses to a given stimulus usually vary, we need to deal with variation and the associated uncertainty. In adult males in the United States, there is some typical number that characterizes weight, for example, ~ 75 kg, but there is variation around that typical value. Although fasting normal blood glucose is ~ 100 mg/dL, not every normal person will have this blood sugar concentration. If we measure the length of 2-in. long 20-gauge needles with great accuracy, there will be variation in length from needle to needle. It is an essential role of statistical thought to deal with the uncertainty caused by variability. One of the major ways to assess uncertainty is by doing an appropriate statistical test, but the test is merely the instrument that produces a result that needs to be interpreted. Performing a statistical test without the correct interpretation may lead to incorrect conclusions.

Populations and samples

It is essential to distinguish between measurements in an entire group with a given characteristic, known as a *population*, and those in a subset known as a *sample* drawn from the larger population. A population consists of all the possible measurements for a given variable; for example, heights of all 10-year-old children in Iowa, or all fasting blood glucose measurements in children aged 5–7 years of age. Because it may be impossible or impractical to make all the measurements in a population, we take a sample of the population, make the measurements, and then make certain inferences about the population from which the sample came. Almost all population symbols are Greek letters, and almost all sample symbols are in our usual Roman alphabet. For example, the population mean is symbolized by μ whereas the sample mean is symbolized by \bar{X} , and the slope of a straight line (e.g., relating height to blood pressure in a population of children) is symbolized by β , but in a sample from that population is symbolized by b .

The effects of variability

How does variability interfere with drawing valid conclusions from the data obtained in a study? In the year 2000 adult Dutch males were 4.9 cm taller than their US counterparts (Komlos and Lauderdale, 2007). Several questions can be asked about this result. First, how the measurements were made and with what accuracy; this is not a statistical question, but one that requires knowledge of the field of study; second, how likely is that difference in height in the samples to be a true measure of the height differences?

One way to answer this question would be to measure the whole *population*. All the adult males in the Netherlands constitute one population, and all adult males in the United States form another population. Heights could be measured in the whole population of adult males in both countries, but this would be an enormous undertaking in time, resources, and money. It is much easier to select a sample of adult males from each country and use the resultant measurements to make inferences about national heights. We need a precise definition of the population being studied. For example, national heights depend on the year of measurement. In 1935 adult males in the Netherlands and the United States had equal average heights that were both less than those measured in the year 2000. Population 1935 and Population 2000 are two different populations.

If the measurements were made in the whole population, the results would be unambiguous, and no statistical inference is needed: in year 2000 adult males in the Netherlands were on average 4.9 cm taller than their US counterparts. But if the measurements are actually made in a sample from each country, another question arises. How well does the difference in the averages of the two samples reflect the true difference in the averages of the two populations? Is it possible that the sample from the Netherlands contained a disproportionate number of tall subjects, and that the two population average heights differ by less than 4.9 cm? Indeed, it is possible, either because of bias in selecting subjects or even at random. The term *random* means “governed by chance,” so that any member of the population has an equal chance of being included in the sample; if the sample taken was only of professional basketball players, then their heights do not represent the heights in the general population. Issues of randomization and selection bias are dealt with in detail in [Chapter 38](#). What statistical analysis can do is to allow us to infer from the samples to the population, and to indicate, based on our samples, the likely values for the population averages.

The same considerations apply to experiments. In a population of people with diabetes mellitus, fasting blood glucose concentrations might range from 110 to 275 mg/dL. If we select two samples of 10 patients each from this population, almost certainly one group will have some concentrations higher or lower than any concentrations in the other group, and the average concentrations in the two groups are unlikely to be the same. If the groups had been given different medications that in reality had no effect on fasting blood glucose concentration, we might conclude incorrectly that the group with the lower mean blood glucose concentration had been given an active glucose-

lowering medication. It is to guard against this type of error (and many other types) that we need to think about the role of variability in statistical inference.

Variables and parameters

A measurement with different values in different people, times, or things is a *variable*. Variables such as weight, height, age, or blood pressure are measured in appropriate units. Others, such as eye color, gender, or presence or absence of illness, are *attributes*, and we count the number of items possessing each attribute. Sometimes we deal with one variable at a time, for example, weight gain of groups of young rats on different diets; these lead to univariate statistical descriptions and inferences. At other times, we relate one variable to another, such as age and height in growing children, and then we use bivariate statistics. Finally, many different variables might be involved (multivariate statistics).

Measured variables have several components. If we measure the weights of different people, there will be differences due to individual genetic and environmental (nutritional) influences. These differences are due to the biological processes that control weight and may be the signals that we are examining. On the other hand, if we measured the weight of one particular person several times on a very sensitive scale, we might get different values. The person's true weight is not changing over a few seconds, but there is variability of the measurement process itself. This source of variability is sometimes known as *noise*, and we try to eradicate or minimize it. Finally, the signal might not represent the true value. For example, if each person wore heavy shoes, their weights would all be greater than they should be. This represents a consistent *bias*, and if possible such biases should be detected and eliminated. These ideas produce a model:

$$\begin{array}{ccccccc}
 \text{Measured value} & = & \text{True value} & + & \text{Bias} & + & \text{Noise} \\
 \downarrow & & \downarrow & & \downarrow & & \downarrow \\
 \text{We get this} & & \text{We want this} & & \text{We do not want these} & &
 \end{array}$$

The more we can reduce the second and third sources of error (bias and noise), the closer will be the measured and true values. Statistics is therefore a method for separating the signal from the noise in which it is embedded. Statistical tests help to allow for noise but do not necessarily eliminate bias.

“Noise” includes not only measurement error but also legitimate variation. In an example of adult male weights in two regions, the true value sought is the average in each region. Any one person can be above or below that average for reasons of genetics or nutrition, so that the model is slightly different. It is:

Measured weight in subject A = Average regional weight \pm individual difference between the weight of subject A and the average. This individual difference from the average is known as “error,” often symbolized by ϵ . This is a form of noise that is biological, not due to measurement error, but it plays the same role in making it difficult to determine the true value of the desired signal.

The term “*parameter*” has different meanings in different fields of study. In Statistics, it is usually used to indicate the numerical characteristic of a population, such as the mean μ (average), slope of a relationship β , proportion of positive results π , and so on.

BASIC USES OF STATISTICS

Description

The first use is *descriptive*: how many people survived therapy A? What was the average length of the femur at any given age in children with growth hormone deficiency with and without treatment? How close is the relation of height to blood pressure? There are many ways of quantifying these descriptions, most of them relatively simple. The results obtained are termed *point estimates*.

These descriptions concern a specific set of observations and are unlikely to be identical for another similar set of data. What is important is how much other similar sets of data from the same population might vary from each other. One way of determining this would be to draw many samples from the population and determine how their averages vary, but there is a simpler way to obtain the same answer. Elementary statistical theory provides limits within which 99%, 95%, 90% (or any other percentage desired) of the averages of all future sets of similar data will fall. Thus the sample average (or mean) height of 97 children of a particular age may be 99 cm and the standard deviation (a measure of variability) may be 3.05 cm. This point estimate of the mean and the observed standard deviation then are used to predict that 95% of all similar sized groups of children of that age will have mean heights varying between 98.4 and 99.6 cm; these are referred to as *95% confidence limits or intervals*, discussed in detail in [Chapter 7](#). (A confidence interval is a range of values based on the sample observations that allows us to predict with a given probability (usually 95%) where the true characteristic value of the population is likely to occur. The interval is determined by using information from the size of the sample—a small sample is less reliable than a large one and gives a bigger interval—and the variability of the sample data as indicated by the standard deviation.) Therefore an estimate of mean height in a single sample (the point estimate) together with the estimate of variability allows prediction of the limits within which the means of other samples of the same size are likely to lie, and therefore within what range the population mean is likely to be. Setting these confidence limits is an essential extension of descriptive analysis. How to estimate the population mean from the sample mean will be discussed in later chapters.

Confidence limits convey the degree of uncertainty that exists. Given the data, there is a 95% probability of being right in setting the limits for that mean value in other samples from the same population, but that goes along with a 5% chance of being wrong. Limits may be widened to 99%, or 99.5%, or 99.9999%, but certainty is never achieved. Could

the randomly chosen sample of adult males have by chance selected the tallest members of the population so that the estimate of mean height, instead of being 175 cm, was 185 cm? Yes, indeed such an aberrant sample could have been chosen, even if it is very unlikely, and if it had been chosen our estimates of the mean population height would be seriously in error. That is the uncertainty that we have to live with.

Statistical inference

When we ask whether the mean heights of treated and untreated children with growth hormone deficiency are different, we invoke the second use of statistics—*statistical inference* that is done with *hypothesis tests*. In general, we generate hypotheses about the possible effects of treatment and then deploy statistical techniques to test the appropriateness of these hypotheses. Often the null hypothesis is selected, that is, we test the possibility that the treatments of the groups (taking growth hormone or a placebo) have not caused a change in an outcome (mean height). If the tests indicate that the differences observed would be unlikely to occur by chance, then we may decide to reject the null hypothesis. One of the advantages of the statistical method is that it can give an approximate probability of correctly rejecting the null hypothesis. It does so by calculating a *P* value that may be defined as the probability of finding an observed difference or even more extreme differences (say in mean heights) if the null hypothesis is true. This subject is dealt with in depth in [Chapter 10](#).

Statistical methods also allow us to investigate *relationships* among variables that may be either causative (*independent*) or responsive (*dependent*). Thus if we give a drug to patients and observe a fall in blood pressure, we hypothesize that the drug (independent cause) produced the fall in pressure (dependent response).

We investigate relationships in two main ways. In the *survey method*, we collect examples from one or more populations and determine if the independent and dependent variables are related. In the survey, there may or may not be a comparison group. Without comparison groups it is a *descriptive* study. With comparison groups it is an *analytical* study ([Grimes and Schulz, 2002](#)). For example, the role of fluoride in preventing dental caries was examined in cities that did or did not supplement their water supplies with fluoride; the amount of dental caries was lower in people living in the cities with higher water fluoride concentrations ([Grimes and Schulz, 2002](#)). The advantages of an analytical study are that large populations can be studied and that individuals do not have to be manipulated and assigned to groups; the disadvantage is that factors other than the fluoride might have caused the observed differences in the incidence of dental caries. Factors extraneous to but related to the factors being studied are termed *confounding* factors; a factor is a confounder if it differs in the groups and affects the outcome. For example, the incidence of dental caries might really be due to deficiency of another factor (*X*) that is inversely correlated with fluoride. If we do not know about *X*, or

do not measure it, then we would rightly conclude that increased fluoride in the water was *associated* with decreased incidence of caries, but would be incorrect in stating that the increased fluoride *caused* the decreased incidence of caries because that decrease was really caused by a deficiency of substance X. X is the confounding factor. Confounding variables are not necessarily unimportant, but they confuse the relationship that is being examined.

Had subjects been allocated at random to a high or a low fluoride group, there would have been fewer possible maldistributed confounding factors, but there would have been fewer people examined (because of the cost), and the investigators would have had to make sure that there were no other sources of fluoride.

There are three main types of analytical studies. One is the *cohort* study, in which people exposed to an agent are compared with people who are not so exposed. For example, a survey might compare a large group of people who took aspirin regularly with another group (matched for age, gender, and anything else that seems to be important) that did not take aspirin (exposed vs nonexposed), and after some years the investigators determine how many in each group had had a myocardial infarction (outcome). This might be a prospective study but could also be retrospective if the outcomes were examined from a database started 10 years ago. Then there are two types of noncohort studies. One is the *cross-sectional* study in which exposure and outcome are determined at the same time in the study population. For example, in a group of hospitalized patients, high-density lipoprotein (HDL) concentrations are measured in those who have had a myocardial infarction and a comparable group who have not. If the first group has low HDL concentrations on average and the second group does not, then there is an association between HDL concentration and myocardial infarction. This type of study is relatively easy and cheap to do, but it does not determine if having a low HDL is a cause of a myocardial infarction. The other type, the *case-control* study, starts with the outcome and then looks back at exposure; for example, taking 100 patients who had had a myocardial infarction and another 100 with no infarction (appropriately matched for age and gender) and determining how many in each group had taken aspirin. The case-control study is often used when investigating rare diseases or outcomes.

The other type of relationship is determined by *experiment* in which the independent variable is deliberately manipulated. For example, two groups of people with normal blood pressures are selected, and one group is exposed to increased stress or a high salt diet. If the blood pressures increase in this group but not in the control group, then stress or the salt content of the diet may be a cause of hypertension, and the mechanisms by which this occurs can then be investigated.

Setting confidence limits and making statistical inferences can be done efficiently only after an effective description of the data. Therefore the initial part of this book is devoted to descriptive methods on which all subsequent developments are based.

DATA

Statistics involves thinking about numbers, but sometimes these are not what they seem. For example, a University announces that its ranking has improved because last year it accepted only 5% of all applicants. This might mean that it is indeed a prestigious University so that many candidates apply, but it might also mean that the University indulges in misleading advertising so that many who have no chance of admission are persuaded to apply. A department might assert that it has greatly improved its teaching because 97% of the class passed with flying colors, but that might conceal the fact that several students who were not doing well were advised to withdraw from the course. Therefore whether the University or the department has improved depends not on the numbers provided (often by people who have a stake in the impression given) but on what the data actually mean. This is true of all data, and it is as well to consider what the numbers mean and where they come from before beginning to analyze them.

Models

All statistical tests are based on a particular mathematical model. A model is a representation of reality, but it should never be mistaken for the reality that underlies it. As Box and Draper wrote “Essentially, all **models** are wrong, but some are useful” (Box and Draper, 1987).

All statistical tests use models, either explicitly or implicitly. For example, consider the heights of a group of people in the United States and a group of Central African pygmies. In each group the heights vary, but on average the pygmies are shorter than the Americans. The model that we use is to regard the pygmies as having an average height of μ_P cm and the Americans as having an average height of μ_A cm. Any one pygmy may differ from the pygmy average by a difference ϵ_i , and any one American may differ from the American average by difference ϵ_j . An assumption is that each of these sets of differences is normally distributed (Chapter 6). To compare the two sets of heights to determine if they really are different from each other, a test such as the t -test is used in which this model is required. If the population distributions fit the model and are indeed normally distributed, then the t -test is an efficient way of comparing the two sets of heights. If, on the other hand, the population heights are far from normally distributed, then using a test that demands a statistical model with normal distributions can produce misleading conclusions. Therefore whenever a statistical test is used, the model being invoked must be considered and the data must be evaluated to find out if they fit the model. If the model chosen is incorrect, then the test based on that model may give incorrect results.

GENERAL APPROACH TO STUDY DESIGN

After studying this book, the reader should be able to design simple studies and tell if someone else's studies have been correctly designed. The following discussion

summarizes what should be done and gives some guidance about selecting the appropriate analyses.

The first requisite for any study is to ask a focused question. It is of little use to ask “What happens if we place a subject in a temperature of 35°C (95°F) for 3 weeks?” Many hundreds of anatomic, physiologic, neurologic, biochemical, and molecular biologic changes occur. You may be able to measure only some of these, and for all you know have not measured the most important changes. Furthermore, if you could measure all the possible changes, the mass of results would be difficult to interpret. That is not to state that the effects of persistent high ambient temperatures are not important and should not be studied, but rather that the study be designed to answer specific questions. For example, asking what mechanisms achieve adequate heat loss under these circumstances, and looking at possible factors such as changes in blood volume, renal function, and heat exchange through the skin are valid and important questions.

This requisite applies to any study. The study might be a laboratory experiment of norepinephrine concentrations in rats fed different diets; a clinical trial of two treatments for a specific form of cancer; a retrospective search of population records for the relation between smoking and bladder cancer; or the relationship between prospective votes for Democrats, Republicans, and Independents related to race and gender.

The next decision is what population to sample. If the question is whether a new anti-hypertensive drug is better than previous agents, decide what the target population will be. All hypertensives or only severe hypertensives? Males and females, or only one gender? All ages or only over 65 years? All races or only Afro-Americans? With or without diabetes mellitus? With or without prior coronary artery disease? And so on, depending on the subject to be studied. These are not statistical questions, but they influence the statistical analysis and interpretation of the data. Therefore inclusion and exclusion criteria must be unambiguously defined, and the investigator and those who read the results must be clear that the results apply at best only to a comparable group.

Define what will be measured and how. Will it be a random blood pressure, or one taken at 8 a.m. every day, or a daily average? Will you use a standard sphygmomanometer or one with a zero-muddling device? Will you measure peripheral blood pressure or measure central blood pressure by using one of the newer applanation devices? Again, these are not statistical questions, but they will affect the final calculations and interpretation. The number of possible variables in this comparatively simple study is large, and this explains in part why different studies with different variables often reach different conclusions.

Consider how to deal with confounders. It is rare to find a perfect one-to-one relationship between two variables in a biomedical study, and there is always the possibility that other factors will affect the results. If we know about them, our study might be made more efficient by allowing for them; for example, including patients with diabetes mellitus when examining the outcome of stenting a stenosed coronary artery. Therefore in

planning to study the outcome of stent implantation, we might want to incorporate potential confounders in the study; for example, diabetes mellitus, hypertension, obesity, elevated LDL concentrations, renal function, racial group, and age distribution. If we can arrange our study so that each group has subgroups each with an identical pattern of confounders, analysis will be easier and more effective. On the other hand, with too many subgroups, either the total numbers will be huge or else each subgroup may have insufficient numbers to allow for secure interpretation. If for practical reasons such balancing of confounders cannot be done, an approach such as Cox regression ([Chapter 35](#)) or Propensity Analysis ([Chapter 38](#)) might allow for the influence of each of these other factors. Either approach would be better than not considering these confounders at all. Finally, there are likely to be confounders that we do not know about. The only way to try to allow for these is to make sure that the various groups are chosen at random, so that it is likely that unknown confounders will be equally represented in all the groups.

The term “simple random sampling” means that each member of the target population has an equal chance of being selected, and that selection of one member has no effect on the selection of any other member. “Randomization” is the process of taking a given sample and dividing it into subgroups by random selection. As stated by [Armitage and Remington \(1970\)](#): “Randomization may be thought of as a way of dealing with all the unrecognized factors that may influence responses, once the recognizable factors, if any, have been allowed for by some sort of systematic balancing. It does not ensure that groups are absolutely alike in all relevant aspects; nothing can do that. It does ensure that they are all unlikely to differ on the average by more than a moderate amount in any given characteristic, and it enables the statistician to assess the extent to which an observed difference in response to different treatments can be explained by the hazards of random allocation.” The hope is that any unrecognized but relevant factors will be equalized among the groups and therefore not confound the results. More details about randomization are given in [Chapter 38](#).

Replication and pseudoreplication

Distributions are characterized by means and variability. If we compare two (or more) distributions, for example, the heights of men and women, it would be impossible to reach a sensible conclusion by comparing the heights of one man and one woman. We can make a sensible comparison only if we measure the heights of several men and several women chosen at random. We can then use the differences between the means and the variability to test the probability that the distributions are drawn from the same population. This is the principal purpose of replication, which may be defined as repetitive random selection of *independent* units from a population.

If the measurements are not independent, then there is *pseudoreplication* in which multiple measurements are made on the same units. For example, a study of the value of

fluoride supplementation of toothpaste in preventing dental caries in children might assign children randomly to two groups, one to receive normal toothpaste and one to receive toothpaste with a fluoride supplement. At the end of the study the control group might have 727 carious teeth and the treated group 207 carious teeth. What is the conclusion? The problem is that the unit is the mouth and not the individual teeth. Whatever the causes of dental caries may be, it is likely that they are similar for all the teeth in any one mouth. It may well happen that the numbers of affected children were less different in the two groups. It would be more correct to treat each child as a unit, and record how many had caries or no caries.

The subject becomes more confused in more complex designs. Consider comparing the effects of two growth factors A and B on the production of a specific protein by cells growing in tissue culture. Prepare 4 Petri dishes in which the cells will be grown. Dishes 1 and 2 receive growth factor A and dishes 3 and 4 get B. Specific cells from aortic endothelium are placed in each dish, and after 24 h the protein is measured in cell samples taken from each dish. Taking one cell from each dish does not provide enough measurement for comparison; instead we take 20 cells from each dish. How should we analyze the data?

If we take 20 samples from one of the dishes, we can estimate their mean value and variability. We have replicated the measurements in one dish. What should we do if we take 20 samples from dish 1 and another 20 from dish 2? If we pool both samples, we are ignoring the possibility that the two dishes may differ in their growth-stimulating properties because for example, one dish might be warmer than the other and temperature might affect protein production. Cells within a dish are not independent of each other. Pooling data in this way may lead to exaggerated degrees of freedom and increased Type I errors ([Hurlbert, 1983](#); [Jenkins, 2002](#)) and has been termed *pseudoreplication* by [Hurlbert \(1983\)](#). In this example, the dish is the experimental unit, and we need to replicate the dishes. Another example of pseudoreplication is making several measurements of each subject and treating them as if they were all independent. This is a problem in many neuroscience studies ([Lazic, 2010](#); [Nakagawa and Hauber, 2011](#)), but occurs frequently in other areas of biology and medicine ([Nikinmaa et al., 2012](#)). Reinhart's excellent little book discusses the subject simply ([Reinhart, 2015](#)).

The way to avoid the problem is either to do a nested design analysis or else a repeated measures analysis, depending on the circumstances (see [Chapter 26](#)).

Defining the units is important also in animal studies. After placing 10 randomly selected rats in cage A and 10 others in cage B, and feeding them different supplements, what is the conclusion if the weight gain is substantially greater in cage A than cage B? The difference might be due to the supplement used but could be due to extraneous factors. If cage A was kept in a warm quiet area whereas cage B was in a cold noisy area, quite possibly all the rats in cage A slept more than the rats in cage B, and therefore burned fewer calories. It is also possible for one or two rats in cage B to eat more than their share

of food so the other rats in that cage are underfed. Here the cage is the unit. To avoid these problems, the rats should be placed in separate cages that are randomly dispersed in the room.

Although the outcomes of an intended experiment are not yet available, make some guesses as to what to expect, because this leads to the calculations of sample size.

Begin analysis of any data set with simple preliminary exploration and description before plunging into hypothesis testing.

Decide in advance what type of statistical analysis will be used. Failure to do this might result in the omission of factors that ought to be considered in the final analysis or compel you to perform an inefficient analysis. Any statistical consultation should be done before beginning the investigation. The eminent statistician Ronald Fisher wrote: “To consult the statistician after an experiment is finished is often merely to ask him to conduct a postmortem examination. He can perhaps say what the experiment died of.”

A BRIEF HISTORY OF STATISTICS

For centuries governments collected demographic data about manpower, births and deaths, taxes, and other details. Early examples of this were an Athenian census in the 16th century BC, a Chinese census in AD2 by the Han dynasty that found 57.67 million people in 12.36 million households, and the tabulation in 1085 by William the Conqueror of details of all the properties in England, as collected in the Domesday Book. However, no analysis of data was done until John Graunt (1620–74) published his “Natural History and Political Observations on the London Bills of Mortality” in 1662, perhaps the first major systematic contribution in the field of what was termed “political arithmetic.” He was the first to publish the fact that more boys than girls are born but that the mortality rate is greater for males, resulting in the population’s being almost evenly divided between males and females. Graunt reported the first time trends for many diseases, he offered the first well-reasoned estimate of London’s population, and he produced early hard evidence about the frequencies of various causes of death.

In 1631 ([Lewin, 2010](#)) the term “Statist” was used to describe a person interested in political arithmetic who “desires to look through a Kingdome,” perhaps the first time this term was used. However, the term “*statistics*” seems to have been used first by a professor of politics in Gottingen in 1749, when he wanted a term to describe numerical information about the *state*: number of kilometers of roads, number of people, number of births and deaths, number of bushels of wheat grown, and so on ([Yule and Kendall, 1937](#)). ([Kaplan and Kaplan, 2006](#) attribute the term to a professor in Breslau.) Today these are termed economic and vital statistics. The items in a group are often referred to as statistics, but there is usually no difficulty deciding whether the term statistics refers to items in a group or to the field of study.

One of the origins of Statistics concerns *probability theory*. Even before the Roman empire, people were interested in the odds that occur in games of chance, as described in delightful books by Weaver (1963) and Kaplan and Kaplan (2006). However, the first specific mathematical theory of probability originated in 1654 with a gambling problem that puzzled Antoine Gombauld, Chevalier de Méré, Sieur de Baussay. In one version of the story, he knew that a player had favorable odds of throwing at least one 6 in four throws of a die but could not calculate the odds of throwing at least one double 6 in 24 throws of a pair of dice. He consulted Blaise Pascal (1623–62), the great French mathematician. Pascal solved the problem, and then went on to investigate other probabilities related to gambling. He exchanged letters with Pierre de Fermat (1601–65), and other mathematicians were drawn into the field that grew rapidly. Although the previous problem was a real problem, it was probably known well before Pascal was involved, and the story may be apocryphal (Ore, 1960).

Egon Pearson (1973) (1895–1980) emphasized that big advances in statistics were nearly always made by a combination of a real problem to be solved and people with the imagination and technical ability to solve the problem. For example, astronomers were making measurements of the heavenly bodies and mathematicians were concerned about the accuracy with which astronomical observations could be made. This was not just for curiosity, but because navigation, commerce, and military actions depended critically on accurate knowledge of time and position, including the errors of making these measurements (Stigler, 1986). Many of the developments were made in response to specific practical questions about the orbits of celestial objects or the best estimates of length and weight. In the first half of the 18th century, mathematicians began to investigate the theory of errors (or variability), with a major contribution from Gauss (1777–1855). This phase culminated when Legendre (1752–1833) introduced the *method of least squares* in 1805.

The introduction of statistical methods into nonphysical sciences came later, and began with data collection and analyses, many performed by those we now term epidemiologists. Adolphe Quetelet (1796–1874), an astronomer, studied populations (births, deaths, and crime rates) and used the method of least squares for assessing errors (variability). In 1852 William Farr (1807–83) studied cholera fatalities during the cholera epidemic of 1848/1849 and demonstrated that the fatality rate was inversely related to the elevation above sea level. Florence Nightingale (1820–1920) analyzed deaths in the Crimean campaign, and in 1858 published what might have been the first modified pie chart (called a coxcomb chart) to demonstrate that most deaths were due to preventable diseases rather than to battle injuries (Joyce, 2008). Because of her revolutionary work, in 1858 she was the first woman to be elected a Fellow of the Statistical Society of London.

Gustav Theodor Fechner (1801–87) was apparently the first to use statistical methods in experimental biology, published in his book on experimental psychology in 1860. Then Francis Galton (1822–1911) began his famous studies on heredity with the

publication in 1869 of his book “Hereditary Genius.” He not only analyzed people in innumerable ways but also did experiments on plants. He appears to have been the first person to take an interest in variability, his predecessors being interested mainly in mean values (Pearson, 1973). Darwin’s theory of evolution predicted increasing variability of population characteristics with time, but in fact over a few generations these characteristics appeared to be stable. Galton’s investigation of what came to be called regression to the mean not only solved this problem but changed the direction of statistical thought (Stigler, 2016). In 1889 Galton published a book, “Natural Inheritance,” in which he summarized his studies, and which influenced mathematicians such as Francis Ysidro Edgeworth (1845–1926) and Karl Pearson (1857–1936) to develop better methods of dealing with the variable data and peculiar distributions so often found in biology.

A notable advance was made when William Sealy Gosset (1876–1937) published his article “The Probable Error of A Mean.” Gosset, who studied mathematics and chemistry at Oxford, was employed by Guinness’ brewery to analyze data about the brewing of beer. Up to that time all statistical tests dealt with large data sets, and Gosset realized that tests were necessary to analyze more practical problems that involved small numbers of measurements. (Boland, 1984; Zilliak, 2008) He worked on this in 1906 while on sabbatical leave, where he was closely associated with Karl Pearson. His publication in 1908 of what became called the *t*-test was a landmark. Because at that time the Guinness Company did not allow its employees to publish the results of any studies done (for fear of leakage of industrial secrets), Gosset published his study under the pseudonym of “Student” to avoid association with his employer.

One final root of modern statistics deserves special mention. In agricultural experiments, efficient experimental design has particular importance. Most crops have one growing season per year, and their growth can be influenced by minor variations in the composition, environment, and drainage of the soil in which they grow. Experiments concerning different cultivars or fertilizers have to be designed carefully so that the greatest amount of information can be extracted from the results. It would be inefficient to test one fertilizer one year, another the next year, and so on. As early as 1770 Arthur Young (1741–1820) published his *Course of Experimental Agriculture* that laid out a very modern approach to experiments (Young, 1770), and in 1849 James Johnson published his book “Experimental Agriculture” to emphasize the importance of experimental design (Owen, 1976). In 1843 John Lawes, an entrepreneur and scientist, founded an agricultural research institute at his estate Rothamsted Manor to investigate the effect of fertilizers on crop yield. Ronald Aylmer Fisher (1890–1962) joined what was then known as the Rothamsted Experimental Station in 1919 to develop the objective methods of evaluating the results, and in doing so made Rothamsted a major center for statistics and genetics. Since that time, many of the world’s major statisticians have been based where there is agricultural research, and many of the best-known textbooks reflect this experience. In parallel with Fisher’s work at Rothamsted, L.H.C. Tippett

(1902–85) in 1925 worked with the British Cotton Institute where his statistical insights produced industrially important improvements in the cotton mills, then a major economic force in Great Britain. Also in 1925, Walter Shewhart (1891–1967) began his studies at the Bell Telephone Laboratories that revolutionized the field of industrial quality control and standardization. When he joined the laboratories the telephone industry was beginning an unprecedented expansion, and large quantities of precision equipment with a high standard of performance and uniform quality were needed. One of his disciples was W. Edwards Deming (1900–93) who was asked by the US Government to help rehabilitate Japan's industry after the Second World War. He was so successful in introducing notions of effective management, statistical analysis, and quality control that he was in part responsible for Japan's renown in high quality innovative products.

A detailed history of the development of medical statistics was provided by Armitage (1985).

REFERENCES

- Altman, D.G., 1992. *Practical Statistics for Medical Research*. Chapman and Hall, London.
- Armitage, P., 1985. Biometry and medical statistics. *Biometrics* 41, 823–833.
- Armitage, P., Remington, R.D., 1970. Experimental design. In: *Statistic in Endocrinology*, The MIT Press, Cambridge, MA, pp. 3–31.
- Boland, P.J., 1984. A biographical glimpse of William Sealy Gosset. *Amer Statist.* 38, 179–183.
- Box, G.E.P., Draper, N.R., 1987. *Empirical Model-Building and Response Surfaces*. Wiley, New York.
- Easterling, R.G., 2015. *Fundamentals of Statistical Experimental Design and Analysis*. John Wiley & Sons, Chichester.
- Grimes, D.A., Schulz, K.F., 2002. An overview of clinical research: the lay of the land. *Lancet* 359, 57–61.
- Hulley, S.B., Cummings, S.R., Browner, W.S., Grady, D.G., Newman, T.B., 2007. *Designing Clinical Research*. Lippincott, Williams & Wilkins, Philadelphia.
- Hurlbert, S.H., 1983. Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* 54, 187–211.
- Jenkins, S.H., 2002. Data pooling and type I errors. *Anim. Behav.* 63, F9–F11.
- Joyce, H., 2008. Florence nightingale: a lady with more than a lamp. *Significance* 5, 181–182.
- Kaplan, M., Kaplan, E., 2006. *Chances Are*. Viking Penguin, New York.
- Komlos, J., Lauderdale, B.E., 2007. The mysterious trend in American heights in the 20th century. *Ann. Hum. Biol.* 34, 206–215.
- Lazic, S.E., 2010. The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neurosci.* 11, 5.
- Lewin, C., 2010. The politic survey of a Kingdom: the first statistical template? *Significance* 7, 36–39.
- Moses, L.E., 1985. Statistical concepts fundamental to investigations. *N. Engl. J. Med.* 312, 890–897.
- Nakagawa, S., Hauber, M.E., 2011. Great challenges with few subjects: statistical strategies for neuroscientists. *Neurosci. Biobehav. Rev.* 35, 462–473.
- Nikinmaa, M., Celander, M., Tjeerdema, R., 2012. Replication in aquatic biology: the result is often pseudoreplication. *Aquat. Toxicol.* 116–117, iii–iv.
- Ore, O., 1960. Pascal and the invention of probability theory. *Amn Math Monthly* 67, 409–419.
- Owen, D.B., 1976. *On the History of Probability and Statistics*. Marcel Dekker, Inc, New York.
- Pearson, E.S., 1973. Some historical reflections on the introduction of statistical methods in industry. *The Statistician* 22, 165–179.
- Reinhart, A., 2015. *Statistics Done Wrong*. No Starch Press, San Francisco.
- Stigler, S.M., 1986. *The History of Statistics*. Harvard University Press, Cambridge, MA.

- Stigler, S.M., 2016. *The Seven Pillars of Statistical Wisdom*. Harvard University Press, Cambridge, MA.
- Weaver, W., 1963. *Lady Luck. The Theory of Probability*. Dover Publications, Inc, New York.
- Young, A., 1770. In: Fenwick, J. (Ed.), *A Course of Experimental Agriculture*. J Dooley, London
- Yule, G.U., Kendall, M.G., 1937. *An Introduction to the Theory of Statistics*. Charles Griffin & Co, London.
- Ziliak, S.T., 2008. Guinnessometrics: the economic foundation of “Student’s” t. *J Econ. Perspect.* 22, 199–216.

CHAPTER 2

Statistical Use and Misuse in Scientific Publications

EARLY USE OF STATISTICS

How well is statistical thought incorporated into existing scientific publications? Do reviewers and editors guarantee that the interpretations of the data are correct?

In the early 20th century most scientific reports were descriptive and numerical analysis was rudimentary. Gradually elementary statistical analysis began to appear in publications. A landmark publication with a major effect on the practice of medical statistics occurred when A. Bradford Hill published in 1937 the first edition of his book “Principles of Medical Statistics,” based on articles that he had written for *The Lancet*; it is now in its 12th edition. (Hill and Hill, 1991) An important book that can still be read with pleasure and profit is “Elementary Medical Statistics,” published by Donald Mainland (1963). A third influential publication was “Clinical Biostatistics” by Alvin Feinstein. Feinstein (1977) that consolidated articles published by him in the journal *Clinical Pharmacology and Therapeutics* between 1970 and 1976.

After 1952 statistical analysis in medical research became more common. Hayden (1983) reviewed the journal *Pediatrics* for 1952, 1962, 1972, and 1982 and observed that the proportion of articles using interpretive statistical techniques increased from 13% in 1952 to 48% in 1982. Furthermore, whereas in 1952 knowledge of the basic *t*-test, chi-square test, and Pearson’s correlation coefficient sufficed to understand 97% of the articles that used statistics, by 1982 only 65% of the articles used these tests, and the others used more sophisticated tests. Hayden pointed out that the increasing use of sophisticated tests often puts the reader at a disadvantage. Hellem et al. (2007) surveyed the same journal for 2005. The proportion of articles using inferential statistics increased from 48% in 1982 to 89% in 2005. The most commonly encountered statistical procedures were “... descriptive statistics, tests of proportions, measures of risk, logistic regression, *t*-tests, non-parametric tests, analysis of variance, multiple linear regression, sample size and power calculation, and tests of correlation.” They remarked that a reader familiar only with these tests would understand the analyses used in only 47% of these articles. Horton and Switzer reached a similar conclusion based on a study of the *New England Journal of Medicine* from 1978 to 2005 (Horton and Switzer, 2005). In another study of surgical publications between 1985 and 2003, the percentage of articles with no statistics declined from 35% to

10%, nonparametric tests increased from 0% to 12% (depending on the journal) to 33%–49%, and the use of more complex tests increased (Kurichi and Sonnad, 2006).

Today, with computers ubiquitous and software statistical packages freely available, complex statistical analyses are within the grasp of the nonstatistician. However, this does not mean that investigators necessarily choose the appropriate analyses or perform them correctly. As Hofacker (1983) stated “The good news is that statistical analysis is becoming easier and cheaper. The bad news is that statistical analysis is becoming easier and cheaper.”

CURRENT-TESTS IN COMMON USE

The statistical tests most often used (80%–90%) between 1978 and 2015 in several major medical journals, in a wide variety of medical fields, are presented in Table 2.1 (Altman, 1991; Emerson and Colditz, 1983; Pilcik, 2003; Reed et al., 2003; Lee et al., 2004; Greenfield et al., 2009; Oliver and Hall, 1989; du Prel et al., 2010; Hellems et al., 2007; Baer et al., 2010).

Table 2.1 List of commonly used statistical procedures

-
- | | |
|-----|---|
| 1. | Descriptive measures of position, dispersion, and shape; confidence interval. |
| 2. | <i>t</i> -test. |
| 3. | Contingency tables (chi-square) and other tests of proportions. |
| 4. | Nonparametric tests. |
| 5. | Transformations, for example, from linear to logarithmic. |
| 6. | Correlation and regression, including multiple and logistic regression. |
| 7. | Analysis of variance. |
| 8. | Multiple comparisons. |
| 9. | Life tables, survival. |
| 10. | Epidemiology statistics, for example, odds ratios, attributable risk. |
| 11. | Sample size and power calculations, but only in a minority of publications. |
-

STATISTICAL MISUSE

Many investigators have examined the use of statistics in both clinical and basic science biomedical publications. Altman reviewed the use and misuse of statistics up to the 1980s (Altman, 1991); more recent studies as summarized by Glantz (2005), Good and Hardin (2009), and Kilkenny et al. (2009) are discussed later. The results have been fairly uniform. Of the published articles in which statistical analyses were used, from 50% to 78% used incorrect tests; this figure has varied little over the past 60 years (Ross, 1951; Badgley, 1961; Schor and Karten, 1966; Schoolman et al., 1968; Gore et al., 1977; Freiman et al., 1978; Glantz, 1980; Sheehan, 1980; Reed and Slaichert, 1981; Emerson and Colditz, 1983; Sheps and Schechter, 1984; Pocock et al., 1987; Williams et al., 1997;

Kilkenny et al., 2009). Some investigators reported that recently statistical usage, especially associated with clinical trials, showed slight improvement (Altman, 1991; Altman and Dore, 1990; Greenfield et al., 2009; Kober et al., 2006; Dar et al., 1994). In some articles the misuse of statistical analyses did not alter the conclusions drawn by the investigators, but in others the incorrect analysis caused the investigators to draw incorrect conclusions, based on their own data and a subsequent correct analysis. As Norman and Streiner (1994) put it “...doing statistics really is easier now than doing plumbing, but unfortunately errors are much better hidden—there is no statistical equivalent of a leaky pipe. Also, there is no building inspector or building code in statistics,.....”

Good statistical practice is still uncommon. For example, Curran-Everett and Benos (2007) surveyed articles published in three high quality biomedical journals: American Journal of Physiology, Journal of Applied Physiology, and the Journal of Neurophysiology for the years 1996, 2003, and 2006. There was slight improvement over time, but even in 2006 only 0%–6% of the articles described confidence intervals, and only 13%–38% gave exact *P* values; both of these omissions indicate poor statistical practice. In Drummond et al. (2011) on behalf of the Physiological and Pharmacological Societies of the United Kingdom, began a series of short articles on statistical procedures to remedy the fact that the quality of data reporting and statistical analysis was still poor. Poor statistical practice has even been found in high impact journals such as Science and Nature (Tressoldi et al., 2013). It seems that little improvement has occurred.

Many types of errors found in the literature, with some representative references, are set out in Table 2.2. The first five are the most important, but any one of them can vitiate a potentially useful study.

Table 2.2 List of some important statistical errors in biomedical publications

1. Failure to state clearly the hypothesis to be tested (Harris et al., 2009; Drummond et al., 2010; Ludbrook, 2008).
2. Failure to check the accuracy of data used for analysis.
3. Failure to describe the statistical tests and software used (innumerable).
4. Failure to understand the prerequisites of statistical tests, leading frequently to serious misinterpretation of the results (Badgley, 1961; Schor and Karten, 1966; Schoolman et al., 1968; Gore et al., 1977; Glantz, 1980; Sheehan, 1980; Hayden, 1983; Sheps and Schechter, 1984; Pocock et al., 1987); failure to use control groups or adequate control groups (Ross, 1951; Badgley, 1961; Schor and Karten, 1966); failure to distinguish between ratio and ordinal numbers, misuse of paired vs unpaired tests; and failure to indicate whether the data are normally distributed or not, with consequent complications of analysis and interpretation (Gore et al., 1977; Kurichi and Sonnad, 2006).
5. Failure to assess effect size or to use a large enough sample size to give adequate power (Freiman et al., 1978; Huang et al., 2002; Kurichi and Sonnad, 2006; George, 1985; Hokanson et al., 1986; Murphy, 1979; Sackett, 1981a; Sheps and Schechter, 1984; Yates, 1983; Williams et al., 1997; Tsang et al., 2009, 2009; Brown et al., 1987; Button et al.,

- 2013; Chung et al., 1998; Nieuwenhuis et al., 2011). This might be the most serious of these errors, because it is likely that most of the work and expense of the study have been wasted.
6. Confusion between standard deviation and standard error (Reed et al., 2003; Oliver and Hall, 1989; Bunce et al., 1980; Gardner, 1975; Weiss and Bunce, 1980; Glantz, 1980) and absence or misuse of confidence limits (Harris et al., 2009, 2009; Hayden, 1983; Huang et al., 2002; Hokanson et al., 1986; Belia et al., 2005; McCormack et al., 2013).
 7. Use of multiple *t* tests without appropriate correction or failure to use techniques such as analysis of variance designed for comparisons of more than two groups (Schor and Karten, 1966; Glantz, 1980; Pocock et al., 1987; Williams et al., 1997; Kusuoka and Hoffman, 2002; Kurichi and Sonnad, 2006).
 8. Incorrect use or definition of sensitivity and specificity (Schor and Karten, 1966; Sheps and Schechter, 1984) and failure to understand when the odds ratio is an unreliable guide to relative risk (Feinstein, 1986; Schwartz et al., 1999; Holcomb et al., 2001; Katz, 2006).
 9. Failure to understand how *P* should be interpreted, and an undue reliance on $P < 0.05$ to assess the null hypothesis (Dar et al., 1994; Oliver and Hall, 1989; Goodman, 1999; Motulsky, 2015; Colquhoun, 2014). Mistaking “statistical significance” for importance. {innumerable}
 10. Pseudoreplication, when nonindependent measurements are treated as if they were independent (Hurlbert, 1983; Jenkins, 2002; Lazic, 2010).
 11. Misuse of graphical displays—innumerable examples, and also see Chapter 6 in Gierliński (2016).
 12. A number of the above errors are common in clinical trials, which may also show failure of or inadequate randomization, failure to describe how patients are included in the trial, failure to use double-blind procedures, failure to define when a trial should be stopped early (Harris et al., 2009; Hayden, 1983; Huang et al., 2002; Hokanson et al., 1986).
-

These errors are discussed in detail by Strasak et al. (2007), Marino (2014), and Thiese et al. (2015).

The failure to use and interpret statistical tests correctly in such a large number of research enterprises, despite the availability of programs that will do the tests, is serious. Chalmers and Glasziou (2009), who included in their list of errors the failure to take cognizance of preexisting studies of a problem, concluded that all these errors might account for 80% wastage of the US\$100 billion spent annually worldwide on biomedical research.

Schoolman et al. (1968) observed in 1968: “Current practices have created an extraordinary and indeterminate risk to the reader if he accepts the authors’ conclusions based on statistical tests of significance.” This statement is true today. The readers of these journals, who for the most part are not statistically sophisticated, frequently cannot tell if the statistical tests have been correctly performed and interpreted (Berwick et al., 1981).

Some editorial boards of medical journals address these issues by having statistical consultants. There is, however, no consistent policy about which submitted manuscripts

receive statistical analysis. Although over the years more and more submitted manuscripts are inspected by statistical reviewers, (George, 1985, 1985; Gardner and Bond, 1990; Goodman et al., 1998) there are differences among journals that correlate roughly with the size of their circulations (Goodman et al., 1998). Journals in the lowest quartile of circulation numbers had about 31% probability of having a statistical consultant on the staff as compared to 82% in journals in the upper quartile. In the lower 3 quartiles only 15% of articles were submitted to statistical review whereas 40% were reviewed in the highest quartile. The reader cannot assume that statistical adequacy of the study has been verified before publication. This is not to discount the value of statistical consultation by the journals. Gardner and Bond (1990) in a small study in the British Medical Journal for 1988 observed that only 5/45 relevant articles were statistically acceptable on submission, but this had increased to 38/45 after consultation and revision. Having statistical consultants on Biomedical Journals, however, does nothing to prevent major errors of planning and analysis by the investigators before the manuscript is submitted for review.

Erroneous conclusions from faulty statistical tests not only produce incorrect information but have major ethical consequences. If more animals or people are used than are needed to establish a statistically valid conclusion, or if the numbers are too few to establish that a real difference between treatments exists, then time, money, and animals are wasted, and some subjects are treated ineffectively when they might have been given a more effective treatment (Gore et al., 1977; Freiman et al., 1978; Altman, 1980, 1994; Yates, 1983; Mann et al., 1991; Williams et al., 1997; Chalmers and Glasziou, 2009). It is incumbent on the investigator to plan the study and analyze its results effectively. As Altman and Simera wrote: "Complete, accurate and transparent reporting should be regarded as an integral part of responsible research conduct. Researchers who fail to document their research study according to accepted standards should be held responsible for wasting money invested in their research project. In addition, researchers have a moral and ethical responsibility to research participants, funders and society at large" (Altman and Simera, 2010).

A study of 635 NIH funded completed clinical trials found that only 46% were published in peer reviewed journals within 30 months of trial completion; about one-third were still not reported after a median of 51 months after completion. (Ross et al., 2012) Another study showed that only 86% of clinical trials had been reported within 5 years of ending Anderson et al. (2015) and Chen et al. (2016) surveying only academic institutions found that only 60% of studies had been reported or published within 5 years of ending. This indicates serious deficiencies of the research system and is not only harmful to the research enterprise but is also unethical and wasteful of public funds.

In addition to problems stemming from incorrect use of statistical tests there is a more pervasive problem of bias in reporting. There is a tendency not to submit or publish negative reports. Turner et al. (2008) examined the results of 74 trials of antidepressants submitted, as required by law, to the FDA. Of 38 studies with positive results, 37 were published. Of 36 with negative or equivocal results, 22 were not published and 11 were

published in a way that suggested better results than were obtained. Other studies support these conclusions (Dwan et al., 2008; Hopewell et al., 2009; Chalmers and Glasziou, 2009). Pica and Bourgeois (2016) analyzed 559 pediatric randomized clinical trials and found that 104 were discontinued early and that 30% of those that were finished were never published. Some clinical trials that were not reported in a timely fashion had repeatedly been submitted to journals but were rejected because of negative results. This is an issue that needs to be addressed because it may lead to unnecessary repetition of the study (Gordon et al., 2013).

There is also evidence of conscious bias in that an unduly high proportion of published studies have found results favoring a given company's product when that company has sponsored the research (Gotzsche, 1989; Montori et al., 2004; Bero et al., 2007; Chalmers and Glasziou, 2009; Ross et al., 2012).

Even the correct use of a statistical test can lead to incorrect conclusions if simple arithmetical or typographical errors are made. One might think that with computer programs the calculations would be correct. That is probably true, but unfortunately people who transcribe the results can misplace decimal points or minus signs. Vickers reported a brief litany of such disasters (Vickers, 2006). Major flaws were uncovered in an important study of the sensitivity of various cell lines to drugs at Duke University (Baggerly and Coombes, 2009). Errors occurred, for example, because of failure to check the numbers used in the various tests, incorrect labeling of samples, and poor documentation. When the authors of this critical report reexamined data from their own institution (M.D. Anderson Cancer Institute) they found similar examples that needed to be corrected. Both these institutions have access to experienced statistical consultants, yet serious errors were made.

Motulsky (2015) has pointed out that investigators have often failed to reproduce accepted preclinical studies or studies of basic cancer biology, and that inadequate understanding of statistical thought and procedures is a frequent cause of this failure. It is essential to assess statistical techniques and interpret results with the greatest of care and be aware of likely problems (Begley and Ellis, 2012; Prinz et al., 2011).

Finally, and perhaps most important of all, the study must be worthwhile and well designed. As Schoolman et al. (1968) emphasized "Good answers come from good questions not from esoteric analysis." Even the best statistical analysis cannot produce useful conclusions from a poorly conceived and poorly executed study.

BASIC GUIDES TO STATISTICS

There are many articles in the literature that can help the reader evaluate important aspects of an experiment or study. One readable article is by Finney (1982). The Department of Clinical Epidemiology and Biostatistics at McMaster University Health Sciences Center at Hamilton, Ontario, Canada has published a useful and detailed series to help

the average reader ask the right questions (Tugwell, 1981, 1984; Sackett, 1981a,b; Trout, 1981; Haynes, 1981). Recently guidelines for statistical use and reporting for animal experiments (Kusuoka and Hoffman, 2002) and randomized trials (Ross et al., 2012; Montori et al., 2004) have been published, and the EQUATOR (Enhancing the Quality And Transparency Of health Research) program published a catalog of reporting guidelines for health research (Simera et al., 2010). A simply written book that details statistical errors is by Reinhart (2015).

Why is it that despite innumerable articles, books, and exhortations about the correct use of statistics there are still so many errors made? One plausible reason is that Statistics is regarded as peripheral to Science, rather than an integral part of it. A physician who would not consider treating a patient for a given ailment without thorough knowledge of the advantages and disadvantages of different treatments usually has no compunction about doing a research project without knowing anything about statistical analysis and interpretation. No one is allowed to practice Medicine without a license to show that a standard of competence has been attained, yet anyone may spend thousands of dollars doing research without having the requisite credentials.

One of the difficulties that we face as nonmathematicians is that we are uncomfortable with the language of mathematics so that even fairly simple reports about statistics appear to be hard to read. We suffer from “statisticophobia.” Nevertheless, we should try to understand what these reports are telling us. We should avoid “.... The tendency for clinicians to respond to the advances of the mathematician with one of two extremes, truculent skepticism or obsequious docility”(Murphy, 1979). It is the responsibility of all of us to familiarize ourselves with the role of statistics in designing and evaluating the results of research.

REFERENCES

- Altman, D.G., 1980. Statistics and ethics in medical research. III. How large a sample? *BMJ* 281, 1336–1338.
- Altman, D.G., 1991. Statistics in medical journals: developments in the 1980s. *Stat. Med.* 10, 1897–1913.
- Altman, D.G., 1994. The scandal of poor medical research. *BMJ* 308, 283–284.
- Altman, D.G., Dore, C.J., 1990. Randomisation and baseline comparisons in clinical trials. *Lancet* 335, 149–153.
- Altman, D.G., Simera, I., 2010. Responsible reporting of health research studies: transparent, complete, accurate and timely. *J. Antimicrob. Chemother.* 65, 1–3.
- Anderson, M.L., Chiswell, K., Peterson, E.D., Tasneem, A., Topping, J., Califf, R.M., 2015. Compliance with results reporting at ClinicalTrials.gov. *N. Engl. J. Med.* 372, 1031–1039.
- Badgley, R.F., 1961. An assessment of research methods reported in 103 scientific articles from two Canadian medical journals. *Can. Med. Assoc. J.* 85, 246–250.
- Baer, H.J., Tworoger, S.S., Hankinson, S.E., Willett, W.C., 2010. Body fatness at young ages and risk of breast cancer throughout life. *Am. J. Epidemiol.* 171, 1183–1194.
- Baggerly, K.A., Coombes, K.R., 2009. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Ann. Appl. Stat.* 3, 1309–1334.
- Begley, C.G., Ellis, L.M., 2012. Drug development: raise standards for preclinical cancer research. *Nature* 483, 531–533.

- Belia, S., Fidler, F., Williams, J., Cumming, G., 2005. Researchers misunderstand confidence intervals and standard error bars. *Psychol. Methods* 10, 389–396.
- Bero, L., Oostvogel, F., Bacchetti, P., Lee, K., 2007. Factors associated with findings of published trials of drug–drug comparisons: why some statins appear more efficacious than others. *PLoS Med.* 4e184.
- Berwick, D.M., Fineberg, H.V., Weinstein, M.C., 1981. When doctors meet numbers. *Am. J. Med.* 71, 991–998.
- Brown, C.G., Kelen, G.D., Ashton, J.J., Werman, H.A., 1987. The beta error and sample size determination in clinical trials in emergency medicine. *Ann. Emerg. Med.* 16, 183–187.
- Bunce III, H., Hokanson, J.A., Weiss, G.B., 1980. Avoiding ambiguity when reporting variability in biomedical data. *Am. J. Med.* 69, 8–9.
- Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14, 365–376.
- Chalmers, I., Glasziou, P., 2009. Avoidable waste in the production and reporting of research evidence. *Obstet. Gynecol.* 114, 1341–1345.
- Chen, R., Desai, N.R., Ross, J.S., Zhang, W., Chau, K.H., Wayda, B., Murugiah, K., Lu, D.Y., Mittal, A., Krumholz, H.M., 2016. Publication and reporting of clinical trial results: cross sectional analysis across academic medical centers. *BMJ* 352, i637.
- Chung, K.C., Kallianen, L.K., Hayward, R.A., 1998. Type II (beta) errors in the hand literature: the importance of power. *J. Hand Surg. [Am]* 23, 20–25.
- Colquhoun, D., 2014. An investigation of the false discovery rate and the misinterpretation of P values. *Roy Soc Open Sci.* 1 <https://doi.org/10.1098/rsos.140216>.
- Curran-Everett, D., Benos, D.J., 2007. Guidelines for reporting statistics in journals published by the American Physiological Society: the sequel. *Adv. Physiol. Educ.* 31, 295–298.
- Dar, R., Serlin, R.C., Omer, H., 1994. Misuse of statistical test in three decades of psychotherapy research. *J. Consult. Clin. Psychol.* 62, 75–82.
- Drummond, G.B., Paterson, D.J., McGrath, J.C., 2010. Arrive: new guidelines for reporting animal research. *J. Physiol.* 588, 2517.
- Drummond, G.B., Paterson, D.J., McLoughlin, P., McGrath, J.C., 2011. Statistics: all together now, one step at a time. *Exp. Physiol.* 96, 481–482.
- Du Prel, J.B., Rohrig, B., Hommel, G., Blettner, M., 2010. Choosing statistical tests: part 12 of a series on evaluation of scientific publications. *Dtsch. Arztebl. Int.* 107, 343–348.
- Dwan, K., Altman, D.G., Arnaiz, J.A., Bloom, J., Chan, A.W., Cronin, E., Decullier, E., Easterbrook, P.J., Von Elm, E., Gamble, C., Ghera, D., Ioannidis, J.P., Simes, J., Williamson, P.R., 2008. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One.* 3, e3081.
- Emerson, J.D., Colditz, G.A., 1983. Use of statistical analysis in *New England Journal of Medicine*. *New Engl J Med* 309, 709–713.
- Feinstein, A.R., 1977. *Clinical Biostatistics*, St Louis. C.V.Mosby, Co.
- Feinstein, A.R., 1986. The bias caused by high values of incidence for p_1 in the odds ratio assumption that $1-p_1$ approximately equal to 1. *J. Chronic Dis.* 39, 485–487.
- Finney, D.J., 1982. The questioning statistician. *Stat. Med.* 1, 5–13.
- Freiman, J.A., Chalmers, T.C., Smith Jr., H., Kuebler, R.R., 1978. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *New Engl J Med* 299, 690–694.
- Gardner, M.J., 1975. Understanding and presenting variation. *Lancet* 1, 230–231.
- Gardner, M.J., Bond, J., 1990. An exploratory study of statistical assessment of papers published in the British Medical Journal. *JAMA* 263, 1355–1357.
- George, S.L., 1985. Statistics in medical journals: a survey of current policies and proposals for editors. *Med. Pediatr. Oncol.* 13, 109–112.
- Gierliński, M., 2016. Understanding Statistical Error. Wiley Blackwell, Chichester, UK.
- Glanz, S.A., 1980. Biostatistics: how to detect, correct, and prevent errors in the medical literature. *Circulation* 61, 1–7.

- Glantz, S.A., 2005. *Primer of Biostatistics*. McGraw-Hill, New York.
- Good, P.I., Hardin, J.W., 2009. *Common Errors in Statistics (and How to Avoid Them)*. John Wiley & Sons, Hoboken, NJ.
- Goodman, S.N., 1999. Toward evidence-based medical statistics. 1: the P value fallacy. *Ann. Intern. Med.* 130, 995–1004.
- Goodman, S.N., Altman, D.G., George, S.L., 1998. Statistical reviewing policies of medical journals: caveat lector? *J. Gen. Intern. Med.* 13, 753–756.
- Gordon, D., Taddei-Peters, W., Mascette, A., Antman, M., Kaufmann, P.G., Lauer, M.S., 2013. Publication of trials funded by the National Heart, Lung, and Blood Institute. *N. Engl. J. Med.* 369, 1926–1934.
- Gore, S.M., Jones, I.G., Rytter, E.C., 1977. Misuse of statistical methods: critical assessment of articles in *BMJ* from January to March 1976. *BMJ* 1, 85–87.
- Gotzsche, P.C., 1989. Methodology and overt and hidden bias in reports of 196 double-blind trials of non-steroidal antiinflammatory drugs in rheumatoid arthritis. *Control. Clin. Trials* 10, 31–56.
- Greenfield, M.L., Mhyre, J.M., Mashour, G.A., Blum, J.M., Yen, E.C., Rosenberg, A.L., 2009. Improvement in the quality of randomized controlled trials among general anesthesiology journals 2000 to 2006: a 6-year follow-up. *Anesth. Analg.* 108, 1916–1921.
- Harris, A.H., Reeder, R., Hyun, J.K., 2009. Common statistical and research design problems in manuscripts submitted to high-impact psychiatry journals: what editors and reviewers want authors to know. *J. Psychiatr. Res.* 43, 1231–1234.
- Hayden, G.F., 1983. Biostatistical trends in *Pediatrics*: implications for the future. *Pediatrics* 72, 84–87.
- Haynes, R.B., 1981. How to read clinical journals. II. To learn about a diagnostic test. *Can. Med. Assoc. J.* 124, 703–710.
- Hellems, M.A., Gurka, M.J., Hayden, G.F., 2007. Statistical literacy for readers of pediatrics: a moving target. *Pediatrics* 119, 1083–1088.
- Hill, A.B., Hill, I.D., 1991. *Bradford Hill's Principles of Medical Statistics*. Hodder Education Publishers, Abingdon, UK.
- Hofacker, C.F., 1983. Abuse of statistical packages: the case of the general linear model. *Am. J. Physiol.* 245, R299–R302.
- Hokanson, J.A., Luttman, D.J., Weiss, G.B., 1986. Frequency and diversity of use of statistical techniques in oncology journals. *Cancer Treat. Rep.* 70, 589–594.
- Holcomb Jr., W.L., Chaiworapongsa, T., Luke, D.A., Burgdorf, K.D., 2001. An odd measure of risk: use and misuse of the odds ratio. *Obstet. Gynecol.* 98, 685–688.
- Hopewell, S., Loudon, K., Clarke, M.J., Oxman, A.D., Dickersin, K., 2009. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst. Rev.* MR000006.
- Horton, N.J., Switzer, S.S., 2005. Statistical methods in the journal. *N. Engl. J. Med.* 353, 1977–1979.
- Huang, W., Laberge, J.M., Lu, Y., Glidden, D.V., 2002. Research publications in vascular and interventional radiology: research topics, study designs, and statistical methods. *J. Vasc. Interv. Radiol.* 13, 247–255.
- Hurlbert, S.H., 1983. Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* 54, 187–211.
- Jenkins, S.H., 2002. Data pooling and type I errors. *Anim. Behav.* 63, F9–F11.
- Katz, K.A., 2006. The (relative) risks of using odds ratios. *Arch. Dermatol.* 142, 761–764.
- Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M.F., Cuthill, I.C., Fry, D., Hutton, J., Altman, D.G., 2009. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One.* 4, e7824.
- Kober, T., Trelle, S., Engert, A., 2006. Reporting of randomized controlled trials in Hodgkin lymphoma in biomedical journals. *J. Natl. Cancer Inst.* 98, 620–625.
- Kurichi, J.E., Sonnad, S.S., 2006. Statistical methods in the surgical literature. *J. Am. Coll. Surg.* 202, 476–484.
- Kusuoka, H., Hoffman, J.I., 2002. Advice on statistical analysis for Circulation Research. *Circ. Res.* 91, 662–671.
- Lazic, S.E., 2010. The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neurosci.* 11, 5.

- Lee, C.M., Soin, H.K., Einarson, T.R., 2004. Statistics in the pharmacy literature. *Ann. Pharmacother.* 38, 1412–1418.
- Ludbrook, J., 2008. The presentation of statistics in clinical and experimental pharmacology and physiology. *Clin. Exp. Pharmacol. Physiol.* 35, 1271–1274 (author reply 1274).
- Mainland, D., 1963. *Elementary Medical Statistics*. W.B. Saunders Company, Philadelphia.
- Mann, M.D., Crouse, D.A., Prentice, E.D., 1991. Appropriate animal numbers in biomedical research in light of animal welfare considerations. *Lab. Anim. Sci.* 41, 6–14.
- Marino, M.J., 2014. The use and misuse of statistical methodologies in pharmacology research. *Biochem. Pharmacol.* 87, 78–92.
- McCormack, J., Vandermeer, B., Allan, G.M., 2013. How confidence intervals become confusion intervals. *BMC Med. Res. Methodol.* 13, 134.
- Montori, V.M., Jaeschke, R., Schunemann, H.J., Bhandari, M., Brozek, J.L., Devereaux, P.J., Guyatt, G.H., 2004. Users' guide to detecting misleading claims in clinical research reports. *BMJ* 329, 1093–1096.
- Motulsky, H.J., 2015. Common misconceptions about data analysis and statistics. *Br. J. Pharmacol.* 172, 2126–2132.
- Murphy, E.A., 1979. *Probability in Medicine*. Johns Hopkins University Press, Baltimore.
- Nieuwenhuis, S., Forstmann, B.U., Wagenmakers, E.J., 2011. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nat. Neurosci.* 14, 1105–1107.
- Norman, G.R., Streiner, D.L., 1994. *Biostatistics. The Bare Essentials*. Mosby, St. Louis.
- Oliver, D., Hall, J.C., 1989. Usage of statistics in the surgical literature and the 'orphan P' phenomenon. *Aust. N. Z. J. Surg.* 59, 449–451.
- Pica, N., Bourgeois, F., 2016. Discontinuation and nonpublication of randomized clinical trials conducted in children. *Pediatrics*. 138.
- Pilcik, T., 2003. Statistics in three biomedical journals. *Physiol. Res.* 52, 39–43.
- Pocock, S.J., Hughes, M.D., Lee, R.J., 1987. Statistical problems in the reporting of clinical trials. *New Engl J Med* 317, 426–432.
- Prinz, F., Schlange, T., Asadullah, K., 2011. Believe it or not: how much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* 10, 712.
- Reed III, J.F., Slaichert, W., 1981. Statistical proof in inconclusive 'negative' trials. *Arch. Intern. Med.* 141, 1307–1310.
- Reed III, J.F., Salen, P., Bagher, P., 2003. Methodological and statistical techniques: what do residents really need to know about statistics? *J. Med. Syst.* 27, 233–238.
- Reinhart, A., 2015. *Statistics Done Wrong*. No Starch Press, San Francisco.
- Ross Jr., O.B., 1951. Use of controls in medical research. *JAMA* 145, 72–75.
- Ross, J.S., Tse, T., Zarin, D.A., Xu, H., Zhou, L., Krumholz, H.M., 2012. Publication of NIH funded trials registered in ClinicalTrials.gov: cross sectional analysis. *BMJ* 344, d7292.
- Sackett, D.L., 1981a. How to read clinical journals. I. Why to read them and how to start reading them critically. *Can. Med. Assoc. J.* 124, 555–558.
- Sackett, D.L., 1981b. How to read clinical journals. V. To distinguish useful from useless or even harmful therapy. *Can. Med. Assoc. J.* 124, 1156–1162.
- Schoolman, H.M., Becktel, J.M., Best, W.R., Johnson, A.F., 1968. Statistics in medical research: Principles versus practices. *J. Lab. Clin. Med.* 71, 357–367.
- Schor, S., Karten, I., 1966. Statistical evaluation of medical journal manuscripts. *JAMA* 195, 145–150.
- Schwartz, L.M., Woloshin, S., Welch, H.G., 1999. Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization. *N. Engl. J. Med.* 341, 279–283 (discussion 286–7).
- Sheehan, T.J., 1980. The medical literature. Let the reader beware. *Arch. Intern. Med.* 140, 472–474.
- Sheps, S.B., Schechter, M.T., 1984. The assessment of diagnostic tests: a survey of current medical research. *JAMA* 252, 2418–2422.
- Simera, I., Moher, D., Hoey, J., Schulz, K.F., Altman, D.G., 2010. A catalogue of reporting guidelines for health research. *Eur. J. Clin. Invest.* 40, 35–53.
- Strasak, A.M., Zaman, Q., Pfeiffer, K.P., Gobel, G., Ulmer, H., 2007. Statistical errors in medical research—a review of common pitfalls. *Swiss Med. Wkly.* 137, 44–49.
- Thiese, M.S., Arnold, Z.C., Walker, S.D., 2015. The misuse and abuse of statistics in biomedical research. *Biochem Med (Zagreb)* 25, 5–11.

- Tressoldi, P.E., Giofre, D., Sella, F., Cumming, G., 2013. High impact = high statistical standards? Not necessarily so. *PLoS One*. 8. e56180.
- Trout, K.S., 1981. How to read clinical journals. IV. To determine etiology or causation. *Can. Med. Assoc. J.* 124, 985–990.
- Tsang, R., Colley, L., Lynd, L.D., 2009. Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized controlled trials. *J. Clin. Epidemiol.* 62, 609–616.
- Tugwell, P.X., 1981. How to read clinical journals. III To learn the clinical course and prognosis of disease. *Can. Med. Assoc. J.* 124, 869–872.
- Tugwell, P.X., 1984. How to read clinical journals. *Can. Med. Assoc. J.* 130, 377–381.
- Turner, E.H., Matthews, A.M., Linardatos, E., Tell, R.A., Rosenthal, R., 2008. Selective publication of antidepressant trials and its influence on apparent efficacy. *N. Engl. J. Med.* 358, 252–260.
- Vickers, A.J., 2006. Look at your garbage bin: it may be the only thing you need to know about statistics. Available: <http://www.medscape.com/viewarticle/546515>.
- Weiss, G.B., Bunce III, H., 1980. Statistics and biomedical literature. *Circulation* 62, 915 (letter).
- Williams, J.L., Hathaway, C.A., Kloster, K.L., Layne, B.H., 1997. Low power, type II errors, and other statistical problems in recent cardiovascular research. *Am. J. Physiol.* 273 (Heart Circ Physiol 42), H487–H493.
- Yates, F.E., 1983. Contributions of statistics to the ethics of science. *Am J PhysiolRegul Integrati Comp Physiol* 244, R3–R5.

CHAPTER 11

Hypothesis Testing: Sample Size, Effect Size, Power, and Type II Errors

BASIC CONCEPTS

Statistical Power

If the null hypothesis is rejected when it is true, we have committed a Type I error, with a probability symbolized by α . On the other hand, accepting the null hypothesis of no difference between two means if they really are from different populations produces a Type II error with a probability symbolized by β .

Consider two populations of means, each population having different grand means, but the normal curves characterizing the distributions of those means overlap (Fig. 11.1).

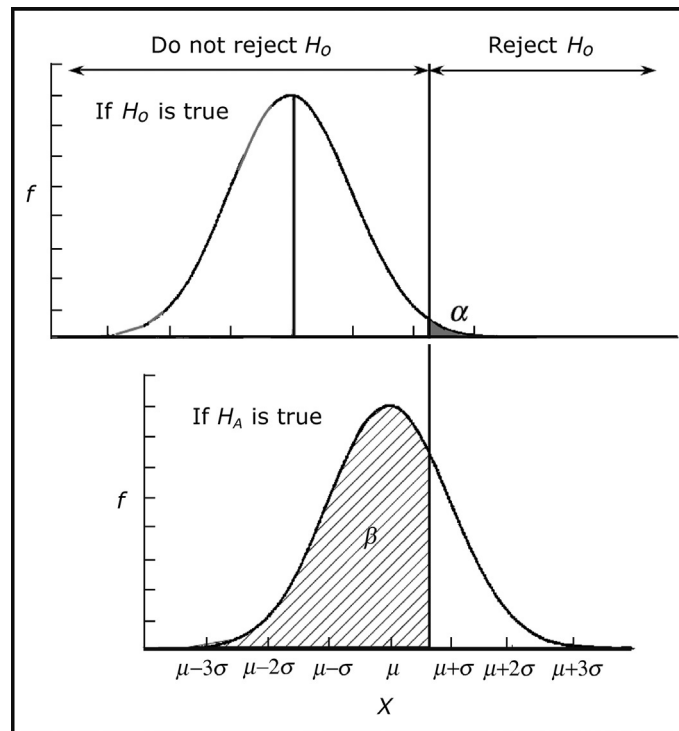


Fig. 11.1 Illustration of Type II error. The Type I error α is the *black* area in the upper panel and is constant and determined by the investigator. The Type II error β is cross-hatched in the lower panel and varies with the value of H_A .

If H_0 is true (i.e., that the two means come from the same population) then a single sample mean falling to the right of the heavy vertical line is unlikely and can lead to rejection of the null hypothesis; the probability of falsely rejecting the null hypothesis is the Type I error, symbolized by α . (Note, however, the criticisms in [Chapter 10](#) about how much the Type I error is.) If, however, H_A is true, as shown in the lower part of the diagram, then $>50\%$ of the time the single sample mean will fall to the left of the heavy vertical line and the null hypothesis would not be rejected at level α . This is the Type II error. The chance of making a Type II error (the probability of accepting the null hypothesis if it is false) is symbolized by β . If the chances of making a Type II error are 0.67 then the chances of not making a Type II error are $1 - 0.67 = 0.33$; there is a 33% probability of making the correct assumption that there are two different groups. This value, $1 - \beta$, is known as the power of the test, that is, its ability to reject the null hypothesis correctly. The power is lowest when the means are close together and highest when they are farthest apart ([Fig. 11.2](#)). Calculating the power of a test in advance is essential, because as pointed out by [Ellis \(2016\)](#) an underpowered study is one designed to fail.

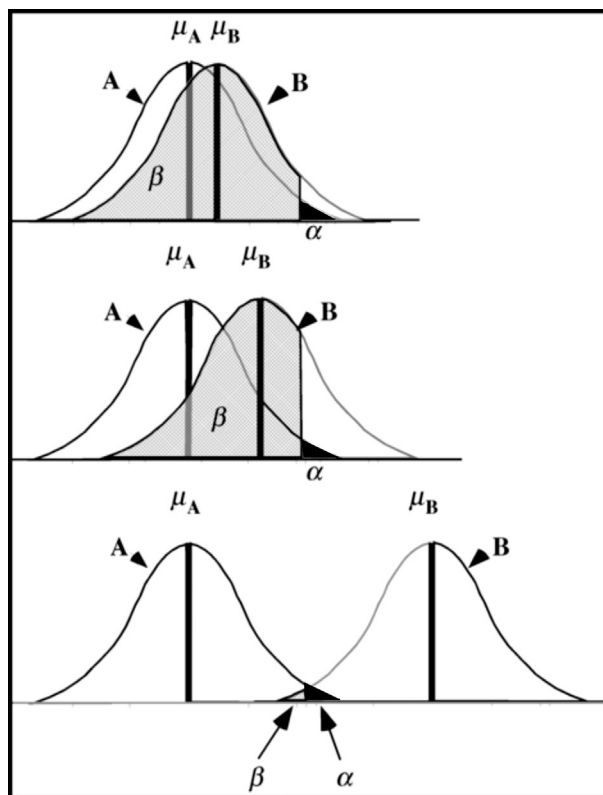


Fig. 11.2 Change in the Type II error as H_A diverges from H_0 .

In this diagram, A is a reference curve showing the distribution of means of samples of size N from a population. B is similar curve derived from means of samples also of size N from a different population. The black triangle indicates the upper tail that includes 0.025 of the A distribution. A decision to accept or reject the null hypothesis that the mean of B differs from the mean of A depends on whether a *single sample mean* from B falls to the right (reject null hypothesis and state that there are probably two different distributions) or left (do not reject null hypothesis that the sample comes from the A distribution) of the shaded area.

In the upper panel, the sample means are close together, and there is about a 90% chance of accepting the null hypothesis as shown by the cross-hatched marking of curve B. Because the null hypothesis is actually wrong, we have committed a Type II error with a probability of about 0.9. In the middle panel, with the two distribution means farther apart, about two-thirds of the B samples fall to the left of the decision line, so that the chances of making a Type II error are about 0.67. In the bottom panel, with the distribution means far apart, there is only a 0.025 chance that the sample from B will be considered as coming from the A distribution, that is, the Type II error is about 0.025. The calculation of power is given in the [Appendix](#).

Effect Size

The effect size is the quantity being measured, whether it is a mean value, a difference between means, a regression coefficient, and so on. Although effect size is not emphasized in text books, it should be regarded as more important than P values ([Cohen, 1990](#); [Coe, 2002](#)).

As [Fig. 11.2](#) shows, the power of the test is closely related to the difference between the means, either between two sample means, or one sample mean and a population mean. The effect size may be classified as:

- a. Absolute effect size, sometimes symbolized as Δ . If in treating a group of patients with hypertension the pressure falls by an average of 10 mmHg, then $\Delta = 10$ mmHg. What we do with the information depends on how important that effect size is. It is unlikely that a pharmaceutical company would spend millions of dollars to produce a new anti-hypertensive drug with an effect size of 3 mmHg, but they might do so for an effect size of 15 mmHg.
- b. Relative effect size, usually symbolized as δ , is the actual effect size relative to the standard deviation. Therefore $\delta = \frac{\Delta}{\sigma}$, where σ is a general estimate of variability that we approximate by the sample values. (Not all texts use these symbols as defined before, and sometimes δ represents the absolute difference.) Relative effect size is used in determining how many measurement or subjects will be needed for an adequate study (see later).

There are several slightly different formulas for calculating relative effect size that take into account the variability of the two (or more) groups that are to be compared. Cohen's d and Hedges' g are the two most often used ([Durlak, 2009](#)) (see [Appendix](#)).

Effect size can be calculated online at <http://www.polyu.edu.hk/mm/effect-sizefaq/calculator/calculator.html>, <https://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-SMD11.php>, https://www.psychometrica.de/effect_size.html, and <http://georgebeckham.com/2016/cohens-d-and-hedges-g-excel-calculator/>.

Despite its importance, effect size is seldom reported in publications (Ellis, 2016).

It is not enough to calculate effect size, but the investigators should put it in context and explain its importance. A small effect size that affects a very large number of people may make a big difference; for example, aspirin may have a small effect size in reducing the number of fatal heart attacks, but when multiplied by the number of people at risk it has a big effect on public health.

How can we reduce the Type II error? For a given biological system the difference between the means of the two populations is determined by the system and not subject to manipulation except perhaps by selecting subgroups and increasing homogeneity. What we can do is to increase the sample size, with the effect shown in Fig. 11.3.

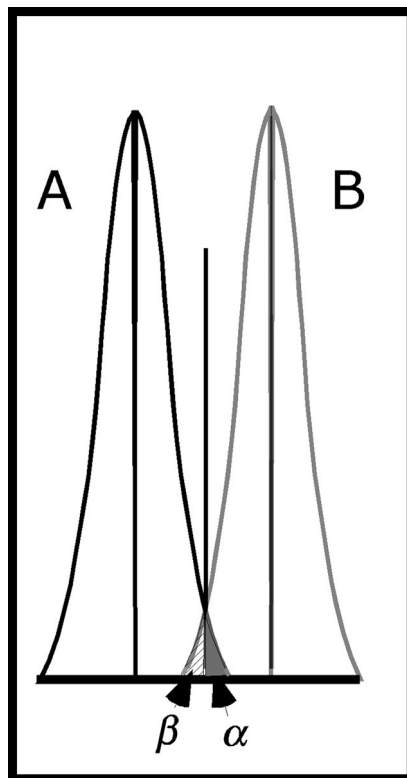


Fig. 11.3 Effect of markedly increasing sample size. For a mean difference similar to that in the middle panel of Fig. 11.7, the Type II error has been reduced from 0.67 to 0.025.

There remains the question of how much to increase the sample size so that the power of the test is high. In theory we would like a power of 0.9, but in practice often settle for 0.8. Whether we can achieve this increased sample size depends on the availability of samples, and the cost and manpower needed to obtain them. In principle, solve the equation

$$t_{0.05} = \frac{\bar{X}_1 - \bar{X}_2}{\frac{S_{\bar{X}_1 - \bar{X}_2}}{\sqrt{N}}}$$

(or whatever other value we want for α) for N by assigning the critical value of t , the difference between the means (the desired effect size), and the standard deviation. From our knowledge of previous studies in a particular field or a pilot study we guess the standard deviation of the population. This can be very wrong, with misleading sample size calculations as a result. It is therefore best to take any sample size calculations as only approximations, and wise to plan for a larger sample size (Schulz and Grimes, 2005).

With an estimated standard deviation decide what difference between means would be important, and calculate the relative effect size δ :

$$\delta = \frac{\bar{X}_1 - \bar{X}_2}{s}.$$

Use the sample standard deviation, not the standard deviation of the mean.

Then use Tables in which the number of subjects is listed for given values of δ , α , and $1 - \beta$ (Beyer, 1966; Cohen, 1988; Kraemer, 1988). Some publications give nomograms to determine these numbers (Gore and Altman, 1982). Alternatively, there are computer programs to make the calculation. An extensive interactive freeware program is termed G*Power at http://download.cnet.com/G-Power/3000-2053_4-141879.html and is very useful for calculating power for all types of statistical tests. Power calculations in the protocol are required by most granting agencies to show that the proposed study is feasible in terms of subjects, time, and money. Another extensive program (Macintosh and Windows) is available at <http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize>. Other simpler programs online are <http://www.statisticalsolutions.net/pssZtest_calc.php>, http://hedwig.mgh.harvard.edu/sample_size/quant_measur/assoc_quant.html, and <http://www.sample-size.net>.

A readable discussion of the steps needed for calculating sample size is provided by Flight and Julious (2016) who have created an application called SampSize for the iPhone and iPad (free from the Apple Store).

Lehr (1992) pointed out that for $\alpha = 0.05$, and power $(1 - \beta) = 0.80$, sample size can be closely approximated by a simple relationship

$$n = \frac{ks^2}{\delta^2},$$

where $k = 8$ for a paired t -test and 16 for an unpaired t -test (see Chapter 22) and δ is the difference to be detected (effect size). This produces numbers very close to the exact number from the more complex calculation; because we have to guess at the value of s , a simple formula seems preferable. For other values of α and β there is a simple table (Table 11.1).

Table 11.1 Values of k for the Lehr equation

	α (Two sample)			α (One sample)
Power $1 - \beta$	0.01	0.05	0.10	0.05
0.80	23.5	16	12.5	8
0.90	30	21	17.5	11
0.95	36	26	22	13
0.975		31		16

Problem 11.1 Determine the sample sizes needed for determining a mean change in myocardial blood flow from 1 to 1.3 mL/g min (paired samples) if the standard deviation is 0.4 with $\alpha = 0.05$ and power of 0.8, 0.85, or 0.9.

Problem 11.2 Repeat the calculations if standard deviation is 0.64 mL/g min.

The Type I error of falsely rejecting the null hypothesis, as previously stated, has nothing to do with the importance of any difference, but is more an issue of consistency of data and the comfort that we feel in deciding to reject the null hypothesis. The degree of certainty is under our control, and we can make the requirement as stringent as we please. On the other hand, the Type II error is more insidious. If we do not have enough power and decide to accept the null hypothesis we may neglect a difference that might be important. If an intervention doubles flow to an ischemic region of the myocardium but because of lack of power of the test we cannot reasonably reject the null hypothesis of no effect, we might be induced to ignore a very useful intervention. That is why if we cannot reject the null hypothesis it is better to regard the effect of the intervention as unproven rather than nonexistent. In fact, Williams et al. (1997) found that failure to achieve a high enough power was the most common statistical error made in publications in the American Journal of Physiology, often because investigators were unaware of the need to assess the power of their negative results. This subject is so important that readers should also go to excellent explanatory writings by Berkowitz that can be downloaded from www.columbia.edu/~mvp19/RMC/M6/M6.doc

A study drawing attention to this problem in Medicine was by [Freiman et al. \(1978\)](#) who examined 71 randomized trials that compared the effects of two drugs or treatments and in which the authors concluded that there was no statistically significant difference between them ([Fig. 11.4](#)).

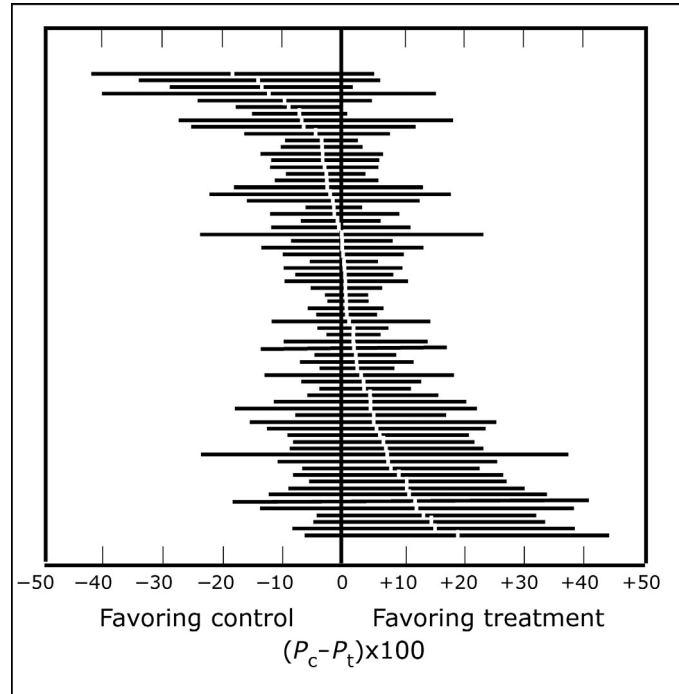


Fig. 11.4 Trials comparing two therapeutic agents. Each *horizontal line* represents one clinical trial. The *white central dot* is the mean and the *black bars* demarcate 90% confidence limits. The *thick vertical black line* indicates no difference. Because the confidence limits include zero difference, P was >0.05 in all these tests. Almost all the mean differences (effect size) are $<15\%$. (Modified from [Freiman et al.](#) by *arranging mean differences in order.*)

They showed that in many of those studies the responses were quite large, but the sample sizes were too small to show a 25% difference between the two treatments, let alone a 50% difference between them. In some instances, this led the investigators to discontinue studying the new treatment and to conclude that it was of no benefit. This is undesirable; a 25% improved cure rate in any disease would be very welcome. This error has been present in many studies with negative results: randomized clinical trials ([Moher et al., 1994](#); [Burbach et al., 1999](#); [Bedard et al., 2007](#); [Tsang et al., 2009](#)), emergency medicine ([Brown et al., 1987](#)), neuroscience ([Button et al., 2013](#); [Nieuwenhuis et al., 2011](#); [Weaver et al., 2004](#)), surgery ([Chung et al., 1998](#); [Freedman et al., 2001](#)). All these studies showed that sample sizes and power were too small to detect a 25%–50% change in outcomes. The cost in time and money of these inadequate studies must have been enormous.

It is possible to make the Type I error as small as you like, for example, reducing it from 0.05 to 0.01 (or even smaller) so that the risk of falsely rejecting the null hypothesis becomes very small indeed. However, the cost of making the Type I error smaller is making the Type II error bigger (Fig. 11.5).

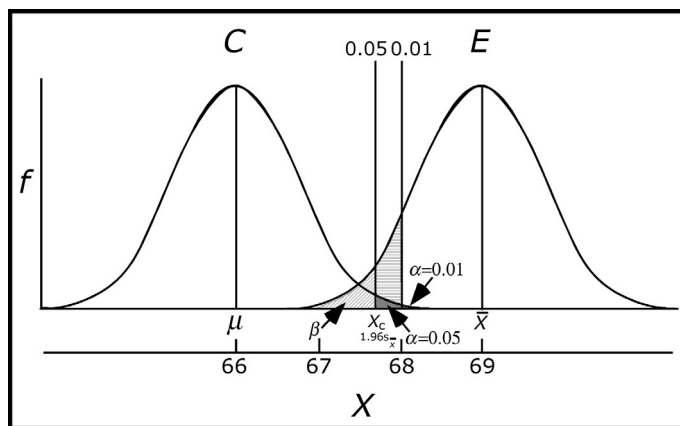


Fig. 11.5 Relation between Type I and II errors.

If for curve *C* the Type I error is set at 0.05, with 0.025 in each tail (solid shaded area), then the Type II error is shown by the cross-hatched area for curve *E*. If the Type I error is made 0.01, with 0.005 in each tail of curve *C*, then the Type II error has increased to include the horizontally shaded area under curve *E*.

Posthoc Power Analysis

Power analysis may be used in several ways. It is best used in planning experiments and is now required by most grant agencies. The investigator decides on the effect size desired, a value for the Type I error α (0.05–0.001), and a value for the power that is $1 - \beta$, the Type II error, and is usually set at 0.8–0.9. Then the number to be used is determined from tables or programs. These numbers are provisional, depending on preliminary observations of means and standard deviations. Julious and Owen (2006) observed that if a variance obtained in a previous study was based on small sample sizes and used as if it were the population variance, its use in the standard formulas could underestimate the future sample size needed to achieve a given power. They provide tables of corrections, but in general it is safe to increase the predicted sample size by about 20%–30%. If some patients or animals are expected to leave the study prematurely, even bigger numbers will be needed. Sometimes, however, there are no previous studies to provide data, and a pilot study might need to be done. A sample size of 12 in each group may be adequate (Julious, 2005, van Belle, 2002).

How should we think about results in which the null hypothesis could not be rejected, with $\alpha > 0.05$, but with a sample size too small to provide adequate power? Many statistical programs allow posthoc calculations of power, but this may not be the best way to assess the data (Kraemer, 1988; Williams et al., 1997; Sterne and Davey Smith, 2001). In fact, by definition a P value > 0.05 indicates that the power was too low to reject the null hypothesis for that effect size. Investigators have argued that a “nonsignificant” result indicates either too small a standardized difference or too small a sample size, and results might or might not be important. In place of the posthoc power analysis, Walters (2009) suggested using confidence intervals to help distinguish statistical significance from clinical importance (Fig. 11.6).

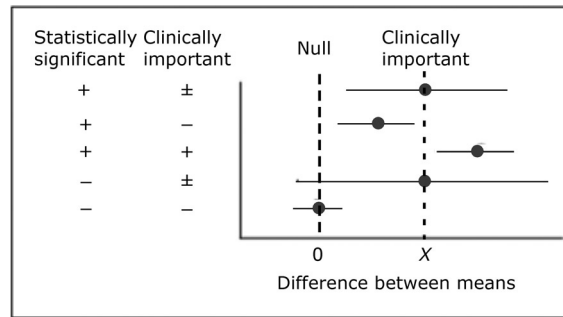


Fig. 11.6 Distinction between statistical significance and importance, showing a useful way of evaluating a “nonsignificant” result by plotting means and confidence limits. + yes; – no; ± possible. Any one of these rows presents information more useful than merely stating that the power was low. (Based on figure published by Walters, S.J., 2009. Consultants’ forum: should post hoc sample size calculations be done? *Pharm. Stat.* 8, 163–9.)

APPENDIX

1. Take care when evaluating relative effect size, because there are different ways of calculating it.

Cohen’s d is $\frac{\overline{X}_1 - \overline{X}_2}{s_p}$, where s_p is the pooled standard deviation

$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2}}$. Hedge’s g is similar. It is $g = \frac{\overline{X}_1 - \overline{X}_2}{s_p}$, where

$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$. To correct for bias the effect size is calculated as

$d_{\text{corr}} = g \left(1 - \frac{3}{4(N_1 + N_2) - 9} \right)$, also written as $d_{\text{corr}} = d \left(1 - \frac{3}{4df - 1} \right)$. If the two sample standard deviations are very different, some people prefer to use the control standard

deviation as the denominator. In a paired design, d for the difference may be related to either the control (or pretest) standard deviation or else to the average of the two standard deviations (Cumming et al., 2012). Always describe which form of d and which denominator you are using.

2. If the null hypothesis (μ_0) is correct, the t distribution is symmetrical, but when an alternative hypothesis is selected ($\mu_A \neq \mu_0$) the t distribution is asymmetrical, the asymmetry increasing as $\mu_A - \mu_0$ (the noncentrality parameter) increases. This is termed the noncentral t distribution. Therefore the confidence limits are asymmetrical, although the differences are not marked except for very small sample sizes. The area under the curve for any value of t can be determined online from <http://www.danielsoper.com/statcalc3/calc.aspx?id=91>, and the confidence limits can be obtained from <http://keisan.casio.com/exec/system/1180573219%3e>.

The reasons for the asymmetrical noncentral t are given by Cumming and Finch (2001).

Calculation of Power

We can calculate the power for any values of μ and \bar{X} if we know N and an estimate of σ derived from the sample standard deviation. As a rule, this is done by tables, formulas, or programs, but the following example shows how power is calculated. Consider the heights of adult European males in the 18th century (Komlos and Cinnirella, 2005) with a mean of 66" and a standard deviation of 4.47" (curve C in Fig. 11.7). We draw at random a group of 30 subjects today and wish to determine if their mean height is consistent with the 18th century population sample. If we drew several samples we might get the distribution of means in curve E in Fig. 11.7.

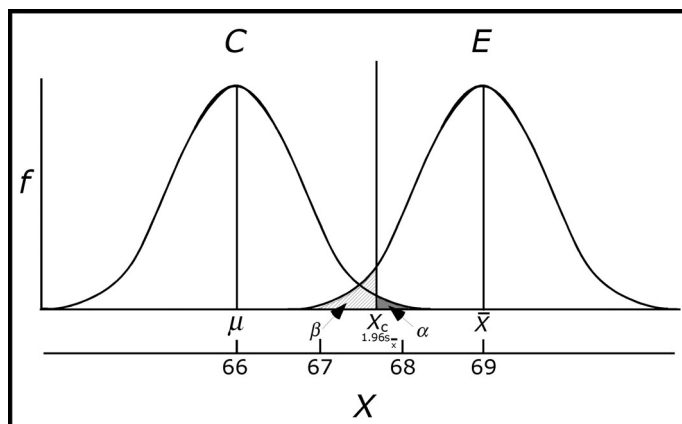


Fig. 11.7 Diagram for power calculation. C is the distribution of the control population means around a "population" mean $\mu = 66$, E is the distribution of our sample means around a single observed sample mean $\bar{X} = 69$. The shaded area to the right of X_c for the C distribution (H_0) is the Type I error, the cross-hatched area to the left of X_c is the Type II or β error, and the total area to the right of X_c for the E distribution (H_A) is the power $1 - \beta$.

The shaded area to the right of the vertical line X_c at $1.96s_{\bar{X}}$ represents $\alpha/2$, or 0.025 of the area under curve C , and represents α , the Type I error. Reject the null hypothesis with $\alpha = 0.05$ if

$$z - \frac{\bar{X} - \mu}{s_{\bar{X}}} \geq 1.96.$$

This can be rearranged to give

$$\bar{X} - \mu > z_{\alpha/2}s_{\bar{X}}.$$

Rearrangement and substitution gives

$$\bar{X} \geq 1.96 \times \frac{4.47}{\sqrt{30}} + 66 \text{ or } \bar{X} \geq 67.60.$$

Therefore any mean value for height in the series of 30 subjects in excess of a critical value of 67.60" leads to rejection of the null hypothesis, and we conclude that mean heights of adult males today are probably $>66''$.

Now consider what happens if HA is true. The cross-hatched area to the left of the critical line X_c is the Type II error, (β) and this is given by 1–area under curve E to the right of X_c . We can calculate this error if we know how many standard error units \bar{X} is from μ .

Consider the true mean difference $\bar{X} - \mu$:

$$\delta = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}} = \frac{\bar{X} - 66}{\frac{4.47}{\sqrt{30}}} = \frac{\bar{X} - 66}{0.8161}.$$

Therefore if \bar{X} is 69, the previous equation becomes

$$\delta = \frac{69 - 66}{0.8161} = 3.676.$$

In other words, if HA is true, a mean value of 69" is 3.676 standard errors from the "population" mean of 66". But X_c is 1.96 standard errors from 66", so the area between X_c and 69" is represented by $z = 3.676 - 1.96 = 1.716$. This is z_b , which represents how many standard error units X_c is below \bar{X} . This area referred to $z = 1.716$ is 0.0431. If HA , the sample mean of 69", is true, there is 0.0431 chance of rejecting the alternative hypothesis. The power is $1 - \beta = 0.9568$.

More generally,

$$z_{\beta} = \delta - z_{\alpha}.$$

Simplify the calculation by using the relationship

$$z_{\beta} = \delta - z_{\alpha} = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}} - \frac{X_c - \mu}{\frac{s}{\sqrt{N}}} = \frac{\bar{X} - X_c}{\frac{s}{\sqrt{N}}}$$

Thus $z_{\beta} = \frac{69-67.6}{0.8161} = 1.7155$, as shown before (with slight difference due to rounding off).

This discussion implies using a 2-tailed test for z , so that $z = 1.96$ includes 0.025 of the area under curve C at each end. To use a 1-tailed test with 0.05 of the area under curve C at one end, use $z = 1.645$ to calculate the β error. (If the sample size is <100 , use the t table instead of the z table, but as long as $N > 10$ the error is under 0.01 (Zar, 2010). It is not worth worrying about this small error in view of the fact that we are guessing at the standard deviation.

If we compare two independent samples, then the standardized deviation δ is

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{2\sigma^2}{N}}},$$

and proceed as before.

REFERENCES

- Bedard, P.L., Krzyzanowska, M.K., Pintilie, M., Tannock, I.F., 2007. Statistical power of negative randomized controlled trials presented at American Society for Clinical Oncology annual meetings. *J. Clin. Oncol.* 25, 3482–3487.
- Beyer, W.H., 1966. Handbook of Tables for Probability and Statistics. The Chemical Rubber Company, Cleveland, OH.
- Brown, C.G., Kelen, G.D., Ashton, J.J., Werman, H.A., 1987. The beta error and sample size determination in clinical trials in emergency medicine. *Ann. Emerg. Med.* 16, 183–187.
- Burback, D., Molnar, F.J., St John, P., Man-Son-Hing, M., 1999. Key methodological features of randomized controlled trials of Alzheimer's disease therapy. Minimal clinically important difference, sample size and trial duration. *Dement. Geriatr. Cogn. Disord.* 10, 534–540.
- Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14, 365–376.
- Chung, K.C., Kallianen, L.K., Hayward, R.A., 1998. Type II (beta) errors in the hand literature: the importance of power. *J. Hand Surg. [Am]* 23, 20–25.
- Coe, R., 2002. It's the effect size, stupid. What effect size is and why it is important. Available, <http://www.leeds.ac.uk/educol/documents/00002182.htm>.
- Cohen, J., 1988. Statistical Power Analysis for Behavioral Sciences. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Cohen, J., 1990. Things I have learned (so far). *Am. Psychol.* 45, 1304–1312.
- Cumming, G., Finch, S., 2001. A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. *Educ. Psychol. Meas.* 61, 532–574.
- Cumming, G., Fidler, F., L'au, J., 2012. Association Publication Manual: effect sizes, confidence intervals, and meta-analysis. *Austral J Psychol* 64, 138–146.
- Durlak, J.A., 2009. How to select, calculate, and interpret effect sizes. *J. Pediatr. Psychol.* 34, 917–928.
- Ellis, P.D., 2016. The Essential Guide to Effect Sizes. Cambridge University Press, Cambridge.

- Flight, L., Julious, S.A., 2016. Practical guide to sample size calculations: an introduction. *Pharm. Stat.* 15, 68–74.
- Freedman, K.B., Back, S., Bernstein, J., 2001. Sample size and statistical power of randomised, controlled trials in orthopaedics. *J. Bone Joint Surg. Br.* 83, 397–402.
- Freiman, J.A., Chalmers, T.C., Smith Jr., H., Kuebler, R.R., 1978. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *N. Engl. J. Med.* 299, 690–694.
- Gore, S.M., Altman, D.G., 1982. *Statistics in Practice*. Devonshire, Torquay, UK.
- Julious, S.A., 2005. Sample size of 12 per group rule of thumb for a pilot study. *Pharm. Stat.* 4, 287–291.
- Julious, S.A., Owen, R.J., 2006. Sample size calculations for clinical studies allowing for uncertainty about the variance. *Pharm. Stat.* 5, 29–37.
- Komlos, J., Cinnirella, F., 2005. European heights in the early 18th century. Available: <http://epub.ub.uni-muenchen.de>.
- Kraemer, H.C., 1988. Sample size: when is enough enough? *Am. J. Med. Sci.* 296, 360–363.
- Lehr, R., 1992. Sixteen s-squared over d-squared: a relation for crude sample size estimates. *Stat. Med.* 11, 1099–1102.
- Moher, D., Dulberg, C.S., Wells, G.A., 1994. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 272, 122–124.
- Nieuwenhuis, S., Forstmann, B.U., Wagenmakers, E.J., 2011. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nat. Neurosci.* 14, 1105–1107.
- Schulz, K.F., Grimes, D.A., 2005. Sample size calculations in randomised trials: mandatory and mystical. *Lancet* 365, 1348–1353.
- Sterne, J.A., Davey Smith, G., 2001. Sifting the evidence—what’s wrong with significance tests? *Br. Med. J. (Clin Res Ed)* 322, 226–231.
- Tsang, R., Colley, L., Lynd, L.D., 2009. Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized controlled trials. *J. Clin. Epidemiol.* 62, 609–616.
- Van Belle, G., 2002. *Statistical Rules of Thumb*. Wiley Interscience, New York.
- Walters, S.J., 2009. Consultants’ forum: should post hoc sample size calculations be done? *Pharm. Stat.* 8, 163–169.
- Weaver, C.S., Leonardi-Bee, J., Bath-Hextall, F.J., Bath, P.M., 2004. Sample size calculations in acute stroke trials: a systematic review of their reporting, characteristics, and relationship with outcome. *Stroke* 35, 1216–1224.
- Williams, J.L., Hathaway, C.A., Kloster, K.L., Layne, B.H., 1997. Low power, type II errors, and other statistical problems in recent cardiovascular research. *Am. J. Physiol.* 273, H487–H493 (Heart and Circulation Physiology 42).
- Zar, J.H., 2010. *Biostatistical Analysis*. Prentice Hall, Upper Saddle River, NJ.

CHAPTER 4

Exploratory Descriptive Analysis

BASIC CONCEPTS

Counting

When counting items, a common technique is to put a vertical stroke opposite the group into which the item falls (Fig. 4.1).

Weight (lbs)		Total
140–149		8
150–159		11
160–169		3

Fig. 4.1 Tallying in fives.

After four vertical strokes, a diagonal line is added to make a cluster of five. After all the items are recorded, the totals for each subgroup are obtained. This method is prone to error. When counting rapidly and putting down the lines carelessly, it is easy to put a fifth vertical line on top of the fourth, and then the oblique line demarcates a group of six, not five. For this reason, [John Tukey \(1977\)](#) introduced a safer counting method (Fig. 4.2).

A square is drawn. For the first four counts, one dot is placed in each corner of the square. For the next four counts, lines join pairs of dots. For the ninth count, a diagonal line joins two dots, and for the tenth count the other diagonal line is drawn. Each completed square registers ten counts. Not only is it marginally easier to count in tens than in fives, but the separation of the points makes it less easy to mistake the number of counts.

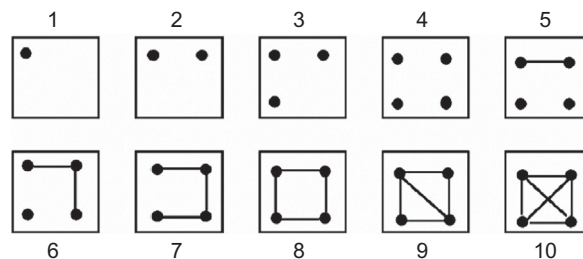


Fig. 4.2 Tallying by tens.

Distribution

Fasting blood sugar concentrations in normal people vary from 80 to 130 mg/dL, so that the values are *distributed* over a range. This is true of all variables. One definition of a distribution is a set of values that a variable can assume, and how often each value occurs. Before analyzing a distribution, we need answers to certain questions:

How symmetrical is the distribution?

How variable are the data?

Are there some unusually small or big values?

Are there clumps of data, with big gaps between subsets of data?

Could the data have come from a normal (Gaussian) distribution?

The questions apply equally to relations among two or more variables.

A Sorting Experiment

There are three boxes, one red, one green, and one blue, and a crate with a mixture of red, green, and blue balls. A blindfolded assistant picks out balls one at a time from the crate and gives them to you. You put each ball into the box with the same color as the ball. After 20 picks, there are 4 red balls, 11 green balls, and 5 blue balls. Therefore the sample of 20 balls drawn from the population has been sorted or *distributed* into three groups or classes based on the color of the balls. The results—4 red, 11 green, and 5 blue—make up a frequency distribution. Sometimes the results are given as proportions rather than as absolute numbers; in the sample of 20 balls, 20% are red, 55% are green, and 25% are blue. Often, the sample size is called 1, and the proportions of 1 made up by each color (the relative frequency distribution) are determined; here they are 0.2, 0.55, and 0.25. These data are displayed in Fig. 4.3, a bar graph designed to display nominal or ordinal data in which each bar represents a category or ordinal value and has a height proportional to the counts or relative frequencies.

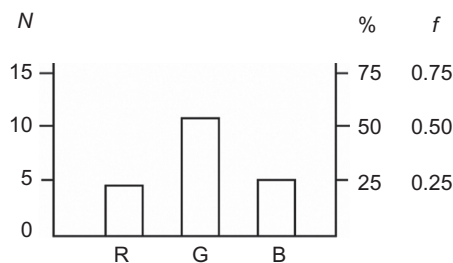


Fig. 4.3 Discrete frequency distribution of colored balls. N —number of balls; %—percentage of each color; f —fraction or proportion of each color; R—red; G—green; B—blue.

The classes make up a probability density distribution. The boxes into which the balls are put are often called *bins* and may contain real or abstract entities; for example, one bin may contain odd numbers and another bin may contain even numbers.

Histograms and Frequency Polygons

The classes in the colored ball experiment are discrete categories, but they need not be. Consider sampling the weights of 2000 people from a larger population. Rather than list all 2000 weights, group them into a *grouped* frequency distribution: <50 , $50\text{--}100$, and $100\text{--}150$ kg. Suppose that there were 400 people <50 kg, 1100 people between 50 and 100 kg, and 500 people between 100 and 150 kg. Then the bins represent the three weight groups or classes and show a distribution of weights from that population. This graph resembles that for the distribution of colored balls, but with the difference that the columns touch each other because the range of weights represents continuous ratio numbers (Fig. 4.4, upper panel). This figure is termed a histogram and the distribution is termed a frequency distribution or, when the areas are taken as proportions of 1, a probability density distribution. A histogram can be used also for interval numbers, but never for ordinal or nominal data.

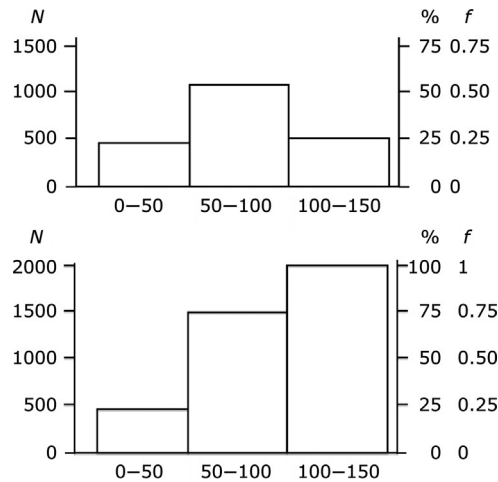


Fig. 4.4 Continuous distribution. N , number of measurements; f , relative frequency or proportion; %, percentage.

With continuous distributions what is done about measurements that fall on a boundary, for example, 100 kg? One solution is to weigh with greater accuracy, so that a weight of 100.1 kg goes into the highest class, and one of 99.9 kg goes into the middle class. If that cannot be done, the convention is for one half to go into one class and one half into the adjacent class.

We can calculate a cumulative frequency or probability density distribution by adding successive results. For the cumulative frequency distribution, there are $400 < 50$ kg, $400 + 1100 = 1500 < 100$ kg, and $400 + 1100 + 500 = 2000 < 150$ kg. For the cumulative probability density distribution, there are $0.20 < 50$ kg, $0.20 + 0.55 = 0.75 < 100$ kg, and $0.20 + 0.55 + 0.25 = 1.00 < 150$ kg (Fig. 4.4, lower panel).

When grouping measurements into classes, such as those in the previous figure, giving the beginning and ending values of the class makes it impossible to do further calculations. It is easier to define each class by a single number. To do this, take the midpoint of the range of values in each class, so that the class 0–50 becomes 25, the class 50–100 becomes 75, and so on. If the values are evenly distributed throughout the range in each class, then the average represents them all. This assumption is reasonable if there are large numbers of measurements in many classes and the distribution is not severely asymmetrical.

We can draw Fig. 4.4 in another way. Take the midpoint of each column and join these points to get Fig. 4.5.

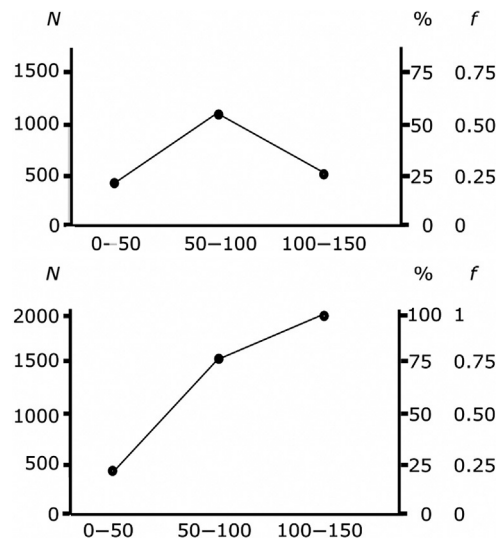


Fig. 4.5 Frequency polygon.

This is a frequency polygon, and it gives the same information as in Fig. 4.4. Cumulative distribution polygons are termed ogives. The frequency polygon may be easier to interpret than the histogram when several groups are compared.

When drawing a histogram to show a grouped frequency distribution, make the area under the column proportional to the frequency. If the base width (the class interval) is constant for all the columns, then the heights are proportional to the frequencies. The probability of being in any one group is the area under the rectangle. Dividing the area by the class size, that is, the width of the base, gives the height; this height is termed the probability density. In the example before, each group (or class) interval is 50 kg, but it could be 5, 8, 15 kg, or whatever is useful, as long as it is constant.

Histograms can be produced using online programs: http://www.wessa.net/rwasp_varia1.wasp#output, <http://easycalculation.com/graphs/create-histogram.php> or <http://www.socscistatistics.com/descriptive/histograms/>.

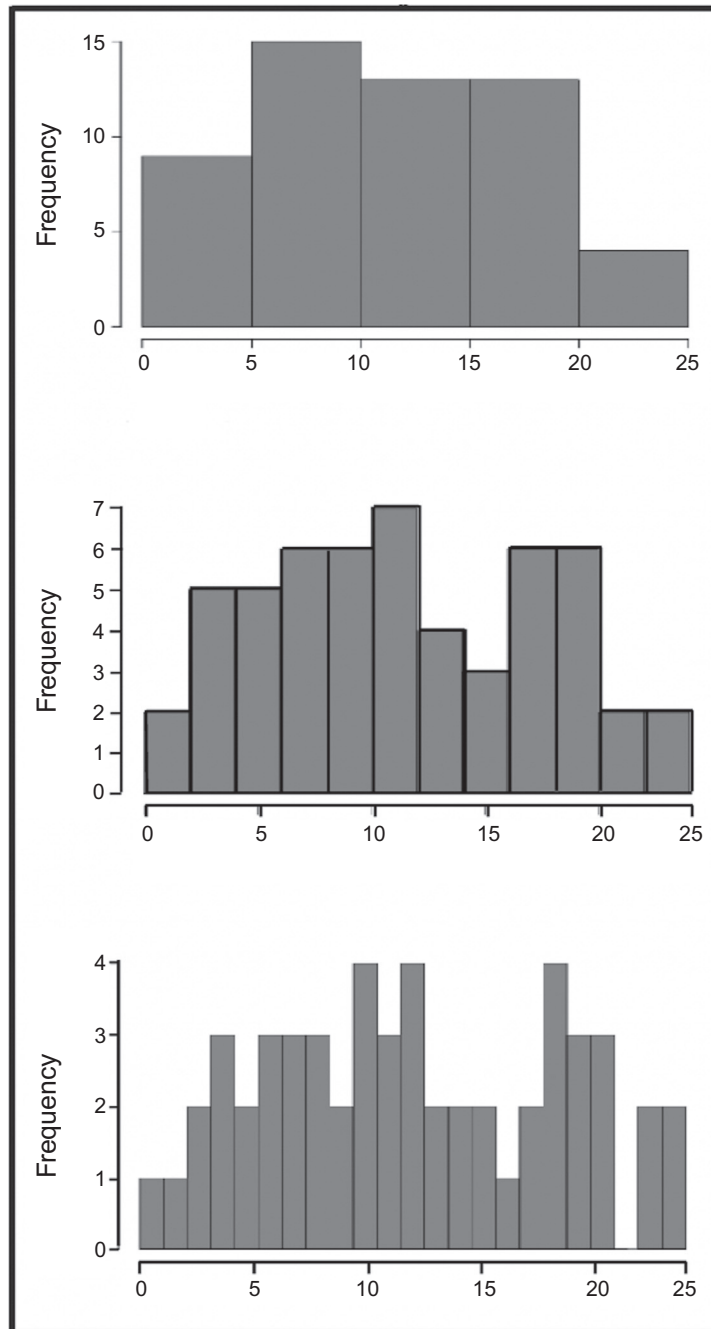


Fig. 4.6 Three histograms drawn with the same data but different starting points and class sizes. These changes produce large changes in shape of the histogram.

Histograms can be misleading. A change in the class interval (bin size) or the origin of the figure can greatly alter the apparent distribution that is indicated by the outline of the columns. This is shown in Fig. 4.6 of the same data formatted in three different ways.

Striking examples of these distortions are given by Scott (2010) who demonstrated that they can be minimized by using an averaged shifted histogram; the method is available in some specialized programs. In principle, the method estimates an optimal bin width and then produces a number of histograms with different starting points and averages them. The optimal bin width can be estimated from $3.5 \times \text{standard deviation} \times N^{-1/3}$.

The online programs listed before allow the user to change the number of columns (bin width) and examine the result.

Sometimes two histograms can be placed back to back so that two distributions can be compared (Fig. 4.7).

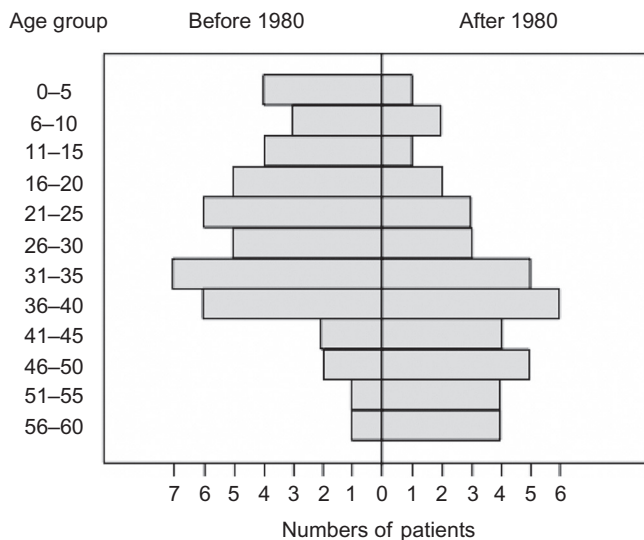


Fig. 4.7 Hypothetical data of age distribution in two time periods at a clinic. The shift to older subjects after 1980 is shown. This type of histogram can be drawn with the online program http://www.wessa.net/rwasp_backtobackhist.wasp#output.

Problem 4.1 The following data are the smiling times in seconds of an 8-week-old baby observed over several hours (Dean and Illowsky, 2012).

10.4	19.6	18.8	13.9	17.8	16.8	21.6	17.9	12.5	11.1	4.9
12.8	13.8	22.8	20.0	15.9	16.3	13.4	17.1	14.5	19.0	22.8
1.3	0.7	8.9	11.9	10.9	7.3	5.9	3.7	17.9	19.2	9.8
5.8	6.9	2.6	5.8	21.7	11.8	3.4	2.1	4.5	6.3	10.7
8.9	9.4	9.4	7.6	10.0	3.3	6.7	7.8	11.6	13.8	16.6

Smiling times (seconds).

Create histograms using different bin widths or numbers of bins.

Shapes of Distributions

Transformations

Certain shapes are common. One shape, discussed in [Chapter 6](#), is the symmetrical bell-shaped, normal, or Gaussian curve that has great theoretical significance. More often, the distribution is asymmetrical and pulled (or skewed) to the right or left by a few unusually high or low values, respectively. Sometimes the skewing is so marked that the greatest frequency is observed at one end of the scale. For the common right-skewed distributions, we need transformations to have a greater effect on large than small numbers; for example, 4 and 100 are 25-fold different, but their square roots of 2 and 10 are only 5-fold different. Logarithms have an even greater effect (0.60 and 2, respectively) and have the advantage of spreading out the low values. (Fig. 4.8) Negative reciprocal square roots and negative reciprocals spread out the lower values even more. If the distribution is left skewed, we want to increase the larger values to move them away from the mean. Squares and cubes will often effect this change.

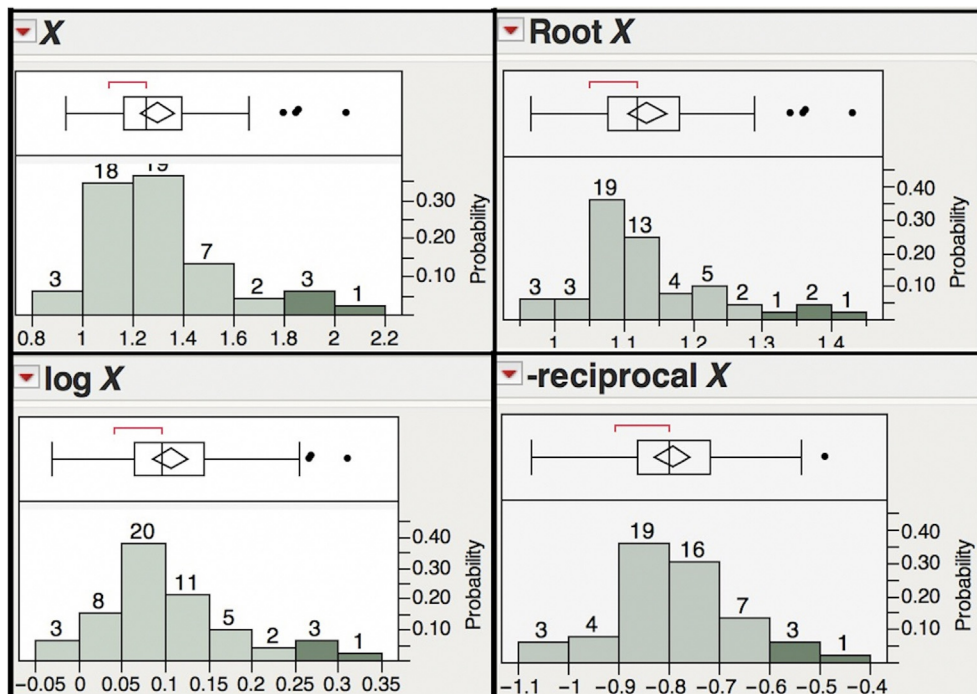


Fig. 4.8 (Upper left) distribution of X presented in [Table 4.1](#) (below) There are four outliers, indicated by dots. (Upper right) square root of X . Four outliers. (Lower left) logarithm (base 10) of X . The distribution is more symmetrical, and the lower values are more spread out. Three outliers. (Lower right) negative reciprocal of X , slightly more symmetrical than for $\log X$. Only one outlier.

All of these transformations can be done on almost any handheld calculator, but usually for only one number at a time. Online programs for logarithmic transformation to any base can be performed online at <http://www.1728.org/logrithm.htm> and http://www.rapidTable4.s.com/calc/math/Log_Calculator.htm. They can also be performed for batches of numbers at <http://vassarstats.net/index.html> that allows other transformations such as reciprocal or square root.

If two logarithmically transformed groups are compared, the difference might be wanted in untransformed numbers. To obtain these, the antilogarithms are computed.

Stem and Leaf Diagrams

Listing all the data in a set preserves the details but makes it difficult to picture the distribution. Consider the data set out as consecutive measurements in Table 4.1.

Table 4.1 Consecutive measurements

1.65	0.93	1.30	1.40	1.30	1.12
2.05	1.05	1.25	1.25	1.17	1.19
1.80	1.20	1.11	1.29	0.95	1.66
1.10	1.17	1.26	1.39	0.95	1.18
1.12	1.45	1.14	1.20	1.20	1.35
1.10	1.28	1.12	1.16	1.25	1.16
1.12	1.25	1.86	1.16	1.40	1.17
1.23	1.32	1.85	1.50	1.18	1.53
1.20	1.50	1.30	1.32	1.50	

The smallest and the largest measurements are displayed in bold type. It is difficult to envision the form of this distribution by inspecting the data.

The next maneuver is to arrange the data in order from smallest to biggest value (Table 4.2).

Table 4.2 Ordered measurements

0.93	0.95	0.95	1.05	1.10	1.10
1.11	1.12	1.12	1.12	1.12	1.14
1.16	1.16	1.16	1.17	1.17	1.17
1.18	1.18	1.19	1.20	1.20	1.20
1.20	1.23	1.25	1.25	1.25	1.25
1.25	1.26	1.29	1.30	1.30	1.30
1.32	1.32	1.35	1.39	1.40	1.40
1.45	1.50	1.50	1.50	1.53	1.65
1.66	1.80	1.85	1.86	2.05	

This gives a better idea of the distribution, but it is still not easy to perceive its shape. Therefore go one further stage and produce a histogram (Fig. 4.9).

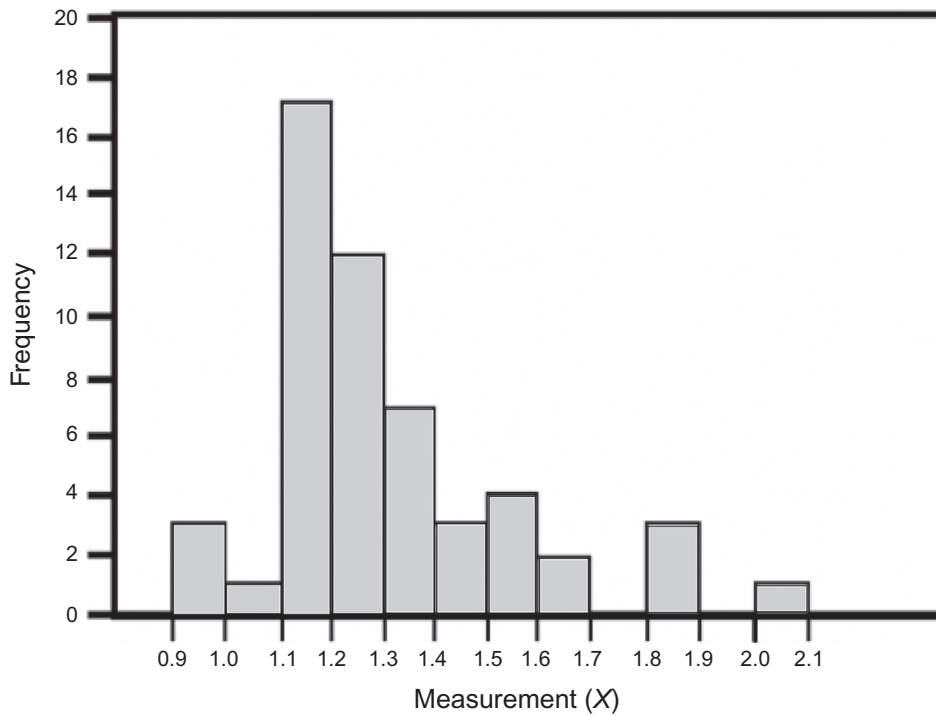


Fig. 4.9 Histogram.

The distribution is skewed to the right; the peak is at a low X value, and there is a long tail at higher X values. However, some detail is lost. In the 17 measurements between 1.1 and 1.2, we do not know if the measurements are evenly distributed throughout the interval or are clustered at the low or the high end. If we characterize this column by the average of 1.15, we have lost detail.

There is, however, a way to have the best of both worlds, and that is to create a stem and leaf diagram, as (Fig. 4.10).

To apply this principle to the data set of Table 4.2, examine the data that ranges from 0.93 to 2.05. Arbitrarily call the numbers up to the first decimal place the stems, and the second decimal place numbers the leaves to get Table 4.3.

When applied to the data in Table 4.2, we get Table 4.4.

Take the first stem of 0.9. The leaves represent the second decimal place. Thus stem 0.9, leaf 3 represents 0.93; stem 1.2, leaf 9 represents 1.29; stem 1.6, leaf 6 represents 1.66; and so on. The column labeled “Sum” indicates the number of measurements on each stem.

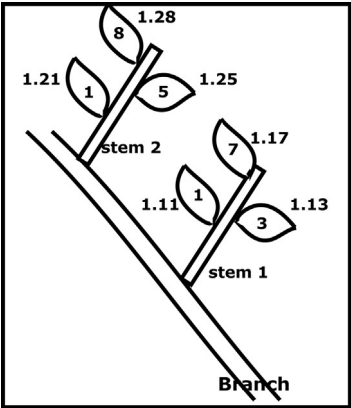


Fig. 4.10 This branch has 2 stems, each with 3 leaves. The lower stem is labeled 1.1, and the 3 leaves attached to it have the numbers 1, 3, and 7. These leaves represent the full numbers shown in parentheses beside them. The leaf with a 3 on stem 1.1 represents 1.13.

Table 4.3 Basis for stems and leaves

	Stem	Leaf
0.93 →	0.9	3
1.05 →	1.0	5

Table 4.4 Stem and leaf diagram

Sum	Stem	Leaf
3	0.9	3, 5, 5
1	1.0	5
17	1.1	0, 0, 1, 2, 2, 2, 2, 4, 6, 6, 6, 7, 7, 7, 8, 8, 9
12	1.2	0, 0, 0, 0, 3, 5, 5, 5, 5, 5, 6, 9
7	1.3	0, 0, 0, 2, 2, 5, 9
3	1.4	0, 0, 5
4	1.5	0, 0, 0, 3
2	1.6	5, 6
0	1.7	
3	1.8	0, 5, 6
0	1.9	
1	2.0	5

In this diagram all the information has been preserved, and it is possible to see the shape of the distribution. In fact, the outline of the numbers is like a histogram on its side.

In the stem and leaf diagram, the order of the measurements has been preserved, and each measurement is separated from the next by a comma for reasons discussed later.

Stem and leaf diagrams can be created online at <http://www.calculatorsoup.com/calculators/statistics/stemleaf.php> and <http://easycalculation.com/statistics/stem-leaf-plot.php>. They can also be created with Excel; see <https://www.qimacros.com/quality-tools/stem-and-leaf-back-to-back-excel/that> will also produce back-to-back stem and leaf diagrams. In practice, these are easier to create by hand.

How many classes (bins) should there be in a histogram or how many stems in a stem and leaf diagram? One class or stem for each number does not give enough data to represent the shape of the distribution accurately (Table 4.5).

Table 4.5 Too many classes

Sum	Stem
1	0.93
2	0.95
1	1.05
2	1.10
1	1.11
4	1.12
1	1.14
3	1.16
3	1.17
2	1.18
1	1.19
4	1.20
1	1.23
5	1.25
1	1.26
1	1.29
3	1.30
2	1.32
1	1.35
1	1.39
2	1.40
1	1.45
3	1.50
1	1.53
1	1.65
1	1.66
1	1.80
1	1.85
1	1.86
1	2.05

On the other hand, dividing the data into too few classes also is inefficient (Table 4.6).

The upper limit of the number of lines for a stem and leaf display or the number of intervals for a histogram can be estimated from either $10\log_{10}n$ or $4\sqrt[3]{n}$, where n is the

Table 4.6 Too few classes

Sum	Class/stem	
3	<1.0	0.93, 0.95, 0.95
49	1.0–1.99	1.05, 1.10, 1.10, 1.11, 1.12, 1.12, 1.12, 1.12, 1.14, 1.16, 1.16, 1.16, 1.17, 1.17, 1.17, 1.18, 1.18, 1.19, 1.20, 1.20, 1.20, 1.20, 1.23, 1.25, 1.25, 1.25, 1.25, 1.25, 1.26, 1.29, 1.30, 1.30, 1.30, 1.32, 1.32, 1.35, 1.39, 1.40, 1.40, 1.45, 1.50, 1.50, 1.50, 1.53, 1.65, 1.66, 1.80, 1.85, 1.86
1	≥2.0	2.05

number of observations; these rules work well for n between 20 and 300 (Emerson and Hoaglin, 1983). If n is 53, as in the example before, the recommended upper number of classes is 17 by the first rule and 15 by the second rule.

The next question is what stems to use. As the diagram is intended for ease of use, select a convenient stem such as 0.1. This matches the “natural” classification that gave 12 intervals, close enough to the recommended upper limit. There were 30 intervals in Table 4.5, and 3 intervals Table 4.6; both of these are of little use.

Sometimes the stem and leaf classification must be adapted when the natural numbering system by ones or tens does not give a reasonable number of classes. Consider having 32 measurements with stems of 1, 2, 3, and 4. Using these stems, one to a line, will not give a good idea of the distribution, because there will be too many leaves on any one line, as in Table 4.7. To get more lines, divide each line into two, each with the same stem, but with leaves 0–4 going on the first line and leaves 5–9 going on the second line, and distinguish the two sets by putting a star (*) next to the stem on the first line and a dot (•) next to the stem on the second line. As an example, consider the hypothetical data in Table 4.7.

Table 4.7 Artificial data set

1.1, 1.1, 1.3, 1.5, 1.7, 1.7, 1.7, 1.9
2.0, 2.0, 2.2, 2.2, 2.2, 2.5, 2.6, 2.6, 2.8, 2.8, 2.9
3.1, 3.1, 3.3, 3.3, 3.5, 3.6, 3.9
4.2, 4.2, 4.6, 4.7, 4.7, 4.9

Splitting the stems has converted 4 lines into 8, and this might be enough (Table 4.8).

If there still are not enough lines, split each unit into 5 lines per stem by taking two subunits at a time, as follows:

Stem * represents 0, 1

T represents 2, 3 (both terms begin with T)

F represents 4, 5 (both terms begin with F)

S represents 6, 7 (both terms begin with S)

• represents 8, 9.

Table 4.8 Doubled stem display

Sum	Stem	Leaf
3	1*	1, 1, 3,
5	1•	5, 7, 7, 7, 9
5	2*	0, 0, 2, 2, 2,
6	2•	5, 6, 6, 8, 8, 9
4	3*	1, 1, 3, 3
3	3•	5, 6, 9
2	4*	2, 2,
4	4•	6, 7, 7, 9

Consider the data set of 73 values presented in Table 4.9.

Table 4.9 Hypothetical set of numbers

0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.2, 0.2, 0.2, 0.2, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.5, 0.5, 0.5, 0.5, 0.6, 0.6, 0.6, 0.6, 0.7, 0.7, 0.7, 0.8, 0.8, 0.8, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 0.9, 1.0, 1.0, 1.0, 1.0, 1.0, 1.1, 1.1, 1.1, 1.2, 1.3, 1.3, 1.3, 1.3, 1.4, 1.4, 1.5, 1.5, 1.5, 1.6, 1.6, 1.7, 1.8, 2.0, 2.5.
--

The upper number of lines needed is $\sim 10\log_{10}73 = 18.6$, or 19 to the nearest whole number. Select 0.2 as a useful unit for the stem and get the modified stem and leaf diagram presented in Table 4.10.

Table 4.10 Five lines per unit stem

Sum	Stem	Leaf
6	0*	1, 1, 1, 1, 1, 1
12	T	2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3,
12	F	4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5
7	S	6, 6, 6, 6, 7, 7, 7
12	0•	8, 8, 8, 9, 9, 9, 9, 9, 9, 9, 9
8	1*	0, 0, 0, 0, 0, 1, 1, 1
6	T	2, 3, 3, 3, 3, 3,
5	F	4, 4, 5, 5, 5
3	S	6, 6, 7
1	1•	8
1	2*	0
0	T	
1	F	5
	S	
	2•	

The choice of the number of stems to use in a stem and leaf diagram depends on the number of data values and the range to be covered but should always be guided by convenience. There is no single way to do a stem and leaf diagram.

If the distribution of measurements is badly skewed so that, for example, the highest number is a million times as big as the smallest number, there would be enormous gaps

between the stems. It would be more useful to take logarithms of the numbers, so that a range from 1 to 1,000,000 becomes a range from 0 to 6 (Mosteller et al., 1983). On the other hand, if the distribution has most of its measurements within a given range, and a few measurements very far from the rest, it would be inappropriate to use logarithms, and unwieldy to use a conventional stem and leaf diagram with dozens of empty stems. Under these circumstances, use a conventional stem and leaf diagram without the extreme observations, and put these extremes as numbers at the bottom of the graph.

Stem and leaf diagrams are impractical with huge databases and should be replaced by histograms or box plots.

Problem 4.2 Use the data set from Problem 4.1 to create stem and leaf diagrams. Compare these with the histograms that you created from the same data set.

Problem 4.3 Use the data from Table 4.2 to construct a stem and leaf diagram.

Measures of Central Tendency (Location)

It is useful to have a single number to summarize a data set. For example, a value of 3.2 kg characterizes the weight of healthy, full-term newborn human infants. Some infants are lighter and some are heavier than 3.2 kg, but the value of 3.2 kg conveys information to someone who did not know what the characteristic weight of a term human infant is. (It is 107 kg for a newborn elephant.) As with other summarizing numbers, what we gain by compactness we lose by ignoring details. Summary numbers require prior attention to the whole distribution.

Arithmetic Mean

What properties should this summarizing number have? It should be near the middle of the distribution. The measure most often used is the *arithmetic* mean of the data set. To calculate this number, add up all the measurements and divide the total by the number of measurements. In formal terms,

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{N},$$

where \bar{X} is the mean, X_i stands for any member of the data set, and N is the number of observations. There are also harmonic means and geometric means that have specific uses to be discussed later. Unless one of these is specified, the term “mean” implies the arithmetic mean¹.

¹ Despite its obvious appeal, it took centuries before the concept of the mean was accepted. Raper, S. 2017. The shock of the mean. *Significance* 14(6), 12–16.

As an example, take the data set in Table 4.11.

Table 4.11 Simple data set

1 2, 3, 4, 5, 6

The mean is $(1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$. This is indeed in the middle of this data set. If these were the weights in kilograms of newborn human infants, each infant has a weight that differs from the weight typical of the group (3.5 kg) by some individual deviation from that mean, be it negative or positive.

If the values are plotted on a line (Fig. 4.11), the mean is the point at which the line with its points balances.

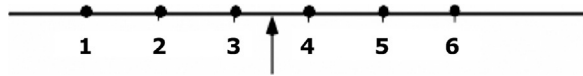


Fig. 4.11 Line diagram of points and their mean. There is no requirement that the mean be an observed value.

The sum of deviations from the mean is always zero.

What other properties should the measure of central tendency have? Instead of considering deviations of individual measurements from the mean, important statistical methodology has revolved around measuring the square of these deviations. One criterion is termed the *Principle of Least Squares*², which states that the sum of the squared deviations of the individual measurements from some number will be the least if that number is the mean. This is exemplified in Table 4.12.

Table 4.12 Principle of least squares

1 X_i	2 $(X_i - 7)^2$	3 $(X_i - 5)^2$
2	25	9
4	9	1
6	1	1
8	1	9
10	9	25
12	25	49
$\sum_{i=1}^{\infty} X_i = 42$	$\sum_{i=1}^N (X_i - \bar{X})^2 = 70$	$\sum_{i=1}^{\infty} (X_i - X_k)^2 = 94$
$\bar{X}_i = 7$		Let $X_k = 5$

² Most statistical tests evaluate a difference between two or more groups by relating that difference to a measure of variability. The less the variability, the easier it is to show that a given difference is not due to chance selection. Therefore, all other things being equal, any test that minimizes variability makes it easier to make statistical inferences. That is why the Principle of Least Squares is used so often in Statistics.

The data set in column 1 has a mean of 7, and the sum of squared deviations from the mean of 7 is 70 (column 2). Deviations from another value, such as 5 (shown in column 3) give a larger sum of squared deviations, namely, 94. This will be true for any value that is not the mean.

Formal proof of the least squares principle is given in the [Appendix](#).

The mean is easily calculated and conforms to the Principle of Least Squares. However, the mean has one major disadvantage—it is not a resistant statistic. Statisticians use the term “*resistant*” to mean that the statistic used is not much affected by unusually large or unusually small observations. Unfortunately, the mean is definitely not a resistant statistic. Consider what happens to the mean by changing the last value in the data set from 6 to 11 kg ([Table 4.13](#)). (According to the Guinness Book of World Records, the biggest human infant to survive weighed 10.2 kg at birth.)

Table 4.13 Modified data set

1 2, 3, 4, 5, 10

Then the mean becomes 4.16 kg, not 3.5 kg. A change in a single value has increased the mean that no longer serves to provide a reasonable indication of where the middle of the distribution is. A line diagram ([Fig. 4.12](#)) shows how the balance point has shifted from the center.

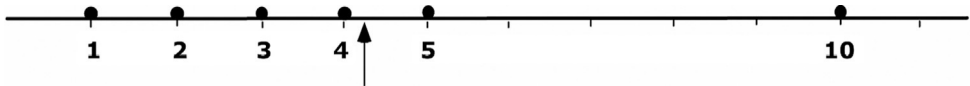


Fig. 4.12 Line diagram for points and their mean.

Median and Quantiles

To deal with the lack of resistance, use another measure of central tendency that is resistant, the *median*. This is the value of the measurement that splits the data set into two equal halves, so that 50% of the values are *below or equal* to the median and 50% are *above or equal* to it. In [Table 4.14](#) the median is 7 because two of the measurements are below it and two are above it.

Table 4.14 New data set

2, 4, 7, 8, 10.

Change the last X value from 10 to 50 and, applying this definition of the median, the median remains 7; the median has been resistant to the huge change in the final measurement. The median has to be used in place of the mean with censored observations and can also be used for ordinal values.

The definition includes the words “equal to.” This is to meet the contingency presented in [Table 4.15](#).

Table 4.15 New data set

2, 4, 6, 6, 6, 8, 10.

Here the median is 6, because half of the measurements (2, 4, 6) are below or equal to 6, and half of the measurements (6, 8, 10) are equal to or > 6 .

What happens to the median with an even number of measurements in the data set? In [Table 4.13](#) no one of these numbers satisfies the definition. The median cannot be 4, because then three measurements will be above it and three below, nor can it be 3 that produces two smaller and two larger measurements. By convention, the median is taken as the average of the two middle measurements, that is, $(3+4)/2=3.5$. Even though this is not one of the measurements, it satisfies the requirement that half of the 6 measurements are below it and half are above it. Values such as 3.3 and 3.9 also fit the definition of the median, but by convention are not so designated. Care is needed in using the values for the median, because if the two numbers used to form the median are far apart, their average might not be a good estimate of the population median.

To make the definition more general, the median is the value of the $(N+1)/2$ th measurement. If N is 5, then the median is the value of the $(5+1)/2=3$ rd measurement, and if N is 6 then the median is the value of the $(6+1)/2=3.5$ th measurement. Distinguish between measurements and ranks. If the measured values are ordered from smallest to largest, the smallest measurement is rank 1. The next smallest is rank 2, and so on. The rank gives the position of the measured value in the array. The median is the measured value corresponding to the $(N+1)/2$ th rank.

In a grouped frequency distribution, the median is calculated in a different way. [Table 4.16](#) presents heights of adult males grouped in class intervals of 2 in.

The median is the value of the observation with rank $(1052+1)/2=526.5$ th rank. By examining the cumulative frequencies, this rank lies within the group labeled with a height of 63 in. The class interval is 2 in., so that this group contains 346 people with heights from 62 to 64 in. The cumulative rank frequency up to that group is 279, so another $526.5-279=247.5$ ranks are required to reach the median value. If the measurements of height in this group are distributed evenly throughout the group, then the median value occurs $247.5/346$ ths ($=0.715$) of the way through that group from its beginning at 62 in. This value is therefore $62+0.715 \times 2=63.43$ in.

Heights (in.)	f	Cf
---------------	-----	------

55	3	3
57	8	11
59	53	64
61	215	279
63	346	625
65	278	903
67	119	1022
69	23	1045
71 and over	7	1052
Total	1052	

f , frequency and Cf , cumulative frequency.

The median is one of the ways in which a set of ordered data can be divided. Dividing an ordered set into equal parts produces *quantiles* (or fractiles). Division of the set into four, five, eight, ten, or one hundred equal parts produces quartiles, quintiles, octiles, deciles, and percentiles (or centiles), respectively. $X_{(i)}$, the p_i th percentile, is estimated as $100\left(\frac{i-0.5}{N}\right)$. By rearrangement, this becomes

$$i_{\text{th percentile}} = \frac{np_i}{100} + 0.5$$

To find the rank of the 50th percentile (the median) calculate

$$\text{Median(50th percentile)} = \frac{50N}{100} + 0.5 = \frac{N}{2} + 0.5 = \frac{N+1}{2}.$$

Each half of the data set (one below and one above the median) can be further subdivided into halves in the same way, so that the total data set can be divided into *quartiles*. The lower quartile (the 25th percentile) is the value of the measurement such that 25% of the measurements are below or equal to it and 75% are above or equal to it. The upper quartile (75th percentile) is the value of the measurement such that 75% of the measurements are below or equal to it and 25% are above or equal to it (Table 4.17).

Table 4.17 Set with quartiles

Lower quartile Median Upper quartile

↓ ↓ ↓

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

The lower quartile is 4.5, because 3 of the 12 measurements are below or equal to it. The median is 7.5, because half the measurements are below (or equal) to it, and half are above it. The upper quartile is 10.5, because 9 of the 12 measurements are below or equal to it, and 3 are equal to or above it.

Similarly, the 10th centile is the value of a measurement in an ordered array that has 10% of the measurements smaller and 90% of the measurements larger than it is. The value may be one of the measurements in the array or may be interpolated.

The calculation of any quantile in a grouped frequency distribution is similar to that for calculating the median.

The stem and leaf diagram is particularly well suited to defining the median and the quartiles. (Table 4.18, based on Table 4.4).

Table 4.18 Complete stem and leaf diagram

	Cumulative total	Sum	Stem	Leaf
53	3	3	0.9	3, 5, 5
50	4	1	1.0	5
49	21	17	1.1	0, 0, 1, 2, 2, 2, 2 , 4, 6, 6 , 6, 7, 7, 7, 8, 8, 9
32	33	12	1.2	0, 0, 0, 0, 3, 5 , 5, 5, 5, 5, 6, 9
20	40	7	1.3	0, 0, 0, 2, 2, 5, 9
13	43	3	1.4	0 0, 5
10	47	4	1.5	0, 0, 0, 3
6	49	2	1.6	5, 6
4	49	0	1.7	
4	52	3	1.8	0, 5, 6
1	52	0	1.9	
1	53	1	2.0	5

The numbers in bold type show the lower quartile, median, and upper quartile respectively.

Two new left-hand columns are added. The second column next to the column labeled Sum is the cumulative total starting from the first line. There are 3 measurements in the first line, 4 in the first two lines, 21 in the first three lines, and so on. The left-hand column is the cumulative sum starting from the last line and working up. The median is the value of the $(53 + 1)/2$ th measurement, that is, the 27th measurement. The number 27 is the *rank* of the measurement. Because the first three lines cumulate to 21 measurements, to reach the 27th rank add another 6 to get to the number 5 in the fourth line. This is indicated in bold and enlarged type. If the 27th measurement is the median, then the lower fourth (or quartile) is the $(27 + 1)/2$ th = 14th measurement from the lowest one, namely, the 6 on the third line indicated by the enlarged number in bold italics. Similarly, the upper fourth or quartile is the 14th measurement back from the highest measurement, or the 9 on the fifth line; it is the 9 and not the 0 on that line, because we are counting backwards from higher to lower numbers. These quartiles can be calculated online at

<http://easycalculation.com/statistics/inter-quartile-range.php> and <http://www.alcula.com/calculators/statistics/dispersion/>.

Mode

Another measure of central tendency is the measurement with the highest frequency—the *mode*. In a symmetrical distribution, the mean, median, and mode are identical. In an asymmetrical distribution, the median is toward the long-tailed side of the mode, and the mean is further away from the mode, being pulled toward one end by the larger measurements in the longer tail (Fig. 4.13).

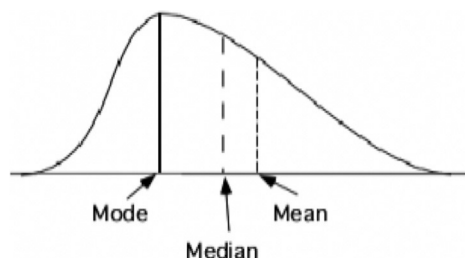


Fig. 4.13 Mode, median, and mean.

The Geometric Mean

For some data the arithmetic mean should be replaced by the geometric mean. Fig. 4.14 is a graph relating pH to hydrogen ion concentration.

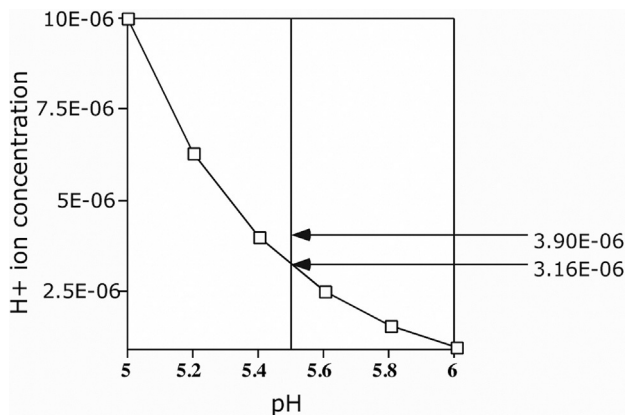


Fig. 4.14 Hydrogen ion concentration versus pH.

The pH scale is linear, but because pH is the negative logarithm of hydrogen ion concentration, the line relating the two sets of values is curved. For a mean pH of 5.5, the corresponding hydrogen ion concentration is $10^{-3.16}$, but averaging the hydrogen ion

concentrations gives a mean hydrogen ion concentration of $10^{-3.90}$, an incorrect value. To avoid this error, calculate the geometric mean. By definition, this is

$$\text{Geometric mean} = \sqrt[n]{\prod_{i=1}^n X_i}.$$

That is, the n th root of the products of all the n values. In the pH example $\text{Geometric mean} = \sqrt[6]{10 \times 6.31 \times 3.98 \times 2.51 \times 1.58 \times 1}$, where the values are the arguments with exponents 10^{-6} . This comes out to be 3.16, a correct value as shown from Fig. 4.15. Another way of calculating the geometric mean is

$$\text{Geometric mean} = \text{antilog} \frac{1}{n} \sum_{i=1}^n \log X_i.$$

Therefore taking the (negative) logarithms of the hydrogen ion concentrations produces the pH values; adding up these and dividing by 6 provides the true mean value. The geometric mean of raw data is the same as the arithmetic mean of the logarithm of the data.

Geometric means are used when data fit a logarithmic or an exponential function. One common application is when titers are measured with serial dilutions, for example, 1, 1/2, 1/4, 1/8..... (Chapter 31). They can also be used with badly skewed distributions, which can often be made reasonably symmetrical by taking the logarithms of the measurements.

Geometric means may be calculated online at <http://www.numberempire.com/statisticscalculator.php>, <http://www.easycalculation.com/statistics/geometric-mean.php>, and <http://www.calculatewhat.com/math/average/geometric-mean-calculator/>.

Harmonic Mean

This is the reciprocal of the arithmetic mean of reciprocals of individual values and is

$$\text{written as } H = \frac{1}{N} \sum X_i^{-1}, \text{ or alternatively as } \frac{1}{H} = \frac{\frac{1}{X_1} + \frac{1}{X_2} + \dots + \frac{1}{X_N}}{N} \text{ so that}$$

$$H = \frac{1}{\frac{1}{N} \left(\frac{1}{X_1} + \frac{1}{X_2} + \dots + \frac{1}{X_N} \right)}.$$

It is used occasionally when rates of change are involved. If a leucocyte moves 3 cm/min for 5 min, 5 cm/min for 5 min, and 7 cm/min for 5 min, its average speed of movement is not 5 cm/min but rather 4.44 cm/min (Table 4.19).

Table 4.19 Calculation of harmonic mean

$$\frac{1}{H} = \frac{\frac{1}{3} + \frac{1}{5} + \frac{1}{7}}{3} = 0.2243, \text{ and } H = 4.44 \text{ cm/min.}$$

To see why it is wrong to average the speeds, examine Table 4.20.

Table 4.20 Derivation of harmonic mean

Speed (cm/min)	Distance (cm)	Time (min)
3	5	1.67
5	5	1.00
7	5	0.71
	15	3.38
	Average = $15/3.38$ = 4.44	

Redrawn from McGraw-Hill.

As another example, if a car travels half the distance from A to B at 30 mph and the other half at 60 mph, what is the average speed? It cannot be 45 mph because the car has spent twice as much time in the first half than the second half. The harmonic mean by the previous formula is 40 mph and is a weighted mean.

In harmonic means every value is weighted by its reciprocal, and these means are used in certain multiple comparison calculations (Chapter 24).

This mean can be calculated online at <http://easycalculation.com/statistics/harmonic-mean.php>, and <http://www.calculatewhat.com/math/average/harmonic-mean-calculator/>.

Measures of Variability

Fig. 4.15 shows two distributions with the same mean but different variability. The terms variability, dispersion, and spread are often used interchangeably, but the term spread that has at least 25 different definitions should not be used.

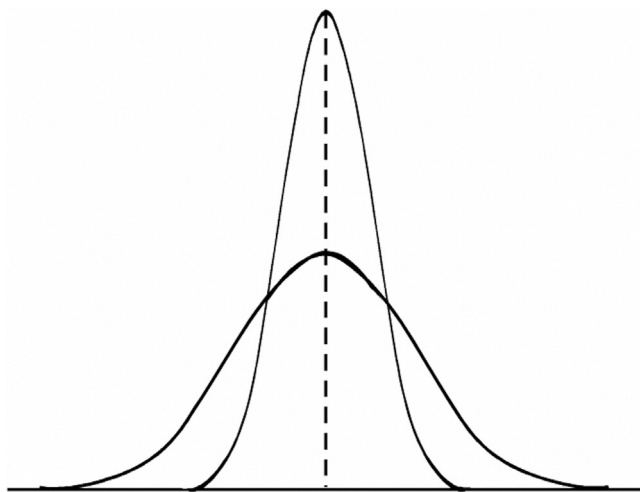


Fig. 4.15 Variability.

Early in the history of statistics in biology attention was concentrated on averages, and it was Francis Galton who emphasized the importance of variation. In his book “Natural Inheritance” (Galton, 1889) he wrote: “It is difficult to understand why statisticians commonly limit their inquiries to Averages and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once. An Average is but a solitary fact, whereas if a single other fact be added to it, an entire Normal Scheme, which nearly corresponds to the observed one, starts potentially into existence.”

Range

The simplest measure of variability is the *range*, the difference between the biggest and the smallest measurements. For the data in Table 4.18, the range is $2.05 - 0.93 = 1.12$. However, the range is not a resistant number. One unusually large or small value at either end of the distribution greatly alters the range, even if the variability of the remaining measurements does not change.

Standard Deviation

Another approach to summarizing the variability is to take the deviations of each measurement from the mean and average them:

$$\frac{\sum_{n=1}^{\infty} [X_i - \bar{X}]}{N}.$$

Unfortunately, this is useless, because

$$\sum_{n=1}^{\infty} \frac{X_i - \bar{X}}{N} = \sum_{n=1}^{\infty} \frac{X_i}{N} - \sum_{n=1}^{\infty} \frac{\bar{X}}{N}.$$

But $\sum_{n=1}^{\infty} \frac{X_i}{N} = \bar{X}$, and $\sum_{n=1}^{\infty} \frac{\bar{X}}{N} = \frac{N\bar{X}}{N} = \bar{X}$, because \bar{X} is a constant for this data set.

Therefore $\frac{\sum_{n=1}^{\infty} [X_i - \bar{X}]}{N} = \bar{X} - \bar{X} = 0$, and this is always true.

Another way of removing negative signs is to square the deviations. Calculate each deviation, square it, add up all the squared deviations, and divide by the number of deviations to get an average squared deviation that also gives a measure of variability known as the *variance* or *mean square*. However, the result gives squared units that may have no physical meaning. What possible meaning, for example, could square pounds have? The estimate of variability and restoration of the original units is completed by taking the square root to give the *standard deviation*:

$$\text{Variance} = \frac{\sum_{i=1}^{\infty} (X_i - \bar{X})^2}{N}, \text{ and standard deviation} = \sqrt{\frac{\sum_{i=1}^{\infty} (X_i - \bar{X})^2}{N}}.$$

In a population, the standard deviation is σ , and the variance is σ^2 . In a sample, the corresponding symbols are s and s^2 , but these are calculated slightly differently (see later).

As an example, consider Table 4.21.

Table 4.21 Calculation of standard deviation

X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
2	$2-6 = -4$	16
4	$4-6 = -2$	4
6	$6-6 = 0$	0
8	$8-6 = 2$	4
10	$10-6 = 4$	16
$\Sigma x = 30$ $\bar{X} = 30/5 = 6$	$\Sigma(X_i - \bar{X}) = 0$	$\Sigma(X_i - \bar{X})^2 = 40$

Variance = $40/5 = 8$, standard deviation = $\sqrt{8} = 2.83$.

Mean and standard deviation can be calculated online at <http://easycalculation.com/statistics/standard-deviation.php>, <http://www.numberempire.com/statisticscalculator.php>, <http://www.calculator.net/standard-deviation-calculator.html>, and <https://www.mathsisfun.com/data/standard-deviation-calculator.html>.

The standard deviation has three problems.

1. It is not a resistant statistic. Change one value in the Table 4.21, for example, 10–50, and the standard deviation changes drastically (Table 4.22).

Table 4.22 Calculation of standard deviation (2)

X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
2	$2-14 = -12$	144
4	$4-14 = -10$	100
6	$6-14 = -8$	64
8	$8-14 = -6$	36
50	$50-14 = 36$	1296
$\Sigma x = 70$ $\bar{X} = 70/5 = 14$	$\Sigma(X_i - \bar{X}) = 0$	$\Sigma(X_i - \bar{X})^2 = 1640$

Variance = $1640/5 = 328$ and standard deviation = $\sqrt{328} = 18.11$.

The calculations of the mean and population standard deviation are correct, but no longer give good estimates of the center of the distribution or the average variability.

2. For the standard deviation of a whole population, dividing by N is correct. Usually, though, we determine the standard deviation of a sample. Statisticians refer to unbiased statistics, by which they mean that with more and more samples the statistic

approaches the population parameter. If the long run value of the statistic does not approach the population parameter, they refer to a biased statistic. Dividing the sum of the squared deviations from the mean by N to get variance of a sample and then taking the square root to get standard deviation provides a biased statistic. Intuitively a sample will usually not have the biggest and smallest measurements of the populations and is likely to have a little less variability than would the whole population. Furthermore, because in general the sample and population means will not be the same, the average deviations will be smaller about the sample mean than the population mean. To correct for this, divide the sum of squared deviations from the mean by some value less than N to compensate for this difference.

What we divide by depends on the *degrees of freedom*. Consider the following problem. I give you a number, say 110, and ask you to find three numbers that add up to 110. Then you can select any number as the first one, any number as the second one, but now the third number is fixed because when added to the other two it has to come to 110. Therefore although there are three numbers to find, you have only two degrees of freedom of choice. Thus $N = 3$ leads to $N - 1 = 2$ degrees of freedom. Dividing the sum of squares of the deviations from the mean by the degrees of freedom yields an unbiased estimate of the standard deviation. In the previous example, because there is one data set (or, to put it another way, one mean or one total sum of all the measurements), the degrees of freedom are $N - 1$. The sample variance is $1640/4 = 410$, and the sample standard deviation is $\sqrt{410} = 20.25$, a little bigger than that calculated before. Another way of thinking about this is that when calculating N deviations from the mean, there are only $N - 1$ independent observations because all the deviations sum to zero, so that any one deviation can be determined by the remaining deviations. Consider the second column in [Table 4.20](#). The sum of the first four deviations is -36 , which but for the sign is the same as the fifth deviation; the fifth deviation is not an independent measurement. The sum of squared deviations from the mean is the sum of one dependent and four independent deviations from the mean, so that it is more appropriate to take an average by dividing by 4 rather than by 5.

The degrees of freedom are not always $N - 1$. In general, they are $N - k$, where k is the number of sample means that are being investigated. In the relationship between height and weight, each data set contributes one mean value, and so the total degrees of freedom are $N - 2$.

3. Calculating accuracy. In the example used before, with a small set of simple whole numbers, it is easy to calculate the sum of squares of deviations from the mean. It would not be as easy, however, with 169 measurements, each to two decimal places, the mean also having two decimal places after rounding off. Then each subtraction risks an arithmetic error. Furthermore, because the mean might not be an exact number, every subtraction produces a small error (due to rounding off) that is exaggerated by squaring it. For these reasons, we need a calculation that avoids the mean, minimizes errors, and is quicker to perform. Such a method is found by considering the following identity.

$$\sum (X_i - \bar{X})^2 = \sum (X_i^2 - 2X_i\bar{X} + \bar{X}^2) = \sum X_i^2 - 2\sum X_i\bar{X} + \sum \bar{X}^2.$$

Substitute $\frac{\sum X_i}{N}$ for \bar{X} . Then the expression becomes

$$\sum X_i^2 - 2\sum X_i \frac{\sum X_i}{N} + \sum \left(\frac{\sum X_i}{N} \right)^2 = \sum X_i^2 - \frac{2\left(\sum X_i\right)^2}{N} + \frac{N\left(\sum X_i\right)^2}{N^2},$$

and collecting like terms gives

$$\sum (X - \bar{X})^2 = \sum X_i^2 - \frac{\left(\sum X_i\right)^2}{N}.$$

This is an identity that holds no matter what X_i and N may be. By using it, we have avoided using the mean. The first part of the equation, $\sum_{i=1}^{\infty} X_i^2$ tells us to square each of the X values and then add up all the squares. The second part tells us to subtract what we get by squaring the sum of the X values, and dividing that squared sum, a single number, by N . This method is accurate and time saving. You still have to divide by the degrees of freedom to get the variance. That step is not included in the identity.

The mean and standard deviation are often listed as $\bar{X} \pm s$; for example, 6 ± 1.5 . Although there is no confusion about this format, it is incorrect, because a standard deviation can never be negative. It makes more sense to write $\bar{X}(s)$ or $\bar{X}(sd)$, and mention in the text that this is the format being used. Many books and journals now use the latter format.

Interquartile Distance

If the standard deviation is not a resistant statistic, what can replace it? Because the extremes of the distribution are more likely to be unrepresentative of the distribution than the central values, it is useful to replace the standard deviation by a number based on the middle 50% of the measurements, namely, the distance between the lower and the upper quartiles—the *interquartile distance* (IQD). The IQD is not affected by alterations in the outermost 25% of measurements in each tail of the distribution and represents the variability of the bulk of the measurements. In a statement attributed to the statistician Charles Winsor (1895–1951) by Tukey “all distributions are normal in the middle.” In [Table 4.18](#), the IQD is $1.39 - 1.16 = 0.23$. The interquartile distance also replaces the standard deviation when some values at one or both ends of the scale are censored (indeterminate) if, for example, the lowest values were listed as “<0.5 ng/mL,” or the upper values were listed as “>200 kg.”

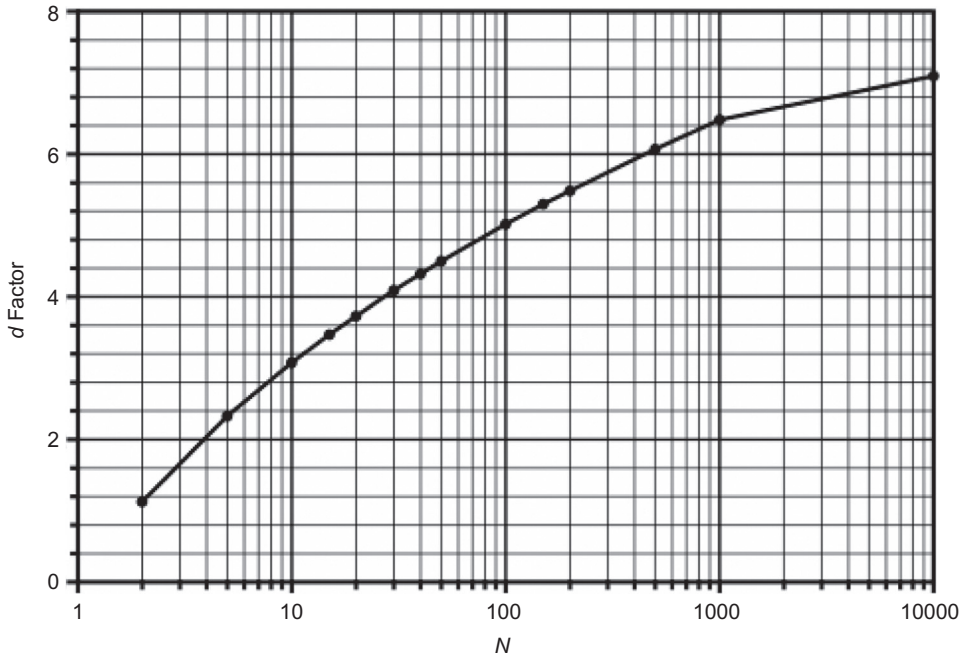


Fig. 4.16 Using sample size N to obtain a factor d that divided into the range approximates the standard deviation. Based on Table 4. A-18 in Mosteller, F., Rourke, R.E.K., 1973. *Sturdy Statistics. Nonparametrics and Order Statistics*, Addison-Wesley Publishing Company, London.

Interquartile distances can be determined online at <http://www.alcula.com/calculators/statistics/interquartile-range/>, <http://easycalculation.com/statistics/inter-quartile-range.php>, <http://www.miniwebtool.com/interquartile-range-calculator/>, and <http://www.statisticshowto.com/calculators/interquartile-range-calculator/>.

There are rough relationships among the range, standard deviation, and IQD. If the distribution is normal, the IQD is theoretically 1.349 times the standard deviation (Chapter 6). The ratio of the interquartile distance to 1.349 is termed the pseudostandard deviation or PSD (Hamilton, 1990). The relationship between range and standard deviation depends on sample size. Dividing range by factor d from Fig. 4.16 approximates the standard deviation.

This relationship is a check to see if the calculated standard deviation is approximately correct, and it allows an investigator to estimate the standard deviation for use in power analysis (see later) by using range data obtained in previous investigations. If the distribution is not normal, these divisors overestimate standard deviation, especially for sample sizes >100 .

Coefficient of Variation

The *coefficient of variation* is a dimensionless unit used for comparing variabilities. If 3 mice had weights 48, 50, and 52 g, their mean weight is 50 g and their standard deviation is 2 g. If 3 dogs had weights 48, 50, and 52 kg, their mean weight is 50 kg, and their standard deviation is 2 kg. The question “Which group has the greater variability?” is meaningless, because small light objects cannot vary as much as large heavy objects; a standard deviation of 2 kg is much more variable than a standard deviation of 2 g. But change the question and ask “Which of the two groups shows more variability relative to its average size?” Then a variability of 2 g out of 50 g is the same variability as 2 kg out of 50 kg. Therefore to determine relative variability, compute the coefficient of variation as

$$\frac{\text{Standard deviation}}{\text{Mean}},$$

and convert this to a percentage by multiplying by 100. This is a useful measurement. Workers in any field soon get familiar with the coefficients of variation that they have to deal with, and if some group has more than the expected variability it needs further examination. The coefficient is also useful when determining sample sizes.

Tables and Graphs

Any publication must present data clearly, providing numerical data, and figures to illustrate the numbers.

Tables

It is acceptable to include a few numbers in the text, for example, the mean and standard deviation of group A versus group B. If, however, the means and standard deviations of many groups are to be compared, writing them out in two or three lines makes it difficult for the reader to grasp the information without having to read the numbers several times. It is better to place the data in a simple Table. The more data there are, the more the need for a Table. Compare two ways of presenting the following data:

1. “In group A the mean and standard deviation were 17.3 and 2.7 mg, in group B they were 19.2 and 5.1 mg, in group C they were 16.0 and 1.1 mg, in group D they were 20.2 and 3.9 mg, and in group E they were 18.8 and 4.4 mg respectively.”
2. [Table 4.23](#)

Table 4.23 Summary of data

Group (age)	Mean (mg)	Standard deviation (mg)
A (0–10)	17.3	2.7
B (11–20)	19.2	5.1
C (21–30)	16.0	1.1
D (31–40)	20.2	3.9
E (41–50)	18.8	4.4

There are many possibilities for such a table. Separation of subgroups by judicious use of space, for example, sets off groups A, B, and C from groups D and E. Where possible, use subheadings, and try to give the groups meaningful descriptions, such as 10–19 years, 20–29 years, and so on, rather than force the reader to find out what A, B, C, D, and E are by looking at the legend.

Larger data sets with more possible comparisons demand a larger [Table 4.24](#).

Table 4.24 Section of table

	Diet alone <i>N</i> = 30 (%)	Statins <i>N</i> = 30 (%)	Placebo <i>N</i> = 30 (%)
Age (years)	62.1	64.6	59.3
Gender (male)	24 (80%)	27 (90%)	25 (83%)
High blood pressure	20 (67%)	24 (80%)	22 (74%)
Diabetes mellitus	0 (0%)	3 (10%)	2 (7%)
Smoking history	17 (57%)	14 (47%)	17 (57%)

Examples of effective tabulation are well shown by [Ehrenberg \(1975\)](#) and [Tufte \(1997\)](#).

Figures and Graphs

The books by [Tufte \(1983, 1990, 1997\)](#) are masterpieces to show good and bad graphic examples, covering almost all fields of human endeavor. [Tukey \(1977\)](#) and [Cleveland \(1984, 1985\)](#) initiated studies of what makes an effective scientific graph. Other excellent references include chapters [Good and Hardin \(2009\)](#) and articles by [Wainer \(1984, 1992\)](#) and [Moses \(1987\)](#). Graphs are used to display data in a more convenient form than the data from which they are derived, and often serve to emphasize and clarify relationships that would be difficult to obtain by examining the original data.

Many people ([Tufte, 1983](#); [Cleveland, 1985](#); [van Belle, 2002](#)) do not regard pie charts as useful in scientific articles. It is difficult to compare the areas of individual sectors without referring to the associated numbers that might just as well have been given without the chart. Almost always, data in a pie chart can be plotted on a linear scale, and it is easier to compare distances on a linear scale than angles in a pie chart. The pie chart has low data density and may not show subsets readily. Historically, however, one of the most influential charts ever published was the coxcomb type of pie chart developed by Florence Nightingale ([Joyce, 2008](#)).

Plain bar graphs, although ubiquitous, are of doubtful value. As usually used in scientific publications they seldom give more information than could be put more accurately in a simple Table. [Weissgerber et al. \(2015\)](#) pointed out that bar graphs should not be used for continuous data, and that scatter plots (that show all the data) and box plots (see later) are more useful, and incorporate a mass of data in a small area. Stacked bar graphs are also not favored. Although they do show more information, with an attempt to display relationships, it is often difficult to extract details that could usually

be better conveyed by line graphs (van Belle, 2002). Three-dimensional bar graphs are poor because it is often difficult to tell the exact height of the bars from the scale and because the eye is diverted from the essential data by the third dimension. The vertical scale in any of these plots should indicate units clearly, and as a rule the lower limit should be zero. If it is not, for example, to emphasize small differences, there should be a break in the axis to indicate that the axis does not start at zero (Fig. 4.17).

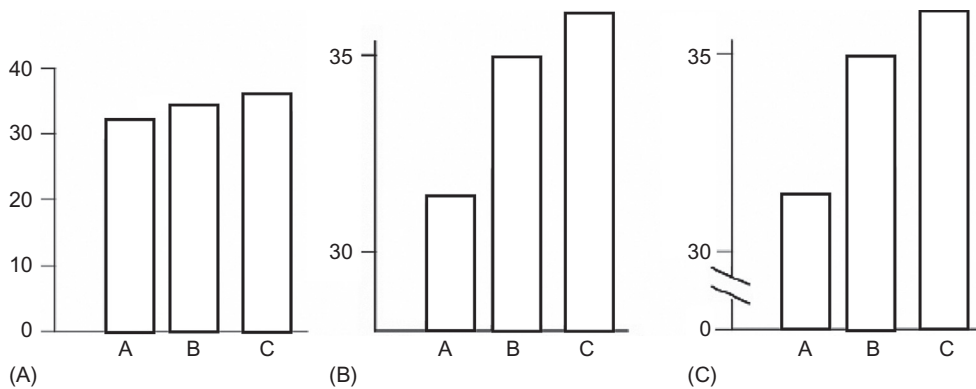


Fig. 4.17 (Left) full scale, (middle) expanded vertical scale, no zero, (right) expanded vertical scale with zero and break indicator.

The middle diagram can be misleading because it gives the appearance of large differences without warning the reader that zero has been suppressed and the scale changed. This warning is given in the right diagram. Other problems with bar charts are discussed by Reese (2007).

XY plots are very useful for showing relationships, and Cleveland (1985) gives a wealth of information about graph construction: the use of symbols, tick marks, color, fonts, spacing of numbers on the scale, and so on. The general advice is to use a rectangular graph with the width about 50%–70% greater than the height, but this may need to be changed so that emphasis is correctly placed on the X variate or the Y variate. Just as in Fig. 4.20, zero should be included or else its absence specified. Examples of how changes in the aspect ratio (vertical axis length/horizontal axis length) influence the reader's impression of the data are shown by Cook and Weisberg (1994).

A plot of two sets of data against time allows useful comparisons. Fig. 4.18 shows the age distribution of population in two countries—one poor and one rich.

Multivariable graphs can be made to display 5 variables as a moving bubble graph (Fig. 4.19). Two numeric variables are on the X and Y axes, bubble size represents another variable such as population, the bubbles are colored in various ways to represent different groups (e.g., countries and continents), and then a time variable can be added to give a moving representation of changes in the variables with time.

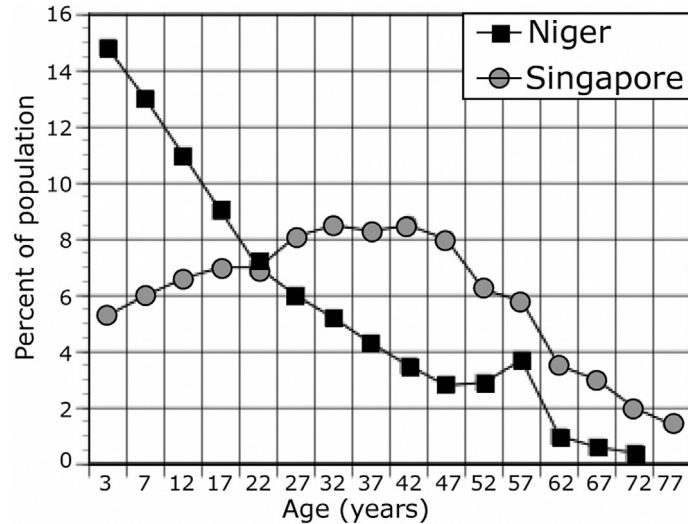


Fig. 4.18 In the poor country (Niger) the bulk of the population are young (due to a high fertility rate), so that the burden of supporting the population falls on a small number of elders. In the rich country (Singapore) the age distribution is more even so that many people from 22 to 57 who are wage earners contribute to the welfare of the country.

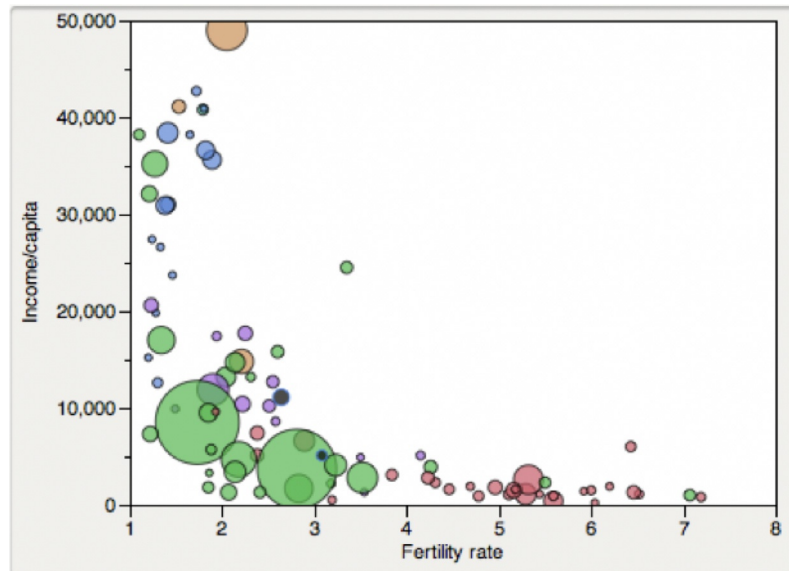


Fig. 4.19 Bubble graph to show relationship between fertility rate and income per capita and major world countries. The size of the bubble represents the relative population of a country; the two largest are China and India. The continents may be color coded, and it is possible to label countries. The figure shows a static graph at a given time, but sequential figures to show the effects of time can be incorporated.

These bubble graphs, developed by Hans Rosling's GapMinder Foundation, are now available at Google under the name of TrendAnalyzer. This is described online at <http://en.wikipedia.org/wiki/Trendalyzer>, and excellent examples can be found at <http://www.gapminder.org/world> that shows where the software can be obtained. An outstanding talk by Rosling in which these moving bubble graphs are used can be seen at http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html.

Box Plots

The data in Table 4.17 can also be set out in the form of a diagram known as a box plot, attributed to Tukey (1977) (Fig. 4.20).

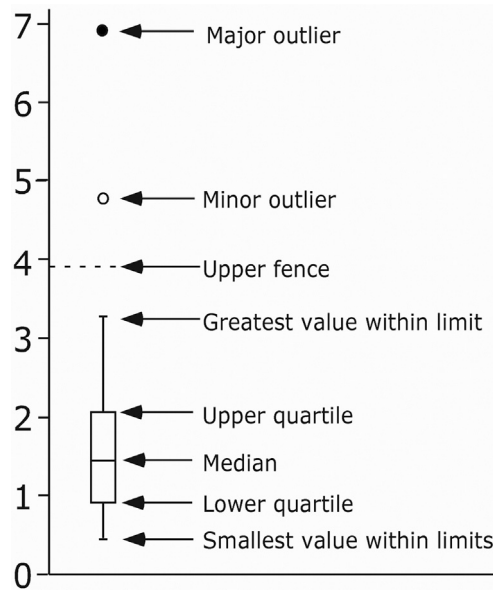


Fig. 4.20 Box plot.

Fig. 4.20 shows some of the summarizing numbers discussed before. The rectangular box (hence the name, box plot) has for ends the values corresponding to the upper and lower quartiles. The line across the middle of the box is the median. (Some programs also put in a symbol to indicate the mean.) Vertical lines ending in small crossbars are termed “whiskers,” and they end with the largest and smallest measurements in the data set, providing that these values fall within normal limits. These limits are determined by taking the interquartile distance (upper minus lower quartile) and multiplying this difference by 1.5 to obtain what is termed a “step.” In this example, the interquartile distance is $2.1 - 0.9 = 1.2$, and $1.2 \times 1.5 = 1.8$, so that the step is 1.8. The normal limits extend from

the upper quartile plus the step (to give the upper “fence”) down to the lower quartile minus the step (to give the lower fence); that is, from $2.1 + 1.8 = 3.9$ to $0.9 - 1.8$; because in this example the measurements cannot be less than zero, the lower fence is zero. The dashed line in Fig. 4.20 shows the upper fence; it does not normally appear in box plots. Any value above the upper fence or below the lower fence is an outlier, and each of these is put into the plot; minor outliers lie between one and two steps beyond the quartiles, and major outliers are $>$ two steps from the quartiles. The reason for selecting these values for minor and major outliers is that, with a normal distribution, only 0.349% of the observations will lie beyond the first fences (i.e., one step beyond the quartiles), and only 0.00012% will lie beyond the outer fences (i.e., more than two steps) beyond the quartiles.

Box plots and their associated numbers can be produced online at <http://www.alcula.com/calculators/statistics/box-plot/>, <http://illuminations.nctm.org/activitydetail.aspx?id=77>, <https://www.easycalculation.com/statistics/box-plot-grapher.php>, and http://nlvm.usu.edu/en/nav/frames_asid_200_g_4_t_5.html?open=instructions.

Box plots are useful for summarizing a lot of data in one figure, especially when two or more groups are compared. These plots may be easier to read and label displayed horizontally rather than vertically (Reese, 2005). Some programs make the box width proportional to the size of the group or its square root. Others place a notch around the median such that if two notches do not overlap the medians are significantly different.

Problem 4.4 Construct a box plot from Table 4.2.

Comparison box plots can be produced online at http://www.wessa.net/rwasp_notcheujmudbox1.wasp#output, which produces notched plots.

Box plots may be misleading if the data come from a bimodal distribution in which two clumps of data at high and low values are separated by a region with sparse intermediate data.

ADVANCED AND ALTERNATIVE CONCEPTS

Classical versus Robust Methods

General Aspects

The mean and standard deviation are the most often calculated summary statistics. Summary, however, is not their main purpose, and a box plot is a more extensive and useful summary of the needed information. The value of the mean and standard deviation is that they allow an estimation of the population values from which they were derived. However, they should not be used mechanically, but thought must be given to when they are or are not useful. If the sample is approximately normal

(Gaussian, symmetrical—[Chapter 6](#)) then the mean and standard deviation lead to excellent estimates of those in the population. Unfortunately, many sample distributions are ill behaved—asymmetrical, or having one or more unusually small or large measurements—and when this happens both the mean and standard deviation can be misleading. When this occurs, alternative measures of location and dispersion are needed.

The first line of defense is to use the median and the interquartile distance as measures of location and dispersion. The median, however, has both advantages and disadvantages. Imagine an array of X values with a given median value of X_{med} . We could change all the values above the median 10-fold and leave X_{med} unaltered. Clearly, to use the median alone as an indicator of the distribution would be inefficient. Further, consider the following array of numbers:

10, 20, 30, 70, 80, 90.

Conventionally, the median of this array is 50, but this might be an inefficient estimator of the location of the distribution, because any value from 31 to 69 could be the population median, so that our judgment could be greatly in error. The two examples given are bizarre, but they can occur and illustrate the need to consider the distribution before deciding how to use and analyze it.

A number of other measures have been used to improve our ability to deal with ill-behaved distributions.

Trimean

The trimean is defined as

$$\text{TRI} = \frac{q_l + 2\text{med} + q_u}{4},$$

where q_l and q_u are the lower and upper quartiles, and med is the median. These values correspond to the central rectangle in the box plot and use its three most important numbers to give a single statistic. The trimean is a resistant value that adds information beyond that provided by the median and gets rid of any difficulties resulting from using a single value.

The trimean can be calculated by using a quartile calculator at <https://www.hackmath.net/en/calculator/quartile-q1-q3>, <http://www.alcula.com/calculators/statistics/quartiles/>, <https://www.easycalculation.com/statistics/inter-quartile-range.php>, and <https://www.miniwebtool.com/quartile-calculator/>.

Trimming and Winsorization

Although the mean and standard deviation are not resistant statistics, they may be made more useful if some of the extreme outlying observations were modified. An example that most people are familiar with occurs in diving competitions. After each dive a panel of judges gives a score from 1 to 10. The highest and lowest scores are discarded, and the remaining scores are averaged. The idea is that extreme scores suggest poor adherence to the standards and should not play any further part. The method also guards in part against bias for or against a particular diver.

In statistics this technique is known as trimming (Mosteller and Rourke, 1973; Koopmans, 1987). A fraction of the data between 5% and 25% is discarded from *each* end of the distribution, and the trimmed mean is calculated. For $x\%$ trimming, calculate $x\%$ of N , the number in the array, and truncate it. If N is 14, 10% gives 1.4 and this is converted into 1 number dropped at each end. For 20% trimming, then 20% of 14 is 2.8, truncate this to 2, and drop off 2 numbers at each end. For example, the data set might be 3, 4, 5, 8, 9, 10, 11, 12, 18, 23, 39. The underlined measurements are regarded as extremes (for more on outlying values, see Chapter 9), and we can discard 1 or 2 values at each end.

For the standard deviation the data are Winsorized by replacing the discarded values by their nearest retained value; for example, after trimming 2 values at each end the new array becomes 5, 5, 5, 8, 9, 10, 11, 12, 18, 18, 18. Now calculate the variance of this Winsorized sample, s_W^2 , and calculate the trimmed standard deviation s_T as

$$s_T = \sqrt{\frac{(N-1)s_W^2}{N_T-1}} \text{ or } s_W \sqrt{\frac{N-1}{N_T-1}}$$

where N is the original sample size and N_T is the size of the trimmed sample.

Trimming brings the mean closer to the median, the biggest incremental change coming from trimming one value at each end of the array (Table 4.25).

Table 4.25 Effect of trimming

Statistic	All data	1-trim	2-trim
Mean	12.91	11.11	10.43
Upper quartile	18	15	12
Median	10	10	10
Lower quartile	5	6.5	8
IQD	13	8.5	4
Standard deviation	10.49	6.06	4.04
s_W		6.95	5.19
s_T		7.77	6.70
Trimean	10.75		

IQD, interquartile distance; s_W , Winsorized standard deviation; s_T , trimmed standard deviation.

Trimming has also markedly reduced the standard deviation. Trimming has given a sample mean and standard deviation that are more representative of the central part of the distribution. The choice between trimming 1 or 2 numbers from each end of the array depends in part on how discrepant the extreme numbers are and is subjective. Either trimming, however, is better than the mean of the original data set. The trimean is closer to the trimmed means than the original mean. A variant of the trimmed mean is the *mid-mean*, the arithmetic mean of the middle 50% of the values; this is equivalent to a 25% trimming fraction. The free online program <http://www.wessa.net/ct.wasp> makes it easy to calculate these various measures of central tendency, as does <http://smallbusiness.chron.com/calculate-trimmed-mean-excel-38559.html>.

Two issues about trimming must be considered. One is that trimming does not work with very asymmetrical distributions. The second is that trimming comes dangerously near to selecting data. It might be more advantageous to ask why some data points are so far from the main mass of data and whether they are really part of that distribution, or even to retain the data but transform the whole data set into a more normal distribution. If trimming is done, the procedure should be accompanied by an adequate justification, and the original data should be given. Finally, think back to the original example of rating divers. Would the average mark be much affected if both the upper and lower values were retained?

Order Statistics

Median Absolute Deviation

A large set of estimators based on order statistics have been used. One is based on the median absolute deviation (MAD)—the median of the absolute deviations of each value X_i from the median. It is the median of $|X_1 - \text{median}|, |X_2 - \text{median}|, \dots, |X_N - \text{median}|$ (Example 4.1).

Example 4.1 Data set is 2, 4, 6, 8, 50. The median is 6. The absolute deviations are $|2-6|=4$; $|4-6|=2$; $|6-6|=0$; $|8-6|=2$; and $|50-6|=44$. These deviations in order are 0, 2, 2, 4, 44 so that their median is 2. (Be careful. The abbreviation MAD is used for both the median and the mean absolute deviation; we use the former here.) It may be calculated online using <http://www.miniwebtool.com/median-absolute-deviation-calculator/>.

There are close relationships between MAD, standard deviation, and the interquartile distance if the distribution is normal:

$$\text{MAD} \approx 0.6745\sigma.$$

$$\sigma \approx 1.4826 \text{MAD}.$$

$$\text{Interquartile distance} \approx 2 \times \text{MAD}.$$

Therefore if the distribution is normal, any value beyond $9 \times \text{MAD}$ is >6 standard deviations from the mean, and a factor of $6 \times \text{MAD}$ excludes residuals above 4 standard deviations from the mean (Mosteller and Tukey, 1977).

Many statisticians recommend that the mean and standard deviation should be accompanied by one of these resistant statistics. Large differences between the classical and robust statistics should lead the investigator to reconsider whether using classical statistics is the best way of proceeding. As an example, the height of American men in one study of 1052 adults had a mean of 62.49," a standard deviation of 2.435," and the interval of 57.72–67.26" included 95% of the sample population. To determine how this compared with the heights of adult Chinese males in China, we could go to China, and by some random selection technique, perhaps based on the Census, select a sample of 30 adult males. Now in Shanghai there is a basketball player named Yao Ming (who played in the United States) who is 90" tall. This is an extremely rare height, but the fact remains that in a random selection, Yao Ming has as much chance of being selected as any

other adult male. If our sample then included 29 people with heights of 57–68" (and the same mean and standard deviation as found in the United States) and one man of 90", and if we went ahead and calculated means and standard deviations by classical methods, the mean height would be $\sim 63.41''$ and the standard deviation $\sim 5.17''$. These results imply that adult Chinese males are about 1" taller with much greater variability in height than we observed in the United States. Any robust method dealing with the one abnormal observation will give us better estimates.

Propagation of Errors

Sometimes we wish to know the standard deviation of the product when two numbers are added, subtracted, multiplied, or divided. If there are two or more sets of measurements made on the same population with the same methods, we can pool the sets and recalculate the mean and standard deviation from the larger pooled set. On the other hand, the measurements might have been made on samples from the same population but with different degrees of precision. The problem then becomes how to combine the data sets with due attention to the differences in variability. There are different approaches to this problem (Barford, 1967; Bevington and Robinson, 1992; Taylor, 1982) but all give similar results.

Combining Experiments (Addition)

Chapter 3 discussed weighting the data, in one example to take account of different numbers in the two samples by taking a weighted average, and in the other to account for differences in variance by dividing by a measure of variability. When combining two or more groups, both of these functions are involved by using the standard error (Chapter 7). It bears the same relationship to the variability of sample means drawn from the same population as standard deviation does to the variability of measurements in a sample, and is estimated by s/\sqrt{N} , the square of which is s^2/N . This statistic when used to weight means includes both the number of items and their variability in one.

To add two sets of independent measurements of the same variable, n measurements of X_A and m measurements of X_B , the precision of each data set being different as assessed by their respective standard errors, then their sum Z will have mean estimated by Barford, 1967, p. 63)

$$\bar{Z}_{A,B} = \frac{1}{se_A^{-2} + se_B^{-2}} \left(\frac{\bar{Z}_A}{se_A^2} + \frac{\bar{Z}_B}{se_B^2} \right)$$

and their combined standard error is

$$se_{AB}^2 = \frac{se_A^2 se_B^2}{se_A^2 + se_B^2}.$$

This formula can be extended to >2 data sets.

As an example, consider three sets of measurements of serum cholesterol from the same population with mean and (standard error) in mg/L being 141(10), 151(12), 161(41).

It would be unreasonable just to average them because we do not know the number in each group, and because the variability of the third group is so much higher than that of the other two groups. Plain averaging gives mean and standard deviation of 151(21). On the other hand, their weighted averages are $\left(\frac{1}{10^{-2} + 12^{-2} + 41^{-2}}\right) \left(\frac{141}{10^2} + \frac{152}{12^2} + \frac{161}{41^2}\right) = \left(\frac{1}{0.017539}\right)(1.41 + 1.0486 + 0.09578) = 145.64$ for the mean and 57.01 for the standard deviation.

The weighted average gives less prominence to the highest group with the largest standard error, as shown by the smallest terms derived from this third set of data.

Simple Multiplication (Scale Factor)

Consider a scale factor “a,” for example, the distances between cities (X_i) on a map in which 1 “represents 10 miles. Then the measured and actual distances are related by $X_1 = az_1$, $X_2 = az_2$, and so on, where z is the actual measurement.

$$\text{Then } \bar{X}_i = \frac{X_1 + X_2 + \dots X_n}{N} = \frac{az_1 + az_2 + \dots az_n}{N} = \frac{a(z_1 + z_2 + \dots z_n)}{N} = a\bar{z}_i$$

By similar reasoning, the population variance is $a^2\sigma^2$, and the sample variance is a^2s^2 .

Multiplying Means (Taylor, 1982)

Assume that quantity Q is the product of two sets of simultaneous measurements of flow rate (F) and time (T), both measured with some error.

Define the fractional error se_F of flow as $\frac{se_F}{F}$ and the fractional error se_T of time as $\frac{se_T}{T}$.

Then the variability of flow and time can be written as $\bar{F}\left(1 \pm \frac{se_F}{F}\right)$ and $\bar{T}\left(1 \pm \frac{se_T}{T}\right)$, respectively.

The variability of $\bar{Q} = \bar{F}\bar{T}$ is

$$\bar{F}\bar{T}\left(1 \pm \frac{se_T}{T}\right)\left(1 \pm \frac{se_F}{F}\right) = \bar{F}\bar{T}\left[1 \pm \left(\frac{se_T}{T} + \frac{se_F}{F} + \frac{se_T se_F}{TF}\right)\right]..$$

However, because the product of two small fractions is smaller still, neglect the final term to get

$$\bar{Q} = \bar{F}\bar{T}\left(1 \pm \frac{se_T}{T}\right)\left(1 \pm \frac{se_F}{F}\right) = \bar{F}\bar{T}\left(1 \pm \left[\frac{se_T}{T} + \frac{se_F}{F}\right]\right)$$

Example 4.2 Let mean flow in l/min be 3 with standard error 0.3 and let mean time in minutes be 5 with standard error 0.1, then the mean quantity in liters is estimated as $3 \times 5 = 15$, with standard error 0.12.

Dividing Means (Taylor, 1982)

By similar reasoning the best estimate of a value obtained by dividing two means \bar{A} and \bar{B} is

$$\bar{Q} = \frac{\bar{A}}{\bar{B}} \left[1 \pm \left(\frac{se_A}{\bar{A}} + \frac{se_B}{\bar{B}} \right) \right].$$

Three Useful Applications of the Variance

1. If each value of X_i is multiplied by a constant k , the variance of kX is

$$\text{Var } kX = \frac{\sum (kX_i - k\bar{X})^2}{N-1} = \frac{\sum [k(X_i - \bar{X})]^2}{N-1} = \frac{k^2 \sum (X_i - \bar{X})^2}{N-1}.$$

2. If there are two groups X_i and Y_i with equal sample sizes, what is the variance of $X_i + Y_i$?

By definition, variance is $\frac{\sum (X_i - \bar{X})^2}{N-1}$. The variance of $X_i + Y_i$ is

$$\text{Var}(X_i + Y_i) = \frac{\sum [(X_i + Y_i) - (\bar{X} + \bar{Y})]^2}{N-1}.$$

This can be rearranged to give

$$\begin{aligned} \text{Var}(X_i + Y_i) &= \frac{\sum [(X_i - \bar{X}) - (Y_i - \bar{Y})]^2}{N-1} \\ &= \frac{\sum [(X_i - \bar{X})^2 - 2(X_i - \bar{X})(Y_i - \bar{Y}) + (Y_i - \bar{Y})^2]}{N-1} \\ &= \frac{\sum (X_i - \bar{X})^2}{N-1} - \frac{2(X_i - \bar{X})(Y_i - \bar{Y})}{N-1} + \frac{\sum (Y_i - \bar{Y})^2}{N-1} \end{aligned}$$

If the two sets of data are independent, the term $\frac{2(X_i - \bar{X})(Y_i - \bar{Y})}{N-1}$ equals zero, involving as it does the sum of deviations from the mean. Therefore the variance of the sum of two variables is the sum of the variances of each variable. This is true if there are more variables than two.

3. By similar argument, the variance of the difference between two variables is also the sum of the variances of each variable.

APPENDIX

Least Squares Principle

We wish to show that the expression $\sum_{i=1}^{\infty} [X_i - k]^2$ is a minimum when k is the mean.

Take this expression and add to and subtract from it the mean, \bar{X}

$$\begin{aligned}\sum (X_i - k)^2 &= \sum (X_i - k + \bar{X} - \bar{X})^2 \\ &= \sum [(X_i - \bar{X}) + (\bar{X} - k)]^2 \\ &= \sum [(X_i - \bar{X})^2 + 2(X_i - \bar{X})(\bar{X} - k) + (\bar{X} - k)^2] \\ &= \sum (X_i - \bar{X})^2 + 2\sum (X_i - \bar{X})(\bar{X} - k) + \sum (\bar{X} - k)^2\end{aligned}$$

Now the sum of deviations from the mean, $\sum_{i=1}^{\infty} [X_i - \bar{X}]$, always equals zero. Therefore

$$\sum (X_i - k)^2 = \sum (X_i - \bar{X})^2 + \sum (\bar{X} - k)^2.$$

The expression $\sum (X_i - k)^2$ will be a minimum when the last term in the right-hand side equals zero, that is, when $\bar{X} = k$.

REFERENCES

- Barford, N.C., 1967. *Experimental Measurements: Precision, Error and Truth*. Addison-Wesley Publishing Co, London.
- Bevington, P.R., Robinson, D.K., 1992. *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill, Inc, New York.
- Cleveland, W.S., 1984. Graphs in scientific publications. *Amer. Stat.* 38, 261.
- Cleveland, W.S., 1985. *The Elements of Graphing Data*. Wadsworth, Monterey, CA.
- Cook, R.D., Weisberg, S., 1994. *An Introduction to Regression Graphics*. John Wiley & Sons, New York.
- Dean, S., Illowsky, B., 2012. *Continuous Random Variables: The Uniform Distribution*. Connexions, May 25, 2012. Available: <http://cnx.org/contents/130e078d-6f27-4bde-8648-a67eae701805@17>.
- Ehrenberg, A.S.C., 1975. *Data Reduction. Analysing and Interpreting Statistical Data*. John Wiley and Sons, London.
- Emerson, J.D., Hoaglin, D.C., 1983. Stem-and-Leaf displays. In: Hoaglin, D.C., Mosteller, F., Tukey, J.W. (Eds.), *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons, Inc., New York.
- Galton, F.I., 1889. *Natural Inheritance*. Macmillan and Company, London.
- Good, P.I., Hardin, J.W., 2009. *Common Errors in Statistics (and How to Avoid Them)*. John Wiley & Sons, Hoboken, NJ.
- Hamilton, L.C., 1990. *Modern Data Analysis. A First Course in Applied Statistics*. Brooks/Cole Publishing Co, Pacific Grove, CA.
- Joyce, H., 2008. Florence Nightingale: a lady with more than a lamp. *Significance* 5, 181–182.
- Koopmans, L.H., 1987. *Introduction to Contemporary Statistical Methods*. Duxbury Press, Boston.
- Moses, L.E., 1987. Graphical methods in statistical analysis. *Annu. Rev. Public Health* 8, 309–353.
- Mosteller, F., Rourke, R.E.K., 1973. *Sturdy Statistics*. In: *Nonparametrics and Order Statistics*. Addison-Wesley Publishing Company, London.
- Mosteller, F., Tukey, J.W., 1977. *Data Analysis and Regression. A Second Course in Statistics*. Addison-Wesley, Reading, CA.

- Mosteller, F., Fienberg, S.E., Rourke, R.E.K., 1983. *Beginning Statistics with Data Analysis*. Addison-Wesley Publishing Company, Menlo Park, CA.
- Reese, R.A., 2005. Boxplots. *Significance* 2, 134–135.
- Reese, R.A., 2007. Bah! Bar charts. *Significance* 4, 41–44.
- Scott, D.W., 2010. Averaged shifted histogram. *Wires Comp. Stat.* 2, 160–164.
- Taylor, J.R., 1982. *An Introduction to Error Analysis*. University Science Books, Mill Valley, CA.
- Tufte, E.R., 1983. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT.
- Tufte, E.R., 1990. *Envisioning Information*. Graphics Press, Cheshire, CT.
- Tufte, E.R., 1997. *Visual Explanation*. Graphics Press, Cheshire, CT.
- Tukey, J.W., 1977. *Exploratory Data Analysis*. Addison-Wesley Publishing Co, Menlo Park, CA.
- Van Belle, G., 2002. *Statistical Rules of Thumb*. Wiley Interscience, New York.
- Wainer, H., 1984. How to display data badly. *Amer. Stat.* 38, 137–147.
- Wainer, H., 1992. Understanding graphs and tables. *Educ. Res.* 21, 14–23.
- Weissgerber, T.L., Milic, N.M., Winham, S.J., Garovic, V.D., 2015. Beyond bar and line graphs: time for a new data presentation paradigm. *PLoS Biol.* 13e1002128.

CHAPTER 5

Basic Probability

INTRODUCTION

We use concepts of probability almost every day. The weatherman states that there is a 40% probability of rain. We talk about the likelihood that interest rates will go down still further and wonder about the chances of our favorite team winning the next game. People ask about the chances that changing our diet will prevent cancer of the colon.

TYPES OF PROBABILITY

Probability is often divided into two forms—objective and subjective. Objective probability is subdivided into classical or a priori probability and empirical or a *posteriori* probability. A priori probability is based on theory. For example, when tossing a coin we do not know whether the result will be a head or a tail, but believe that in the long run both heads and tails will occur half the time; this can be expressed symbolically as

$P(\text{head}) = P(\text{tail}) = 0.5$, where P stands for probability.

This hypothesis has been verified experimentally.

For another example, when tossing a six-sided die (one cube is a die, and the plural is dice) each of the numbers of dots from one to six should appear one-sixth of the time. This can be written symbolically as.

$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$, where the numbers in parentheses indicate the number of dots.

These expectations apply only if the coin is a normal coin and the die is not loaded on one side.

On the other hand, there is no theoretical value for the probability of dying from a heart attack in a given year. Instead, make a ratio out of the number of people who die from heart attacks to the total number of people who are alive (frequently subdivided into subgroups based on age, sex, race, etc.); then the probability of dying from a heart attack in that subgroup is, for example, 250 out of 10,000. This can be expressed as

$$P(\text{dying from heart attack}) = \frac{\text{Number dying from heart attack in 1 year}}{\text{Number alive during that year}}$$

More generally, empirical probability is derived from the relative frequency of a certain outcome. If an experiment is performed N times, and if m of these produce outcome E_i , then the relative frequency of E_i is $\frac{m}{N}$ and the probability of the occurrence of E_i is defined as

$$P(E_i) = \lim \left(\frac{m}{N} \right) \text{ as } N \rightarrow \infty.$$

This ratio should be obtained when N is infinitely large, but in practice as long as N is reasonably large, the estimated probability approximates the true value.¹

Subjective probability is vaguer. Consider the statement that the probability of finding a cure for AIDS in the next 5 years is 50%. There is no theory that will provide this information, and no numbers with which to make an empirical ratio. Instead, the statement provides a measure of confidence that indicates how advances in the field are going and thus the likelihood of a cure. There is more to subjective probability than in this simplified example, and section 3.5 of the book by [Barnett \(1999\)](#) discusses subjective probability in more detail.

BASIC PRINCIPLES AND DEFINITIONS

An experiment is a process that produces a definite outcome, for example, a head after tossing a coin, a 5 on throwing a die, a cure after a given treatment. The outcomes must be unique and mutually exclusive, that is, in a single experiment one outcome precludes any others being present. If we throw a die, the outcome can be any one of 6 numbers of dots, but no two numbers can occur in a single throw. The sample space $[S]$ for an experiment is the set of all the possible experimental outcomes. A set is a collection of objects that are termed elements or members of the set.

For coin tossing, $S = [\text{head, tail}]$, and for rolling a die, $S = [1, 2, 3, 4, 5, 6]$. The rectangle below ([Fig. 5.1A](#)) indicates the sample space and the set of outcomes for $S = [1, 2, 3, 4, 5, 6]$. The sample space for tossing two coins is $S = [\text{head, head; head, tail; tail, head; tail, tail}]$, and the rectangle in [Fig. 5.1B](#) indicates the sample space for these outcomes. The sample space for a pack of cards (not illustrated) is the 52 cards in the pack.

A	1	2	3	4	5	6
B	HH	HT	TH	TT		

Fig. 5.1 Sample spaces. H, head; T, tail.

¹ Probability and Odds are related. Probability is the proportion (or percentage) of times a given event can occur; for example, in a horse race with 10 horses, the probability of horse A winning is 1/10. Odds (in favor) is a ratio of the number of times an event occurs related to the number of times it does not occur; for example, the odds of horse A winning are 1–9.

Odds = Probability/(1 – Probability).

Probability = Odds/(1 + Odds)

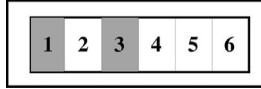


Fig. 5.2 Adding probabilities.

The three major axioms of probability [attributed to Andrey Kolmogorov (1903–87)] are as follows:

1. If an experiment has n possible mutually exclusive outcomes, namely, E_1, E_2, \dots, E_n , then the probability P of any given outcome E_i is a nonnegative number: $P(E_i) \geq 0$. It is impossible to conceive of a negative probability.
2. The sum of the probabilities of all mutually exclusive outcomes is 1:

$$P(E_1) + P(E_2) + \dots + P(E_n) = 1.$$

In [Fig. 5.1A](#) the sample space indicates a probability of 1, and each outcome has its own individual probability of $1/6$; the sum of all these probabilities is 1. Furthermore, consider n occurrences of outcome E_i in m experiments. If all the outcomes of the experiment were E_i , then

$$P(E_i) = \frac{m}{m} = 1, \text{ in conformity with axiom 2.}$$

From these axioms, $0 \leq P(E_i) \leq 1$; the probability of any outcome must range from zero (never occurring) to 1 (always occurring); usually an intermediate fraction occurs.

3. The probability of occurrence of either of two mutually exclusive outcomes E_i and E_j is the sum of their individual probabilities

$$P(E_i \text{ or } E_j) = P(E_i) + P(E_j).$$

The probability of getting a 1 or a 3 on a throw of a die is $1/6$ for a 1 and $1/6$ for a 3. The probability of getting one or the other is $1/6 + 1/6 = 1/3$ (the shaded area in [Fig. 5.2](#)).

This argument can be extended to more outcomes:

$$P(E_i \text{ or } E_j \text{ or } E_k) = P(E_i) + P(E_j) + P(E_k).$$

Additional definitions

4. Two sets are equal only if they contain the same elements.
5. If set A has one or more elements from set B , and every element of set A is also an element of set B , then set A is termed a subset of set B ; this is written $A \subset B$. For example, the shaded set $[1, 3, 5]$ is a subset of the universal set $U[1, 2, 3, 4, 5, 6]$ ([Fig. 5.3](#)).
6. A collection of elements from a set or sample space is a subset or an event.

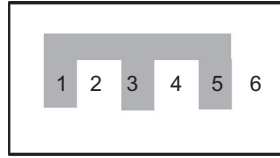


Fig. 5.3 Subsets.

Consider that the number of patients admitted with a myocardial infarction to an emergency room in 24 h may be 0, 1, 2, 3, 4, or 5. Then

$U = [0, 1, 2, 3, 4, 5]$ (U is the universal or total set of events)

$2 =$ the event that two or less patients are admitted with myocardial infarction $= [0, 1, 2]$. This is a subset of the universal set U .

Other definitions

1. The multiplication rule

The probability of getting outcomes E_i and E_j in two experiments is

$$P(E_i \text{ and } E_j) = P(E_i) \times P(E_j).$$

What is the probability of getting a 2 and a 4 in two throws of a die (or one throw of two dice)? For the first throw, there are 6 possible outcomes, each with a probability of $1/6$. Each of these possible outcomes can be associated with 6 possible outcomes from the second throw as long as the two throws are independent, and each of these throws has a probability of $1/6$. Therefore there are 36 possible combinations of outcomes involved in two throws of a die, so that any pair of outcomes has a probability of $1/6 \times 1/6 = 1/36$. Because the two dice are identical, a 2 on one die and a 4 on the other can occur in two ways: 2 on die 1 and 4 on die 2, with a probability of $1/36$, and a 2 on die 2 with a 4, also with a probability of $1/36$. Therefore if the order of throwing the dice is irrelevant, the probability of one 2 and one 4 is $1/36 + 1/36 = 1/18$.

Notice the difference between the addition and multiplication rules. If you bet that horse A or horse B will win a race, then your chance of winning is the sum of each probability; you have doubled your chances of success. If you bet that horse A will win one race and horse B will win the next race, then your chances of winning both races are the product of each individual probability and are less than either.

2. The *union* of two subsets A and B , symbolized by \cup , is another set that consists of all the elements belonging to A or B , or both A and B . This is termed the *inclusive* or (Ash, 1993).

Fig. 5.4 shows a Venn diagram, in which the sample space is represented by a rectangle and the probabilities of various subsets are represented by geometric forms with different areas. (John Venn (1824–1923) was an eminent logician and student

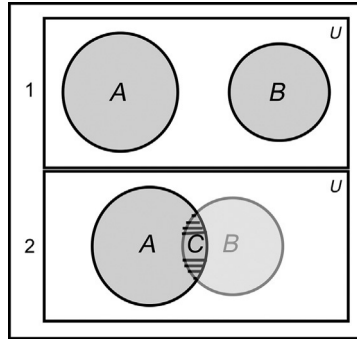


Fig. 5.4 Venn diagrams.

of probability. He did not invent this type of diagram, but his publication in 1880 popularized it.)

The letter U in the sample space represents all the outcomes. The probabilities of the two subsets of outcomes, A and B , are represented by the areas in the circles. In panel 1, the union of A and B includes all the outcomes in A plus all the outcomes in B . In panel 2, the union of A and B is not the same as the sum of the two areas because some outcomes are common to both, as shown by the cross-hatched area C ; the total shaded area in panel 2 is less than that in panel 1. In panel 1, the two subsets are *disjoint*, because they have no elements in common, whereas in panel 2 the two subsets are *conjoint* because they have at least one element in common.

3. The intersection (symbolized by \cap) of two subsets A and B is another subset that consists of all the elements common to both A and B . In panel 2 above, the area C represents the intersection of A and B . Two events are mutually exclusive if $A \cap B = 0$. This is the *exclusive or* (1).
4. If subset A is part of the universal set U , then the complement of A consists of the elements in U that are not part of A . They form another subset that can be symbolized by \overline{A} or A^c (Example 5.1).

$$U - A = \overline{A}, \text{ and } P(A) = 1 - P(\overline{A})$$

Example 5.1

Two dice are thrown. What is the probability that (a) both dice have an uneven number or (b) one even and one uneven number?

- (a) Die 1 can have an uneven number in $3/6$ ways (1,3,5), and so can die 2. By the multiplication rule both will be uneven in $3/6 \times 3/6 = 9/36$ ways.
 - (b) Die 1 has a $1/2$ chance of having an even number and die 2 has a $1/2$ chance of having an odd number. Therefore the combined probability is $1/2 \times 1/2 = 1/4$. But this is also true if Die 1 has the odd numbers and Die 2 has the even numbers, so that the total probability is $1/2$.
-

5. In a disjoint pair of subsets, as in Fig. 5.4 panel one, the probability of the union of A and B is the sum of their two probabilities:

$$P(A \cup B) = P(A) + P(B).$$

If this relationship is correct, it means that the two subsets are independent of each other, and that the events in A and B are mutually exclusive.

In a conjoint pair of subsets, as in Fig. 5.4 panel two, the probability of the union of A and B is the probability of A (the area in circle A) plus the probability of B (the area in circle B) minus the area of overlap (cross-hatched area C) that otherwise would be counted twice. Therefore

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

This is a more general form of expression than the previous one because it allows for the possibility of an intersection of the two subsets. If there is no intersection, then the term $P(A \cap B)$ is zero, and the two expressions are the same (Examples 5.2 and 5.3).

Example 5.2

In a school with 100 students: 80 are athletes (A), 60 play basketball (B), and 50 play chess (C).

Then the probability that a student plays basketball is $P(B) = \frac{60}{100} = 0.6$; the probability of a

student being an athlete is $P(A) = \frac{80}{100} = 0.8$, and the probability of a student playing chess is

$P(C) = \frac{50}{100} = 0.5$. The probability of a student playing basketball or chess or both is

$P(B \cup C) = P(B) + P(C) - P(B \cap C) = 0.6 + 0.5 - P(B \cap C)$, the last term being unknown.

It would be wrong to omit the last term, because then the probability would be 1.1, an impossibility.

Example 5.3

- a. What is the probability of drawing a heart (H) from a deck of cards, replacing it, and then drawing another heart?

Each probability is $13/52 = 1/4$, so that having both occur has a probability of $1/4 \times 1/4 = 1/16$.

- b. What is the probability that at least one of the two cards drawn is a spade? This can be symbolized by

$$\begin{aligned} P(S_1 \cup S_2) &= P(S_1) + P(S_2) - P(S_1 \cap S_2) \\ &= 1/4 + 1/4 - 1/16 = 7/16. \end{aligned}$$

Problem 5.1. What is the probability of drawing a king from a pack of cards?

CONDITIONAL PROBABILITY

Table 5.1 presents the relationship between age and three different bacterial causes of fevers in patients with cyanotic heart disease: brain abscess (BA), infective endocarditis (IE), and other bacterial infections (OB).

Table 5.1 Age and causes of fever

	<2Years	2–10Years	>10Years	Total
Other bacterial infections (OB)	40	15	5	60
Brain abscess (BA)	2	8	25	35
Infective endocarditis (IE)	1	3	7	11
Total	43	26	37	106

From this table, many probabilities can be determined. For example, what are the chances that a febrile patient will be under 2 years of age? There are 43 patients under 2 years of age out of a total of 106 patients, so that

$$P(< 2) = \frac{n(< 2)}{n(U)} = \frac{43}{106} = 0.4507.$$

However, there are more narrowly defined questions. What is the probability of a patient under 2 years of age having a brain abscess? This probability, from the definitions given before, is the intersection of being under 2 years of age and having a brain abscess. There are 2 patients in this category, so that the probability becomes

$$P(BA \cap < 2) = \frac{n[BA \cap < 2]}{n(U)} = \frac{2}{106} = 0.0189.$$

In these probabilities, the denominator is the total number of observations. Frequently, however, we want to determine a probability of the occurrence of an outcome in a subset of the total, that is, we wish to calculate the conditional probability, one that is conditional on the numbers in the marginal total. For example, we may want to ask the question: Among febrile patients with cyanotic heart disease, what is the probability of having a brain abscess and being under 2 years of age?

This probability is symbolized by $P(A|B)$, where the vertical line indicates that the probability of A is to be determined conditional on B being present. We can write

$$P(BA|< 2) = \frac{n[BA \cap < 2]}{n(< 2)} = \frac{2}{43} = 0.0465.$$

More generally, $P(A|B) = \frac{P(A \cap B)}{P(B)}$, providing that $P(B) \neq 0$.

What is the probability of selecting the King of Spades from a pack of cards? Because it is one of 52 cards, the probability is obviously $\frac{1}{52}$. However, work this out on the

basis of conditional probability. Let $P(B)$ be the probability of drawing a king and let $P(A)$ be the probability of drawing a spade. Then the probability of drawing the King of Spades is

$$\begin{aligned} P(B \cap A) &= P(A) \times P(B|A) \\ &= \frac{13}{52} \times \frac{1}{13} = \frac{1}{52} \end{aligned}$$

This relationship can be extended to three or more events.

$$\begin{aligned} P(A \cap B \cap C) &= P([A \cap B] \cap C) = P(A \cap B) \times P(C|[A \cap B]) \\ &= P(A) \times P(B|A) \times P(C|[B]) \end{aligned}$$

(or)

$$P(A \cap B \cap C) = P(A \cap [B \cap C]) = P(B \cap C) \times P(A|[B \cap C])$$

The way in which multiple intersections can be combined is arbitrary and does not affect the outcome.

Problem 5.2. What is the probability of drawing a king, a queen, and a jack in three successive draws from a pack of cards?

BAYES' THEOREM

Thomas Bayes (1701–61) was an English Nonconformist clergyman who was skilled in mathematics, and in 1742 he was elected a Fellow of the Royal Society (Bellhouse, 2004). He became interested in problems of scientific inference, but his publications were obscure and delayed by some doubts that he had about his conclusions. His classical work was first presented to the Royal Society 2 years after his death by his friend Richard Price who probably contributed to its completion (Hooper, 2013), and its importance was not recognized for >100 years. In Medicine, the theorem is used to determine if adding new evidence to existing knowledge alters the probability of a disease being present. For example, the incidence of chest pain due to coronary artery disease in the total population is ~2%, but if we add information that the subject is a 50-year-old obese male with diabetes then the probability of coronary artery disease becomes much higher. A simplified version of his theorem appears below.

Consider the equation that shows how to calculate $P(A|B)$ and the comparable equation for calculating $P(B|A)$, that is, the probability of B conditional on A being present

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(B \cap A)}{P(A)}$$

Multiply the first equation by $P(B)$ and the second equation by $P(A)$:

$$P(A|B) \times P(B) = P(A \cap B) \quad \text{and} \quad P(B|A) \times P(A) = P(B \cap A).$$

Because $P(B \cap A) = P(A \cap B)$, that is, the intersection of A with B is the same as the intersection of B with A , the right-hand sides of these two equations are equal. Therefore the two left-hand sides are also equal, and

$$P(B|A) \times P(A) = P(A|B) \times P(B).$$

Therefore $P(B|A) = \frac{P(A|B) \times P(B)}{P(A)}.$

In words, the conditional probability of B given A (the posterior probability) equals the conditional probability of A given B (also called the likelihood) multiplied by the probability of B (the prior probability) divided by the probability of A .

Bayes' theorem helps you to answer the question "Does new evidence help you to change your views about a problem." For example, does finding a positive test for a given disease increase the probability that the patient has that disease? Replace A by T^+ , standing for a positive result in a diagnostic test for a particular disease, and replace B by D^+ , that particular disease. Then $P(D^+|T^+)$ is the conditional probability of having a certain disease if there is a positive test, and $P(T^+|D^+)$ is the conditional probability of having a positive test, given that the patient has the disease. In clinical practice, we often do a diagnostic test, and if it is positive then ask how likely is it that the patient has a certain disease. Because tests are subject to false positives and false negatives, it is customary to take a number of patients with a given disease that is proven by some "gold standard," be it a biopsy, the results of surgery, an autopsy, or some other definitive test. This information yields $P(T^+|D^+)$. Then, if we want to know $P(D^+|T^+)$, the probability that the patient has the disease if the test is positive, we use Bayes' theorem:

$$P(D^+|T^+) = \frac{P(T^+|D^+) \times P(D^+)}{P(T^+)}$$

Returning to [Table 5.1](#), the rows indicate the numbers associated with different diseases (D^+) and the columns show the numbers associated with a category that can be regarded as a test outcome (T^+). These data are used to estimate $P(D^+|T^+)$. Without knowledge about the marginal totals, the probability of a patient under 2 years of age having a brain abscess [$P(BA \cap <2)$] is $2/106 = 0.0189$. This basic probability is referred to as the prior or pretest probability—the probability prior to having more information. Once we restrict the age group to <2 years (equivalent to a positive test result) and then ask about the probability of having a brain abscess, calculate

$$P(D^+ | T^+) = \frac{P(T^+ | D^+) \times P(D^+)}{P(T^+)}.$$

$$\text{Thus } P(BA | < 2) = \frac{P(< 2 | BA) \times P(BA)}{P(< 2)} = \frac{0.0572 \times 0.3302}{0.4057} = 0.0466.$$

By including knowledge of the outcome of some pertinent test the prior probability of 0.0189 has been increased to 0.0466; this latter probability is known as the posterior or posttest probability—the probability obtained after more information utilized. All that is required is knowledge of the particular $P(T^+ | D^+)$ involved; this is usually determined empirically.

Excellent books for further reading are by [Ash \(1993\)](#), [Hogg and Tanis \(1977\)](#), [Murphy \(1979\)](#), [Mosteller et al. \(1970\)](#), and [Ross \(1984\)](#). They give large numbers of examples, some complex, which can be used as templates for specific problems. There is an easy self-teaching guide by Koosis 10).

The clinical application of diagnostic tests is described in detail in [Chapters 20 and 21](#).

WORKED PROBLEMS

Many probability problems are related to gambling.

Consider two problems posed by the Chevalier de Méré to Blaise Pascal ([Examples 5.4 and 5.5](#)).

Example 5.4

What is the probability of throwing at least one 6 in four throws of a die?

With one throw, $P(6) = 1/6$ and the probability of a number not 6 = 5/6.

Probability of a number not 6 in 4 throws = $(5/6)^4 = 0.48225$.

Probability of at least one 6 in 4 throws = $1 - 0.48225 = 0.51775$.

Example 5.5

What is the probability of throwing two 6s in 24 throws with two dice?

With one throw of two dice, the probability $P(6,6) = (1/6)^2 = 1/36$, and the probability of not having two 6s = 35/36.

With 24 throws, the probability of having at least two 6s is $1 - (35/36)^{24} = 0.49140$.

If there is one more throw, the probability of having at least two 6s is $1 - (35/36)^{25} = 0.5055$.

Example 5.6

Perhaps the most famous modern probability problem is the Monty Hall door problem. On his guest show *Lets Make a Deal*, Monty Hall would show the contestant 3 closed doors, and state that behind one was a new automobile and behind each of the other two was a goat. The contestant was asked to pick a door. After this was done, Monty Hall then opened one of the two remaining doors to reveal a goat. The contestant was then asked to decide whether or not to change the original choice. The point at issue is whether changing the choice improves the chances of winning after knowing that one of the three choices has been removed. This problem caused enormous interest, with vast numbers of people advocating changing the choice and others advocating no change. Mathematical proofs of the wisdom of both recommendations were supplied, some by University Professors. The correct answer is to switch, but this seemed counterintuitive to many people.

There are numerous ways of describing the correct answer. One simple approach is to point out that initially the contestant has $1/3$ chance of picking the car and $2/3$ chance of picking a goat. After Monty has opened a door to show a goat, there are now two doors to choose from—one hides a goat and one hides a car. Switching is bad if the contestant initially picked the door with the car ($1/3$ of the time) but is good if the contestant had picked the goat that happens $2/3$ of the time. Therefore the chances of winning have doubled by switching choices (Shermer, 2009). Short video clips at <http://www.youtube.com/watch?v=mhlc7peGlGg> and <https://www.quora.com/How-do-I-solve-the-Monty-Hall-Problem-using-Bayes-Theorem> illustrate this approach.

To formalize this with Bayes' theorem

1. The car may be behind Door 1, 2, or 3, each with a prior probability

$$P(\text{Car1}) = P(\text{Car2}) = P(\text{Car3}) = 1/3$$

2. The car is behind Door1.
3. Assume that you pick Door1.
4. The likelihood that Monty opens door 3 to show a goat $= P(\text{Door3} | \text{Car1}) = 1/2$, because the car is behind one of the remaining two doors.
5. The posterior probability of the car being behind Door1 is

$$P(\text{Car1} | \text{Door3}) = \frac{P(\text{Door3} | \text{Car1}) \times P(\text{Car1})}{P(\text{Door3})} = \frac{1/2 \times 1/3}{1/2} = 1/3.$$

This is the probability of winning a car without switching.

6. If the car is behind door 2, the likelihood that Monty will open door 3 is 1 (because you have already picked door 1).
7. The posterior probability of the car being behind door 2 is

$$P(\text{Car2} | \text{Door3}) = \frac{P(\text{Door3} | \text{Car2}) \times P(\text{Car2})}{P(\text{Door3})} = \frac{1 \times 1/3}{1/2} = 2/3.$$

This is the probability of winning the car by switching.

A longer version is given at <https://sc5.io/posts/how-to-solve-the-monty-hall-problem-using-bayesian-inference/>.

Another way of tackling this problem is to set out a table, often helpful in solving probability problems. Based on solutions presented by [Vos Savant \(1990/1991\)](#) and [Everitt \(1999\)](#).

No switch Prize	First choice	Host opens door	No switch
Door 1	Door 1	2 or 3	Win
Door 2	Door 1	3	Lose
Door 3	Door 1	2	Lose

Switch Prize	First choice	Host opens door	Switch to
Door 1	Door 1	2	3 Lose
Door 2	Door 1	3	2 Win
Door 3	Door 1	2	3 Win

To begin, there are three possible winning choices (doors 1, 2, or 3). There is a 1 in 3 probability of being right with any of the doors picked and that does not change after the host opens a door because the choice has already been made. On the other hand, switching after added information has been given (i.e., which door does not conceal the prize) doubles the chances of winning.

The problem has an extensive history, summarized in a comprehensive article in Wikipedia ([Wikipedia Monty Hall, 2009](#)). It has recently been reexamined at length but simply by [Gill \(2011\)](#) ([Example 5.7](#)).

Example 5.7

The birthday problem.

There are 30 people in a room. What is the probability that at least two of them will have the same birthday (day and month, not year)? At first sight it appears to be low.

Start with calculating how many times none will have the same birthday. Person 1 could have a birthday on any one of 365/365 days (excluding leap years). Then if person 2 has a different birthday, his or her birthday must be on one of the remaining 364 days available. That is

$$\begin{aligned}
 &P(\text{Person 1 and Person 2 have different birthdays}) \\
 &= P(\text{Person 1 has a birthday on one of the 365 days in the year}) \\
 &\quad \times P(\text{Person 2 has a birthday on any of the 364 remaining days}) \\
 &= (365/365) \times (364/365)
 \end{aligned}$$

Continuing this approach for all 30 persons, all with different birthdays, leads to

$$\begin{aligned}
 P(\text{no birthdays in common}) &= (365/365) \times (364/365) \times (363/365) \times \dots (336/365) \\
 &= \frac{365 \times 364 \times 363 \dots 336}{365^{30}} \\
 &= \frac{(365 \times 364 \times 363 \dots 336) \times (335 \times 334 \times \dots 1)}{365^{30} \times (335 \times 334 \times \dots 1)} \\
 &= \frac{365!}{365^{30} \times 335!} = 0.2937
 \end{aligned}$$

Therefore the probability of not having all different birthdays is $1 - 0.2937 = 0.7063$.

This approach assumes that the chances of having a birthday are the same for each day in the year. This is incorrect, but differences from ignoring this assumption are small (Borja and Haigh, 2007).

Applets for running repeated trials of this problem can be found at <http://mste.illinois.edu/reese/birthday/>.

Problem 5.3. Use the logic of the birthday problem to solve the following problem.

A group of 20 people are asked to choose at random any number between 1 and 100.

What is the probability that two of them will have chosen the same number?

(a) <10%, (b) 10%–25%, (c) 25%–50%, (d) 50%–75%, or (e) >75%

REFERENCES

- Ash, C., 1993. *The Probability Tutoring Book. An Intuitive Course for Engineers and Scientists*. IEEE Press, New York, p. 470.
- Barnett, V., 1999. *Comparative Statistical Inference*. John Wiley & Sons, Ltd, Chichester.
- Bellhouse, D.R., 2004. The Reverend Thomas Bayes, FRS: a biography to celebrate the tercentenary of his birth. *Stat. Sci.* (1), 3–43.
- Borja, M.C., Haigh, J., 2007. The birthday problem. *Significance* 4, 124–127.
- Everitt, B.S., 1999. *Chance Rules. An Informal Guide to Probability, Risk, and Statistics*. Copernicus/Springer-Verlag, New York.
- Gill, R.D., 2011. The Monty hall problem is not a probability puzzle. (It's a challenge in mathematical modelling). *Stat Neerland* 65, 58–71.
- Hogg, R.V., Tanis, E.A., 1977. *Probability & Statistical Inference*. Macmillan Publishing Company, Inc., New York.
- Hooper, M., 2013. Richard Price, Bayes' theorem, and god. *Significance* 10, 36–39.
- Mosteller, F., Rourke, R.E.K., Thomas Jr., G.B., 1970. *Probability With Statistical Applications*. Addison-Wesley Publishing Company, Menlo Park, CA.
- Murphy, E.A., 1979. *Probability in Medicine*. Johns Hopkins University Press, Baltimore.
- Ross, S., 1984. *A First Course in Probability*. Macmillan Publishing Company, New York.
- Shermer, M., 2009. Prize probabilities. *Sci. Am.* 12.
- Vos Savant, M., 1990/1991. Game Show Problem. <http://marilynvosavant.com/game-show-problem/>.
- Wikipedia Monty Hall, 2009. http://en.wikipedia.org/wiki/Monty_Hall_problem.

SECTION II

Continuous Distributions

CHAPTER 6

Normal Distribution

INTRODUCTION

What is a normal distribution, and why is it important? The term “normal” means different things in everyday conversation and in statistics. In conversation, it implies something that is usual and, if appropriate, healthy.

To determine the “normal” resting blood pressure in healthy 10-year-old girls, take about 1000 healthy girls and measure resting blood pressure, and set the results out as percentiles, just as in the well-known growth charts for children. One percent of these children will have pressures >99th percentile, 5% of them will have pressures >95th percentile, 10% of them will have pressures >90th percentile, and so on. This percentile distribution is not statistically normal (see later), although it might be fairly close to it. Although these measurements are made in apparently healthy children, it is not clear that those at the extremes of the distribution are necessarily healthy. Children with resting blood pressures in the upper part of the percentile chart may have essential hypertension when they are adults. If this is true, then being above the 95th percentile, for example, may mean illness in the future, even if there is no illness now. This dilemma has been emphasized in relation to the standard growth charts for children (Cole, 2010). Because children are heavier now than they were 20 years ago, growth charts for the weights of “healthy” children at any age have a higher value for a given percentile now than they did earlier. A child overweight on a 1990 chart is in the normal range on a 2010 chart, but possibly destined to a variety of diseases in adult life. The World Health Organization (WHO) distinguishes between a *normal* chart and a *standard* chart, the latter involving only children whose weights indicate future health (not an easy task).

NORMAL OR GAUSSIAN CURVE

In statistical usage, the term “normal” is applied to a distribution specified by the equation

$$f_i = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(X_i - \mu)^2}{2\sigma^2}}$$

f_i is the height of the curve at value X_i , μ is the mean of the distribution, and σ is its standard deviation; π and e are constants; μ and σ are parameters; and X_i is a variable. All normal distribution curves have the same bell shape (Fig. 6.1) but differ in their means and standard deviations. The whole expression defines the Gaussian curve. Changing the values of the

parameters changes the position and width of the curve, but not its bell shape. Changing the value of X_i gives an estimate of the function of the expression at that value.

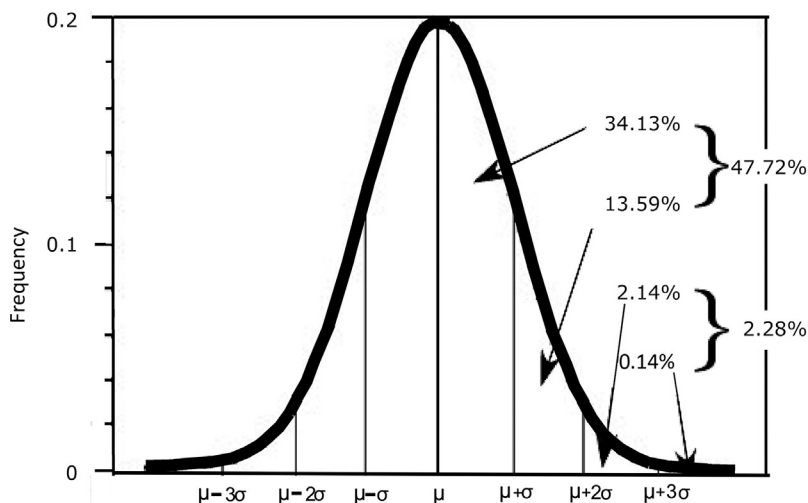


Fig. 6.1 Normal probability density plot of X variate against frequency. The areas under the curve as related to deviations from the mean (μ) in units of standard deviation (σ) are shown. 2.28% of the area under the curve is $>$ two standard deviations from the mean, and 0.14% of the area is $>$ three standard deviations from the mean. 2.5% of the area under the curve is $>$ 1.96 standard deviations from the mean.

The first description of the normal distribution was given in 1733 by Abraham de Moivre (1667–1754), a French mathematician who left France for England after the persecution of the Huguenots. He started with probability theory, in particular the binomial theorem (see [Chapter 16](#) and <http://www.mathsisfun.com/data/quincunx-explained.html>), and developed the formula so that it would be possible to compute the binomial distribution when the number of binary events (e.g., coin tosses) was very large. Later, scientists, particularly astronomers, began to be concerned about how to allow for measurement errors in celestial mechanics, and Laplace, Gauss, Galton, and others began to associate the distribution of these errors with the normal curve; the “true” value was the mean, and the errors distributed around the mean produced a normal curve.

This distribution produces the well-known bell shaped or Gaussian curve ([Fig. 6.1](#)):

Why is the X -axis labeled in standard deviation units? Different Gaussian curves have different means and standard deviations ([Fig. 6.2](#)).

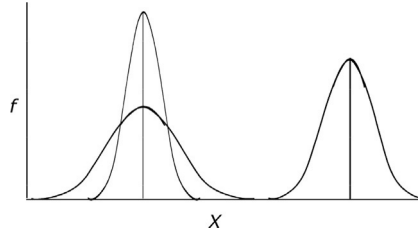


Fig. 6.2 Normal Gaussian curves with different means and standard deviations.

Between any two points on the X -axis, expressed in terms of number of standard deviations from the mean, the area under the curve is the same for all normal curves. Subtracting the mean μ from each value of X_i and then dividing by the standard deviation σ produces *standard deviates* symbolized by z .

$$z_i = \frac{X_i - \mu}{\sigma}$$

This is one type of linear transformation and it achieves two goals. First, by subtracting μ from every value of X_i , the numerator becomes zero. The normal curves are shifted so that each curve has a mean of zero (Fig. 6.3). Second, by dividing the numerator by the standard deviation, every normal curve has a unit standard deviation. Therefore wide curves with big standard deviations and narrow curves with small standard deviations each assume a standard shape with a mean of zero and a standard deviation of 1 (Fig. 6.3).

Once the z transformation has been performed, the new normal curve will be that shown in Fig. 6.1. The z transformation is based on population values for the mean and the standard deviation.

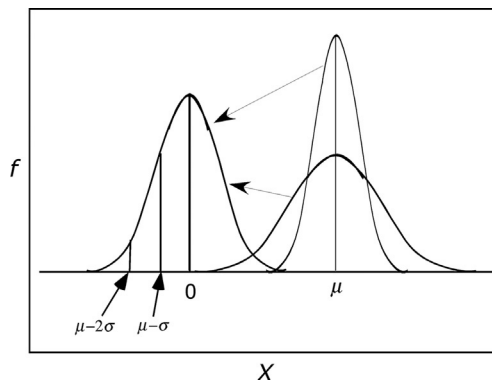


Fig. 6.3 z Transform. The curve on the left has a mean of zero and a standard deviation of 1 unit and is what both the curves on the right would look like after the z transformation.

The Quincunx

A quincunx is a device resembling a pinball machine that was developed by Sir Francis Galton (1822–1911) to illustrate the theory of random errors (Figs. 6.4 and 6.5).

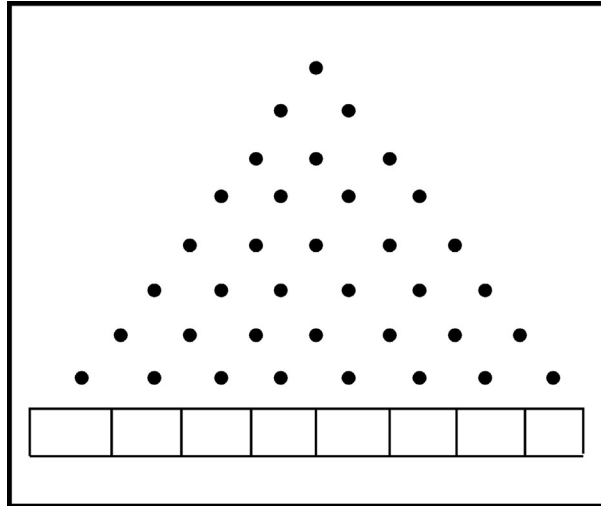


Fig. 6.4 Quincunx consisting of regularly placed pins arranged vertically on a board. Steel balls much smaller than the space between the pins are dropped from the top, bounce off the pins, and end up in one of the bins at the bottom. If balls are dropped into bins, the following distributions might occur (Fig. 6.5).

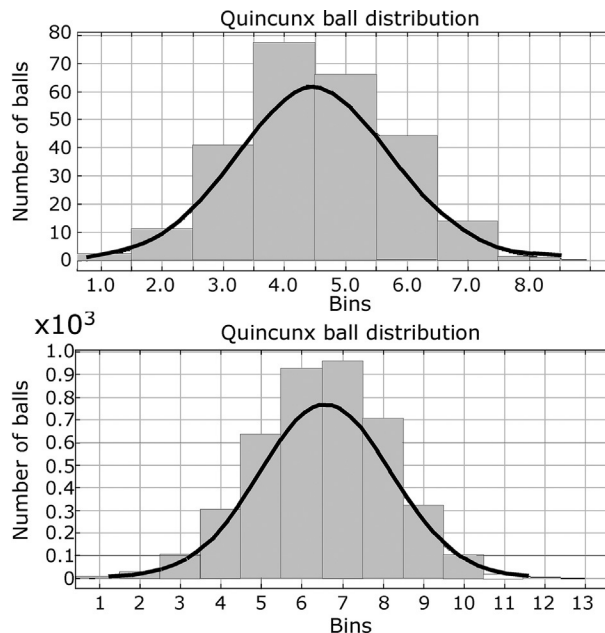


Fig. 6.5 (Upper): distribution of 256 balls into 8 bins. The normal Gaussian curve is superimposed on the histogram. The fit is fair. (Lower): results after dropping 4096 balls into 12 bins. The histogram is more symmetrical.

Constructed with applet from <http://www.jcu.edu/math/iseq/quincunx/quincunx.html>.

Each pin deflects a falling ball either to the left or the right. Despite the most careful construction of pins and balls, tiny imperfections or even air currents make each move at random to either side. Intuitively it is very unlikely for any ball to move always to the left or always to the right, so that the bins at the ends have very few balls. Most balls tend to have roughly equal numbers of leftward or rightward deflections, accounting for the peaks in the central bins. The more pins, bins, and balls there are, the closer the distribution matches the normal Gaussian distribution: a Gaussian distribution results from the addition of large numbers of positive and negative independent values. A right skewed distribution, as in Fig. 4.8, is often lognormal and often results from the multiplication of many independent numbers, which is why logarithmic transformation of multiplication to addition makes the distribution more symmetrical.

Fascinating pictures of balls moving through the pins and into the bins can be seen at <http://www.mathsisfun.com/data/quincunx.html> and <http://www.jcu.edu/math/iseq/quincunx/quincunx.html>.

Properties of the Normal Curve

The normal distribution curve is symmetrical about a mean (μ) of zero, and 68.26% of the measurements are within the limits of one standard deviation (σ) below to one standard deviation above the mean. The mean value is the most frequent measurement: 0.025 (or 2.5%) of the area under the curve is above a value of $X = \text{mean} + 1.96$ times the standard deviation; because the curve is symmetrical, there is an equal area below a value of $X = \text{mean} - 1.96$ times the standard deviation. These two areas added together give 0.05 (5%) of the area under the curve that is beyond the limits set by the mean ± 1.96 times the standard deviation. Fig. 6.6 shows how the areas under the curve are often represented.

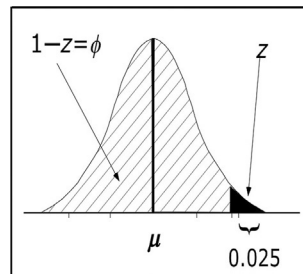


Fig. 6.6 Components of normal curve. z is the area beyond some value on the horizontal X -axis.

Most texts and tables give the values of the total area under the curve beyond any given value of z (black area). The remaining area (cross-hatched area) to the left of the z demarcation is $1 - z$, often termed ϕ ; this is the cumulative area from the left-hand end of the curve to the value of z . Occasionally some tables have different shaded areas, and the reader should check to see what the listed values refer to. Cumulative areas can be calculated at <http://stattrek.com/online-calculator/normal.aspx> and <http://www.danielsoper.com/statcalc3/calc.aspx?id=2>.

For a continuous distribution of this type it makes little sense to ask about the probability of obtaining a given X value. The probability of an adult male human weighing 68.83997542001 kg is virtually zero. What does matter is the area under the curve between different values of X . As examples:

- What proportion of the area under the curve lies between the mean μ and one standard deviation σ below the mean? From Fig. 6.1, the area between these limits is 0.3413 or 34.13% of the total area.
- What proportion of the area under the curve lies between the μ and 0.5 standard deviations below the mean? From tabulated areas under the curve, the area under the curve for $\mu - 0.5\sigma$ is 0.1914 (Fig. 6.7A).

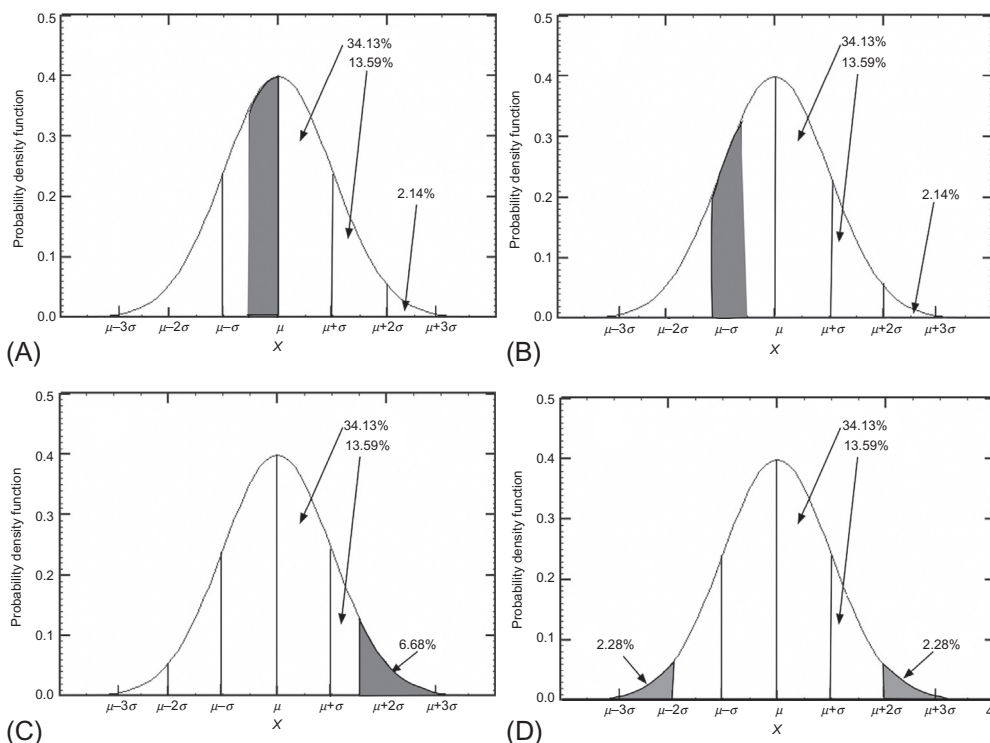


Fig. 6.7 Areas under the curve.

- c. What proportion of the area under the curve lies between 0.5 and 1 standard deviations below the mean? From tabulated areas under the curve, the area under the curve for $\mu - 0.5\sigma$ is 0.1914, and for $\mu - \sigma$ is 0.3413. Therefore the required area is $0.3413 - 0.1914 = 0.1499$ (Fig. 6.7B).
- d. What proportion of the area under the curve is >1.5 standard deviations above the mean? From tabulated areas under the curve, the area under the curve for $\mu \pm 1.5\sigma$ is 0.4332. The whole area above the mean is 0.5. Therefore the required area is $0.5 - 0.4332 = 0.0668$ (Fig. 6.7C).
- e. What proportion of the area under the curve is >2 standard deviations above and below the mean? From tabulated areas under the curve, the area under the curve for $\mu \pm 2\sigma$ is 0.04560 (Fig. 6.7D).
- f. 99.73% of the area under the curve lies between $\mu \pm 3\sigma$, 99.9937% between $\mu \pm 4\sigma$, and 99.999942 between $\mu \pm 5\sigma$.

Problem 6.1 Use the online calculator to determine the area under the normal curve between the limits of 0.75σ below the mean to 1.5σ above the mean.

All these areas can be calculated easily using http://davidmlane.com/hyperstat/z_table.html, or <http://psych.colorado.edu/~mcclella/java/normal/accurateNormal.html>.

The normal curve can also be converted into a cumulative probability density curve with its characteristic S or sigmoid shape (Fig. 6.8).

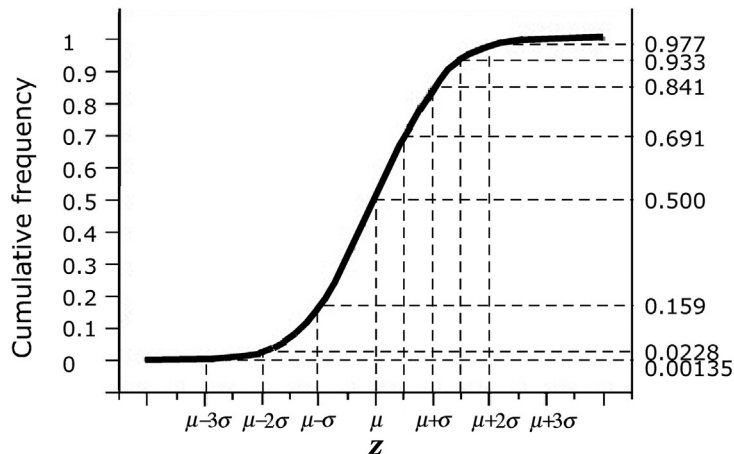


Fig. 6.8 Cumulative probability density curve (frequency curve). Because of the changing slope of the curve, an increase in z from μ to $\mu + 1\sigma$ changes the cumulative frequency from 0.500 to 0.691 for a difference of 0.191, whereas a one standard deviation increase from $\mu + 2\sigma$ to $\mu + 3\sigma$ changes the cumulative frequency from 0.977 to 0.99865 (not shown) for a difference of 0.0265. These cumulative frequencies can be obtained easily online. Cumulative distributions when transformed into the standard deviation scale on the X-axis are termed normal equivalent deviates or NEDs.

POPULATIONS AND SAMPLES

Initially, attention was paid to the sampling features of the mean of large samples; the sample standard deviation and the population standard deviation were assumed to be virtually identical. Gosset realized that some correction was required if inferences about small samples were to be made from the normal Gaussian curve. He introduced a statistic termed t , similar to z but with one important difference. Whereas for means

$$z = \frac{\bar{X}_i - \mu}{\sigma_{\bar{X}}}, \text{ where } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} \text{ and } \sigma = \frac{\sum (X_i - \mu)^2}{N},$$

Gosset used an analogous expression

$$t = \frac{\bar{X}_i - \mu}{s_{\bar{X}}}, \text{ where } s_{\bar{X}} = \frac{s}{\sqrt{N}} \text{ and } s = \frac{\sum (X_i - \bar{X})^2}{N - 1}.$$

There were two added features to his “Student’s” t distribution. One was that $N - 1$ is a specific example of $N - k$, where k is the number of degrees of freedom, and second that the areas under the normal curve varied with the degrees of freedom. For large sample size, >200 , the t and z distributions were identical, but for a sample of 11 with 10 degrees of freedom 95% of the area under the normal curve lies within the limits of $\mu \pm 2.228s_{\bar{X}}$, not $\pm 1.96\sigma_{\bar{X}}$ as for the z table.

DESCRIPTION OF THE DISTRIBUTION SHAPE

The mean μ gives the average size of the measurements, a measure of central tendency, and the square root of the variance—the standard deviation σ —is a measure of variability. Unfortunately, most real distributions are not exactly normal and may even be very far removed from it. One common form of nonnormality is to have a few very large measurements, with the effect shown diagrammatically in Fig. 6.9.

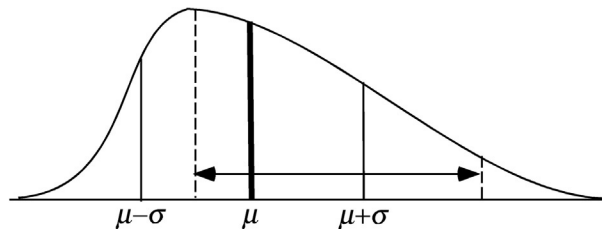


Fig. 6.9 The dotted lines and double-headed arrow indicate the range within which two-thirds of the measurements lie.

The curve is no longer symmetrical but is pulled to the right toward the larger measurements; it is *skewed* to the right. The mean is no longer near the most frequent measurement but to the right of it. More importantly, the range from one standard deviation below to one standard deviation above the mean is not symmetrical and does not include a known proportion of the measurements (as shown by the arrowheads in the figure). For some distributions with extreme skewing, the mean gives no useful information.

Many other types of nonnormality exist. One might guess that the distribution of serum electrolyte concentrations in healthy people approximates a normal distribution, but Elveback et al. (1970) and Elveback (1972) showed that the distributions of several commonly obtained laboratory biochemical values are not normal in the Gaussian sense. Furthermore, they pointed out that to calculate mean and standard deviations from the data in the hope that the 2.5% with the highest values and the 2.5% with the lowest values could be declared abnormal (and therefore unhealthy) would lead to serious underdiagnosis of illness. In their studies the distributions of commonly determined biochemical values were often leptokurtotic (excessively peaked), with >68% of the area between the limits of $\mu - \sigma$ and $\mu + \sigma$ and with excessively long tails (Fig. 6.10). The standard deviation of the leptokurtotic distribution is wider than for a normal distribution, and this can have serious clinical consequences. For example, if the (leptokurtotic) distribution of serum calcium is regarded as normal, the “normal” limits are too wide, and the upper critical normal value based on the calculated standard deviation would lead investigators to miss about 20% of patients with hyperparathyroidism. The problems of using the “normal range” to make clinical diagnoses have been discussed many times (Murphy and Abbey, 1967; Mainland, 1971). The International Federation of Clinical Chemistry recommends the term “reference range” rather than the poorly defined term “normal range” (Strike, 1981).

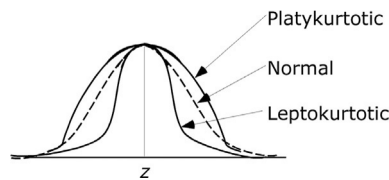


Fig. 6.10 Normal, Leptokurtotic (for long-tailed) and Platykurtotic (for flat or plateau top) curves.

Other symmetrical curves are short-tailed and flat topped (platykurtotic) so that fewer than 68% of the measurements are between the limits of $\mu - \sigma$ and $\mu + \sigma$ (Fig. 6.10).

One approach to defining shape is by computing *moments*. The first moment about the mean is $\frac{\sum(X_i - \mu)}{N}$, the average value of the deviations from the mean, and this is zero. The second moment about the mean is $\frac{\sum(X_i - \mu)^2}{N}$, the average of the squared

deviations of X_i from the population mean, that is, the population variance. The third moment about the mean is $\frac{\sum(X_i - \mu)^3}{N}$, the average value of the sum of the cubed deviations from the mean; it is designated as m_3 or k_3 . The fourth moment about the mean is $\frac{\sum(X_i - \mu)^4}{N}$, the average value of the sum of the deviations about the mean to the fourth power; it is designated as m_4 or k_4 .

Skewness

With perfect symmetry, the third moment is zero, and if the distribution is nearly symmetrical the cubed numerator yields a small number because pluses and minuses almost cancel out. If the curve is skewed to the left, there will be more negative than positive numbers and k_3 will be negative. Conversely, skewing to the right yields a positive result. Because this expression has cubed units, it is customary to divide by the cube of the standard deviation to obtain a dimensionless measurement termed γ_1 : $\gamma_1 = \frac{k_3}{\sigma^3}$. With sample data, the expression becomes $g_1 = \frac{k_3}{s^3}$.

Standard computer programs perform the calculations. Free online calculators can be found at <https://www.wessa.net/skewkurt.wasp>, <https://www.easycalculation.com/statistics/skewness.php>, and <https://calculator.tutorvista.com/skewness-calculator.html>, and they can be calculated by the Skew function in Excel. The program gives the probability that $g_1 = 0$, or tables of g_1 can be consulted (see <http://mvpprograms.com/help/mvpstats/distributions/SkewnessCriticalValues> or <http://www.englunt.edu/~leubank/researchmethods/appendicesa&b.html>) to indicate if a positive or negative value of g_1 is large enough that the hypothesis that $g_1 = 0$ can be rejected.

Kurtosis

If the distribution is symmetrical, then kurtosis (degree of peaking) can be assessed with the fourth moment about the mean to calculate k_4 or m_4 , and this can be made dimensionless and standardized to the standard deviation by $\gamma_2 = \frac{k_4}{\sigma^4}$: the equivalent sample values are $g_2 = \frac{k_4}{s^4}$. γ_2 for a normal curve should be 3. It is customary to subtract 3 from the value of g_2 ; if $g_2 - 3$ is significantly below zero (low) then the curve is platykurtotic (low peak) and if it is significantly above zero (high) then it is leptokurtotic (high peak). If either of these distorted normal curves is encountered, it is worth considering why they have occurred. A leptokurtotic curve might indicate the superimposition of two normal curves with the same mean but different standard deviations, and a platykurtotic curve might indicate the superimposition of two normal curves with similar standard deviations but different means. Platykurtosis often occurs when batches of data are collected at different times, with

a slight shift in mean from one batch to the other. A free online calculator can be found at http://www.wessa.net/rwasp_skewness_kurtosis.wasp#output or <http://www.calculatorsoup.com/calculators/statistics/descriptivestatistics.php>. Kurtosis can also be calculated by the KURT function in Excel.

A rough assessment of kurtosis can be made with the ratio interquartile distance/1.35 (pseudostandard deviation or PSD). If the standard deviation \ll PSD (low), the distribution is platykurtotic, and if the standard deviation \gg PSD (high), the distribution is leptokurtotic (Hamilton, 1990).

Problem 6.2 Take the data from Table 4.17 and test it for skewness and kurtosis. Also calculate the pseudostandard deviation and decide if the curve is leptokurtotic or platykurtotic.

Small sample sizes produce wide confidence limits.

DETERMINING NORMALITY

Inspect the histogram, stem-and-leaf diagram or box plots for asymmetry and outliers. Now that these graphics are incorporated into standard programs, there is no excuse for not determining if a set of observations appears to be normal.

Skewing is easy to detect, but symmetrical distributions with straggling of the highest and lowest measurements are more difficult to judge by eye. The PSD described before is an easy way of assessing this. In addition, the mean $\pm 0.25s$ should include about 40% of the measurements. If many more are included, this suggests that the upper part of the curve is narrower than it should be from a Gaussian distribution, Straggling tail values can grossly distort the standard deviation and make comparisons between groups inefficient.

Some calculations yield a number that can be used to determine the likelihood that a given data set could represent a normal distribution. Shapiro and Wilk's test, Lilliefors test, and D'Agostino's test are available in most statistical programs. (See <http://scistatcalc.blogspot.com/2013/10/shapiro-wilk-test-calculator.html> or <http://sdittami.altervista.org/shapirotest/ShapiroTest.html> for the Shapiro-Wilk's test, <http://www.wessa.net/test.wasp> for D'Agostino's test, and <http://in-silico.net/statistics/lillieforrestest> or <http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Normality.htm> for the Lilliefors test.) These tests are easy to do and interpret but may not indicate where the distribution has departed from normality and thus may not indicate what to do about the problem.

A test for both skewness and kurtosis combined is the Jarque-Bera test. The test value

JB is calculated from
$$JB = \frac{N}{6} \left(s^2 + \frac{(k-3)^2}{4} \right),$$
 where s is the sample skewness and k is the

sample kurtosis. The statistic JB has an asymptotic chi-square distribution with 2 degrees of freedom (Chapter 7) and tests the assumption that the data come from a normal distribution. It can be performed online at http://www.wessa.net/rwasp_skewness_kurtosis.wasp#output, or <https://www.easycalculation.com/statistics/jarque-bera-test-calculator.php>

Graphic tests. Some tests are graphic: for example, the use of probability paper, or normal quantile plots. In Fig. 6.8, the typical sigmoid cumulative frequency curve was shown. If data points could be plotted on such a curve with an excellent fit it is reasonable to conclude that the distribution from which those points came was normal. It is, however, difficult to assess S-shaped curves and easier to assess straight lines. Fortunately, the S-shaped curve can be made straight by plotting its points on probability paper (Fig. 6.11).

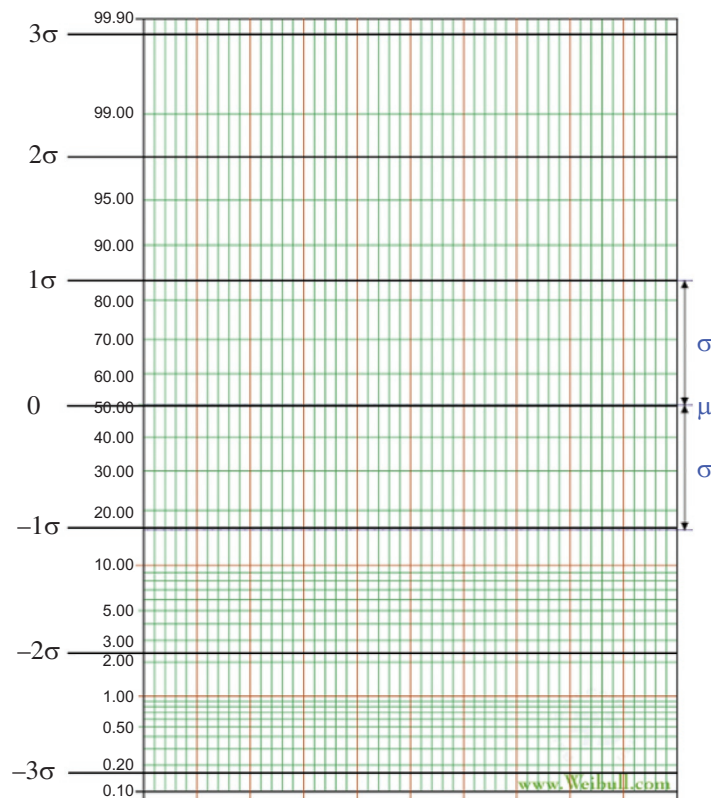


Fig. 6.11 Normal probability paper. The X-axis is linear, but the Y-axis becomes magnified as one goes further away from the mean. The *thick horizontal lines* are placed 1, 2, and 3 standard deviations above and below the mean, as defined by the areas under the *normal curve*. Graphs for this and other distributions may be obtained free online from the ReliaSoft Corporation at <http://www.weibull.com/GPaper/>, <https://incompetech.com/graphpaper/logarithmic/>, or <http://www.printfreegraphpaper.com>.

There are other ways of plotting and assessing normal distributions. Normal quantile plots are standard on most computer programs. Fig. 6.8 showed how for equal standard deviation increments the incremental change in the cumulative percentages became progressively less as the distance from the mean increased. This was rectified in the probability paper featured in Fig. 6.11 where the heavy horizontal lines demarcated standard deviation units rather than cumulative percentages. The vertical Y-axis is linear in standard deviation units, sometimes called standard normal deviates or normal equivalent deviates (NED). [The transformation of the S-shaped curve to the linear NED scale is attributed to the pharmacologist J.H. Gaddum (1900–65).] The cumulative percentages corresponding to various z values are presented in Table 6.1.

Table 6.1 Normal equivalent deviates

Cumulative percentage	Normal equivalent deviate
0.00135	−3
0.02775	−2
0.1589	−1
0.5	0
0.6915	0.5
0.8413	1
0.933	1.5
0.977	2
0.9986	3

The NED is the same as the area defined as $1 - z = \phi$ in Fig. 6.4.

With this information plot the data on probability graph paper. The X-axis is the measurement of the variable, and on the Y-axis plot the observed cumulative percentages of the X variable. As an example, Table 6.2 gives data on the distribution of heights of 18th century English soldiers in America (Komlos and Cinnirella, 2005). The resultant frequency and cumulative frequency distributions are shown in Fig. 6.12.

The mean of grouped data can be calculated using <https://www.easycalculation.com/statistics/group-arithmetic-mean.php>, and the mean and standard deviation from <http://www.ltcconline.net/greenl/java/Statistics/FrequencyData/FrequencyData.html>.

One way of testing the normality of the distribution is to plot the cumulative percentage against the height on probability paper (Fig. 6.13).

Another method that does not involve probability paper is to plot the NEDs on the Y-axis and the height on the X-axis. This involves calculating the cumulative percentage, transforming these values into NEDs (as presented in Table 6.1), from detailed tables, from Fig. 6.6, or from online calculators <http://stattrek.com/online-calculator/normal.aspx>, or http://davidmlane.com/hyperstat/z_table.html (Fig. 6.14). The calculators are more accurate but the results are similar.

Table 6.2 Distribution of heights

Height (in.)	Frequency	Cumulative frequency	% Cumulative frequency	NED
59	10	10	0.99	-2.330
60	14	24	2.15	-2.024
61	36	60	5.38	-1.609
62	50	110	9.87	-1.289
63	98	208	18.65	-0.891
64	172	380	34.08	-0.410
65	174	554	49.69	-0.008
66	184	738	66.19	0.418
67	119	857	76.86	0.734
68	127	984	88.25	1.188
69	60	1044	93.63	1.524
70	31	1075	96.41	1.800
71	22	1097	98.39	2.142
72	14	1111	99.64	2.687
73	4	1115	100.00	
Mean = 65.6 SD = 2.54				

Discussion of columns in text later.

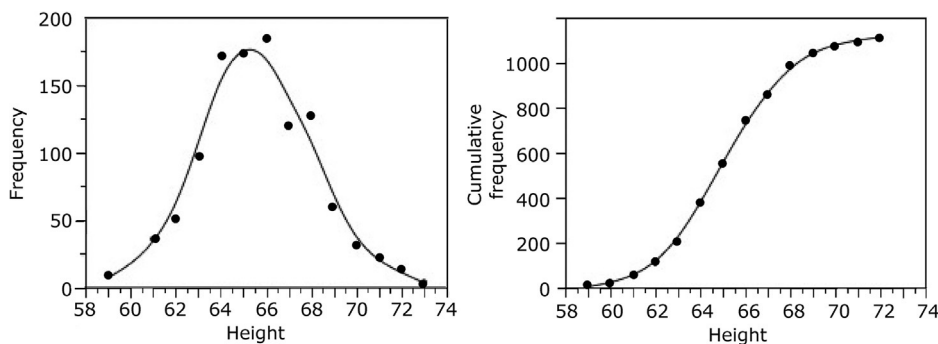


Fig. 6.12 Frequency and cumulative frequency distributions of heights. The cumulative curve evens out the irregularities shown in the original distribution.

The Q-Q plot is a way to compare two distributions, most often an observed distribution and the normal distribution. The values in the sample of data, in order from smallest to largest, are denoted $x(1)$, $x(2)$, ..., $x(n)$. For $i = 1, 2, \dots, n$, $x(i)$ is plotted as $(i - 0.5)/n$ on the horizontal X-axis. For normality, the normal distribution is used in a special way and plotted on the vertical Y-axis. Details for doing this by hand are given by David Scott at http://onlinestatbook.com/2/advanced_graphs/q-q_plots.html, and QQ plots can be done online at https://www.wessa.net/rwasp_varia1.wasp, <http://scistatcalc.blogspot.com/2013/11/q-q-plotter-for-gaussian-distribution.html> and <http://facweb.cs.depaul.edu/cmiller/it223/normQuant.html>.

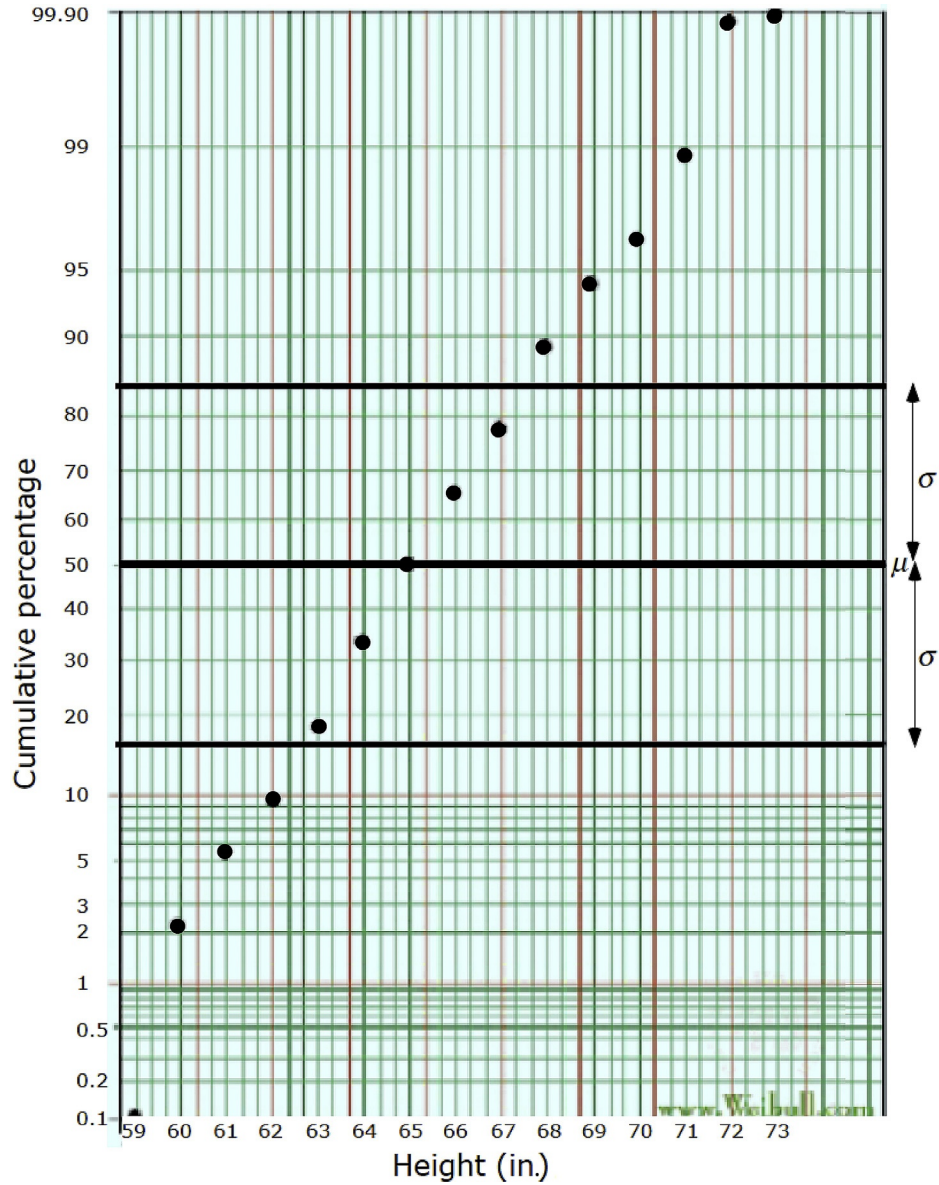


Fig. 6.13 Probability paper showing cumulative heights. Apart from the ends, which are based on very small numbers, the distribution is reasonably linear. This suggests that the distribution is approximately normal, although it cannot be truly normal because it is truncated at each end.

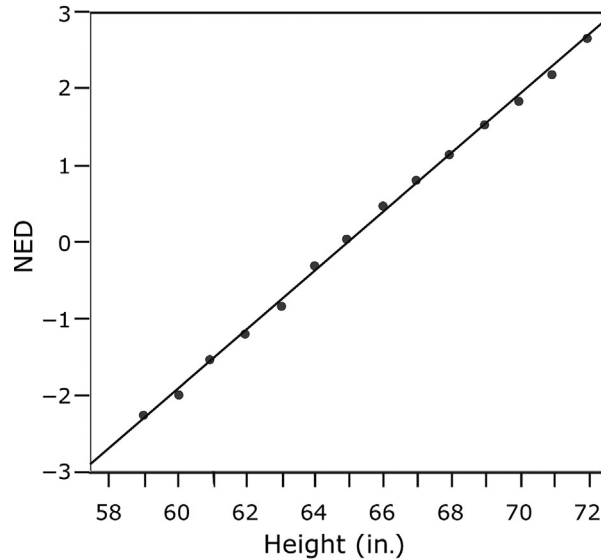


Fig. 6.14 Using NED to determine normality on ordinary graph paper. This is the equivalent of using probability paper. It is similar to the quantile-quantile (Q-Q) plot (see later).

UNGROUPEd DATA

Table 6.3 presents the weight of 13 dogs. For ungrouped data, the individual frequencies are 1,1,1,...,1; the cumulative frequencies are 1,2,3,...,N; and the relative cumulative frequencies are $1/N$, $2/N$,..., N/N . Because in Table 6.3 $N = 13$, the first relative cumulative frequency is $1/13 = 0.077$, the second is $2/13 = 0.154$, and so on.

Table 6.3 Dog data

Number	Weight (LK)	Cum rel f	NED
1	17.2	0.077	-1.426
2	20.8	0.154	-1.019
3	21.0	0.231	-0.736
4	21.2	0.308	-0.502
5	21.5	0.385	-0.292
6	24.2	0.462	-0.095
7	24.3	0.539	0.098
8	25.6	0.616	0.295
9	27.7	0.693	0.504
10	31.0	0.77	0.739
11	34.5	0.857	1.067
12	38.7	0.924	1.433
13	39.2	1.00	-
$\sum X_i = 346.9$ $\bar{X} = 26.68$ $s = 7.12$			

Plotting the NED in column 4 against the actual values in column 2 gives Fig. 6.15.

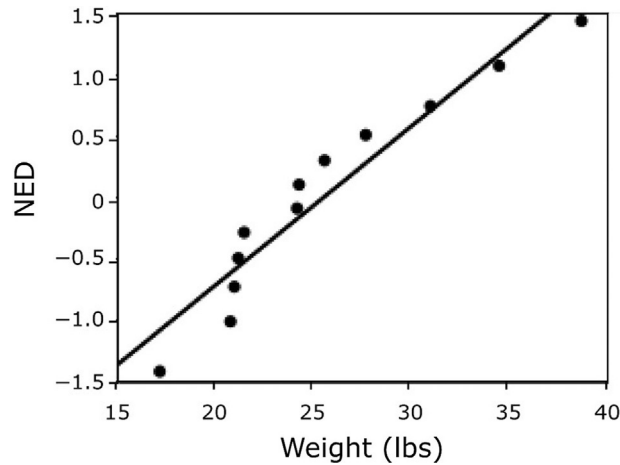


Fig. 6.15 Dog data in NED plot.

This is not as good a fit as shown in Fig. 6.14, and some points are quite far from the theoretical line of normality. To determine if these data are consistent with a normal distribution enter them into the programs <http://stattrek.com/online-calculator/normal.aspx> or http://www.wessa.net/rwasp_skewness_kurtosis.wasp#output. These show that skewness and kurtosis are not abnormal enough to allow rejection of the null hypothesis.

To determine if these data are still compatible with a normal distribution, some programs, insert 95% confidence limits (Fig. 6.16). Quantile plots can be implemented

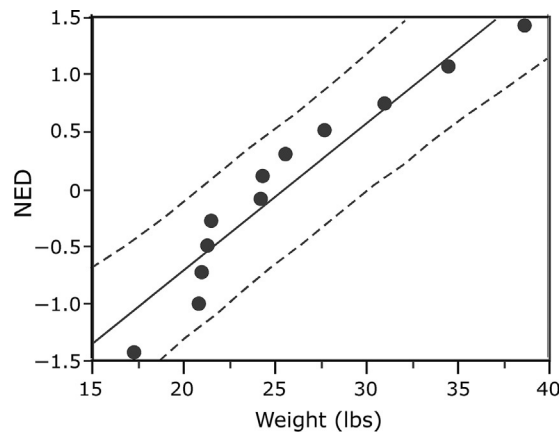


Fig. 6.16 The *diagonal line* indicates perfect normality. The vertical scale shows NEDs, and the horizontal scale shows the weight. The *dashed outer lines* show the 95% confidence limits.

online by http://www.wessa.net/rwasp_harrell_davis.wasp#output, but without confidence limits. They can also be produced in Excel at <http://facweb.cs.depaul.edu/cmiller/it223/normQuant.html>. Confidence limits can be produced by entering the data into a regression plot (see Chapter 27).

If the distribution had been perfectly normal, then all the points would have been on the line. These plots show exactly where the deviations from normality occur and give a probability that can be used to decide if the distribution is sufficiently far from a normal distribution.

Problem 6.3 Draw a quantile plot of the data from Table 4.2.

HOW IMPORTANT IS NORMALITY?

The normal distribution curve plays a central part in statistical thinking and modeling. If the distribution is markedly abnormal, then even though a test statistic can be calculated, inferences drawn from it may be wrong. Furthermore, because statistics based on the normal distribution are very efficient, try to use normalizing transformations so that conventional statistics can be used. Alternatively there are tests to use when distributions are markedly abnormal; these are called nonparametric tests.

How necessary is it to test normality and how much abnormality of the distribution can be tolerated? If a stem and leaf diagram, a box plot, or a histogram appears roughly normal and symmetrical by eye, then, as long as there are no extreme outliers, the distribution is normal enough, and more elaborate tests may not be needed. This attitude is supported by an article by Sall (with the appropriate title “Leptokurtophobia: irrational fear of non-normality”) that appeared in the technical publication for JMP users, JMPer Cable (Sall, 2004). He pointed out that: “In large samples it is easy to detect non-normality, but it doesn’t matter. In small samples, non-normality may matter, but you can’t detect it.” Sall concluded, however, that graphical testing, even if of limited use for detecting nonnormality, was of value looking for anomalies or a pattern that might be a clue to some hidden structure of the distribution.

REFERENCES

- Cole, T.J., 2010. Babies, bottles, breasts: Is the WHO growth standard relevant? *Significance*, Virtual Medical Issue, 6–10.
- Elveback, L., 1972. A discussion of some estimation problems encountered in establishing “normal” values. In: Gabrieli, E.R. (Ed.), *Clinically Oriented Documentation of Laboratory Data*. Academic Press, New York.
- Elveback, L.R., Guillier, C.L., Keating Jr., F.R., 1970. Health, normality, and the ghost of Gauss. *J. Am. Med. Assoc.* 211, 69–75.

- Hamilton, L.C., 1990. *Modern Data Analysis. A First Course in Applied Statistics*. Brooks/Cole Publishing Co, Pacific Grove, CA.
- Komlos, J., Cinnirella, F., 2005. European Heights in the Early 18th Century. Available: https://epub.ub.uni-muenchen.de/572/1/european_heights_in_the_early_18th_century.pdf.
- Mainland, D., 1971. Remarks on clinical “norms” *Clin. Chem.* 17, 267–274.
- Murphy, E.A., Abbey, H., 1967. The normal range—a common misuse. *J. Chronic Dis.* 20, 79–88.
- Sall, J., 2004. Leptokurtophobia: irrational fear of non-normality. *JMPer Cable.* 15.
- Strike, P.W., 1981. *Medical Laboratory Statistics*. John Wright and Sons, Ltd., Bristol.

CHAPTER 7

Statistical Inference: Confidence Limits and the Central Limit Theorem

CENTRAL LIMIT THEOREM

In a 1966 publication, the mean height of 24,404 males in the US Army was reported as 68.4 in., with a standard deviation of 2.5 in. (Newman and White, 1966). These figures are valuable for understanding biology and, more practically, for deciding what sizes to make uniforms! The question is whether the sample mean of 68.4" is a good or a bad estimate of the theoretical population mean μ .

Imagine a theoretical population of heights of young adult males. In a sample S_1 of N subjects, the mean height might be 68.1". This is one estimate of the population mean, μ . A second sample S_2 of N subjects might have a mean value of 69.4". Which is the better estimate of μ ? One way to answer this question would be to draw thousands of samples, each with N subjects, at random from the population. For each sample calculate the mean, and end up with thousands of mean heights taken from samples of size N . These means are themselves numbers, and the mean of all these means is the grand mean. With enough samples the grand mean will be very close to the population mean and can be termed μ with minimal inaccuracy. The standard deviation of these means can be calculated and will closely approximate σ , the population standard deviation of the mean. Because this standard deviation is not of individual measurements but of means, it is symbolized by a subscript: $\sigma_{\bar{x}}$ and called the standard deviation of the mean. This is sometimes termed the standard error of the mean.

An important theorem in Statistics is the Central Limit Theorem. The theorem states that even if a distribution from which samples are drawn is not normal, the means of samples drawn at random from this population will be normally distributed (Fig. 7.1A).

Another example that shows how the mechanism of these curves comes from drawing random samples from the skewed distribution is shown in Fig. 7.1B.

Even for extreme distortions of the sample frequency curves, the sampling distribution of the mean is normal for sample sizes over 30, and for smaller sample sizes if the basic frequency distribution is closer to normal.

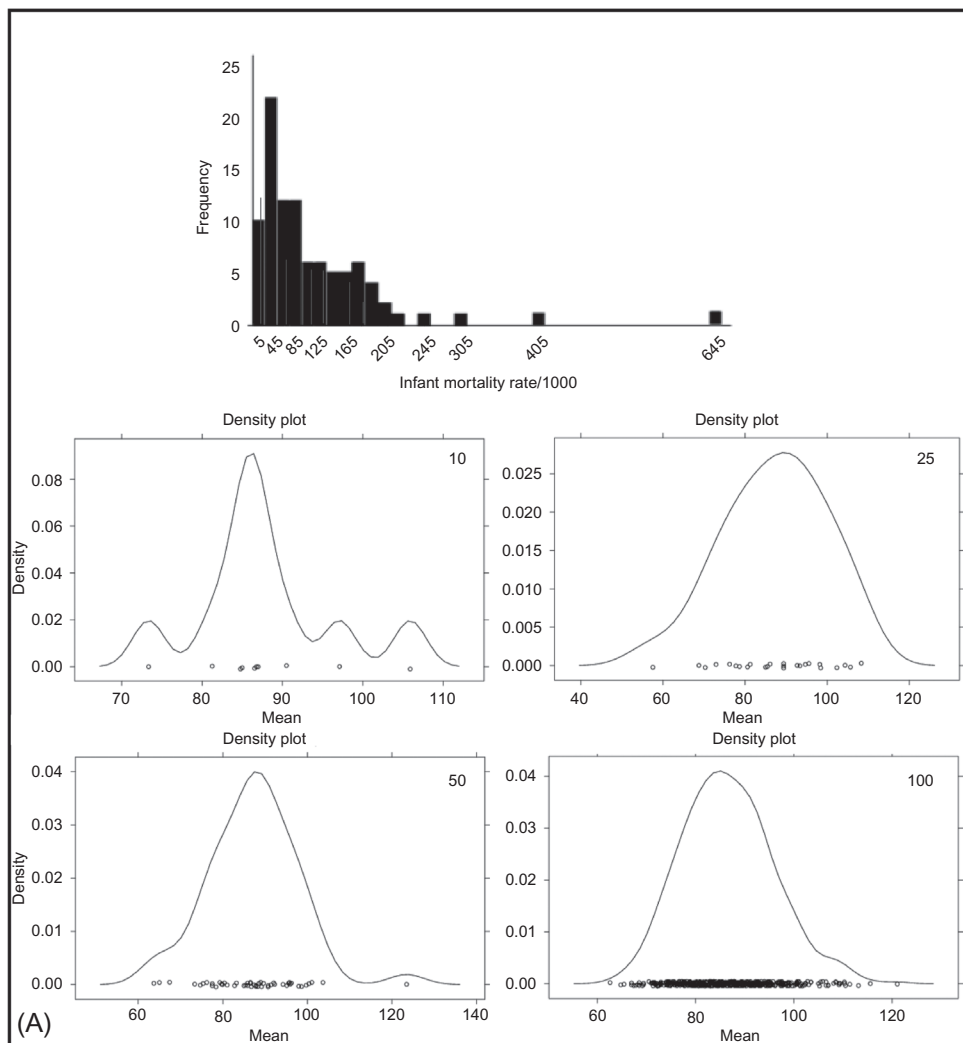


Fig. 7.1 (A) Upper panel shows a grossly abnormal distribution. Below it are the distributions of means of repeated samples from this distribution, the numbers in the upper right-hand corners showing how many samples were taken. Despite the abnormal distribution, even 10 samples have a roughly normal distribution of means, and this distribution becomes more symmetrical as the sample size increases.

(Continued)

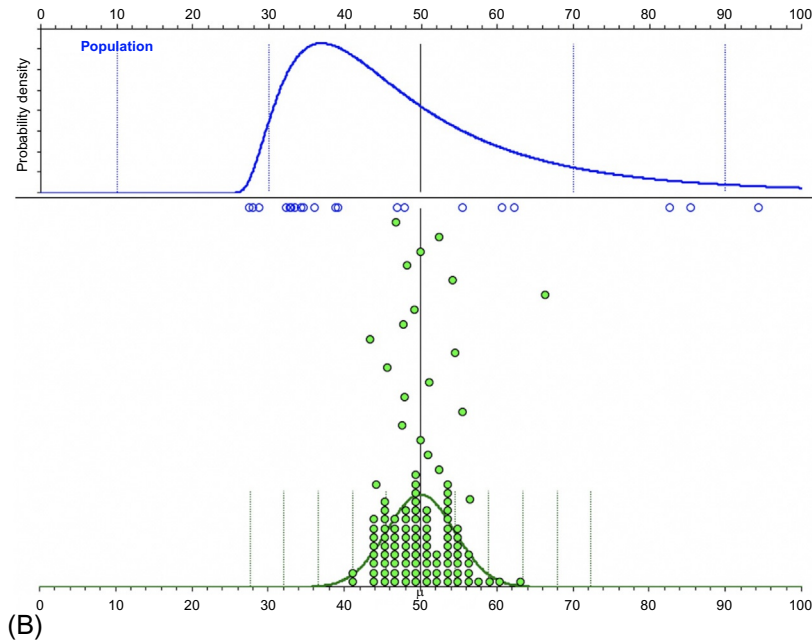


Fig. 7.1—Cont'd (B) Skewed distribution. The row of *circles* below the curve shows one sample. The *vertical dots* show the means of successive samples. The pile of *dots* at the bottom is the accumulation of the individual means which are fitted to a normal Gaussian curve. Vertical lines on upper curve are standard deviations and on lower curve are standard errors. Graph constructed with Exploratory Software for Confidence Intervals (ESCI) created by Cumming and available at <http://thenewstatistics.com/itns/esci/>.

Therefore the sample of very large numbers of means of heights should be normally distributed, and it is possible to draw a normal curve (of mean heights) (Fig. 7.2).

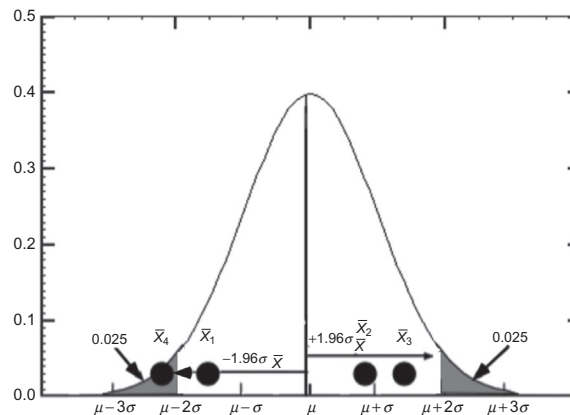


Fig. 7.2 Normal Gaussian curve of sample means from samples of size N with population mean μ and standard deviation $\sigma_{\bar{x}}$. The *shaded area* in the tails each include 2.5% of the area under the curve, that is, the area more than $1.96\sigma_{\bar{x}}$ above and below the mean. The *unshaded area* within the limits $\mu - 1.96\sigma_{\bar{x}}$ and $\mu + 1.96\sigma_{\bar{x}}$ (indicated by the two offset horizontal lines and arrow heads) includes 95% of the area under the curve.

In this example, the large black dots indicate the means of 4 samples, each of size N , drawn at random from this population. The population mean μ is approximated from the grand mean of the thousands of samples and $\sigma_{\bar{X}}$ from those thousands of means. Because this experiment is a thought experiment and we have actually not drawn these thousands of samples, we do not know what μ and $\sigma_{\bar{X}}$ are. We can, however, argue that 95% of these sample means (e.g., \bar{X}_1, \bar{X}_2 , and \bar{X}_3) fall within the limits of $\mu \pm 1.96\sigma_{\bar{X}}$, and only 5% of them (e.g., \bar{X}_4) fall outside those limits. In other words, from the properties of the normal curve, the probability is 2.5% that a given mean is more than $1.96\sigma_{\bar{X}}$ above μ , 2.5% that it is more than $1.96\sigma_{\bar{X}}$ below μ , and 95% that it is between these limits.

Now if there is 5% probability that \bar{X}_i is more than $\pm 1.96\sigma_{\bar{X}}$ from μ , there is the same 5% probability that μ is $\pm 1.96\sigma_{\bar{X}}$ from \bar{X}_i (Fig. 7.3A–D).

In samples 1 (Fig. 7.3A), 2 (Fig. 7.3B), or 3 (Fig. 7.3C), the value of μ lies between the limits demarcated by the arrowheads. On the other hand, occasionally the limits set by $\bar{X} \pm 1.96\sigma_{\bar{X}}$ do not include μ (Fig. 7.3D). However, this event should occur $<2.5\%$ of the time or $<5\%$ of the time if the excessive deviation could be above or below the mean.

To recapitulate, if we draw thousands of samples of size N from a population with a mean of μ and a standard deviation of the mean of σ , then 95% of the means of those samples will be within $\mu \pm 1.96\sigma_{\bar{X}}$, so that there is a 95% probability that the

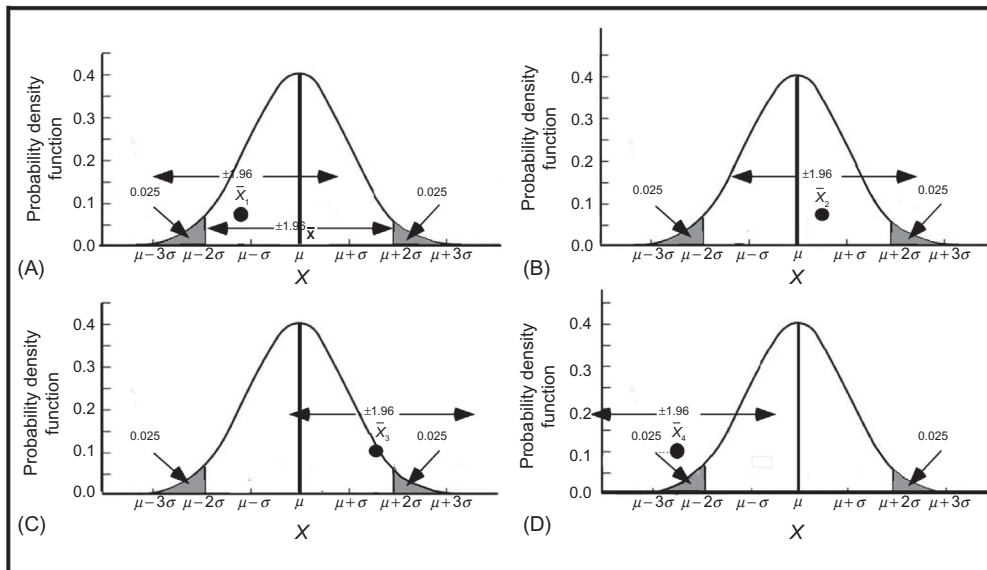


Fig. 7.3 Reverse inference. See text.

unknown value of μ will be within those limits. Only 2.5% of the time will a sample mean be too small or too big for this prediction to be correct. Therefore a single sample of size N and its mean allows us to predict that the unknown value of μ would lie between $\bar{X}_1 \pm 1.96\sigma_{\bar{X}}$ 95% of the time. These limits are termed confidence limits.

There is still one problem if we wish to avoid drawing thousands of samples. The confidence limits described before are calculated using $\sigma_{\bar{X}}$, and this was obtained from the thousands of samples. Fortunately, there is an easy way of predicting $\sigma_{\bar{X}}$ from the single sample standard deviation:

$$\sigma_{\bar{X}} \approx \frac{s}{\sqrt{N}},$$

where s is the sample standard deviation and N is the sample size.

Therefore all that is needed is a single sample, calculate its mean and standard deviation, estimate the standard deviation of the mean as $\frac{s}{\sqrt{N}}$, and multiply this value by 1.96 to obtain

the 95% confidence limits about the mean. (Remember that s itself has variability from sample to sample, so that it is important that the sample be representative of the population. If the sample has outliers they may inflate s . An estimate of the confidence limits of the variance (s^2) may be determined from http://www.wessa.net/rwasp_hypothesisvariance4.waspori, <http://www.wolframalpha.com/widgets/view.jsp?id=5327a43b4c8b784329da9e81caee6987>, and <https://home.ubalt.edu/ntsbarsh/business-stat/otherapplets/Esteem.htm>. For example, with $N=20$ and $s=20$, the 95% CL of s are about 15–29.)

GENERALIZATIONS BASED ON A SINGLE SAMPLE

Our sample of measurements allows us to determine point estimates of central tendency and variation, but what use are these estimates? They are unlikely to be repeated in other samples drawn at random from the same population, but nevertheless are useful in making hypotheses about the parameters of the populations from which the samples were drawn. This is the beginning of scientific inference that often begins with determining confidence (or interval) estimates.

Setting confidence limits

An interval estimate is a range that we believe will include the parameter being estimated with a given degree of confidence. The two ends of the range are called the confidence

limits for that parameter. The confidence interval is the range between the upper and lower confidence limits; the terms are used interchangeably.

Confidence limits may be set for any given level of significance (Chapter 10). They are often specified as $(1 - \alpha)$ or $(100 - \alpha)\%$ where α is the significance level. Usually 0.95 (95%) limits are determined, corresponding to a level α of 0.05 (5%), but at times other confidence limits are more useful. (α is conditional on the null hypothesis being true.)

The formula for confidence limits (CL) may be written as

$$(1 - \alpha)\text{CL} = \bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{N}}$$

because $z_{\alpha/2}$ gives the area beneath one tail of the curve, and twice that value is α . In Fig. 7.2, the lines demarcating the unshaded area show the confidence limits.

Limits are not restricted to 95% confidence. In some data such as survival after cardiac surgery, 70% limits have been used, even though they are wide. Alternatively, 99% confidence limits are derived by multiplying $S_{\bar{X}}$ by 2.588. This allows more certainty about the limits within which μ lies, but the disadvantage is that the limits have become wider. Returning to the population mean for height, by using a big enough value for z we could even be 99.9999999999% sure that the limits are between 20" and 120", but those limits are so wide as to be useless.

For estimating results from small samples (<100), use an adjusted value of z that is called t . Confidence limits for the mean are then determined by multiplying the standard error by the appropriate value selected from the t table, and then subtracting that product from and adding it to the mean (Gardner and Altman, 1995). Thus if the mean of a series of 10 measurements is 57.5 and the standard deviation is 12.57, then the standard error of the mean is $\frac{12.57}{\sqrt{10}} = 3.98$. $t_{0.05}$ for 9 degrees of freedom is 2.262. Therefore the 95% confidence limits for the mean are $57.5 \pm 2.262 \times 3.98 = 57.5 \pm 9.00 = 48.5$ to 66.5. These limits are symmetrical about the observed mean. If the distribution is not normal, then the limits are symmetrical about the mean but will be wider because of the larger standard deviations. Confidence limits can be calculated online at <https://easycalculation.com/statistics/confidence-limits-mean.php>, <http://ncalculators.com/statistics/confidence-interval-calculator.htm>, <http://www.danielsoper.com/statcalc/calculator.aspx?id=96>, and <https://www.mathsisfun.com/data/confidence-interval-calculator.html>. Extensive manipulations of confidence intervals can be done with Exploratory Software for Confidence Intervals (ESCI) for use by Apple or PC, accessed at the web site for Cumming's book "The New Statistics: Estimation for better research" at <http://thenewstatistics.com/itns/esci/>.

Many journals recommend that confidence intervals be placed within square brackets after the mean; 6 [3,9] For their first use the percentage limits—95%, 90%, and so on—should be specified, but may be omitted thereafter.

Problem 7.1. A data set has mean 93.2 and standard deviation of 13.7.
Set 95% confidence limits for $N=14$ and $N=77$.

Confidence intervals with $N < 12$ are usually wide, and for larger N decrease rapidly (van Belle, 2002). In order to obtain narrow confidence limits that allow efficient comparisons between two mean values, for example, we need a big value for N . What that value needs to be will be discussed in Chapter 11.

If there is concern about outliers and a trimmed mean $\overline{X}_{\text{trim}}$ is used, estimate the standard error of the trimmed mean as

$$S_{\overline{X}_{\text{trim}}} = \frac{1}{1 - 2\gamma} \frac{s_w}{\sqrt{N}}.$$

Here γ is the trimming fraction, usually $0.1N$ or $0.2N$, removed at each end of the distribution, and s_w is the Winsorized standard deviation (Chapter 4). The confidence limits are narrower for any given value of α because eliminating the outliers has reduced the standard error. The reasons for the outliers need careful examination (Chapter 9).

An assumption is that the sample standard deviation represents the population standard deviation, and this is one good reason for inspecting the distribution for abnormalities and unusually large outlying observations that might falsely inflate the estimate of the population standard deviation and the confidence limits. What can we do to avoid this source of error other than inspecting the distribution carefully? One way is to increase the number of observations because this gives a more secure basis for assessing the standard deviation (Fig. 7.4).

As sample size increases the standard deviation changes, sometimes exceeding and sometimes falling below the population standard deviation σ , but the swings about the mean value become smaller with increased N . On the other hand, as N increases the value of $s_{\overline{X}}$ becomes smaller because s is divided by greater values for \sqrt{N} . Because of this square root function, to halve the value of $s_{\overline{X}}$ for a given value of s we must increase sample size fourfold. To make $S_{\overline{X}}$ one-tenth as large we must increase sample size 100-fold. Although having a larger sample size gives a smaller value for $S_{\overline{X}}$ and decreases the confidence interval, the cost in time and money may not justify the increased size.

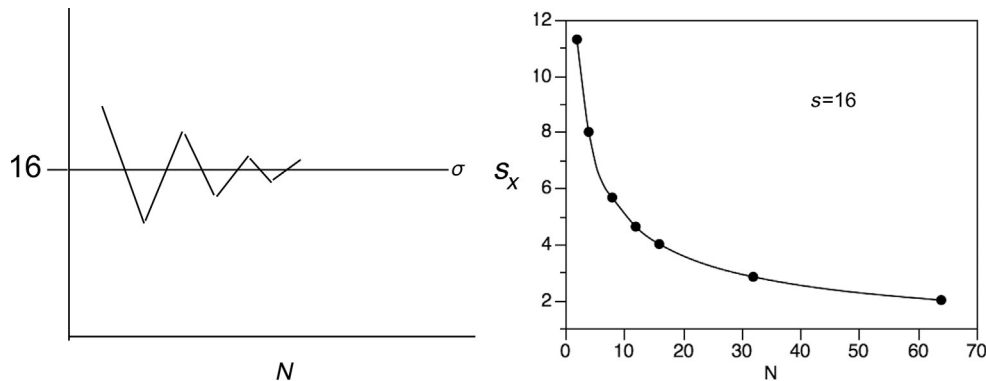


Fig. 7.4 Left panel: effect of increasing sample size on estimated population standard deviation of 16. Right panel: effect of increasing sample size on standard error of the mean.

Confidence limits, as an indicator of the uncertainty of a test, provide only minimum estimates unless the distribution of errors is perfectly normal—an unlikely event. Furthermore, a given confidence interval is specific to a particular sample and will differ for other samples from the same population. Cumming and Maillardet (2006) showed that even though 95% confidence intervals had a 95% probability of including the population mean from which the sample was drawn, there was only an 83% chance that confidence interval would include future replication means. Fig. 7.5 shows 25 random samples from a normal and a skewed population with $\mu = 50$ and $N = 20$ (two upper panels) an $N + 80$ (two lower panels).

Although a single confidence interval may or may not include the true population mean, 95% of the multiple confidence intervals drawn at random from the population will include the population mean. A single confidence interval does not give as much assurance as we usually believe it does.

Confidence limits can be calculated for other statistics such as proportions, slopes, standard deviations, and so on, and will be discussed in the appropriate chapters.

TOLERANCE AND PREDICTION LIMITS

Confidence intervals are determined for means by the formula $\bar{X} \pm t_{\alpha} s^{\alpha} \bar{X}$, and this formula is often incorrectly modified as $\bar{X} \pm t_{\alpha} s^{\alpha}$ to set confidence limits for individual data values. Such limits might be needed to evaluate when, for example, a single value for the concentration of a biological substance (glucose, calcium, etc.) is outside the usual limits and hence should be considered as possibly pathological. Unfortunately, the standard

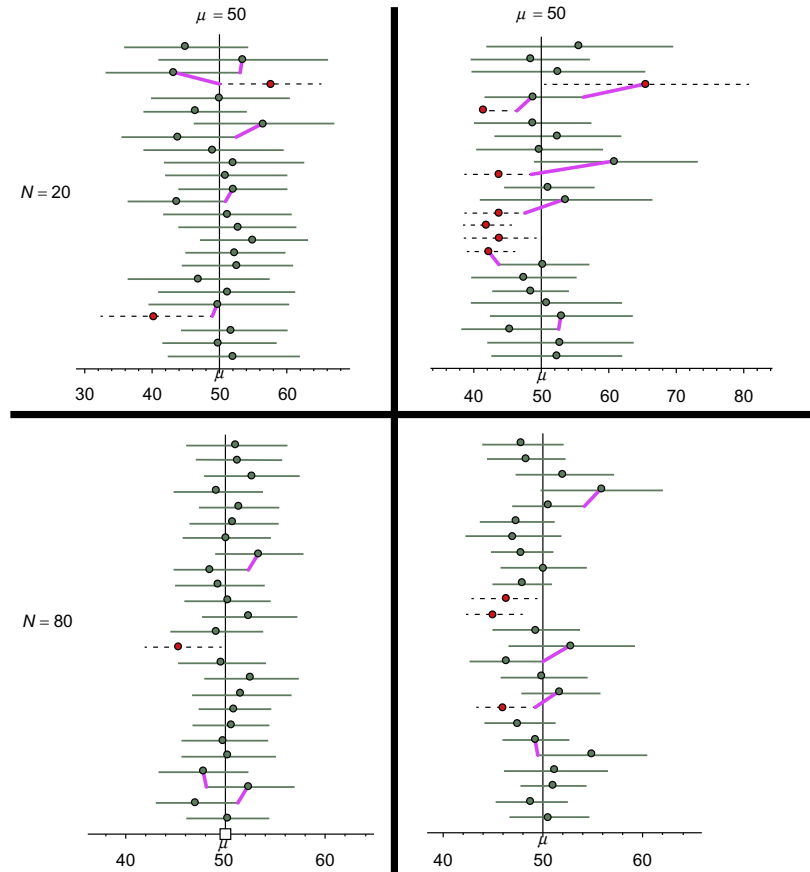


Fig. 7.5 Ninety five percent confidence limits from 25 random samples from a normal distribution (left) and a right-skewed distribution (right); both distributions have a population mean of 50. The *dashed* horizontal confidence limits exclude the known mean. Confidence intervals vary because random samples have different standard deviations. The *oblique lines* show that the next sample (the higher of the pair) has a mean beyond the confidence limits of the previous (lower) sample (no p-capture). These examples show that two investigators may perform the same experiment in identical ways and yet reach different conclusions; a single experiment that has not been repeated and confirmed may be misleading. Increasing N from 20 to 80 (two lower panels) has decreased the confidence intervals but some intervals still do not include μ or capture the next mean. This variability is one of the reasons behind Ioannidis' article entitled "Why most published research findings are false" (Ioannidis, 2005).

formula was developed to examine the distribution of sample means around a population mean by making use of an estimate of the population standard deviation of the mean. When a single data point is to be evaluated from a sample, however, neither the population mean nor the population standard deviation is known (Fig. 7.6).

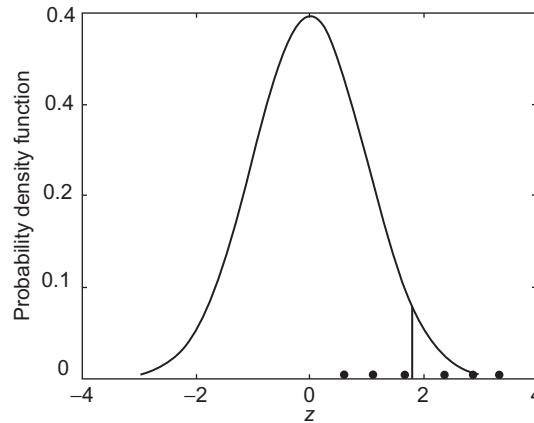


Fig. 7.6 The Gaussian curve shows the distribution of means drawn at random from samples of the same size from a population; the horizontal scale is in standard units, with mean = 0 and standard deviation = 1. The short vertical line at 1.96z indicates the upper 2.5th percentile. As shown, the individual values that make up that percentile are themselves scattered about the percentile, and the upper 2.5th percentile for some points has a z value > 1.96. It is to allow for this added source of variability that prediction and tolerance limits were devised.

The prediction interval is a range that is likely to contain the value of a single new observation from the population. A tolerance interval is a range that is likely to contain a specified proportion of the population.

A prediction interval for a single future observation can be obtained by modifying the formula for the confidence interval ([Ramirez, 2009](#))

$$\bar{X} \pm t_{1-\frac{\alpha}{2}, n-1} s \sqrt{1 + \frac{1}{n}}.$$

Tolerance limits define an interval that includes a designated proportion of the population (p) with a specified level of confidence (γ).

As described by [Hahn and Meeker \(1991\)](#) “A tolerance interval provides limits that one can claim, with a specified degree of confidence (e.g., $\gamma = 90\%$), contains the (measured) diameters of at least a specified proportion (e.g., $p = 95\%$) of units from the sampled population. The two percentages in the preceding statement should not create any confusion when one recognizes that one (95% in the example) refers to the percentage of the population to be contained, and the other (90%) deals with the degree of confidence associated with the claim that the interval encloses (at least) 95% of the population.”

Tolerance intervals can examine three types of questions:

1. What interval will contain p percent of the measurements? (two-sided interval)

This is calculated as $\bar{X} \pm k_2 s$, where k_2 is the two-sided tolerance factor.

2. What interval guarantees that p percent of measurements will not be below a lower limit? (one-sided interval)

The lower limit X_L is calculated as $\bar{X} - k_1 s$, where k_1 is the one-sided tolerance factor.

3. What interval guarantees that p percent of measurements will not exceed an upper limit? (one-sided interval)

The upper limit X_U is calculated as $\bar{X} + k_1 s$, where k_1 is the one-sided tolerance factor.

Some computer programs make these calculations, and both two-sided and one-sided limits can be determined online at <http://statpages.org/tolintvl.html>, or by consulting a Table of one-sided tolerance factors, for example, Table 6 in the book by Goldstein (1964) or Table A7 by Natrella (1963).

Formulas for calculating these limits may be found at <http://www.itl.nist.gov/div898/handbook/prc/section2/prc263.htm>. These limits are accurate only for reasonably normal distributions (Handbook, n.d.).

The distinction between these intervals is discussed simply at <http://blog.minitab.com/blog/adventures-in-statistics-2/when-should-i-use-confidence-intervals-prediction-intervals-and-tolerance-intervals> and <https://www.graphpad.com/support/faqid/1506/>.

Consider the weights of the dogs from Table 6.3. They have mean weight 26.68 kg and standard error 1.97 kg (range 17.2–39.2 kg). The 95% confidence limits of the mean are 22.38 to 30.99 kg. The 95% prediction limits for a single dog drawn at random from this population are 10.58–42.79 kg, considerably wider because individual measurements vary more than means. The tolerance interval with a confidence of 95% that the interval includes 90% of the population is 8.16–45.21. The tolerance and prediction intervals are of similar magnitude and are much greater than the confidence interval. The difference between tolerance, prediction, and confidence limits decreases as sample size decreases, but always persists.

REPORTING RESULTS

Most published reports give the point estimates for mean and standard deviation and should also give the sample size. However, that is not sufficient. At the very least the confidence limits (usually 95%) should be given, and many journals now insist on this because it gives the reader essential information. The presentation of confidence limits

is so important for all point estimates that everyone should read a simple book by Gardner and Altman entitled “Statistics with Confidence” (Gardner and Altman, 1995). Confidence limits should be provided not only because many journals require them, but because they provide added information (that investigators often ignore) about the data (Healy, 1992). Healy (1992) gives the example of a sample mean of 0.22 with a 95% confidence interval of 0.02–0.42. Because the confidence interval does not include zero it implies statistical significance at the 5% level (see Chapter 10), but the lower limit is consistent with an effect that is too small to be of importance. Under these circumstances the investigator would be wise to put less emphasis on the significance of the test and more on the possible magnitude of the effect.

GRAPHIC REPRESENTATION

The question about what to include in a bar graph is not settled, and as pointed out in Chapter 4 the bar graph is an inefficient use of space. Usually in such a graph the mean is a bar with its height proportional to the mean, and a measure of variability is added as a vertical line (Fig. 7.7, right panel). As shown in the figure, a box plot (left panel) presents far more useful data.

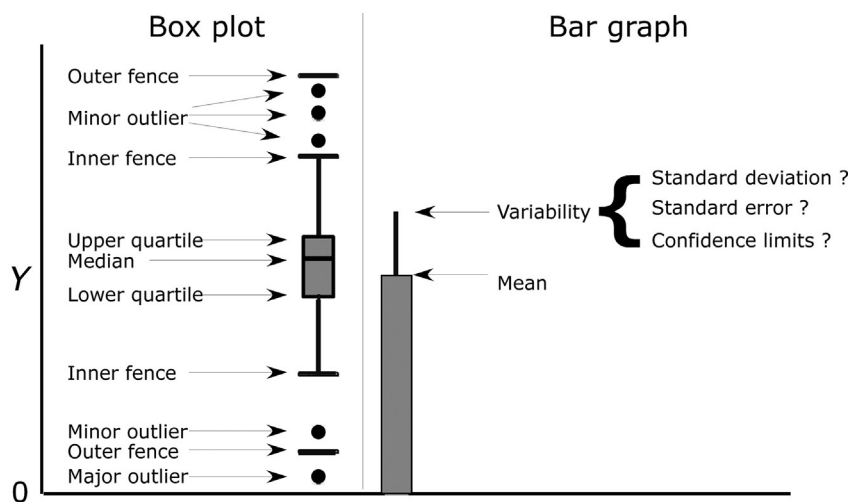


Fig. 7.7 Bar graph on right with height representing the value on the Y-axis. A vertical line extending up from the top of the bar indicates variability, but there is little consistency about whether it represents standard deviation, standard error, or confidence limits. Furthermore, the authors may not make this information available. Box plot (Velleman and Hoaglin, 1981) on left shows vastly more information for about the same space. (Other programs give slightly different box plots.)

The box plot is more informative. Furthermore, the values for the mean and standard deviation are almost always present in the text and do not need to be shown graphically. The subject is discussed at length by [Cumming et al. \(2007\)](#).

Whether to give standard error as well as or instead of standard deviation is the most frequently asked question that I get. Some authors have commented on the confusion that this issue causes ([Gardner, 1975](#); [Bunce et al., 1980](#); [Bartko, 1985](#)). In one sense the distinction is unimportant, because any two values of N , s , and $s_{\bar{x}}$ allow the remaining value to be calculated. When presenting data in the form of a bar graph (which is inferior to a box plot) there is a tendency to give the mean and standard error for two reasons. One is that it makes it easier to compare two or more means visually to determine if they are close together or far apart in terms of standard error, and this visual impression certainly reinforces any statistics that appear in the text. However, the visual impression is of value only if the numbers in the groups are similar. Furthermore, if N is large then the small value of the standard error tends to conceal what might be large variability (standard deviation) of the sample ([Fig. 7.8](#)). Several issues about presenting data are discussed clearly in recent publications ([Altman et al., 1983](#); [Curran-Everett and Benos, 2004, 2007](#); [Ludbrook, 2008](#); [Morton, 2009](#); [Weisserger et al., 2015](#)). There is a tendency for journals to prefer the standard deviation or confidence intervals ([Fidler et al., 2004](#); [Cumming et al., 2007](#)). It is possible to present standard deviation, standard error, and confidence limits in the same figure, for example, by making the standard error of the mean a thicker part of the line, or side by side, and by making the confidence limits a dashed line. If used, the various lines should be described.

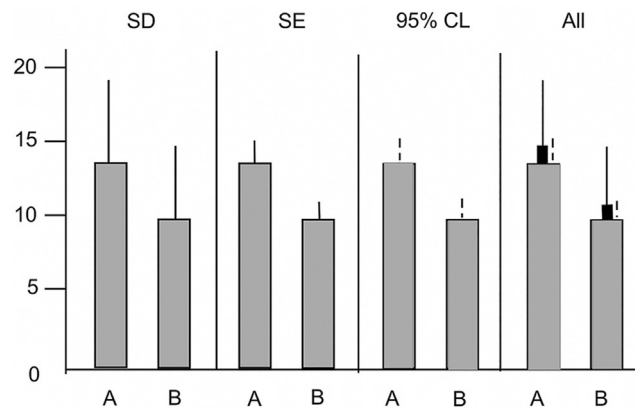


Fig. 7.8 Comparison of bar graphs with standard deviation (SD), standard error (SE), and confidence limits (CL) separately and combined. A and B indicate two different groups of data.

Panel A emphasizes the variability of the data and even suggests that groups with these large standard deviations might have outliers. Panel B allows the reader to make judgments about differences of means among groups and is useful if sample sizes are comparable. Panel D shows one way to include both numbers as well as the confidence limits. There is no one right answer to the question about which deviation to show, but authors should make the basic data available to the reader (Brown, 1982). Variability is as much a feature of the data set as is the mean.

Using confidence limits has advantages. These limits incorporate sample size and standard deviation in one omnibus number. A large interval indicates lack of precision due to a large standard deviation, a small sample size, or both.

REFERENCES

- Altman, D.G., Gore, S.M., Gardner, M.J., Pocock, S.J., 1983. Statistical guidelines for contributors to medical journals. *BMJ* 286, 1489–1493.
- Bartko, J.J., 1985. Rationale for reporting standard deviations rather than standard errors of the mean. *Am. J. Psychiatry* 142, 1060.
- Brown, G.W., 1982. Standard deviation, standard error. Which 'standard' should we use? *Am J Dis Child* 136, 937–941.
- Bunce III, H., Hokanson, J.A., Weiss, G.B., 1980. Avoiding ambiguity when reporting variability in biomedical data. *Am. J. Med.* 69, 8–9.
- Cumming, G., Maillardet, R., 2006. Confidence intervals and replication: where will the next mean fall? *Psychol. Methods* 11, 217–227.
- Cumming, G., Fidler, F., Vaux, D.L., 2007. Error bars in experimental biology. *J. Cell Biol.* 177, 7–11.
- Curran-Everett, D., Benos, D.J., 2004. Guidelines for reporting statistics in journals published by the American Physiological Society. *Am. J. Physiol. Endocrinol. Metab.* 287, E189–E191.
- Curran-Everett, D., Benos, D.J., 2007. Guidelines for reporting statistics in journals published by the American Physiological Society: the sequel. *Adv. Physiol. Educ.* 31, 295–298.
- Fidler, F., Thomason, N., Cumming, G., Finch, S., Leeman, J., 2004. Editors can lead researchers to confidence intervals, but can't make them think: statistical reform lessons from medicine. *Psychol. Sci.* 15, 119–126.
- Gardner, M.J., 1975. Understanding and presenting variation. *Lancet* 1, 230–231.
- Gardner, M.J., Altman, D.G., 1995. *Statistics with Confidence—Confidence Intervals and Statistical Guidelines*. British Medical Journal, London.
- Goldstein, A., 1964. *Biostatistics, an Introductory Course*. Macmillan Company, New York.
- Hahn, G.J., Meeker, W.Q., 1991. *Statistical Intervals. A Guide for Practitioners*. John Wiley and Sons, Inc, New York.
- Handbook, E.S., n.d. Tolerance intervals for a normal distribution. <http://www.itl.nist.gov/div898/handbook/prc/section2/prc263.htm>.
- Healy, M.J.R., 1992. Statistics from the inside. 3. Estimation. *Arch. Dis. Child.* 67, 149–150.
- Ioannidis, J.P., 2005. Why most published research findings are false. *PLoS Med.* 2, e124.
- Ludbrook, J., 2008. The presentation of statistics in clinical and experimental pharmacology and physiology. *Clin. Exp. Pharmacol. Physiol.* 35, 1271–1274 (author reply 1274).
- Morton, J.P., 2009. Reviewing scientific manuscripts: how much statistical knowledge should a reviewer really know? *Adv. Physiol. Educ.* 33, 7–9.
- Natrella, M.G., 1963. *Experimental Statistics*. Dover Publications, Inc, Mineola, NY.
- Newman, R.W., White, R.M., 1966. In: Damon, A., Stoudt, H.W., McFarland, R.A. (Eds.), *The Human Body in Equipment Design*. Harvard University Press, Cambridge, MA.

- Ramirez, B.A., 2009. Statistical intervals: confidence, prediction, enclosure. http://www.sascommunity.org/mwiki/images/3/3c/Statistical_Intervals-Confidence%2C_Prediction%2C_Enclosure.pdf.
- Van Belle, G., 2002. Statistical Rules of Thumb. Wiley Interscience, New York.
- Velleman, P.F., Hoaglin, D.C., 1981. Applications, Basics and Computing of Exploratory Data Analysis. Duxbury Press, Boston, MA.
- Weissgerber, T.L., Milic, N.M., Winham, S.J., Garovic, V.D., 2015. Beyond bar and line graphs: time for a new data presentation paradigm. PLoS Biol. 13, e1002128.

CHAPTER 8

Other Continuous Distributions

CONTINUOUS UNIFORM DISTRIBUTION

In a uniform distribution, continuous or discrete, the random variable assumes all its values with equal probability. This distribution is used, for example, to determine if the incidence of a disease is constant month by month throughout the year.

If the continuous interval from which x is chosen extends from a to b , then the mean $\mu = \frac{a+b}{2}$, and the variance $\sigma^2 = \frac{(b-a)^2}{12}$ (Hogg and Tanis, 1977).

The probability of selecting a value of x in the range a to x is $\frac{x-a}{b-a}$.

The distribution can be used to assess waiting times. A physician sees one patient every 30 min from 8 a.m. until 5 p.m. What are the chances that a randomly arrived patient will have to wait less than 5 min? In each 30-min period the last 5 min meet the requirements, so that the probability is $5/30 = 0.167$. More formally, calculate $(30-25)/(30-0)$ to get the same answer.

EXPONENTIAL DISTRIBUTION

The times to failure, for example, of a heart valve or a transplanted kidney, or the survival time of a cancer patient often closely follow a curve defined as the exponential distribution, characterized by the quantity being studied growing or decaying at a rate proportional to its current value.

The continuous random variable X has an exponential distribution if its probability density function is given by

$$f(X) = \frac{1}{\beta} e^{-\frac{X}{\beta}} \text{ when } x > 0 \text{ and } \beta > 0.$$

It can also be written as:

$$f(x) = \lambda e^{-\lambda x}.$$

Because $\lambda = \frac{1}{\beta}$, either formulation can be used. λ is sometimes termed the rate parameter, and β is termed the scale parameter.

The mean of this distribution is β , and the variance is β^2 , or $\frac{1}{\lambda}$ and $\frac{1}{\lambda^2}$, respectively. (The mean and the standard deviation are the same.) Examples of exponential curves are given in Fig. 8.1.

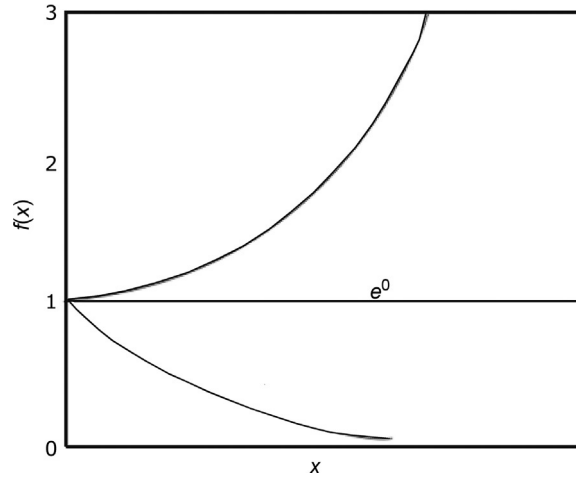


Fig. 8.1 Exponential curves with positive, zero, and negative exponents.

A negative rate constant estimates the time to failure, for example, of a heart valve, a transplanted kidney, or the whole person (death). It also estimates radioactive decay, where the amount of radioactivity emitted decays exponentially with time. In addition to playing a part in survival analysis (Chapter 35) it appears in Chapter 18 (Poisson distribution). Whereas the Poisson distribution deals with the number of random events that occur, the exponential distribution deals with the waiting time between those events.

As an example, patients arrive at a clinic randomly at an average rate of 20/h. Assume that this is a Poisson process in which the arrival times are independent of each other (Chapter 18). What is the probability that the first patient arrives more than 5 min after the clinic opens? An average of 20 patients per hour is equivalent to $\lambda = 1/3$ patients/min. Because $\beta = 1/\lambda$, the function may be written as

$$f(X) = \beta e^{-\beta X} = \frac{1}{3} e^{-\left(\frac{1}{3}\right)X}.$$

Therefore

$$P(X > 5) = \int_5^{\infty} \frac{1}{3} e^{-\left(\frac{1}{3}\right)X} = e^{-\frac{5}{3}} = 0.189.$$

When the rate constant is positive, the exponential distribution characterizes growth processes such as the increase in bacteria or cancer cells with time.

LOGARITHMIC DISTRIBUTION

Many distributions are skewed to the right, making analysis awkward (Limpert et al., 2001; Limpert and Stahel, 2017). One widely used solution is to use the logarithms of

the observations, rather than the original observations themselves because these transformed values often resemble a normal distribution. This distribution is particularly useful in epidemiological studies as atmospheric pollutants, and so on, cannot be negative (van Belle, 2002). Motulsky (2009) emphasized that an exponential decay curve could not be linearized by a logarithmic transformation unless it went to zero. He recommended using these transformations also for ratios (such as the odds ratio) and for growth curves. They should not be used for bar graphs.

CHI-SQUARE DISTRIBUTION

Draw at random a variable X_i from a normal distribution and transform it into a standard normal variate z_i by

$$z_i = \frac{X_i - \mu}{\sigma},$$

then if z_1, z_2, \dots, z_n are independent standard normal deviates, the distribution of the random variable $\chi_n^2 = z_1^2 + z_2^2 + \dots + z_n^2$ is called the χ^2 distribution with n degrees of freedom.

The distribution of z^2 is called a χ^2 distribution with one degree of freedom, often written as χ_1^2 . If there are two or more independent normal distributions of X , then $\chi_2^2 = z_1^2 + z_2^2$, termed χ^2 with 2 degrees of freedom, and in general $\chi_n^2 = z_1^2 + z_2^2 + \dots + z_n^2$ is termed χ^2 with n degrees of freedom.

- a. If χ_n^2 and χ_m^2 are two independent χ^2 random variables, their sum $\chi_n^2 + \chi_m^2$ has a χ^2 distribution with $n + m$ degrees of freedom.
- b. The mean value of χ^2 is ν , the degrees of freedom, and its variance is 2ν .
- c. The possible values of χ^2 run from 0 to infinity.
- d. When ν is greater than 30, $\sqrt{2\chi^2}$ is distributed approximately normally with unit standard deviation and mean $\sqrt{2\nu - 1}$.

The distributions of χ^2 for different degrees of freedom are shown in Fig. 8.2.

Because this distribution involves the square of z it is always positive. The probability is 0.05 that z is bigger than +1.96 or -1.96. If either of these events occurs, then z^2 exceeds 3.84. The right tail of the curve is the same for negative and positive values of z .

The probabilities of any given value of χ^2 for given degrees of freedom can be obtained from standard tables or from online computer programs listed in Chapter 14.

This distribution can be used to determine the confidence limits for the variance by using the online program <https://www.easycalculation.com/statistics/confidence-interval-variance.php> or <http://www.wolframalpha.com/widgets/view.jsp?id=5327a43b4c8b784329da9e81caee6987>.

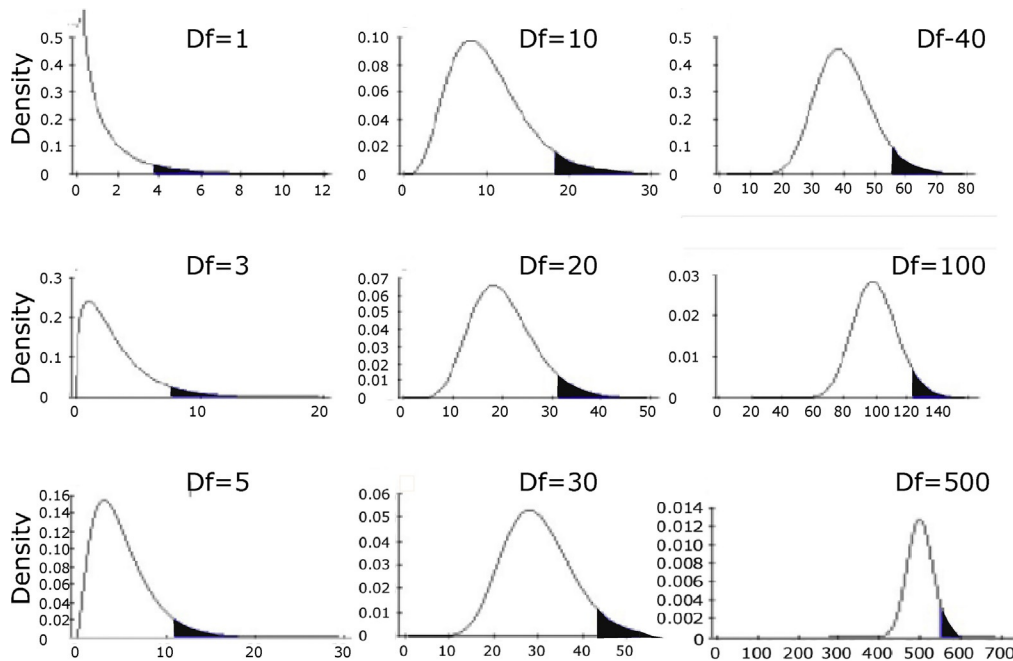


Fig. 8.2 Chi-square for various degrees of freedom. The area under the curve representing 0.05 of the total area is shaded. The curves become more symmetrical as the degrees of freedom (Df) increase, and the peak of the curve is approximately at the corresponding degrees of freedom. If we draw at random one normal variate and square it, the probability of getting a value near the mean of zero is very high, and the probability of getting a large deviation from the mean is very low. Therefore we get the curve as shown by the upper left for 1 degree of freedom. If we draw three normal variates at random and square them, the probability that they sum to a very small value near zero is lower, as shown by the curve for 3 degrees of freedom. By adding more squared normal variates, the sum tends to move further away from zero, as shown for increasing degrees of freedom.

Another use of this distribution is to analyze categorical data, as discussed in detail in [Chapter 14](#).

VARIANCE RATIO (F) DISTRIBUTION

If we draw two random samples of size n_1 and n_2 from a normal distribution and calculate their variances s_1^2 and s_2^2 , they will not be identical. The distribution of the ratio of the two variances $\frac{s_1^2}{s_2^2}$ was determined by R.A. Fisher and it is possible to determine the probability of the ratio exceeding any given value. The general form of the distribution is shown in [Fig. 8.3](#).

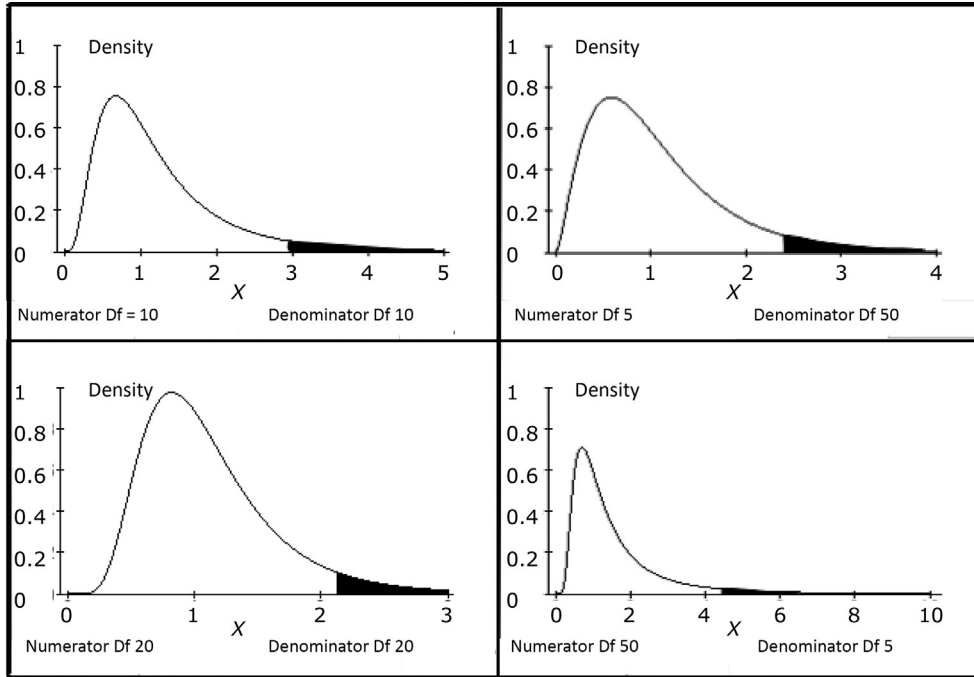


Fig. 8.3 Typical F distributions. There is a family of curves. Each curve depends on the degrees of freedom in the numerator and the denominator, for example, 50, 5, respectively. The shaded areas indicate the F value beyond which 0.05 of the curve occurs (one sided).

As generally used, this is a one-sided test, and we are usually interested only if the bigger variance exceeds the smaller variance by some critical ratio. Rarely we need to calculate values to the left of the distribution; for example, what is the value of F above which 95% of the distribution lies.

The distribution was discovered by R. A. Fisher in the 1920s. George Snedecor (1881–1974) named it the F distribution in Fisher's honor. The variance ratio is used extensively in analysis of variance and in regression statistics.

The distribution depends on the degrees of freedom ν_1 and ν_2 for the two variances. All statistical texts give tables of the critical values, and they can be found online at <http://www.itl.nist.gov/div898/handbook/eda/section3/eda3673.htm>, http://vassarstats.net/textbook/apx_d.html, <http://stattrek.com/online-calculator/f-distribution.aspx>, and http://www.tutor-homework.com/statistics_tables/f-table-0.001.html. Online probabilities for given values of F and the degrees of freedom of the two components can be found at http://www.socr.ucla.edu/Applets.dir/F_Table.html, <http://vassarstats.net/tabs.html>, <http://www.statdistributions.com/f/>, and <http://danielsoper.com/statcalc3/calc.aspx?id=4>.

REFERENCES

- Hogg, R.V., Tanis, E.A., 1977. Probability & Statistical Inference. Macmillan Publishing Company, Inc, New York.
- Limpert, E., Stahel, W.A., 2017. The log-normal distribution. *Significance* 14, 8–9.
- Limpert, E., Stahel, W.A., Abbt, M., 2001. Log-normal distribution across the sciences: keys and clues. *Bioscience* 51, 352.
- Motulsky, H.J., 2009. Statistical Principles: The Use and Abuse of Logarithmic Axes. <https://s3.amazonaws.com/cdn.graphpad.com/faq/1910/file/1487logaxes.pdf>.
- Van Belle, G., 2002. Statistical Rules of Thumb. Wiley Interscience, New York.

CHAPTER 9

Outliers and Extreme Values

OUTLIERS

In an example of the effect of raw and roasted peanuts on the growth of rats, ([Chapter 22](#)) one rat in one pair of littermates showed an unusually small weight gain. This discrepant observation had the effect of reducing the difference in average weight gain on the two diets and making it impossible to reject the null hypothesis. Was it really an outlier and if so what should we do about it?

As defined by [Grubbs \(1969\)](#), “An outlying observation, or “outlier”, is one that appears to deviate markedly from other members of the sample in which it occurs.” There may be more than one outlier. First, eliminate errors of observation or recording. It is not uncommon to enter weight instead of height or to misplace a decimal point. For the peanut example, a check of the weighing procedures may show that for this rat the scale had not been correctly calibrated or checking the feeding logs may show that by mistake this rat missed one or more feedings. The rat might appear ill, and examination might disclose a lung infection. If errors of measurement or recording are detected, they are corrected if possible, and if not correctable (e.g., a scale that was incorrectly calibrated by some unknown and unrecoverable factor) the measurement is omitted with an explanation. If the rat missed a feed or was ill that also justifies removal, with an explanation. There are legitimate reasons for excluding data, but it is essential to describe why the data were removed.

If no apparent cause for the discrepancy is found, how discrepant is the measurement? One simple method is to make a box plot, looking for minor and major outliers, as defined by [Tukey \(1977\)](#) ([Fig. 4.20](#)). By definition, a minor outlier is one that is $>1.5 \times$ interquartile distance above the 75th or below the 25th centile, and this comes to a z value of 2.698 (one sided). For a normal curve, only 0.35% of the area under the curve (about 1 in 285) is beyond this value. A major outlier is one that is $3 \times$ interquartile distance above the 75th or below the 25th centile, and this comes to a z value of 4.7215 (one sided). For a normal curve, only 0.00012% of the area under the curve (about 1 in 833,333) is beyond this value.

Should the distribution be very skewed, there is a strong tendency for too many of the measurements to be declared outliers based on the Tukey limits as defined before. This bias can be reduced by using a skewness adjusted box plot ([Hubert and Vandervieren, 2008](#)) that requires special programs for its use. A normalizing transformation can also be used.

Grubb's test

Other methods for making a decision depend on what we know of the variate that is being measured. *If there is reason to believe that the data should have been normally distributed* there are tests to disclose outliers. A simple test is known as Grubbs' test or the extreme studentized deviate test (Grubbs, 1969). If we draw a sample of size N from a normal population, we can examine the difference between the mean and the extreme value of X , divided by the standard deviation:

$$T = \frac{\max |\bar{X} - X_i|}{s}$$

where $\max |\bar{X} - X_i|$ is the greatest deviation either above or below the mean. The standard deviation is calculated from all the data, including the suspected outlier. This ratio is called T . In a large number of random selections it is possible to state that T exceeds a certain value 5% of the time, or 2.5% of the time, or 1% of the time, and so on. The relevant table is provided by Grubbs (1969) and appears also on the web at <http://www.graphpad.com/quickcalcs/Grubbs1.cfm> or <http://contchart.com/outliers.aspx>. Grubb's test allows only one outlier to be examined per sample. Sometimes, however, there may appear to be two or three outliers, and then a sequential test can be done. The greatest outlier is examined by Grubbs' test. If it is too extreme it is removed, and the next greatest outlier is examined, but assessed against a modified ratio. <http://contchart.com/outliers.aspx> can be used for multiple outliers. Alternatively, if there is suspicion about the two largest or the two smallest observations, Grubbs (1969) calculated the ratio $\frac{s^2}{s_{12}^2}$ where s^2 is the variance of the whole data set and s_{12}^2 is the variance of the data set after removing the two presumptive outliers. This ratio is then referred to a table of critical values (Grubbs, 1969) or at free online programs at http://www.statistics4u.com/fundstat_eng/ee_grubbs_outliertest.html and <http://graphpad.com/quickcalcs/Grubbs1.cfm>.

The difficulty with Grubb's test is that is based on a presumed normal distribution.

Slippage and contamination

Even with a normal distribution, outliers may appear if there is contamination or slippage. In a study of the ability of people to identify the range of a distant object, the investigators were unaware that some people have no depth perception. Most people had errors that were normally distributed with a small deviation, and the others had errors with a much larger standard deviation (Fig. 9.1).

The people without depth perception are represented by the dashed curve. A point from the upper (or lower) region of this curve (x) appears as an outlier in the main curve shown in solid lines.

Slippage implies that the data may come from a normal distribution in which the mean is larger than the control mean by some factor. For example, Karl Pearson recorded

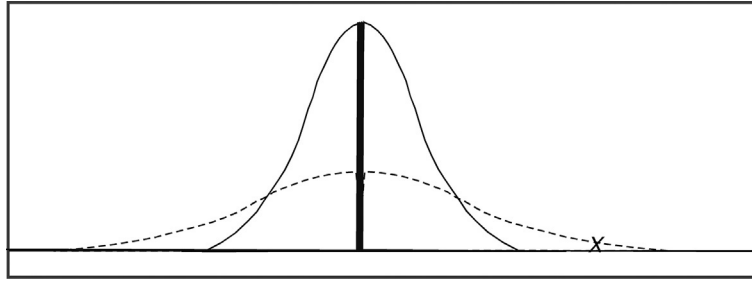


Fig. 9.1 Contaminated distribution.

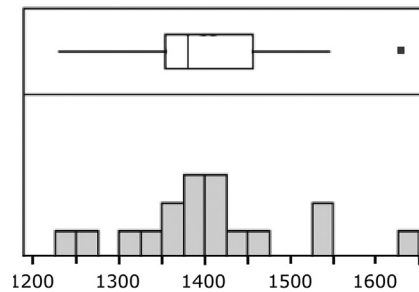


Fig. 9.2 Cranial capacities.

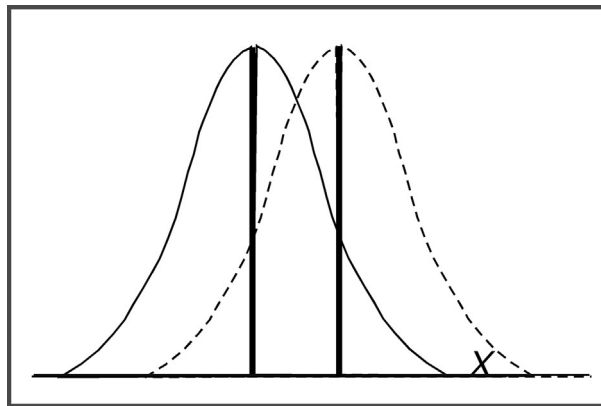


Fig. 9.3 Slippage.

cranial capacities of 17 skulls found in an archeological site (Barnett and Lewis, 1984). Sixteen of these ranged from 1230 to 1540 ccs, but one was 1630 ccs (Fig. 9.2).

The value at 1630 ccs is an outlier. If it is an outlier, what might have been the cause? Did it come from a diseased subject, or did the site from which the skulls were retrieved contain skulls from two different species (Fig. 9.3)?

Here the second species with a larger cranial capacity is represented by the dashed curve. An individual from the upper part of this curve (X) appears as an outlier from the primary curve in solid lines. Slippage may also occur when a chemical or physical process becomes distorted.

Robust tests

Many tests are based on order statistics, one of the simplest being to use the median absolute deviation. The median absolute deviation (MAD, [Chapter 3](#)) is the median of the deviations from the median. In a *normal* distribution, there is a relationship among MAD, standard deviation, and the interquartile distance ([Chapter 4](#)). From these relationships, $M - 2\text{MAD} < x_i < M + 2\text{MAD}$ is a normal range, and values outside this range are outliers. A coefficient of 3 instead of 2 makes this more conservative.

Dealing with outliers

An outlier that is not due to an error requires careful consideration rather than exclusion based on a statistical test. Sometimes these outliers allow the investigator to develop new concepts. For example, in a study in which a low concentration of a gas is given to a group of people with normal lungs, all but one shows minimal changes in airway resistance but one subject had marked bronchoconstriction. Rather than discarding the data from this subject the investigator looks further and finds that this subject has a medium-sized ventricular septal defect. This might lead to the hypothesis that children with congenital heart disease and a left to right shunt may have unusually reactive airways.

There are four ways of dealing with such an outlier. (1) Retain it but use robust distribution-free methods of analysis; for example, using the median instead of the mean so that the size of the outlier has minimal effect. (2) Transform the data, for example, by logarithm or square root, to remove outliers. (3) Retain the outlier, and use conventional parametric analysis, accepting the loss of power that this will produce. (4) Omit the outlier, but report what you have done. An eminent statistician Kruskal wrote: “As to practice, I suggest that it is of great importance to preach the doctrine that apparent outliers should always be reported, even when one feels that their causes are known or when one rejects them for whatever good rule or reason. The immediate pressures of practical statistical analysis are almost uniformly in the direction of suppressing announcements of observations that do not fit the pattern; we must maintain a strong seawall against these pressures” ([Barnett and Lewis, 1984](#)). After all, if the observations do not fit the expected pattern, perhaps it is the expected pattern that is wrong. Do not look at data as if you were wearing blinkers adjusted to ± 2 standard deviations.

EXTREME VALUES

Sometimes we wish to know how large or how small an extreme value may be. An engineer might want to know, when designing a bridge, what the highest water level might be, a level perhaps reached only once in a 100 years, or what the highest possible wind force could be. A hospital administrator might want to know the maximal number of hospital beds required in a severe influenza epidemic. Alternatively, the minimum breaking strain of a metal bar or a surgical suture is needed. It was this last problem as applied to cotton threads that was the starting point for Tippett's investigations of this subject. Subsequently, Fisher and particularly Gumbel (1891–1966) contributed to the field.

Consider the extreme high value. Some type of distribution needs to be assumed, based on prior knowledge. One common distribution for this purpose is the cumulative distribution function of Gumbel: $F(y) = e^{-e^{-y}}$. In this equation, $F(y)$ is the probability of getting a value of y or less. If $y = 3$, then $F(y)$ is 0.9514, the probability of getting 3 or less. (There are two other general distributions—Fréchet and Weibull—with different features.)

This distribution is one of a family of distributions. In one form of analysis, the **maximum** value in each of several members of a set is noted, and several sets are inspected; the sets may be rainfall per year, red cell diameter per sample of blood, number of patients with severe influenza per year, and so on. These maximal values are ranked from smallest to largest, assigned ranks 1 to n , and then transformed into cumulative frequencies $P_i = i/(n+1)$, where i is the rank of the i -th observation starting from the smallest. The data are plotted on special extreme-value probability paper. Fig. 9.4 shows an example using the largest size of a red blood cell from 14 blood samples selected from the data of Chen and Fung (1973).

Similar methods can be used to determine extreme minima. These are required in fatigue studies; for example, what is the lowest number of stress cycles will a prosthetic valve or an orthopedic prosthesis tolerate before breaking?

Graph papers for log-normal plots can be obtained online at <http://www.printablepaper.net/>, <http://www.weibull.com/GPaper/> and <http://www.printfreegraphpaper.com/>. It is possible for the curves to be a linear and based on other distributions, and then other functions and graph papers are needed. A readable introduction to these analyses is given by Santner who gives clear directions for constructing these graphs (Santner, 1973).

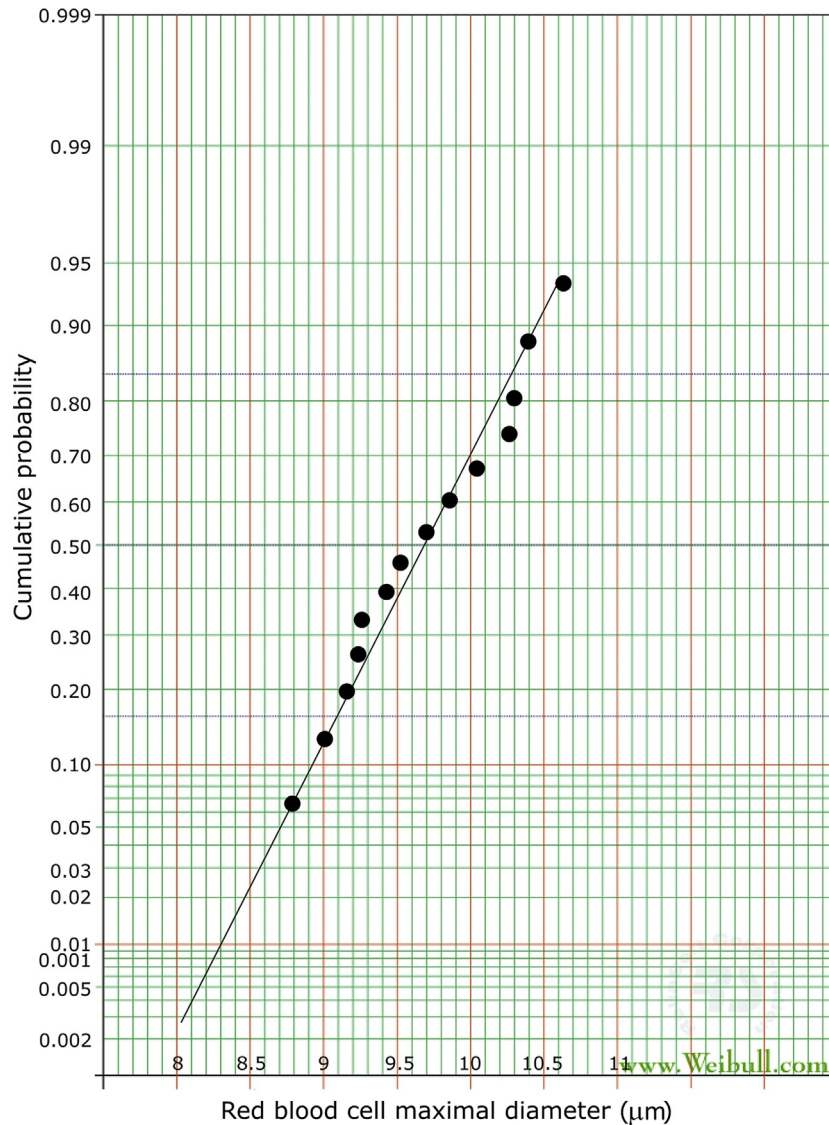


Fig. 9.4 Extreme value of red cell diameters, measured in several samples each with 100 red blood cells. Ninety percent of samples have maximal red cell diameters $<10.6\mu\text{m}$.

REFERENCES

- Barnett, V., Lewis, T., 1984. *Outliers in Statistical Data*. John Wiley & Sons, New York.
- Chen, P.C., Fung, Y.C., 1973. Extreme-value statistics of human red blood cells. *Microvasc. Res.* 6, 32–43.
- Grubbs, F.E., 1969. Procedures for identifying outlying observations in samples. *Technometrics* 11, 1–21.
- Hubert, M., Vandervieren, E., 2008. An adjusted boxplot for skewed distributions. *Comput Stat Data Analysis* 52, 5186–5201.

- Santner, J.F., 1973. An introduction to Gumbel, or extreme-value probability paper. U.S.Environmental Protection Agency, Water Program Operations, National Training Center, Cincinnati, OH. Available <http://nepis.epa.gov/Exe/Zynet.exe/900N0900.TXT?ZyActionD=ZyDocument&Client=EPA&Index=Prior+to+1976&Docs=&Query=&Time=&EndTime=&SearchMethod=1&TocRestrict=n&Toc=&TocEntry=&QField=&QFieldYear=&QFieldMonth=&QFieldDay=&IntQFieldOp=0&ExtQFieldOp=0&XmlQuery=&File=D%3A%5Czyfiles%5CIndex%20Data%5C70thru75%5CTxt%5C00000006%5C900N0900.txt&User=ANONYMOUS&Password=anonymous&SortMethod=h%7C-&MaximumDocuments=1&FuzzyDegree=0&ImageQuality=r75g8/r75g8/x150y150g16/i425&Display=p%7Cf&DefSeekPage=x&SearchBack=ZyActionL&Back=ZyActionS&BackDesc=Results%20page&MaximumPages=1&ZyEntry=1&SeekPage=x&ZyPURLs>.
- Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley Publishing Co, Menlo Park, CA.

SECTION III

Hypothesis Testing

CHAPTER 10

Hypothesis Testing: The Null Hypothesis, Significance and Type I Error

HYPOTHESES

Statistical inference is often based on a test of significance, “a procedure by which one determines the degree to which collected data are consistent with a specific hypothesis...” (Matthews and Farewell, 1996). Hypotheses may be specific, for example, that the slope relating two variables is 1, the line of identity, or that the mean height is 150 cm. Often the hypothesis is that two (or more) sample statistics could have been drawn from the same population. This is the null hypothesis or H_0 .

The null hypothesis holds special importance in statistical inference. In many experiments two (or more) samples drawn from a population are given different treatments: two groups of patients with leukemia are given different agents to determine if one is better at prolonging life; a protein kinase inhibitor is given to one group of cells to find out if it decreases the production of an angiogenic factor; two groups of rats are fed different quantities of two types of fat to find out if the relationship of fat intake to serum cholesterol differs for the types of fat. Because of variation it is almost certain that the means or the slopes will differ even if the treatment had zero effect, so the question to be solved is how to calculate the effects of chance on the outcome. If the differences observed are likely to be due to chance, we would conclude that the treatment was unlikely to have been effective (providing the experiment had sufficient power—see Chapter 11). If the differences are unlikely to be due to chance, then we might want to reject the null hypothesis.

How do we test H_0 ? The principles to be discussed are illustrated by referring to sample and population means but apply equally to all other types of statistics. Conventionally today we apply a test of significance, determine a P value, and if the P value is very small we reject the null hypothesis. This is known as null hypothesis significance testing (NHST), but the approach has been controversial since its inception.¹

¹ If we reject the null hypothesis but it is really correct, we have committed a Type I error, and we would like to keep this low. Conversely if we accept the null hypothesis when it is incorrect, we have committed a Type II error—discussed in the next chapter.

The first person to attempt to determine if the difference between samples was or was not due to random error was Gosset (Ziliak, 2008). This was the impetus to developing the “Student” t -test. He did not, however, think it was appropriate to set a number for the probability of rejecting the null hypothesis. In 1904 he wrote: “Results are only valuable when the amount by which they probably differ from the truth is so small as to be insignificant for the purpose of the experiment. What the odds should be depends:

1. On the degree of accuracy which the nature of the experiment allows, and
2. On the importance of the issues at stake”

He was followed by Fisher who recommended calculating the probability, termed p or P , that a particular difference between two means allowed us to reject the null hypothesis of no difference. A decision of this sort has many implications, so we try to minimize the probability of rejecting the null hypothesis (of no difference) when it is actually correct. He calculated P by a “significance test”; the smaller the value of P the higher the significance. Fisher originally suggested a critical P value < 0.02 , but then settled on $P < 0.05$ as more practical. As used in this way, P represents the probability of obtaining an experimental mean value that differs from the observed control mean value *if the null hypothesis is true*. Fisher used the P value to assess the weight of evidence against the null hypothesis, not as an absolute criterion.

As proposed by Fisher, the statistical test is only one of the factors allowing us to decide whether or not to reject the null hypothesis. The magnitude of the difference and the plausibility of the hypothesis are other important factors. Fisher thus introduced a degree of subjectivity into the assessment, something quite the opposite of what happens when we slavishly accept a given level of significance. Fisher wrote: “If p is between 0.1 and 0.9 there is no reason to suspect the hypothesis tested. If it is below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be led astray if we draw a conventional line at 0.05 and consider that higher values of χ^2 indicate a real discrepancy.” Fisher (1956) was here referring to the null hypothesis.² That he was not rigid in his approach is shown by his statement that “the state of opinion derived from a test of significance is provisional, and capable, not only of confirmation, but of revision” and “A test of significance... is intended to aid the process of learning by observational experience” (Fisher, 1956).² It is worthy of note

² The statement “Nothing is certain except death and taxes” applies here.

Consider the numbers from 1 to 100. Draw two sets of 3 numbers, replace them, draw another two sets, and so on, and compare the two sets. Now you are just as likely to get 1, 2, 3 and 98, 99, 100 as to get 44, 46, 48 and 45, 47, 49. Each combination will be repeated many times. By t -tests some P values will be high, others low. Five percent of them will have high t -values, 1% will have higher t -values, 0.1% will have still higher t -values. The higher the t -value, and the smaller the P value the more confidence you will have in rejecting the null hypothesis, but no matter how small P is you can never be certain. That is why Fisher insisted on biological plausibility and being able to repeat the experiment.

that Gosset, one of Fisher's mentors, disdained the use of a fixed value for significance and advised Fisher against using it.

An alternative approach was recommended in 1928 by Neyman and Pearson who selected in advance the value of α , the long-term risk of falsely rejecting the null hypothesis, usually 0.05 or 0.01, and postulated an alternative hypothesis— H_A . (They introduced the term Type I error.) The differences between the two approaches are given in Table 10.1. More detailed but very readable accounts of the differences are given by Goodman (1999b), Lew (2012), and Hubbard and Bayarri (2003). Goodman used the analogy of a legal trial. In Fisher's approach the accused person (who is actually innocent) would be judged guilty or not guilty with a 5% probability of being incorrectly found guilty. In the Neyman and Pearson approach, only 5% of a large group of such accused people would be incorrectly found guilty.

Table 10.1 Comparing hypotheses

	Fisher	Neyman-Pearson
P value	Indicates weight of evidence against H_0	If <0.05 , reject H_0
Added support H_A	Required	Not required
Number of tests	Unspecified	Specified
	One test only	Probability applies to several repetitions of the test

There is no certainty, merely an expression of confidence that one might accept the results and move on to the next experiment.

I add three more objections.

1. For a given difference between the means, P is a function of sample variability and size. Consequently, two experiments with similar results but different P values are not necessarily in conflict. (See Fig. 4 in the excellent article by Motulsky, 2015.)
2. A given P value, even if low, is no guarantee that a similar P value will be obtained when repeating that experiment. Cumming (2008) has shown that, for example, a P value of 0.05 has an 80% chance of being from 0.00008 to 0.44 on subsequent replications of the experiment.
3. The P value for significance is not a fixed number. If the experiment is trivial and the outcome not important, $P < 0.05$ might suffice. If, however, the outcome is unexpected and—if true—of great importance, then a much smaller value for P is required.

As Carl Sagan stated “Extraordinary claims require extraordinary evidence.”

There is much confusion in textbooks and practice about the relationships among P and α . For Fisher, P (defined before) is the result of a single experiment, and it is not the same as α as defined by Neyman and Pearson because α represents the long-term error

determined in advance. α might be set to be 0.05, but in a single experiment P might be 0.035; they are obviously not the same. In addition, α is a conditional probability, conditional on the null hypothesis being correct.

John Tukey (1991) wrote: “The worst, i.e., most dangerous feature of “accepting the null hypothesis” is the giving up of explicit uncertainty: the attempt to paint with only the black of perfect equality and the white of demonstrated direction of inequality. Mathematics can sometimes be put in such black-and-white terms, but our knowledge or belief about the external world never can.

The black of “accept the null hypothesis” is far too black. It treats “between -101 and $+1$,” “between -101 and $+101$,” and “between -1 and $+1$ ” all alike, when their practical meanings are often very, very different.

The white of demonstrated direction of inequality is too white. On its face, it treats “between $+1$ and $+101$,” “between $+1$ and $+3$,” and “between $+99$ and $+101$ ” as if they were the same when their practical meaning is quite different.”

He continued: “Long ago, Fisher (1926, foot of p. 504) recognized that truly solid knowledge did not come from analyzing a single experiment—even when that gave a confident direction with a very, very small error rate, like one in a million—but rather that solid knowledge came from a demonstrated ability to repeat experiments, each of which showed confident direction at a reasonable error rate, like 5%. This is unhappy for the investigator who would like to settle things once and for all, but consistent with the best accounts we have of scientific method, which emphasize repetition, preferably under varied circumstances.”

An example of this caveat was reported by Crease (2010) in discussing the search for dark matter by astrophysicists. Physicists collect over time the outputs from a detector displayed in different energy levels (bins). If one particular bin shows an increased number of hits compared to what was expected at that energy level, it may provide evidence for a new particle. In one such experiment, the DAMA/LIBRA experiment at the Gran Sasso National Laboratory in Italy, the excess of energy in a particular bin had a confidence level of 8.2σ . This indicates a very low probability of falsely rejecting the null hypothesis, with a probability of 1 in ten million billion. Nevertheless, subsequent experiments failed to confirm this finding.

As a corollary to this, we cannot ignore specific circumstances when assessing the results of a study. Millard (Borenstein et al., 2009) provided an example by citing a report from The Guardian newspaper at <http://www.guardian.co.uk/football/2010/jul/12/paul-psychic-octopus-wins-world-cup> about the ability of an octopus named Paul who in 8 successive years correctly predicted the winner of the World Cup (soccer). Although the probability of this occurring by chance alone was $1/256 = 0.0039$, to draw the conclusion that the octopus was able to predict the results of a soccer match defies credibility, no matter what the P value was.

Ioannidis (2005) published a provocative article entitled “Why most published research findings are false.” He pointed out that when the prior probability of a

relationship is low (and this is true for many relationships) a value of $P < 0.05$ for the Type I error α would almost certainly lead to an exaggerated tendency to reject the null hypothesis. From his analysis he set out some important corollaries, including:

1. The smaller the size of the studies conducted in a scientific field, the less likely the research findings are to be true.
2. The smaller the effect sizes in a scientific field, the less likely the research findings are to be true.
3. The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true. Although this is an ethical and not a statistical matter, the bias is so pervasive ([Chapter 2](#)) that we should be cautious in rejecting the null hypothesis based on a borderline value of 0.05 for the Type I error.

[Schuemie et al. \(2014\)](#) introduced an empirical calibration to correct the error, and estimated that up to 54% of tests deemed significant at the 0.05 level were actually not in favor of rejecting the null hypothesis. The exact proportion of erroneous statistical conclusions is disputed, (see the Symposium in the journal “Biostatistics”, volume 15, issue 1, 2014) but it is almost certainly much $>5\%$. This bias has recently been emphasized ([Colquhoun, 2014](#)) who showed that the Type I error of 5% now in vogue is $\sim 36\%$, and that to keep it below 5% one should use 3σ deviation from the mean ($P \sim < 0.001$), not 2σ ($P \sim < 0.05$) as is now used (See [Appendix](#)). A recent recommendation from a large group of statisticians has advised using $P < 0.005$ as the critical value for rejecting the null hypothesis and has provided additional arguments in favor of this choice ([Benjamin et al., 2017](#)).

[Sterne and Davey Smith \(2001\)](#) agreed that the smaller the value of P the less the chance of a Type I error, but they (along with many others), regarded 0.05 as unsafe per se, and preferred to see a value of $P < 0.001$ for safety [Fig. 10.1](#) summarizes their view.

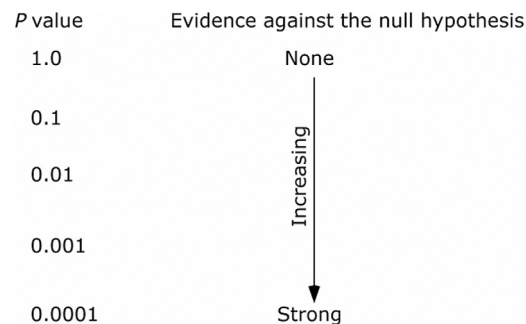


Fig. 10.1 Degree of evidence against null hypothesis. (Based on [Sterne, J.A., Davey Smith, G., 2001. Sifting the evidence-what's wrong with significance tests? BMJ 322, 226–31.](#))

It is commonplace to see in the Methods section of a scientific publication the phrase: “Statistical significance was set at $P < 0.05$.” All that this means is that the reader is given the user’s definition of the term “statistical significance” and little of substance is

provided. It would be more useful to abandon the term “statistical significance” and provide the reader with important information such as the exact P value *and* the confidence limits of the mean or other relevant parameter.

The value of $\alpha = 0.05$ is arbitrary, and values of 0.01 or 0.10 may be used. The apparent importance of 0.05 and 0.01 was in part due to the lack of computing facilities in the 1930s. Statistical tables were difficult and expensive to produce, so that initially only the 0.05 and 0.01 values were provided for certain tests (Barnard, 1992).

Much of the time, however, we wish to evaluate how much uncertainty there is in rejecting the null hypothesis, and it is preferable to give the exact P value, whether it be 0.036 or 0.12, and the confidence limits associated with it, and do not even need to mention the term “significant” (Colquhoun, 1971). In fact, quotations from statisticians indicate the unpopularity of the current usage of the term “statistical significance”:

...it is better to regard the level of significance as a measure of the strength of the evidence against the null hypothesis rather than showing whether the data are significant or not at a certain level.

(Sterne and Davey Smith, 2001)

...the function of significance tests is to prevent you from making a fool of yourself, and not to make unpublishable results publishable.

(Colquhoun, 1971)

Statistical “significance” by itself is not a rational basis for action.

(Deming, 1943)

In biologic experimental work, for instance, a.... common abuse is to use a statistical test to try to “prove” a hypothesis..... The scandal is that the “significant” results are published as though they had meaning.

(Feller, 1969)

Because of these and other caveats, it is better to treat p values as nothing more than useful exploratory tools or measures of surprise. In any search for new physics, a small p value should only be seen as a first step in the interpretation of the data, to be followed by a serious investigation of an alternative hypothesis. Only by showing that the latter provides a better explanation of the observations than the null hypothesis can one make a convincing case for discovery.

(Demortier, 2008)

P values of 0.036, 0.047, or 0.063 are all important pieces of information, and an arbitrary designation of significance may conceal information. As the sports commentator Vin Scully remarked: “Statistics are used much like a drunk uses a lamppost: for support, not illumination.”

A final point to consider was raised by Binder (1963) who wrote “It is surely apparent that anyone who wants to obtain a significant difference enough can obtain one—if his only consideration is obtaining that significant difference.” (Significance can always be attained by a large enough sample size—see Chapter 11.)

In practice, the 0.05 value of P is used in at least two ways. It may be used as a critical value for making a decision by (deductive inference) to accept or reject the null hypothesis, in keeping with the Neyman-Pearson approach. If an investigator is testing 500 chemical

agents for their ability to reduce bacterial growth, there is no doubt about the population mean value that is the control value of growth. What the investigator seeks are those agents worth pursuing in more detail. Therefore in a large batch of experiments, the 0.05 cutoff allows the investigator to concentrate on the more promising agents, knowing that less than 5% of the time the null hypothesis will be rejected falsely (Type I error). Furthermore, although the null hypothesis might not be true, the difference is unlikely to be important. In this usage there are multiple samples, and no inverse logic is used to determine a population mean. However, in most experiments investigators use P as support for a hypothesis, and the exact value of P does not have the same critical importance.

In summary, those who understand statistical methods and philosophy attach only minor importance to the calculated P values, whereas those who merely make the calculations and do not think about them overestimate their importance. Recently the American Psychological Association wrote: “Estimation based on effect sizes, confidence intervals, and metaanalysis usually provides a more informative analysis of empirical results than does statistical significance testing, which has long been the conventional choice in psychology (Cumming et al., 2011).”

This cautious approach to deciding about statistical significance has even been endorsed by the US Supreme Court (Ziliak, 2011). A pharmaceutical company was sued for failing to warn investors about adverse effects that were not statistically significant. The Supreme Court found on behalf of the plaintiffs. At one point in the judgment Justice Sotomayor wrote:

“... argument rests on the premise that statistical significance is the only reliable evidence of causation. This premise is flawed“, and again “medical professionals and researchers do not limit the data they consider to the results of randomized clinical trials or to statistically significant evidence.... The FDA similarly does not limit the evidence it considers for purposes of assessing causation and taking regulatory action to statistically significant data. In assessing the safety risk posed by a product, the FDA considers factors such as “strength of the association,” “temporal relationship of product use and the event,” “consistency of findings across available data sources,” “evidence of a dose-response for the effect,” “biologic plausibility,” “seriousness of the event relative to the disease being treated,” “potential to mitigate the risk in the population,” “feasibility of further study using observational or controlled clinical study designs,” and “degree of benefit the product provides, including availability of other therapies.” ...[The FDA] “does not apply any single metric for determining when additional inquiry or action is necessary”.

CRITICISM OF NHST

Fisher considered only the null hypothesis, but if the null hypothesis is rejected an alternative hypothesis must be considered, as suggested by Neyman and Pearson. Goodman (1993, 1999a) and Goodman and Royall (1988) discussed the philosophical differences

between these two approaches and emphasized that Fisher's use of significance testing required confirmation by repeated experiments and biological plausibility. A low P value suggested that it was worthwhile to repeat the study, a high P value suggested that it was not worth repeating. Fisher did not regard the P value as indicating the frequency of error (Type I error) if the experiment were to be repeated, nor did he consider an alternate hypothesis.

Now it is true that the single sample on which we base our decision to accept or reject the null hypothesis is either from the control population or else from another population. What the rest of the area under the curve from that point out to infinity is has no real meaning. Whether the mean of another sample from the population being examined will be close to or far from the mean of the first sample is unknown. The best that we can do is to make a statement of belief that the sample does or does not come from the control population. [Cohen \(1994\)](#) brought up another problem in the current usage of the P value. He pointed out that what we wanted of the test was "Given these data, what is the probability that H_0 is true?" But as most of us know, what it tells us is "Given that H_0 is true, what is the probability of these (or more extreme) data?" Our test gives us $P(D|H_0)$ (D =data, H_0 =null hypothesis), but what we really want is $P(H_0|D)$. These are not the same. By invoking Bayes' theorem ([Chapter 5](#)),

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

Criticisms of the NHST approach abound. The P value is often used inappropriately. It quantifies the discrepancy between a given data set and the null hypothesis, but it does not give the probability that the null hypothesis is true or that it quantifies a particular error. *Failing to reject the null hypothesis does not mean that the null hypothesis is correct.* It is better to deal with the null hypothesis the way the Scottish law deals with trial sentences. There are three possible verdicts. 1. Guilty—the accused did it. 2. Not guilty—the accused did not do it. 3. Not proven—not enough evidence to make a decision. Similarly, the null hypothesis should be assessed as rejected (very small P value), accepted (large P value, small effect size), or not rejected (P above critical value and moderate effect size); in the last of these the null hypothesis cannot be rejected but is not accepted.

Furthermore, the value of P depends on sample size; the same effect size gives different P values for different sample sizes.

This important subject has been written about simply in several publications ([Weinberg, 2001](#); [Sterne and Davey Smith, 2001](#); [Poole, 2001](#); [Moran and Solomon, 2004](#); [Goodman, 2008](#)). Recently the American Statistical Association set out its views on the subject ([Wasserstein and Lazar, 2016](#)). Among their conclusions were:

1. P values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
2. Scientific conclusions and business or policy decisions should not be based only on whether a P value passes a specific threshold.
3. A P value, or statistical significance, does not measure the size of an effect, or the importance of the result.
4. A P value does not provide a good measure of evidence for a model or hypothesis. On the other hand, P values (but not the term “significance”) provide information if used carefully. Frost (2014) gave five guidelines for using P :
 1. The exact P value matters. The lower the P value, the lower the error rate.
 2. Replication matters.
 3. The effect size matters.
 4. The alternative hypothesis matters.
 5. Subject area knowledge matters.

HYPOTHESIS TESTING AND MAXIMUM LIKELIHOOD; THE BAYES FACTOR

One way out of this dilemma is to use confidence limits (Cumming et al., 2012) and another is to use mathematical likelihood, a concept also introduced by Fisher (Goodman, 1993). We can be more comfortable if we can compare our belief that the sample comes from the control population with our belief that it comes from another specified population.

Consider Fig. 10.2

The Bayes factor, also known as the likelihood ratio, is the ratio:

$$\frac{\text{Probability of data (null hypothesis)}}{\text{Probability of data (alternative hypothesis)}}$$

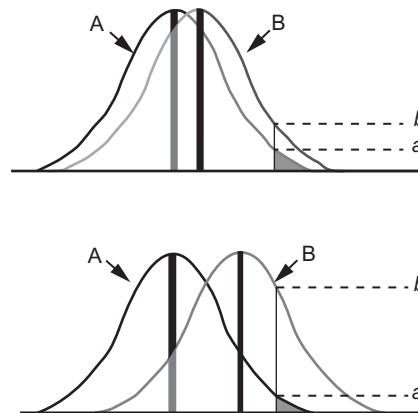


Fig. 10.2 Likelihood ratios.

In the upper panel, with the two means close together, the probability of “a” is about 50% the probability of “b.” The exact probability is used, not the area beyond the critical value. In the lower panel, “a” is about 16% of “b.” Curve B has a better likelihood of being correct in the lower panel, even though in terms of a Type I error the null hypothesis would not be rejected. The calculations of likelihood are presented in Goodman’s article and in the [Appendix](#).

Many experts recommend using the confidence interval to indicate what possible values occur if the null hypothesis is true. This approach meets the objections made by Tukey and provides a range of values for the investigator to consider. It is better than giving a single value. As [Tukey \(1969\)](#) observed, “The physical sciences have learned much by storing up amounts, not just directions. If, for example, elasticity had been confined to.” When you pull on it, it gets longer “Hooke’s law, the elastic limit, plasticity, and many other important topics could not have appeared.”

APPENDIX

Bayes Factor

We can relate *P* values to a minimal Bayes’ likelihood, defined as $e^{-z^2/2}$ ([Table 10.2](#); [Fig. 10.3](#), modified from [Goodman, 1999b](#)).

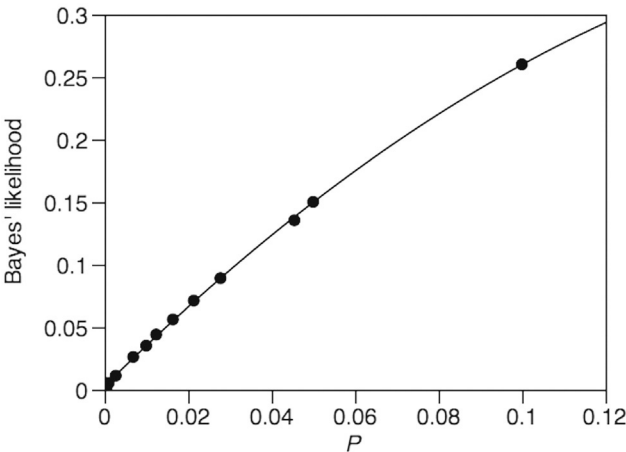


Fig. 10.3 Relationship between *P* and Bayes’ likelihood.

Table 10.2 Bayes factors		
<i>P</i> value (z score)	Minimum Bayes’ likelihood	Strength of evidence
0.10 (1.64)	0.26	Weak
0.05 (1.96)	0.15	Moderate
0.01 (2.58)	0.036	Moderately strong
0.001 (3.28)	0.005	Strong

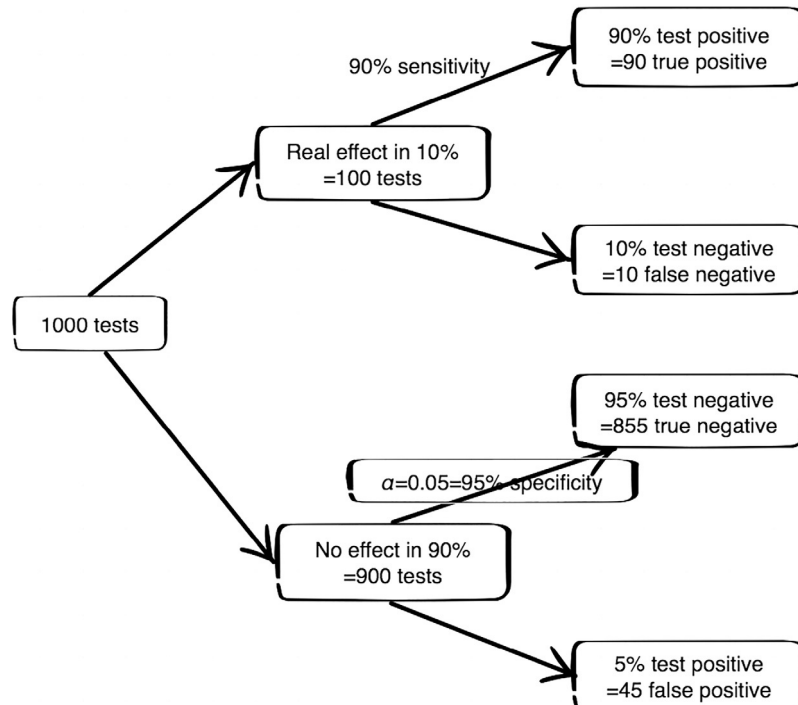
This can be interpreted as showing that when $P = 0.05$, the null hypothesis gets 15% as much support as does the alternative hypothesis. Therefore the evidence against the null hypothesis is not as strong as it might seem, given the alternative hypothesis. Even if $P = 0.01$ the null hypothesis is not clearly rejected. For details, see [Goodman \(1999b\)](#).

Terminology

If the term “statistical significance” has little meaning, how do we convey the idea that the difference between two (or more) groups is not just chance variation? Even if we redefine it, the term “significance” is too tainted and controversial to be retained. One alternative is to use a term to indicate that we are referring to a difference that is likely to be more than chance variation and suggests rejecting the null hypothesis without specifying a specific level of α ; the terms “substantial” and “not due to chance” are appropriate. It is up to the investigator to decide what P value should lead to rejection of the null hypothesis. This level is often set at 0.05, although arguments presented before suggest that 0.001 is safer.

Type I Error and α

Why can the Type I error be so much greater than α ? An example based on Fig. 2 from [Colquhoun \(2014\)](#) provides an explanation.



The assumptions here are that 1000 tests are done to compare control and experimental groups. Of these, 10% or 100 have differences that are large and real (i.e., confirmed by subsequent work) whereas 90% or 900 have differences that are due to chance variation. Consider the 100 real differences. Because almost all tests can produce false results, assume that these tests are 90% sensitive (Chapter 21), that is, they will identify the real difference in 90 (true positives) and mistakenly fail to reject the null hypothesis in 10 (false negatives). Then consider the 900 chance differences. If $\alpha = 0.05$, then 5% or 45 of these tests will reject the null hypothesis; these are false positives. The remaining 95% or 855 will be correctly identified as unable to reject the null hypothesis; these are true negatives. Note that $\alpha = 0.05$ is equivalent to a specificity of 95%—see Chapter 21. Therefore in our 1000 tests there are 45 false positives out of a total of $45 + 90 = 135$ positives, for a percentage of 33% error.

The actual results will depend on the prevalence of real differences and the sensitivity and specificity, but the Type I error will always exceed α . The current usage that expects 5% Type I errors if $\alpha = 0.05$ applies only to the lower arm of the diagram. In practice, however, we do not know if we have real or null differences, and so must include the upper arm as well.

If $\alpha = 0.01$, then the Type I error will be 9/99 or 9.1%. If $\alpha = 0.001$, then the Type I error will be 1/91 \approx 1%. If $\alpha = 0.0048$, the Type I error is \approx 5%.

Colquhoun gives alternate derivations for this error, and his results are similar to those obtained by others by completely different methods.

REFERENCES

- Barnard, G.A., 1992. Review of statistical inference and analysis: Selected correspondence of R. A. Fisher by J T Bennett. *Stat. Sci.* 7, 5–12.
- Benjamin, D.J., et al., 2017. Redefine Statistical Significance. *PsyArXiv*, July 22.
- Binder, A., 1963. Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychol. Rev.* 70, 107–115.
- Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein, H.R., 2009. *Introduction to Meta-Analysis*. John Wiley & Sons, Chichester.
- Cohen, J., 1994. The Earth Is Round ($p < 0.05$). *Am. Psychol.* 49, 997–1003.
- Colquhoun, D., 1971. *Lectures on Biostatistics*. Clarendon Press, Oxford.
- Colquhoun, D., 2014. An investigation of the false discovery rate and the misinterpretation of p-values. *R Soc Open Sci* 1, 140216 (15 pages).
- Crease, R.P., 2010. *Discovery with Statistics*. Available: <http://physicsworld.com/cws/article/indepth/43309>.
- Cumming, G., 2008. Replication and p intervals: P values predict the future only vaguely, but confidence intervals do much better. *Perspect. Psychol. Sci.* 3, 286–300.
- Cumming, G., Fidler, F., Kalinowski, P., Lai, J., 2011. The statistical recommendations of the American Psychological Association Publication Manual: effect sizes, confidence intervals, and meta-analysis. *Austral. J. Psychol.* 177, 7–11.
- Cumming, G., Fidler, F., Lai, J., 2012. Association Publication Manual: effect sizes, confidence intervals, and meta-analysis. *Austral. J. Psychol.* 64, 138–146.
- Deming, W.E., 1943. *Statistical Adjustment of Data*. John Wiley & Sons, Inc., New York.

- Demortier, L., 2008. P values and nuisance parameters. In: Prosper, H.B., Lyons, L., De Roeck, A. (Eds.), PHYSTAT LHC Workshop on Statistical Issues for LHC Physics, 2007. CERN, Geneva.
- Feller, W., 1969. Are life scientists overawed by statistics? *Sci. Res.* 4, 24–29.
- Fisher, R.A., 1956. *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh.
- Frost, J., 2014. Five guidelines for using P values. <http://blog.minitab.com/blog/adventures-in-statistics-2/five-guidelines-for-using-p-values>.
- Goodman, S.N., 1993. P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am. J. Epidemiol.* 137, 485–496 (discussion 497–501).
- Goodman, S.N., 1999a. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann. Intern. Med.* 130, 995–1004.
- Goodman, S.N., 1999b. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann. Intern. Med.* 130, 1005–1013.
- Goodman, S., 2008. A dirty dozen: twelve p-value misconceptions. *Semin. Hematol.* 45, 135–140.
- Goodman, S.N., Royall, R., 1988. Evidence and scientific research. *Am. J. Public Health* 78, 1568–1574.
- Hubbard, R. & Bayarri, M. J. 2003. *P-values are not error probabilities* [online]. Available: <http://ftp.stat.duke.edu/WorkingPapers/03-26.pdf>.
- Ioannidis, J.P., 2005. Why most published research findings are false. *PLoS Med.* 2, e124.
- Lew, M.J., 2012. Bad statistical practice in pharmacology (and other basic biomedical disciplines): you probably don't know P. *Br. J. Pharmacol.* 166, 1559–1567.
- Matthews, D.E., Farewell, V.T., 1996. *Using and Understanding Medical Statistics*. Karger, Basel.
- Moran, J.L., Solomon, P.J., 2004. A farewell to P-values. *Crit. Care Resusc.* 6, 130–137.
- Motulsky, H.J., 2015. Common misconceptions about data analysis and statistics. *Br. J. Pharmacol.* 172, 2126–2132.
- Poole, C., 2001. Low P-values or narrow confidence intervals: which are more durable? *Epidemiology* 12, 291–294.
- Schuemie, M.J., Ryan, P.B., Dumouchel, W., Suchard, M.A., Madigan, D., 2014. Interpreting observational studies: why empirical calibration is needed to correct p-values. *Stat. Med.* 33, 209–218.
- Sterne, J.A., Davey Smith, G., 2001. Sifting the evidence—what's wrong with significance tests? *BMJ* 322, 226–231.
- Tukey, J.W., 1969. Analyzing data: Sanctification or detective work? *Am. Psychol.* 24, 83–91.
- Tukey, J.W., 1991. The philosophy of multiple comparisons. *Stat. Sci.* 6, 1.
- Wasserstein, R. L. & Lazar, N. A. 2016. The ASA's statement on p-values: context, process, and purpose. *Amer Stat*, 70, 129–133.
- Weinberg, C.R., 2001. It's time to rehabilitate the P-value. *Epidemiology* 12, 288–290.
- Ziliak, S.T., 2008. Guinnessometrics: The economic foundation of “Student's” t. *J Econ Perspect.* 22, 199–216.
- Ziliak, S.T., 2011. Matrixx v. Siracusano and Student v. Fisher. Statistical significance on trial. *Significance* 8, 131–134.

CHAPTER 11

Hypothesis Testing: Sample Size, Effect Size, Power, and Type II Errors

BASIC CONCEPTS

Statistical Power

If the null hypothesis is rejected when it is true, we have committed a Type I error, with a probability symbolized by α . On the other hand, accepting the null hypothesis of no difference between two means if they really are from different populations produces a Type II error with a probability symbolized by β .

Consider two populations of means, each population having different grand means, but the normal curves characterizing the distributions of those means overlap (Fig. 11.1).

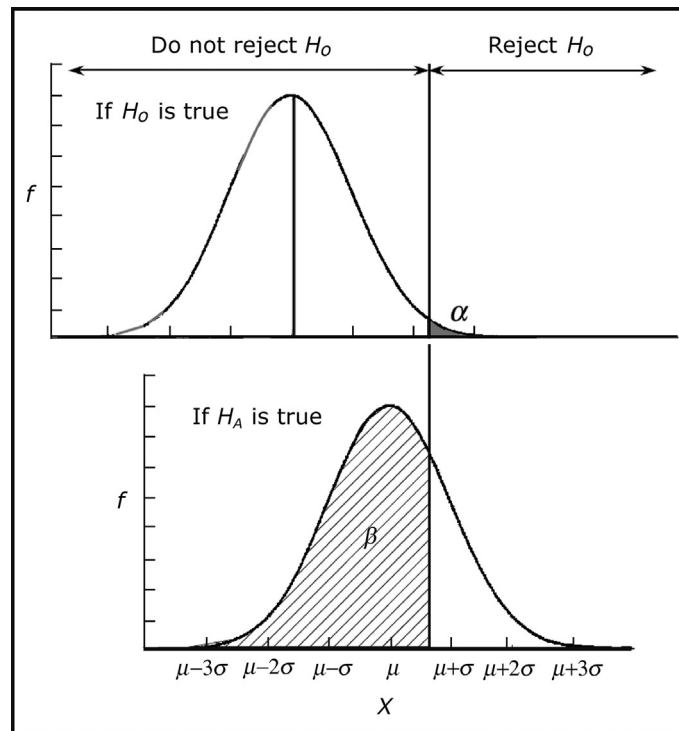


Fig. 11.1 Illustration of Type II error. The Type I error α is the *black* area in the upper panel and is constant and determined by the investigator. The Type II error β is cross-hatched in the lower panel and varies with the value of H_A .

If H_0 is true (i.e., that the two means come from the same population) then a single sample mean falling to the right of the heavy vertical line is unlikely and can lead to rejection of the null hypothesis; the probability of falsely rejecting the null hypothesis is the Type I error, symbolized by α . (Note, however, the criticisms in [Chapter 10](#) about how much the Type I error is.) If, however, H_A is true, as shown in the lower part of the diagram, then $>50\%$ of the time the single sample mean will fall to the left of the heavy vertical line and the null hypothesis would not be rejected at level α . This is the Type II error. The chance of making a Type II error (the probability of accepting the null hypothesis if it is false) is symbolized by β . If the chances of making a Type II error are 0.67 then the chances of not making a Type II error are $1 - 0.67 = 0.33$; there is a 33% probability of making the correct assumption that there are two different groups. This value, $1 - \beta$, is known as the power of the test, that is, its ability to reject the null hypothesis correctly. The power is lowest when the means are close together and highest when they are farthest apart ([Fig. 11.2](#)). Calculating the power of a test in advance is essential, because as pointed out by [Ellis \(2016\)](#) an underpowered study is one designed to fail.

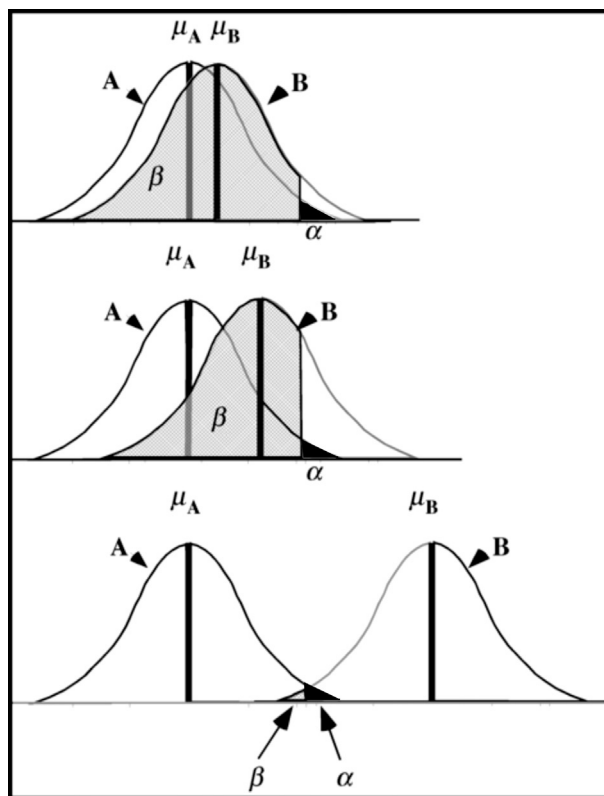


Fig. 11.2 Change in the Type II error as H_A diverges from H_0 .

In this diagram, A is a reference curve showing the distribution of means of samples of size N from a population. B is similar curve derived from means of samples also of size N from a different population. The black triangle indicates the upper tail that includes 0.025 of the A distribution. A decision to accept or reject the null hypothesis that the mean of B differs from the mean of A depends on whether a *single sample mean* from B falls to the right (reject null hypothesis and state that there are probably two different distributions) or left (do not reject null hypothesis that the sample comes from the A distribution) of the shaded area.

In the upper panel, the sample means are close together, and there is about a 90% chance of accepting the null hypothesis as shown by the cross-hatched marking of curve B. Because the null hypothesis is actually wrong, we have committed a Type II error with a probability of about 0.9. In the middle panel, with the two distribution means farther apart, about two-thirds of the B samples fall to the left of the decision line, so that the chances of making a Type II error are about 0.67. In the bottom panel, with the distribution means far apart, there is only a 0.025 chance that the sample from B will be considered as coming from the A distribution, that is, the Type II error is about 0.025. The calculation of power is given in the [Appendix](#).

Effect Size

The effect size is the quantity being measured, whether it is a mean value, a difference between means, a regression coefficient, and so on. Although effect size is not emphasized in text books, it should be regarded as more important than P values ([Cohen, 1990](#); [Coe, 2002](#)).

As [Fig. 11.2](#) shows, the power of the test is closely related to the difference between the means, either between two sample means, or one sample mean and a population mean. The effect size may be classified as:

- a. Absolute effect size, sometimes symbolized as Δ . If in treating a group of patients with hypertension the pressure falls by an average of 10 mmHg, then $\Delta = 10$ mmHg. What we do with the information depends on how important that effect size is. It is unlikely that a pharmaceutical company would spend millions of dollars to produce a new anti-hypertensive drug with an effect size of 3 mmHg, but they might do so for an effect size of 15 mmHg.
- b. Relative effect size, usually symbolized as δ , is the actual effect size relative to the standard deviation. Therefore $\delta = \frac{\Delta}{\sigma}$, where σ is a general estimate of variability that we approximate by the sample values. (Not all texts use these symbols as defined before, and sometimes δ represents the absolute difference.) Relative effect size is used in determining how many measurement or subjects will be needed for an adequate study (see later).

There are several slightly different formulas for calculating relative effect size that take into account the variability of the two (or more) groups that are to be compared. Cohen's d and Hedges' g are the two most often used ([Durlak, 2009](#)) (see [Appendix](#)).

Effect size can be calculated online at <http://www.polyu.edu.hk/mm/effect-sizefaq/calculator/calculator.html>, <https://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-SMD11.php>, https://www.psychometrica.de/effect_size.html, and <http://georgebeckham.com/2016/cohens-d-and-hedges-g-excel-calculator/>.

Despite its importance, effect size is seldom reported in publications (Ellis, 2016).

It is not enough to calculate effect size, but the investigators should put it in context and explain its importance. A small effect size that affects a very large number of people may make a big difference; for example, aspirin may have a small effect size in reducing the number of fatal heart attacks, but when multiplied by the number of people at risk it has a big effect on public health.

How can we reduce the Type II error? For a given biological system the difference between the means of the two populations is determined by the system and not subject to manipulation except perhaps by selecting subgroups and increasing homogeneity. What we can do is to increase the sample size, with the effect shown in Fig. 11.3.

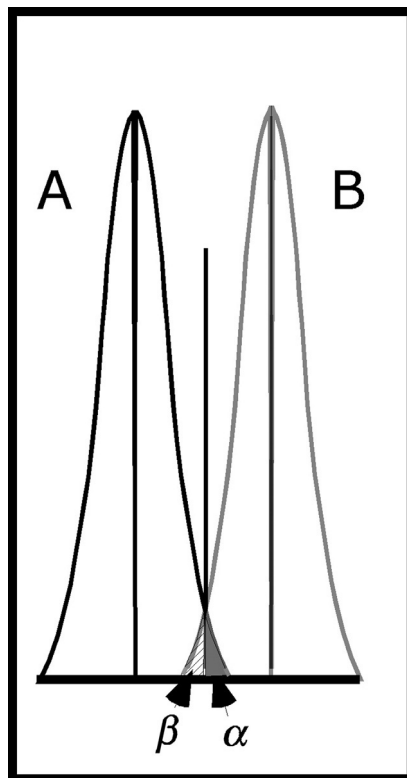


Fig. 11.3 Effect of markedly increasing sample size. For a mean difference similar to that in the middle panel of Fig. 11.7, the Type II error has been reduced from 0.67 to 0.025.

There remains the question of how much to increase the sample size so that the power of the test is high. In theory we would like a power of 0.9, but in practice often settle for 0.8. Whether we can achieve this increased sample size depends on the availability of samples, and the cost and manpower needed to obtain them. In principle, solve the equation

$$t_{0.05} = \frac{\bar{X}_1 - \bar{X}_2}{\frac{S_{\bar{X}_1 - \bar{X}_2}}{\sqrt{N}}}$$

(or whatever other value we want for α) for N by assigning the critical value of t , the difference between the means (the desired effect size), and the standard deviation. From our knowledge of previous studies in a particular field or a pilot study we guess the standard deviation of the population. This can be very wrong, with misleading sample size calculations as a result. It is therefore best to take any sample size calculations as only approximations, and wise to plan for a larger sample size (Schulz and Grimes, 2005).

With an estimated standard deviation decide what difference between means would be important, and calculate the relative effect size δ :

$$\delta = \frac{\bar{X}_1 - \bar{X}_2}{s}.$$

Use the sample standard deviation, not the standard deviation of the mean.

Then use Tables in which the number of subjects is listed for given values of δ , α , and $1 - \beta$ (Beyer, 1966; Cohen, 1988; Kraemer, 1988). Some publications give nomograms to determine these numbers (Gore and Altman, 1982). Alternatively, there are computer programs to make the calculation. An extensive interactive freeware program is termed G*Power at http://download.cnet.com/G-Power/3000-2053_4-141879.html and is very useful for calculating power for all types of statistical tests. Power calculations in the protocol are required by most granting agencies to show that the proposed study is feasible in terms of subjects, time, and money. Another extensive program (Macintosh and Windows) is available at <http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize>. Other simpler programs online are <http://www.statisticalsolutions.net/pssZtest_calc.php>, http://hedwig.mgh.harvard.edu/sample_size/quant_measur/assoc_quant.html, and <http://www.sample-size.net>.

A readable discussion of the steps needed for calculating sample size is provided by Flight and Julious (2016) who have created an application called SampSize for the iPhone and iPad (free from the Apple Store).

Lehr (1992) pointed out that for $\alpha = 0.05$, and power $(1 - \beta) = 0.80$, sample size can be closely approximated by a simple relationship

$$n = \frac{ks^2}{\delta^2},$$

where $k = 8$ for a paired t -test and 16 for an unpaired t -test (see Chapter 22) and δ is the difference to be detected (effect size). This produces numbers very close to the exact number from the more complex calculation; because we have to guess at the value of s , a simple formula seems preferable. For other values of α and β there is a simple table (Table 11.1).

Table 11.1 Values of k for the Lehr equation

	α (Two sample)			α (One sample)
Power $1 - \beta$	0.01	0.05	0.10	0.05
0.80	23.5	16	12.5	8
0.90	30	21	17.5	11
0.95	36	26	22	13
0.975		31		16

Problem 11.1 Determine the sample sizes needed for determining a mean change in myocardial blood flow from 1 to 1.3 mL/g min (paired samples) if the standard deviation is 0.4 with $\alpha = 0.05$ and power of 0.8, 0.85, or 0.9.

Problem 11.2 Repeat the calculations if standard deviation is 0.64 mL/g min.

The Type I error of falsely rejecting the null hypothesis, as previously stated, has nothing to do with the importance of any difference, but is more an issue of consistency of data and the comfort that we feel in deciding to reject the null hypothesis. The degree of certainty is under our control, and we can make the requirement as stringent as we please. On the other hand, the Type II error is more insidious. If we do not have enough power and decide to accept the null hypothesis we may neglect a difference that might be important. If an intervention doubles flow to an ischemic region of the myocardium but because of lack of power of the test we cannot reasonably reject the null hypothesis of no effect, we might be induced to ignore a very useful intervention. That is why if we cannot reject the null hypothesis it is better to regard the effect of the intervention as unproven rather than nonexistent. In fact, Williams et al. (1997) found that failure to achieve a high enough power was the most common statistical error made in publications in the American Journal of Physiology, often because investigators were unaware of the need to assess the power of their negative results. This subject is so important that readers should also go to excellent explanatory writings by Berkowitz that can be downloaded from www.columbia.edu/~mvp19/RMC/M6/M6.doc

A study drawing attention to this problem in Medicine was by [Freiman et al. \(1978\)](#) who examined 71 randomized trials that compared the effects of two drugs or treatments and in which the authors concluded that there was no statistically significant difference between them ([Fig. 11.4](#)).

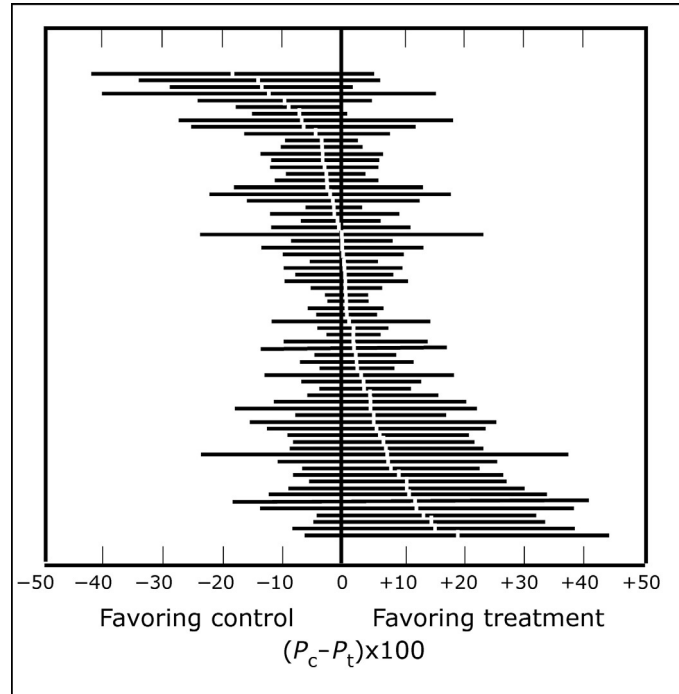


Fig. 11.4 Trials comparing two therapeutic agents. Each *horizontal line* represents one clinical trial. The *white central dot* is the mean and the *black bars* demarcate 90% confidence limits. The *thick vertical black line* indicates no difference. Because the confidence limits include zero difference, P was >0.05 in all these tests. Almost all the mean differences (effect size) are $<15\%$. (Modified from [Freiman et al.](#) by *arranging mean differences in order.*)

They showed that in many of those studies the responses were quite large, but the sample sizes were too small to show a 25% difference between the two treatments, let alone a 50% difference between them. In some instances, this led the investigators to discontinue studying the new treatment and to conclude that it was of no benefit. This is undesirable; a 25% improved cure rate in any disease would be very welcome. This error has been present in many studies with negative results: randomized clinical trials ([Moher et al., 1994](#); [Burbach et al., 1999](#); [Bedard et al., 2007](#); [Tsang et al., 2009](#)), emergency medicine ([Brown et al., 1987](#)), neuroscience ([Button et al., 2013](#); [Nieuwenhuis et al., 2011](#); [Weaver et al., 2004](#)), surgery ([Chung et al., 1998](#); [Freedman et al., 2001](#)). All these studies showed that sample sizes and power were too small to detect a 25%–50% change in outcomes. The cost in time and money of these inadequate studies must have been enormous.

It is possible to make the Type I error as small as you like, for example, reducing it from 0.05 to 0.01 (or even smaller) so that the risk of falsely rejecting the null hypothesis becomes very small indeed. However, the cost of making the Type I error smaller is making the Type II error bigger (Fig. 11.5).

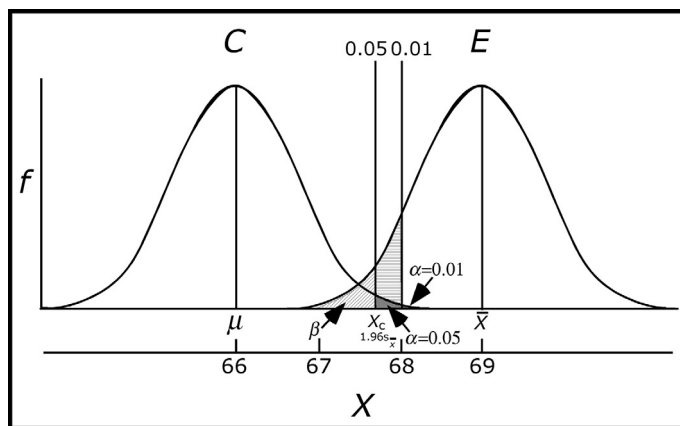


Fig. 11.5 Relation between Type I and II errors.

If for curve *C* the Type I error is set at 0.05, with 0.025 in each tail (solid shaded area), then the Type II error is shown by the cross-hatched area for curve *E*. If the Type I error is made 0.01, with 0.005 in each tail of curve *C*, then the Type II error has increased to include the horizontally shaded area under curve *E*.

Posthoc Power Analysis

Power analysis may be used in several ways. It is best used in planning experiments and is now required by most grant agencies. The investigator decides on the effect size desired, a value for the Type I error α (0.05–0.001), and a value for the power that is $1 - \beta$, the Type II error, and is usually set at 0.8–0.9. Then the number to be used is determined from tables or programs. These numbers are provisional, depending on preliminary observations of means and standard deviations. Julious and Owen (2006) observed that if a variance obtained in a previous study was based on small sample sizes and used as if it were the population variance, its use in the standard formulas could underestimate the future sample size needed to achieve a given power. They provide tables of corrections, but in general it is safe to increase the predicted sample size by about 20%–30%. If some patients or animals are expected to leave the study prematurely, even bigger numbers will be needed. Sometimes, however, there are no previous studies to provide data, and a pilot study might need to be done. A sample size of 12 in each group may be adequate (Julious, 2005, van Belle, 2002).

How should we think about results in which the null hypothesis could not be rejected, with $\alpha > 0.05$, but with a sample size too small to provide adequate power? Many statistical programs allow posthoc calculations of power, but this may not be the best way to assess the data (Kraemer, 1988; Williams et al., 1997; Sterne and Davey Smith, 2001). In fact, by definition a P value > 0.05 indicates that the power was too low to reject the null hypothesis for that effect size. Investigators have argued that a “nonsignificant” result indicates either too small a standardized difference or too small a sample size, and results might or might not be important. In place of the posthoc power analysis, Walters (2009) suggested using confidence intervals to help distinguish statistical significance from clinical importance (Fig. 11.6).

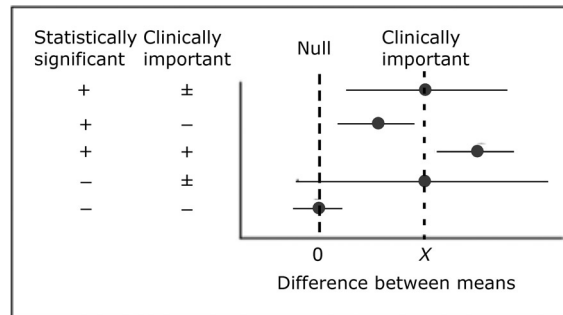


Fig. 11.6 Distinction between statistical significance and importance, showing a useful way of evaluating a “nonsignificant” result by plotting means and confidence limits. + yes; – no; ± possible. Any one of these rows presents information more useful than merely stating that the power was low. (Based on figure published by Walters, S.J., 2009. Consultants’ forum: should post hoc sample size calculations be done? *Pharm. Stat.* 8, 163–9.)

APPENDIX

1. Take care when evaluating relative effect size, because there are different ways of calculating it.

Cohen’s d is $\frac{\overline{X}_1 - \overline{X}_2}{s_p}$, where s_p is the pooled standard deviation

$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2}}$. Hedge’s g is similar. It is $g = \frac{\overline{X}_1 - \overline{X}_2}{s_p}$, where

$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$. To correct for bias the effect size is calculated as

$d_{\text{corr}} = g \left(1 - \frac{3}{4(N_1 + N_2) - 9} \right)$, also written as $d_{\text{corr}} = d \left(1 - \frac{3}{4df - 1} \right)$. If the two sample standard deviations are very different, some people prefer to use the control standard

deviation as the denominator. In a paired design, d for the difference may be related to either the control (or pretest) standard deviation or else to the average of the two standard deviations (Cumming et al., 2012). Always describe which form of d and which denominator you are using.

2. If the null hypothesis (μ_0) is correct, the t distribution is symmetrical, but when an alternative hypothesis is selected ($\mu_A \neq \mu_0$) the t distribution is asymmetrical, the asymmetry increasing as $\mu_A - \mu_0$ (the noncentrality parameter) increases. This is termed the noncentral t distribution. Therefore the confidence limits are asymmetrical, although the differences are not marked except for very small sample sizes. The area under the curve for any value of t can be determined online from <http://www.danielsoper.com/statcalc3/calc.aspx?id=91>, and the confidence limits can be obtained from <http://keisan.casio.com/exec/system/1180573219%3e>.

The reasons for the asymmetrical noncentral t are given by Cumming and Finch (2001).

Calculation of Power

We can calculate the power for any values of μ and \bar{X} if we know N and an estimate of σ derived from the sample standard deviation. As a rule, this is done by tables, formulas, or programs, but the following example shows how power is calculated. Consider the heights of adult European males in the 18th century (Komlos and Cinnirella, 2005) with a mean of 66" and a standard deviation of 4.47" (curve C in Fig. 11.7). We draw at random a group of 30 subjects today and wish to determine if their mean height is consistent with the 18th century population sample. If we drew several samples we might get the distribution of means in curve E in Fig. 11.7.

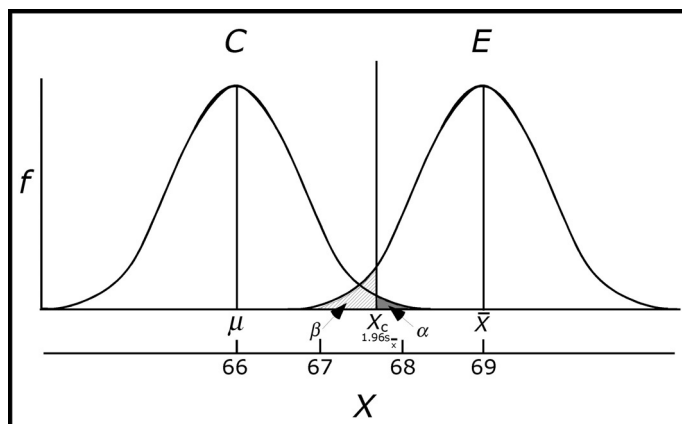


Fig. 11.7 Diagram for power calculation. C is the distribution of the control population means around a "population" mean $\mu = 66$, E is the distribution of our sample means around a single observed sample mean $\bar{X} = 69$. The shaded area to the right of X_c for the C distribution (H_0) is the Type I error, the cross-hatched area to the left of X_c is the Type II or β error, and the total area to the right of X_c for the E distribution (H_A) is the power $1 - \beta$.

The shaded area to the right of the vertical line X_c at $1.96s_{\bar{X}}$ represents $\alpha/2$, or 0.025 of the area under curve C , and represents α , the Type I error. Reject the null hypothesis with $\alpha = 0.05$ if

$$z - \frac{\bar{X} - \mu}{s_{\bar{X}}} \geq 1.96.$$

This can be rearranged to give

$$\bar{X} - \mu > z_{\alpha/2}s_{\bar{X}}.$$

Rearrangement and substitution gives

$$\bar{X} \geq 1.96 \times \frac{4.47}{\sqrt{30}} + 66 \text{ or } \bar{X} \geq 67.60.$$

Therefore any mean value for height in the series of 30 subjects in excess of a critical value of 67.60" leads to rejection of the null hypothesis, and we conclude that mean heights of adult males today are probably $>66''$.

Now consider what happens if HA is true. The cross-hatched area to the left of the critical line X_c is the Type II error, (β) and this is given by 1–area under curve E to the right of X_c . We can calculate this error if we know how many standard error units \bar{X} is from μ .

Consider the true mean difference $\bar{X} - \mu$:

$$\delta = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}} = \frac{\bar{X} - 66}{\frac{4.47}{\sqrt{30}}} = \frac{\bar{X} - 66}{0.8161}.$$

Therefore if \bar{X} is 69, the previous equation becomes

$$\delta = \frac{69 - 66}{0.8161} = 3.676.$$

In other words, if HA is true, a mean value of 69" is 3.676 standard errors from the "population" mean of 66". But X_c is 1.96 standard errors from 66", so the area between X_c and 69" is represented by $z = 3.676 - 1.96 = 1.716$. This is z_b , which represents how many standard error units X_c is below \bar{X} . This area referred to $z = 1.716$ is 0.0431. If HA , the sample mean of 69", is true, there is 0.0431 chance of rejecting the alternative hypothesis. The power is $1 - \beta = 0.9568$.

More generally,

$$z_{\beta} = \delta - z_{\alpha}.$$

Simplify the calculation by using the relationship

$$z_{\beta} = \delta - z_{\alpha} = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}} - \frac{X_c - \mu}{\frac{s}{\sqrt{N}}} = \frac{\bar{X} - X_c}{\frac{s}{\sqrt{N}}}$$

Thus $z_{\beta} = \frac{69-67.6}{0.8161} = 1.7155$, as shown before (with slight difference due to rounding off).

This discussion implies using a 2-tailed test for z , so that $z = 1.96$ includes 0.025 of the area under curve C at each end. To use a 1-tailed test with 0.05 of the area under curve C at one end, use $z = 1.645$ to calculate the β error. (If the sample size is <100 , use the t table instead of the z table, but as long as $N > 10$ the error is under 0.01 (Zar, 2010). It is not worth worrying about this small error in view of the fact that we are guessing at the standard deviation.

If we compare two independent samples, then the standardized deviation δ is

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{2\sigma^2}{N}}},$$

and proceed as before.

REFERENCES

- Bedard, P.L., Krzyzanowska, M.K., Pintilie, M., Tannock, I.F., 2007. Statistical power of negative randomized controlled trials presented at American Society for Clinical Oncology annual meetings. *J. Clin. Oncol.* 25, 3482–3487.
- Beyer, W.H., 1966. Handbook of Tables for Probability and Statistics. The Chemical Rubber Company, Cleveland, OH.
- Brown, C.G., Kelen, G.D., Ashton, J.J., Werman, H.A., 1987. The beta error and sample size determination in clinical trials in emergency medicine. *Ann. Emerg. Med.* 16, 183–187.
- Burback, D., Molnar, F.J., St John, P., Man-Son-Hing, M., 1999. Key methodological features of randomized controlled trials of Alzheimer's disease therapy. Minimal clinically important difference, sample size and trial duration. *Dement. Geriatr. Cogn. Disord.* 10, 534–540.
- Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14, 365–376.
- Chung, K.C., Kallianen, L.K., Hayward, R.A., 1998. Type II (beta) errors in the hand literature: the importance of power. *J. Hand Surg. [Am]* 23, 20–25.
- Coe, R., 2002. It's the effect size, stupid. What effect size is and why it is important. Available, <http://www.leeds.ac.uk/educol/documents/00002182.htm>.
- Cohen, J., 1988. Statistical Power Analysis for Behavioral Sciences. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Cohen, J., 1990. Things I have learned (so far). *Am. Psychol.* 45, 1304–1312.
- Cumming, G., Finch, S., 2001. A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. *Educ. Psychol. Meas.* 61, 532–574.
- Cumming, G., Fidler, F., L'ai, J., 2012. Association Publication Manual: effect sizes, confidence intervals, and meta-analysis. *Austral J Psychol* 64, 138–146.
- Durlak, J.A., 2009. How to select, calculate, and interpret effect sizes. *J. Pediatr. Psychol.* 34, 917–928.
- Ellis, P.D., 2016. The Essential Guide to Effect Sizes. Cambridge University Press, Cambridge.

- Flight, L., Julious, S.A., 2016. Practical guide to sample size calculations: an introduction. *Pharm. Stat.* 15, 68–74.
- Freedman, K.B., Back, S., Bernstein, J., 2001. Sample size and statistical power of randomised, controlled trials in orthopaedics. *J. Bone Joint Surg. Br.* 83, 397–402.
- Freiman, J.A., Chalmers, T.C., Smith Jr., H., Kuebler, R.R., 1978. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *N. Engl. J. Med.* 299, 690–694.
- Gore, S.M., Altman, D.G., 1982. *Statistics in Practice*. Devonshire, Torquay, UK.
- Julious, S.A., 2005. Sample size of 12 per group rule of thumb for a pilot study. *Pharm. Stat.* 4, 287–291.
- Julious, S.A., Owen, R.J., 2006. Sample size calculations for clinical studies allowing for uncertainty about the variance. *Pharm. Stat.* 5, 29–37.
- Komlos, J., Cinnirella, F., 2005. European heights in the early 18th century. Available: <http://epub.ub.uni-muenchen.de>.
- Kraemer, H.C., 1988. Sample size: when is enough enough? *Am. J. Med. Sci.* 296, 360–363.
- Lehr, R., 1992. Sixteen s-squared over d-squared: a relation for crude sample size estimates. *Stat. Med.* 11, 1099–1102.
- Moher, D., Dulberg, C.S., Wells, G.A., 1994. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* 272, 122–124.
- Nieuwenhuis, S., Forstmann, B.U., Wagenmakers, E.J., 2011. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nat. Neurosci.* 14, 1105–1107.
- Schulz, K.F., Grimes, D.A., 2005. Sample size calculations in randomised trials: mandatory and mystical. *Lancet* 365, 1348–1353.
- Sterne, J.A., Davey Smith, G., 2001. Sifting the evidence—what’s wrong with significance tests? *Br. Med. J. (Clin Res Ed)* 322, 226–231.
- Tsang, R., Colley, L., Lynd, L.D., 2009. Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized controlled trials. *J. Clin. Epidemiol.* 62, 609–616.
- Van Belle, G., 2002. *Statistical Rules of Thumb*. Wiley Interscience, New York.
- Walters, S.J., 2009. Consultants’ forum: should post hoc sample size calculations be done? *Pharm. Stat.* 8, 163–169.
- Weaver, C.S., Leonardi-Bee, J., Bath-Hextall, F.J., Bath, P.M., 2004. Sample size calculations in acute stroke trials: a systematic review of their reporting, characteristics, and relationship with outcome. *Stroke* 35, 1216–1224.
- Williams, J.L., Hathaway, C.A., Kloster, K.L., Layne, B.H., 1997. Low power, type II errors, and other statistical problems in recent cardiovascular research. *Am. J. Physiol.* 273, H487–H493 (Heart and Circulation Physiology 42).
- Zar, J.H., 2010. *Biostatistical Analysis*. Prentice Hall, Upper Saddle River, NJ.

SECTION IV

Discrete Distributions

CHAPTER 12

Permutations and Combinations; Logarithms

PERMUTATIONS

Consider selecting k objects from a set of n objects and arranging them in order; we can select all the objects in the population ($k = n$) or a subset of them. Such an arrangement is a permutation of n objects, taken k at a time. For example, a license plate with 7 characters (numbers, letters) is a permutation of length 7 out of 36 characters (0, 1, ..., 9, A, B, ... Z). The order of the objects is important. License plate 2NNP386 is not the same as license plate 2PNN836, even though they use the same characters. The number of possible permutations depends on whether the same selection can appear more than once (“with replacement” into the selection pool), or is excluded from future selection (“without replacement,” or withdrawn from the selection pool).

For the license plate, both letters and numbers can be selected with repetition. The number of possible permutations of n objects, taken k at a time, is nk .

Proof: The first position can be filled in n ways. The second position can be filled in n ways, and because these choices are independent, the first two positions can be filled in $n \times n = n^2$ ways. Similarly, the first three positions can be filled in n^3 ways, and so on, so that the number of possible license plate permutations is $36^7 = 78,364,164,096$.

If the objects are chosen without repetition, then the number of permutations is given by $\frac{n!}{(n-k)!}$.

Proof: The first position can be chosen in n ways, leaving $n - 1$ objects. The second position can therefore be filled in $(n - 1)$ ways, so that the first two positions can be filled in $n(n - 1)$ ways. The third position can be filled in $(n - 2)$ ways, so that the first 3 positions can be filled in $n(n - 1)(n - 2)$ ways. Therefore the first k positions can be filled in $n(n - 1)(n - 2) \dots (n - k + 1)$ ways. This can be rewritten as

$$\begin{aligned} & [n(n - 1)(n - 2) \dots (n - k + 1)] \times \frac{[(n - k)(n - k - 1) \dots 1]}{[(n - k)(n - k - 1) \dots 1]} \\ &= \frac{[n(n - 1)(n - 2) \dots (n - k + 1)] \times [(n - k)(n - k - 1) \dots 1]}{[(n - k)(n - k - 1) \dots 1]} = \frac{n!}{(n - k)!} \\ & [(n! = n(n - 1)(n - 2)(n - 3) \dots (1). 5! = 5 \times 4 \times 3 \times 2 \times 1 = 120).] \end{aligned}$$

If all the objects are chosen, then $k = n$, and the formula reduces to $n!$ because by definition $0! = 1$.

COMBINATIONS

Sometimes the order of the arrangements has no meaning, and we are interested only in how many of each type of object we have. Such an unordered selection is called a combination of n objects taken k at a time.

The number of combinations of k items chosen from a total of n items is less than the number of permutations. Because any k items can be arranged in $k!$ ways, dividing the number of permutations of n objects taken k at a time by the number of permutations of k objects, which is $k!$, gives

$$\frac{n!}{(n-k)!} \text{ divided by } k! = \frac{n!}{k!(n-k)!}. \text{ This is denoted by } \binom{n}{k} \text{ or } nCk.$$

These problems and many variations on them are described well by Ash (1993) and Ross (1984).

Note that $\binom{n}{k} = \binom{n}{n-k}$, because both sides of the equation can be written $\frac{n!}{k!(n-k)!}$.

If we choose a committee of 4 people out of 12 applicants, we can do this in

$$\frac{12!}{4!(12-4)!} = \frac{12 \times 11 \times 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(4 \times 3 \times 2 \times 1)(8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1)} = \frac{12 \times 11 \times 10 \times 9}{4 \times 3 \times 2 \times 1} = 495$$

ways. Therefore there are also 495 ways in which the other 8 applicants can be not chosen. These calculations can be done online at <http://stattrek.com/online-calculator/combinations-permutations.aspx>, <http://www.mathsisfun.com/combinatorics/combinations-permutations-calculator.html>, <http://www.calctool.org/CALC/math/probability/combinations>, <http://www.statisticshowto.com/calculators/permutation-calculator-and-combination-calculator/>, <https://www.calculatorsoup.com/calculators/discretemathematics/permutations.php>, and <http://www.calculatorsoup.com/calculators/discretemathematics/combinations.php>.

In some permutation problems there may be sets of identical objects; for example, A, A, A, B, B, C. To know how many permutations there are, calculate

$$\frac{N!}{n_1!n_2!\dots n_k!},$$

where N is the total, and n_1, n_2, \dots, n_k are the numbers of identical objects n_i in each set over k sets.

For the alphabetic problem before the number of permutations is $\frac{6!}{3!2!1!} = 60$ (Example 12.1).

Example 12.1

How many permutations can be made from the word Tennessee?, $n_1 = 1, n_2 = 4, n_3 = 2,$

$n_4 = 2$, and $N = 9$, so that the number of permutations is $\frac{9!}{1!4!2!2!} = 7560$.

To see the effect of this calculation, consider two five member sets: a,e,i,o,u and b,e,e,b,e. The first of these, with no duplications, has $5! = 120$ permutations. The second has fewer permutations because of the duplications, and there are $5!/(2!3!) = 10$ permutations.

LOGARITHMS

Exponents

$10 \times 10 \times 10$ can be written as 10^3 , where the superscript 3 shows the number of times 10 is multiplied by itself. The superscript is called the exponent.

Logarithms

A logarithm answers the question: $10^? = 1000$.

Because $10^3 = 1000$, we can write $\log_{10} 1000 = 3$. Ten is the base.

We could also ask: $2^? = 8$. Because $2 \times 2 \times 2 = 8$, we can write $\log_2 8 = 3$. The subscript 2 is the base, 8 is the answer, and 3 is the number of times the base must be multiplied by itself to get the answer.

Often the base is the value of $e = 2.71828$. This is written as $\log_e X$ or $\log_e X$ (for natural logarithm).

Antilogarithms

Obtain the antilogarithm of the number from tables or calculators. On the calculators, the antilogarithm symbol is e^x for logarithms to base e and 10^x for logarithms to base 10.

Taking antilogarithms is often called exponentiation.

These conversions may be done online at https://www.rapidtables.com/calc/math/Log_Calculator.html, <http://www.1728.org/logrithm.htm>, <https://www.calculator.net/log-calculator.html>, <https://ncalculators.com/number-conversion/log-logarithm-calculator.htm>, and <https://ncalculators.com/number-conversion/anti-log-logarithm-calculator.htm>. Almost all hand calculators will calculate logarithms and antilogarithms.

WORKED PROBLEMS

1. The dean wishes to select a committee of 5 from a senior faculty of 30 members. In how many ways can this be done?

Answer:

Because the order of selection is not important, we need the combinations.

The answer is $\frac{30!}{5!25!} = 142,506$.

2. 10 of the 30 senior faculty are women. The dean wants to select 3 women and two men. How many combinations are there?

3 women can be selected in $\frac{10!}{3!7!} = 120$ ways.

2 men can be selected in $\frac{20!}{2!18!} = 190$ ways.

These are independent combinations, so that the final set of 5 people can be selected in $120 \times 190 = 22,800$ ways.

3. Of the 10 women, 6 are clinicians and 4 are from basic sciences. Of the 20 men, 15 are clinicians and 5 from basic sciences. How many combinations of 2 female clinicians, 1 female basic scientist, 1 male clinician, and 1 male basic scientist are there?

This is equivalent to selecting 2 female clinicians out of 6, 1 female basic scientist out of 4, 1 male clinician out of 15, and 1 male basic scientist out of 5. The numbers of combinations are, respectively:

$\frac{6!}{2!4!} = 15$, $\frac{4!}{1!3!} = 4$, $\frac{15!}{1!4!} = 15$, and $\frac{5!}{1!4!} = 5$. As these are independent, the total number of combinations is $15 \times 4 \times 15 \times 5 = 4500$.

4. A poker hand consists of any 5 cards from a deck of 52 cards. What are the chances of having the ace of spades?

The number of combinations of 5 cards is $\frac{52!}{5! \times 47!} = 2,598,960$.

The number of combinations of any 4 cards that do not include the ace of spades is $\frac{51!}{4! \times 47!} = 249,900$. Therefore, the number of 5 card hands that include the ace of spades is determined from $\frac{249,900}{2,598,960} = 0.096$, or $\sim 1/10.4$.

This can be calculated more simply as 5 chances of drawing an ace of spades out of 52 cards = $\frac{5}{52} = 0.096$.

Problem 12.1. What are the chances of a flush (all the 5 cards of the same suit)?

Problem 12.2. What are the chances of having a straight, that is, a numerical sequence, no matter what the suits are?

Problem 12.3. What are the chances of getting a straight flush, that is, a numerical sequence of the same suit?

REFERENCES

- Ash, C., 1993. *The Probability Tutoring Book. An Intuitive Course for Engineers and Scientists*. IEEE Press, New York, p. 470.
- Ross, S., 1984. *A First Course in Probability*. Macmillan Publishing Company, New York.

CHAPTER 13

Hypergeometric Distribution

INTRODUCTION

If we select 5 marbles from a jar containing 15 red and 10 green marbles, then the probability that the first marble is red is $15/25 = 0.600$. If that marble is put back in the jar, the jar is shaken, and another marble is picked, the probability that the next marble will be red is still 0.600. On the other hand, if the marble is not replaced, then the jar will contain 14 red and 10 green marbles. The chances that the next marble picked will be red are now $14/24 = 0.5833$. To answer the question “What is the probability of getting 2 red and 3 green marbles from the above population of marbles?”, use the hypergeometric distribution.

GENERAL FORMULA

The formula envisages a total population of N objects, with X having characteristic A (success) and $N - X$ not having it. Choose a sample of n objects from the total population, and ask what is the probability that r objects have A . That is, what is $P(X = r, n)$?

$$P(X = r) = \frac{\binom{X}{r} \binom{N - X}{n - r}}{\binom{N}{n}} = \frac{\left(\frac{X!}{r!(X - r)!} \right) \left(\frac{(N - X)!}{(n - r)!(N - X - n + r)!} \right)}{\frac{N!}{n!(n - N)!}}$$

This is the probability function.

If $N = 25$, $X = 15$, $n = 5$, $r = 2$. Therefore

$$\begin{aligned} P(X = 2, n = 5) &= \frac{\left(\frac{15!}{2!(15 - 2)!} \right) \left(\frac{(25 - 15)!}{(5 - 2)!(25 - 15 - 5 + 2)!} \right)}{\frac{25!}{5!(25 - 5)!}} \\ &= \frac{\left(\frac{15!}{2!13!} \right) \left(\frac{10!}{3!7!} \right)}{\frac{25!}{5!20!}} = 0.2372. \end{aligned}$$

The mean value is

$$\text{Mean} = \frac{nX}{N} = \frac{5 \times 15}{25} = 3$$

and

$$\text{Variance} = \frac{nX(N-X)}{N^2} \left(1 - \frac{n-1}{N-1}\right) = \frac{5 \times 15(25-15)}{25^2} \left(1 - \frac{5-1}{25-1}\right) = 1.$$

The calculation can be done online at <http://stattrek.com/online-calculator/hypergeometric.aspx>, <http://www.adsciengineering.com/hpdcalc/>, <https://www.emathhelp.net/calculators/probability-statistics/hypergeometric-distribution-calculator/?pn=25&pk=15&sn=5&sk=2>, or <https://easycalculation.com/statistics/hypergeometric-distribution.php>. The calculator at <https://calculator.tutorvista.com/hypergeometric-distribution-calculator.html> also gives mean and variance.

Problem 13.1. A chest physician selects at random 6 patients out of a group of 16 patients, 5 of whom are hyperreactors to inhaled particles. What is the probability that the sample contains.

(a) Two hyperreactors? (b) Five hyperreactors?

(Because of the risk of arithmetic errors, always do this type of calculation by online calculator. Some of these calculators supply added information, such as cumulative probabilities.)

FISHER'S EXACT TEST (FISHER-IRWIN TEST)

Fisher's exact test is based on the hypergeometric distribution. Consider sampling a population of size N that has c_1 objects with A and c_2 with not A . Draw a sample of r_1 objects and find a with A . Then

	A	Not A	Total
In sample	a	b	r_1
Not in sample	c	d	r_2
	c_1	c_2	N

There are $\binom{N}{r_1}$ possible samples. Of these, $\binom{c_1}{a}$ is the number of ways of choosing A in a sample of size c_1 , $\binom{c_2}{b}$ is the number of ways of choosing not A in a sample

of size $N - c_1 = c_2$; and because these are independent, there are $\binom{c_1}{a} \binom{c_2}{b}$ ways of choosing a A_s and b not A_s . Therefore the probability of choosing a

$$A_s = \frac{\binom{c_1}{a} \binom{c_2}{b}}{\binom{N}{r_1}} = \frac{\frac{c_1!}{a!c!} \times \frac{c_2!}{b!d!}}{\frac{N!}{r_1!r_2!}} = \frac{c_1!c_2!r_1!r_2!}{N!a!b!c!d!};$$

the last form is the way in which the Fisher's

exact test formula is usually given. Doing this calculation for all combinations more extreme than the one presented produces the probability given in Fisher's exact test (Example 13.1).

Table 13.1 Basis of Fisher's test

	Women	Men	Total
In sample	3	2	5
Not in sample	17	8	25
	20	10	30

Example 13.1

A medical clinic has 30 patients, 20 women and 10 men. A random sample of 5 patients is drawn. What is the probability that there will 2 men?

A sample of 5 patients out of 30 can be chosen in $\binom{30}{5}$ ways = 142,506 ways.

A sample of 2 men and 3 women can be drawn in $\binom{10}{2} \times \binom{20}{3}$ ways = 51,300 ways.

$$\text{Therefore } P(2 \text{ men, 3 women}) = \frac{\binom{10}{2} \times \binom{20}{3}}{\binom{30}{5}} = 51,300/142,506 = 0.359985.$$

Alternatively (Table 13.1), the probability in Fisher's exact test = 0.359985.

This test can be done online at <http://www.danielsoper.com/statcalc3/calc.aspx?id=29> or <http://www.quantitativeskills.com/sisa/statistics/fisher.php?n11=3&n12=2&n21=17&n22=8>. The probability of 0.359985 is the probability of that particular distribution given the marginal totals. Often we need to know the probability of that distribution and more extreme disparities (Chapter 15), and <http://www.quantitativeskills.com/sisa/statistics/fisher.php?n11=3&n12=2&n21=17&n22=8> and <http://www.langsrud.com/fisher.htm> give these accumulated probabilities as well.

Fisher's exact test can be performed for 2×3 , 2×4 , and $3 \times$ tables at <http://www.vassarstats.net> (see frequency distribution).

MULTIPLE GROUPS

More than two groups can be involved (Hogg and Tanis, 1977). If there are $n_1, n_2, n_3, \dots, n_k$ objects in each class, and $n_1 + n_2 + n_3 + \dots + n_k = n$, then $P(X_1 = x_1, X_2 = x_2, \dots,$

$$X_n = x_n) = \frac{\binom{n_1}{x_1} \binom{n_2}{x_2} \dots \binom{n_k}{x_k}}{\binom{n}{r}}, \text{ where } x_1 + x_2 + x_3 + \dots + x_k = r.$$

REFERENCE

Hogg, R.V., Tanis, E.A., 1977. Probability & Statistical Inference. Macmillan Publishing Company, Inc., New York.

CHAPTER 14

Categorical and Cross-Classified Data: Goodness of Fit and Association

BASIC CONCEPTS

Introduction

Members of a population may be classified into different categories: for example, improved, the same, or worse after treatment. The frequency with which members of the sample representing the population of interest occurs in each category can be determined, so that each category shows a count (Table 14.1).

Table 14.1 Table of categories

	Improved	Unchanged	Worse	Total
Treatment A	17	7	2	26

These categories are *mutually exclusive*, so that no one patient can be in two different categories at the same time, and they are also *exhaustive*, that is, there are enough categories to account for all the members of that particular population. This is not essential, but if the categories are not exhaustive, interpretation may be difficult. If an investigation of the effects of a treatment used only the categories improved and the same, it would be difficult to make sensible judgments without knowing how many became worse. The categories represent *nominal* or *qualitative* variables, also termed *attributes*, and do not represent measurements or even a particular order of the variables.

Goodness of Fit

We may wish to compare a distribution of counts in different categories with some theoretical distribution. In a classical genetics experiment, tall pea plants are crossed with short plants to provide an F1 generation, and these are crossed with others of the F1 generation to provide a second (F2) generation. Counts of 120 plants of the F2 generation are listed in Table 14.2a.

Table 14.2a Hypothetical Mendelian experiment

	Tall	Short	Total
Observed counts	94	26	120

In classical Mendelian genetics with the phenotype determined by dominant and recessive alleles, there should be a 3:1 ratio in which one-quarter of the F2 generation have two recessive genes and are short, whereas three-quarters of them have at least one dominant gene and are tall. Is the result of F2 crosses in our series of 120 plants consistent with the hypothesis that we are dealing with classical Mendelian inheritance? The observed ratio of Tall: Short is 3.62:1 which differs from 3:1, but if the 3:1 ratio is true in the population, the relatively small sample might by chance have a ratio as discrepant as 3.62:1.

To analyze this, adopt the null hypothesis that the sample is drawn from a population in which the Tall: Short ratio is 3:1. If this is true, then a sample of 120 F2 crosses is expected to provide $0.75 \times 120 = 90$ tall plants and $0.25 \times 120 = 30$ short plants (Table 14.2b).

Table 14.2b Chi-square analysis of Mendelian experiment

	Tall	Short	Total
Observed (<i>O</i>)	94	26	120
Expected (<i>E</i>)	90	30	120
Deviation (<i>O</i> − <i>E</i>)	+4	−4	0
(<i>O</i> − <i>E</i>) ²	16	16	
χ^2	0.178	0.533	$\chi^2_T = 0.711$

The expected frequencies (symbolized by *fe* or *E*) appear below their respective observed frequencies (symbolized by *fo* or *O*), and the deviations (*O*−*E*) appear in the line below the expected frequencies. Whether a given deviation is small or big depends not on its absolute size but how big the deviation is relative to the numbers used in the experiment. A deviation of 10 is large and perhaps important with 30 counts but small and probably unimportant with 1000 counts. To evaluate the relative size of the deviation, it is squared and then divided by the expected value in that column; the ratio $\frac{(O-E)^2}{E}$ is termed χ^2 , also sometimes written as chi-square. The values of χ^2 for each column are added up to give a total $\chi^2_T = 0.711$. This is a measure of the overall discrepancies between *O* and *E* for each cell, and the larger the discrepancy the larger will be the value of χ^2_T .

Does this value of 0.711 help to support or refute the null hypothesis? If the probability of getting $\chi^2_T = 0.711$ is low, we would tend to reject the null hypothesis, but if the probability is high, then we would not be able to reject the null hypothesis. We can estimate this probability because of the similarity of the distribution of chi-square to the χ^2 distribution described in “Chi-square Distribution” section. (There is some possibility for confusion in the use of symbols here. Most but not all texts distinguish between these two distributions.) A brief explanation of the equivalence is given by Altman (1992).

For the previous example, the $\chi^2_T = 0.711$ is referred to a table of the χ^2 distribution, and with one degree of freedom $P = 0.399$, so that if the null hypothesis is true, then

about 40% of the time, samples of 120 plants drawn from this population could have ratios as deviant from 3:1 as 3.62:1 or even more. Such an estimate would not allow us to be comfortable in rejecting the null hypothesis; we conclude that there is an acceptable fit between the observed and expected results.

Continuity Correction

Because the χ^2 distribution is continuous and if we examine only two groups, in a large series of experiments in which the null hypothesis is known to be true, the values obtained cause us to reject the null hypothesis more than the expected number of times for any critical value of χ^2_T (Type I error). To reduce the error, Yates' correction for continuity is often advised, especially if the actual numbers are small [Yates (1902–94) was Fisher's assistant at Rothamsted and became head of the unit in 1933 when Fisher moved to University College London.] To make this correction, the absolute value of the deviation (written as $|O - E|$) is made smaller by 0.5: +4 becomes +3.5, and -4 becomes -3.5. The result is to make χ^2_T smaller than it would have been without the correction (χ^2_T becomes 0.544 in Table 14.2b), and the excessive number of Type I errors is abolished. Yates' correction for continuity is made also with 2×2 tables but *should not be used* for larger tables. The correction is used only when there is one degree of freedom (see later).

The need for such a correction is disputed. Yates' correction certainly increases the risk of accepting the null hypothesis falsely (Type II error). If, however, the decision about statistical significance or not depends on whether or not Yates' or some other correction is used, it is better to consider the results of the test as borderline or, better still, to use another test such as Fisher's exact test (see later).

The chi-square test is not restricted to two categories. Continuing with the genetic example, with two pairs of dominant-recessive alleles—one for Tall vs Short, one for Green vs Yellow—the expected ratios for the F2 plants are 9 tall green, 3 tall yellow, 3 short green, and 1 short yellow, or 9:3:3:1. Assume that an experiment gives the results presented in Table 14.2c:

Table 14.2c Expanded Mendelian experiment

	Tall green	Tall yellow	Short green	Short yellow	Total
Observed (<i>O</i>)	94	22	33	11	160
Expected (<i>E</i>)	90	30	30	10	160
Deviation (<i>O</i> − <i>E</i>)	4	−8	3	1	
(<i>O</i> − <i>E</i>) ²	16	64	9	1	
χ^2	0.178	2.133	0.300	0.100	2.711 = χ^2_T

The calculations show a value of χ^2_T of 2.711 with 3 degrees of freedom, and from the χ^2 table $P = 0.438$, so that we would not on this basis reject the null hypothesis.

In general, we are interested in large values of chi-square, because it is these that answer the question about whether or not to reject the null hypothesis, and so pay attention to the area on the right-hand side of the chi-square curve where values above certain critical values are found. On the other hand, even if the null hypothesis is true, there should be a certain degree of variation between the observed and experimental data, and if the two sets of data are too much alike, we might be suspicious about why that occurred. Fisher once reviewed a series of experiments reported by Mendel, and after combining their individual chi-squares obtained a total chi-square of 42 with 84 degrees of freedom. The area under the chi-square curve on the left-hand side gives the probability of getting this value, and it is about 0.00004. Therefore either a very unusual event had occurred, or else someone had manipulated the data to make it agree so closely with theory. There is a suggestion that this was done by a gardener who knew what answers Mendel wanted to get.

Degrees of Freedom

The concept of degrees of freedom may be understood by examining the observed data in [Table 14.2a](#). If the marginal total of 120 plants is fixed, then by choosing any number N_T (not negative and < 120) for the counts in Tall, the counts in Short must be $120 - N_T$. We are free to pick only one of the two numbers and so have lost one degree of freedom. With more observed categories, for example, four categories (tall green, tall yellow, short green, short yellow) then we could choose any numbers for the counts in any three of these categories, but the counts in the fourth category must be such that all four counts add up to 160 ([Table 14.2c](#)). Once again, we have lost one degree of freedom, and we look up the value of chi-square in the table for $4 - 1 = 3$ degrees of freedom.)

2×2 Contingency Tables

Both the previous examples compared one observed and one theoretical distribution. Frequently, however, we wish to compare two or more observed distributions. [Table 14.3a](#) presents the binary outcomes of two forms of treatment for a disease set out in a 2×2 table, also known as a fourfold table; this is the most common contingency table. [Table 14.3b](#) presents some symbols commonly used to indicate different components of the table. a, b, c, d are the counts for each *cell*, defined as the intersection of a row category and a column category. In larger contingency tables, the cells may be indicated by $a_1, a_2, \dots, a_m; b_1, b_2, \dots, b_m; n_1, n_2, \dots, n_m$, or may be indicated by the row and column that characterize that cell; r_2c_3 indicates the cell in the second row and the third column, and $r_i c_j$ indicates the cell in the i -th row and the j -th column. By adopting the convention of citing rows before columns the symbols can be shortened to n_{ij} or f_{ij} , where i indicates the i -th row and j indicates the j -th column. Therefore the cell counts in 2×2 table may be defined as n_{11}, n_{12}, n_{21} , and n_{22} , where the first number in the subscript indicates the row

and the second number indicates the column. The totals of each row or column, known as marginal totals, are referred to as R_1 and R_2 (for the first and second rows) and C_1 and C_2 (for the first and second columns), or may be referred to as $n_{1\bullet}$ and $n_{2\bullet}$, where the dot indicates summation of all the columns in that row, and as $n_{\bullet 1}$ and $n_{\bullet 2}$, where the dot indicates summation of all the rows in that numbered column. The grand total of all the cell counts is the same as the sum of the two marginal row totals or the two marginal column totals and is symbolized by N or $n_{\bullet\bullet}$. In larger $m \times n$ contingency tables, the marginal totals in the rows and columns are designated as $R_1, R_2, \dots, R_i, \dots, R_m$ and $C_1, C_2, \dots, C_j, \dots, C_n$, respectively, or else as $n_{1\bullet}, n_{2\bullet}, \dots, n_{i\bullet}, \dots, n_{m\bullet}$ and $n_{\bullet 1}, n_{\bullet 2}, \dots, n_{\bullet j}, \dots, n_{\bullet n}$, respectively.

Table 14.3a Example of 2×2 layout

	Alive	Dead	Total
Treatment A	60	40	100
Treatment B	30	20	50
Total	90	60	150

Table 14.3b General example of 2×2 layout

	Alive	Dead	Total
Treatment A	$a = n_{11}$	$b = n_{12}$	$R_1 = n_{1\bullet}$
Treatment B	$c = n_{21}$	$d = n_{22}$	$R_2 = n_{2\bullet}$
Total	$C_1 = n_{\bullet 1}$	$C_2 = n_{\bullet 2}$	$N = n_{\bullet\bullet}$

Table 14.3c 2×2 layout as proportions

	Alive	Dead	Total
Treatment A	0.40	0.27	0.67
Treatment B	0.20	0.13	0.33
Total	0.60	0.40	1.00

Table 14.3c presents each of the counts as a proportion of the total; thus the 60 patients alive after treatment A represent $60/150 = 0.40$ of the total number of 150 patients. The proportion of survivors after treatment A ($60/100 = 0.6$) is the same as the proportion of survivors after treatment B ($30/50 = 0.6$), so that the two different treatments did not affect the survival rate. Survival was *independent* of the type of treatment used; other words used to express the same idea are that survival was *not associated* with the type of treatment or was *not contingent* on the type of treatment. Table 14.4a presents a different set of results obtained by treating this disease with two different treatments:

Table 14.4a Another 2×2 layout

	Alive	Dead	Total
Treatment A	60	40	100
Treatment B	70	30	100
Total	130	70	200

Table 14.4b 2×2 layout as proportions

	Alive	Dead	Total
Treatment A	0.30 ($=p_{11}$)	0.20 ($=p_{12}$)	0.50 ($=p_{1\bullet}$)
Treatment B	0.35 ($=p_{21}$)	0.15 ($=p_{22}$)	0.50 ($=p_{2\bullet}$)
	0.65 ($=p_{\bullet 1}$)	0.35 ($=p_{\bullet 2}$)	1.00 ($=p_{\bullet\bullet}$)

In [Table 14.4b](#), the proportions p (in parentheses) have subscripts that indicate what they refer to: p_{11} indicates the proportion in row 1 and column 1 (the upper left-hand cell), whereas p_{21} indicates the proportion in the second row, first column (the lower left-hand cell). The proportion $p_{1\bullet}$ indicates the proportion of the marginal total in row 1; the dot indicates that all the columns in row 1 are summed.

A higher proportion of patients survived after treatment B than after treatment A, 0.35 vs 0.30, respectively, but is there a true association between the type of treatment and the outcome, or are the observed differences due to chance variation in relatively small samples from a population in which a larger study would show no association between treatment and outcome? These tables allow us to evaluate association or contingency, and are known as *contingency tables*, defined as *tables that show the association between variables where the variables have been classified into mutually exclusive categories and the cell entries are frequencies*. There are at least two requirements to be fulfilled for the analysis to be valid. There should be no systematic change in the proportion of survivors in each group in different periods throughout the study, and the outcome for any one patient should not affect the outcome for any other patient.

To answer the question, begin by adopting the null hypothesis (H_0) that the two groups *do* come from the same population, and that any differences in the observed proportions in the two samples are due to chance variation. Then calculate how likely those differences are to occur if the null hypothesis is true; if the probability of such differences is very low, then it might be better to reject the null hypothesis. Computer programs carry out all the necessary calculations, but this is a test that can be done readily by hand, with the merit of showing how the contributions to the total chi-square are derived ([Table 14.5a](#)).

Table 14.5a Layout for chi-square analysis

	Category 1	Category 2	Total
Treatment 1	O_{11} E_{11} $(O-E)_{11}$ $(O-E)_{11}^2$ χ_{11}^2	O_{12} E_{12} $(O-E)_{12}$ $(O-E)_{12}^2$ χ_{12}^2	$O_{11} + O_{12} (=E_{11} + E_{12})$
Treatment 2	O_{21} E_{21} $(O-E)_{21}$ $(O-E)_{21}^2$ χ_{21}^2	O_{22} E_{22} $(O-E)_{22}$ $(O-E)_{22}^2$ χ_{22}^2	$O_{21} + O_{22} (=E_{21} + E_{22})$
Total	$O_{11} + O_{21}$	$O_{12} + O_{22}$	$O_{11} + O_{21} + O_{12} + O_{22}$

$$\chi_T^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \chi_{11}^2 + \chi_{12}^2 + \chi_{21}^2 + \chi_{22}^2.$$

Put the observed count (O) in the upper left-hand corner of each cell, the expected count (E) in the upper right-hand corner of each cell, the deviation ($O-E$) below each observed count, the squared deviation below the expected count, and then divide the squared deviation by the expected count (which is just above it) to give the value of χ^2 for that cell. The values of χ^2 for all the four cells are added up to give χ_T^2 , and this total value is looked up in the tables of the distribution of χ^2 for the appropriate degrees of freedom. The first step is to calculate the expected counts in each cell. Given the null hypothesis, the subtotals in the two columns represent better estimates of the proportion surviving or dying in the population of patients with that disease (Table 14.5b).

Table 14.5b Basic analysis

	Alive	Dead	Total
Treatment A	60 65 -5 25 <i>0.385</i>	40 35 5 25 <i>0.714</i>	100
Treatment B	70 65 5 25 <i>0.385</i>	30 35 -5 25 <i>0.714</i>	100
Total	130	70	200

Observed numbers in bold type, chi-square in italics. $\chi_T^2 = 2.198$, 1 d.f., $P = 0.1382$.

These computations can also be performed on freeware designed for larger tables (see later). If out of a total of 200 patients, 130 survive, then in a sample of 100 patients $130 \times 100/200$ are expected to survive. More generally, the expected value in any cell is (Row subtotal \times Column subtotal)/ N , and this applies to any sized chi-square table. Another way of obtaining the same result is to point out that the proportion surviving in the pooled results is $130/200 = 0.65$. Then this proportion multiplied by the marginal

total in a sample (100) gives an expected value of $100 \times 0.65 = 65$. This value is the expected count in the upper left-hand cell. In a similar fashion, the expected counts in the other three cells can be calculated. Once these expected values are obtained, it is easy to obtain the deviations, the squared deviations, and the χ^2 values for each cell, as well as the χ^2_T .

Probability for any chi-square total and degrees of freedom may be found online at <http://www.fourmilab.ch/rpkp/experiments/analysis/chiCalc.html>, <http://vassarstats.net/tabs.html#csq> and <http://stattrek.com/online-calculator/chi-square.aspx>.

The chi-square computation itself may be performed online for 2×2 tables at: http://www.obg.cuhk.edu.hk/ResearchSupport/StatTools/ChiSqTest_Pgm.php, <http://statpages.orgg/ctab2x2.html>, <http://graphpad.com/quickcalcs/chisquared1.cfm>, all of which require you to calculate and insert the expected numbers, and <http://www.socscistatistics.com/tests/chisquare/Default2.aspx>. This last website calculates the expected numbers and provides the chi-square values for each cell, so that it is possible to tell which cell makes the major contribution to the total chi-square.

<http://www.vassarstats.net/newcs.html> also does not require expected numbers and shows for each cell the percentage excess or deficit.

Problem 14.1 The following table shows data for maternal age and the babies' birth weight. Are these independent?

Maternal age	Birth weight (g)	
	<2500	>2500
<25 years	64	216
>25 years	47	273

Once the expected counts in one cell have been obtained, the expected counts in the other three cells may be obtained by subtracting the expected counts from the marginal totals; in a 2×2 table, the absolute deviations are the same in all four cells, and the sum of the deviations in any row or column is zero.

The degrees of freedom can be calculated using the argument described before for the first goodness-of-fit test. If the four marginal totals are given, then the count in any one of the four cells could be any number (provided it was not negative and was less than the total counts in the whole study). Once that count was chosen, however, then the other three counts are forced by the marginal totals.

A more general way of calculating degrees of freedom is $(\text{Number of Rows}-1)(\text{Number of Columns}-1)$.

Therefore in Table 14.5a the degrees of freedom are $(2-1)(2-1) = 1$.

The value for total chi-square in Table 14.5a is 2.198 that, with one degree of freedom, does not allow us to reject the null hypothesis with confidence. What would

happen if we increased the sample size? Assume that we examine another 100 patients in each treatment group, and that the proportions alive and dead do not change. Adding the two sets together gives [Table 14.5c](#).

Table 14.5c Change in sample size

	Alive		Dead		Total
Treatment A	120	130	80	70	200
	-10	100	10	100	
	<i>0.769</i>		<i>1.429</i>		
Treatment B	140	130	60	70	200
	10	100	-10	100	
	<i>0.769</i>		<i>1.429</i>		
Total	260		140		400

Observed numbers in enlarged bold type, χ^2 in italics.

$$\chi^2_{\text{T}} = 4.3960, 1 \text{ d.f.}, P = 0.036.$$

Comparing [Tables 14.5a and 14.5c](#), all the observed and expected numbers, the deviation between observed and expected numbers, the marginal totals, and the total number of patients have been doubled, so that the proportion surviving in each treatment group is the same for both sample sizes; for simplicity, Yates' correction has been omitted. However, because the deviations are squared, the χ^2 contributions for each cell and thus the χ^2_{T} have doubled (apart from small rounding off errors), so that the null hypothesis can be rejected at the 0.05 level for the larger sample size even though it could not be rejected for the smaller sample size. This is reasonable. If a difference between two treatments is consistent as samples get bigger, then we feel more comfortable about rejecting the null hypothesis that there is no difference between the two treatments. For this reason, chi-square tests must always be done with absolute numbers and not with percentages or numbers per unit. On the other hand, expressing the individual proportions as percentages allows the investigator to assess important differences, even if they are not used in subsequent calculations. Avoid the temptation to increase sample sizes repeatedly until a small P value is obtained. This is termed “P hacking” and is illegal ([Motulsky, 2015](#)). It is possible to estimate the sample size needed to have a desired probability of rejecting the null hypothesis for any given difference; the way to do this is given later.

Practical Matters

1. Because the expected values are usually not whole numbers, calculate them and the chi-square to three decimal places to minimize rounding off errors.

2. For 2×2 tables, many statisticians use Yates' correction for continuity; that is, decrease the absolute size of the deviation by 0.5. Table 14.5a should therefore be:

	Alive		Dead		Total
Treatment A	60	65	40	35	100
	-4.5	20.25	4.5	20.25	
	<i>0.312</i>		<i>0.579</i>		
Treatment B	70	65	30	35	100
	4.5	20.25	-4.5	20.25	
	<i>0.312</i>		<i>0.579</i>		
Total	130		70		200

Observed numbers in enlarged bold type, χ^2 in italics.

$$\chi^2_{\text{T}} = 1.78, 1 \text{ d.f.}, P = 0.18.$$

This correction has made the total chi-square smaller, so that the null hypothesis is even less likely to be rejected.

3. None of the expected frequencies should be too small. As a rule of thumb, Cochran suggested that 80% of the cells should have expected frequencies >5 and that none should be below 1. A very small expected value could lead to a big squared deviation that, divided by the small expected value, gives a very large contribution to the χ^2_{T} . This tends to inflate the χ^2_{T} and we should hesitate to accept a conclusion based on a single large value of χ^2 . In addition, a very small expected value makes the theoretical basis for using the χ^2 table suspect. It would be better to use Fisher's exact test (see later). If the expected value is <5 , we can use the chi-square technique, but should be cautious about interpreting the results. If $>20\%$ of the cells (in larger contingency tables) have expected values <5 , either combine adjacent rows or columns to increase the size of the expected numbers (if that makes sense) or do not use the chi-square test. These criteria are frequently used, but not all statisticians agree with them.
4. Problems about whether to use Yates' correction or about too small an expected value can be dealt with by using Fisher's exact test. Computer programs can calculate the probability by this test for any sample size, so that this test may be preferred for any 2×2 table.

Odds Ratio

The chi-square test gives the probability of rejecting the null hypothesis falsely. If this probability is low, then we may elect to conclude that the differences between the two groups are not likely to have occurred by chance. What the test does not indicate is the magnitude of the difference between the proportions (effect size) in the two treatment groups. To find out the magnitude of the difference between the proportions in the two treatment groups consider the proportions themselves in each group (Fleiss, 1981).

As an example, in Table 14.5b, the ratio of alive to dead in group A is $60:40 = 1.5$, and in group B is $70:30 = 2.33$. How much worse are the results for group A? Divide 1.5 by $2.33 = 0.64$, so that survival in group A is 64% of that in group B.

Define the odds (risk) of dying with treatment A as Ω_A :

$$\Omega_A = \frac{\Pi(\text{Dead} | A)}{\Pi(\text{Alive} | A)}.$$

In other words, the risk of dying with treatment A is the ratio of two conditional probabilities, the probability (Π) of dying with treatment A and the probability of not dying with treatment A. Ω_A is the odds that death will occur if treatment A is present.

$\Pi(\text{Dead} | A)$ may be estimated by $\frac{p_{12}}{p_{1\cdot}}$,
and $\Pi(\text{Alive} | A)$ may be estimated by $\frac{p_{11}}{p_{1\cdot}}$.
 Ω_A may be estimated by

$$O_A = \frac{p_{12}/p_{1\cdot}}{p_{11}/p_{1\cdot}} = \frac{p_{12}}{p_{11}}.$$

From Table 14.4b, $O_A = \frac{0.20}{0.30} = 0.667$.

In a similar fashion, define Ω_B , the risk of dying with treatment B, as:

$$\Omega_B = \frac{\Pi(\text{Dead} | B)}{\Pi(\text{Alive} | B)}.$$

This ratio can be estimated by

$$O_B = \frac{p_{22}/p_{2\cdot}}{p_{21}/p_{2\cdot}} = \frac{p_{22}}{p_{21}} = \frac{0.15}{0.35} = 0.429.$$

These two ratios (Ω_A, Ω_B) can be compared in several ways, but most often one is divided by the other to give ω , the *odds ratio*. This ratio $\omega = \frac{\Omega_A}{\Omega_B}$ can be estimated as.

$$o = \frac{O_A}{O_B} = \frac{p_{12}/p_{11}}{p_{22}/p_{21}} = \frac{p_{12} \times p_{21}}{p_{11} \times p_{22}} = \frac{0.20 \times 0.35}{0.30 \times 0.15} = 1.56, \text{ where } o \text{ is the sample odds ratio.}$$

This may also be symbolized by OR.

This ratio can be interpreted as the risk of dying with treatment A is 1.56 times the risk of dying with treatment B.

Because the proportions given in Table 14.4b are the same as the individual cell values given in Table 14.4a divided by the total number of patients, this ratio can be determined simply by using the original cell counts a, b, c, d . Thus.

$$\begin{aligned} o &= \frac{b \times c}{a \times d} \\ &= \frac{40 \times 70}{60 \times 30} = 1.56. \end{aligned}$$

The ratio is also termed the cross-product ratio, because to obtain it we multiply the counts in two cells on one diagonal and divide that product by the product of the counts in the two cells on the other diagonal (see Table 14.6). In Table 14.6, the thick line indicates which product goes in the numerator; in this instance, emphasizing the risk of dying on treatment A. To focus on the “risk” of surviving on treatment A, then the other product would go in the numerator.

Table 14.6 Illustration of cross-product ratio

	Alive	Dead	Total
Treatment A	60	40	100
Treatment B	70	30	100
Total	130	70	200

To focus on survival,

$$o = \frac{a \times d}{b \times c} = \frac{60 \times 30}{40 \times 70} = 0.64.$$

This answer is the reciprocal of the one obtained before. For either form of calculation, the chances of dying are about two-thirds for treatment B.

Odds versus risk ratios and their confidence limits are discussed in detail in Chapter 20. Odds ratios can be calculated online at <http://statpages.org/ctab2x2.html>, <http://www.hutchon.net/ConfidOR.htm>, and <http://vassarstats.net/odds2x2.html>.

Problem 14.2 What is the odds ratio for the data in Problem 14.1?

Cautionary Tales

Selecting Samples

We should not be so involved with the mechanics of doing these tests that we forget to think about the data and what they mean. Very often there is a concealed or lurking variable that alters the results of our analysis.

In an article published in the *New England Journal of Medicine* in 1965, Binder et al. (1965) investigators in the departments of Internal Medicine and Radiology at the Yale University School of Medicine noted a lack of association between achalasia of the esophagus and hiatus hernia, and to see if this impression was correct they reviewed

the records of all patients with a diagnosis of achalasia of the esophagus in the previous 11 years. Forty-three of these patients were found. The X-ray films of a control group of 43 patients of similar age and sex distribution who had a radiological examination of the upper gastrointestinal tract in May 1964 were reviewed by a radiologist with no knowledge of the patients' clinical history. The results are set out in [Table 14.7a](#).

Table 14.7a Achalasia data

	Hernia	No hernia	Total
Achalasia	1	42	43
No achalasia	9	34	43
Total	10	76	86

The total chi-square for this example is 5.54 with 1 d.f., $P = 0.0093$. The cross-product ratio to evaluate the association between achalasia and hernia is $\frac{1 \times 34}{9 \times 42} = 0.0899$. These results show that the null hypothesis of no difference in the incidence of hiatus hernia whether there is or is not achalasia of the esophagus can be rejected, and that the chance of someone with achalasia of the esophagus also having a hiatus hernia is only 9% of that of someone without achalasia. The investigators concluded that "Hiatus hernia is found very much less frequently in patients with achalasia than in the normal hospital population."

In a letter to the journal a few weeks later, a note of warning was introduced ([Muench, 1965](#)). Dr. Muench (from Harvard!) pointed out that patients having a barium examination of their upper gastrointestinal tract probably did so because they had some upper gastrointestinal tract symptoms and could not be taken as representative of the whole hospital population. He hypothesized that if the investigators had done a barium examination of the upper gastrointestinal tract of every hospitalized patient, they might have found a large number with neither achalasia nor hernia, and the resultant fourfold table might resemble [Table 14.7b](#):

Table 14.7b Hypothetical extension of achalasia data

	Hernia	No hernia	Total
Achalasia	1	42	43
No achalasia	9	340	349
Total	10	382	392

The rationale for increasing the number only in the lower right-hand cell is that patients with symptoms that lead to a radiological examination of the upper gastrointestinal tract will already have been included in the study, and the others are unlikely to have either of the two diseases in question. It is true that asymptomatic disease might be revealed by the X-ray examination, but in general most people do not have either of these relatively uncommon diseases. The choice of the actual

Continued

Selecting Samples—cont'd

number of 340 was arbitrary, but it does correspond to what one might find in a medium-sized hospital. From this new table of data, χ^2_T at 0.0099 is not much different from zero, and the cross-product ratio is $\frac{1 \times 340}{9 \times 42} = 0.8995$. There does not seem to be an association between achalasia and hiatus hernia when one takes account of the whole hospital population.

Dr. Muench went one more step. He considered what might have happened if the whole population of a small town, both those in and out of hospital, had been given a radiological examination of the upper gastrointestinal tract. Once again, he used the reasonable assumption that most of the people outside hospital would have neither disease, and postulated a table such as [Table 14.7c](#):

Table 14.7c Further hypothetical extension of achalasia data

	Hernia	No hernia	Total
Achalasia	1	42	43
No achalasia	9	3400	3409
Total	10	3442	3452

Now χ^2_T is still small, partly because of the huge number who have neither disease, but the cross-product ratio is $\frac{1 \times 3400}{9 \times 42} = 8.9947$, indicating a tendency for achalasia to be associated with hiatus hernia in the whole population!

This is an instructive example. People who have any given investigation are not representative of all patients in a hospital, and patients in hospital are not representative of all people in a population. The number of people who have neither of two diseases is almost certainly very large, even if not well known, and attempts made to define associations without including them may well lead to incorrect conclusions. These missing people are the concealed confounders.

Simpson's Paradox

In 1951 Simpson described an apparent paradox, in that successes of individual groups seem reversed when the groups are combined. He showed that it was possible for two subgroups each to have $A_1 > B_1$ and $A_2 > B_2$, yet on pooling them $A_{1+2} < B_{1+2}$. This paradox was the basis for a suit for gender discrimination against the University of California in Berkeley. The plaintiffs alleged that the University discriminated against admitting women to graduate studies ([Bickel et al., 1975](#)). A simplified set of data based on that publication illustrates the problem ([Table 14.8](#)).

In each department the admission rate for men and women is the same. When the two departments are combined, however, the fact that more women apply for the department with the lower admission rate explains the apparent gender discrimination.

Similar examples are scattered throughout the literature. In 1934 Cohen and Nagel examined deaths from tuberculosis in New York City and Richmond, VA for white and nonwhite residents ([Table 14.9](#)).

Table 14.8 University selection data

	Total applications	Admit	% admitted
Department A			
Men	400	200	50
Women	200	100	50
Department B			
Men	150	50	33
Women	450	150	33
Totals			
Men	550	250	45
Women	650	250	38

Table 14.9 An early example of Simpson's paradox

Race	Total population		Deaths		Deaths/100,000	
	NYC	Richmond	NYC	Richmond	NYC	Richmond
White	4,675,174	80,895	8365	131	179	162
Nonwhite	91,709	46,733	513	155	560	332
Total	4,766,883	127,628	8878	286	186	226

For both whites and nonwhites, the death rate per 100,000 population was higher in New York than Richmond, but for the total population the rates were higher in Richmond than New York: 226 vs 187. The cause of this paradox lies in the greater risk of tuberculosis in nonwhites who made up a bigger percentage of the population in Richmond than they did in New York City.

Charig et al. compared the effects of A. open surgery versus B. percutaneous lithotomy for kidney stones (Charig et al., 1986). The success rate for treatment A was 273/350 (78%) and for treatment B was 289/350 (83%), so that treatment B was slightly more effective. When, however, the data were separated into the results for large and small kidney stones, the results were as presented in Table 14.10.

Table 14.10 Renal calculi

	Treatment A	Treatment B
Small stones	81/87 (93%) <i>a</i>	234/270 (87%) <i>b</i>
Large stones	192/263 (73%) <i>c</i>	55/80 (69%) <i>d</i>
Both	273/350 (78%)	289/350 (83%)

Here, each individual group (stone size) showed that treatment A was better yet combining them showed the reverse. What happened was that doctors preferred to offer the better treatment (open surgery) to patients with the larger stones and offered treatment B to more patients with smaller stones. As a result, the total outcome for treatment B depends mainly on the larger number of patients with small stones (who on the whole do well) in cell b as compared with the number for treatment A (who do less well) in cell c.

Continued

Selecting Samples—cont'd

This phenomenon of a lurking confounder is common and must be sought. Other examples are the paradox that low birth weight babies of smoking mothers have a lower perinatal mortality than those from nonsmoking mothers, despite the fact that exposure to smoking during pregnancy is known to be harmful, ([Hernandez-Diaz et al., 2006](#); [Wilcox, 2006](#)) drawing incorrect conclusions from a study on treatment of meningococcal disease ([Perera, 2006](#)), in evaluating trends for the Scholastic Aptitude Test (SAT) ([Bracey, 2003](#)).

Spurious Associations (Berkson's Fallacy)

[Berkson \(1946\)](#) studied the apparent association between diabetes and cholecystitis that had been found at the Mayo Clinic and concluded that the association could be spurious if the occurrence of two disorders in the same person increases the probability that they will be admitted to a hospital or clinic, and if the proportions of these patients are not the same in the hospital and the general population.

Berkson began with four assumptions:

1. In the whole population, three diseases (he chose cholecystitis, diabetes mellitus, and refractive errors) are unrelated.
2. It is possible to have more than one of these diseases.
3. The chances of a disease resulting in hospital admission (or death) vary from disease to disease.
4. Having more than one disease increases the chances of hospital admission.

[Fig. 14.1](#) shows the prevalence of three (unspecified) diseases in a large population area, based on descriptions by [Berkson \(1946\)](#), [Mainland \(1953\)](#), and [Roberts et al. \(1978\)](#).

Consider a population of 1,000,000, of whom 3% (30,000) have disease X, 5% (50,000) have disease A, and 7% (70,000) have disease B. There is no relationship among these diseases. Then we can construct a Venn diagram, based on the approach used by [Roberts et al. \(1978\)](#). [Fig. 14.1](#) shows the numbers.

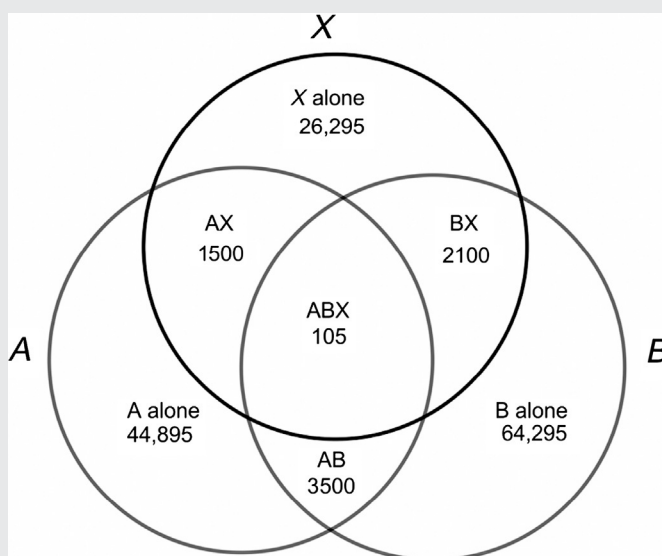


Fig. 14.1 Venn diagram showing components of three diseases, each represented by a circle.

To obtain these numbers, consider the 30,000 who have disease X. Five percent of the population have disease A, irrespective of whatever else they might have, so that $0.05 \times 30,000 = 1500$ have AX. Seven percent of the population also have B, so that $1500 \times 0.07 = 105$ have ABX. The other numbers are calculated in similar fashion.

To test if X is differentially associated with A or B, set up a fourfold table (Table 14.11a).

Table 14.11a 2×2 table for population

	A	B
X present	1,695	2,205
X absent	48,395	67,795

The cross-product ratio is 1.02, showing no association between X and A or B, as expected from the initial assumptions.

Now consider what the hospital numbers will be if the fractional admission rates for X are 0.2%, for B are 0.1%, and for A are 0.03%. For A, B, and X alone we apply the fractional admission rates. For a combination such as AX, we use the formula for the union of two probabilities (Chapter 5): $P(A \cup X) = P(A) + P(X) - P(A \cap X)$, the last term being the product of the two individual probabilities. Thus the admission rate for 1500 population with AX is $0.03 + 0.2 - 0.2 \times 0.03 = 0.224$. Then $1500 \times 0.224 = 336$ of AX are admitted. Similar calculations are done for the remaining combinations.

The numbers admitted are shown in Fig. 14.2.

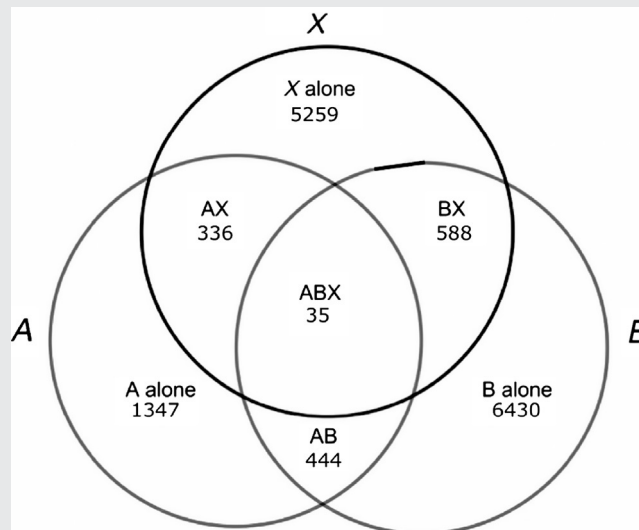


Fig. 14.2 Components of hospital admissions.

Continued

Selecting Samples—cont'd

Now if we set up the fourfold table (Table 14.11b) we get.

Table 14.11b 2×2 table showing Berkson's fallacy

	A	B	Percent with A
X present	371	623	37
X absent	1,791	6,874	25

The chi-square is 141.41, $P < 0.0001$, and when X is present more people have A than if it is absent, even though we know that this does not exist in the whole population.

If instead of hospital admission rates we think about rates of admission to an autopsy room, a clinic, or physician's office, then it is likely that different diseases have different probabilities of being included in an investigation of an association between different diseases. Unless these associations are studied in populations in which this selection bias does not occur, little faith can be placed in the results. Some examples of this error was found in about 25% of studies surveyed by Roberts et al. (1978).

Larger Contingency Tables

Contingency tables need not be restricted to a 2×2 format. For example, patients with migraine were treated with sumatriptan during attacks (Group, 1991). As part of the study, investigators evaluated the effect of placebo and three different regimens of sumatriptan on recurrence of the headache within 24h. The results are set out in Table 14.12a.

Table 14.12a Extended contingency table

	P/P		6mgS/P		6mgS/6mgS		8mgS/P		Total
No recurrence	85	74.576 <i>14.424</i>	194	199.119 <i>−5.119</i>	190	195.390 <i>−5.390</i>	103	102.915 <i>0.085</i>	572
		<i>1.457</i>		<i>0.132</i>		<i>0.149</i>		<i>0.000</i>	
Recurrence	15	25.424 <i>−14.424</i>	73	67.881 <i>5.119</i>	72	66.610 <i>5.390</i>	35	35.085 <i>−0.085</i>	195
		<i>4.274</i>		<i>0.386</i>		<i>0.436</i>		<i>0.000</i>	
Total		100		267		262		138	767

Observed numbers in bold enlarged type, χ^2 in italics. $\chi^2_T = 6.833$, 3 d.f. $P = 0.0774$. P/P, placebo on two occasions; 6mgS/P, 6 mg sumatriptan plus placebo; 6mgS/6mgS, 6 mg sumatriptan on both occasions; 8mgS/P, 8 mg sumatriptan and placebo.

The sum of deviations in any row or any column equals zero, but unlike the 2×2 table, the deviations in different rows need not be the same. Yates' correction is not made for any table larger than a 2×2 table.

This is a 2×4 Table (2 rows, 4 columns). Therefore the degrees of freedom are (2−1)(4−1)=3. The value for χ^2_T indicates that if the null hypothesis is true, then 7.74% of the

time one might get results with the differences shown here by random variation, so that we might not feel safe in concluding that there was an association between the treatment regimen and the duration of freedom from headache; nevertheless, the possibility of an effective treatment should be pursued, perhaps with greater numbers of patients.

These larger chi-square tables can be analyzed online at <http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Catego.htm>, <http://www.quantpsy.org/chisq/chisq.htm>, http://www.physics.csbsju.edu/cgi-bin/stats/contingency_form.sh?nrow=3&ncolumn=3, <https://easycalculation.com/statistics/goodness-of-fit.php>, <https://www.graphpad.com/quickcalcs/chisquared1.cfm>, and <http://vassarstats.net/index.html> (see frequency data).

The first three sites do not require entry of expected values.

Problem 14.3 Assume that the data on maternal age and birth weight are:

Maternal age (year)	Birth weight (g)		
	<2500	2500–3000	>3000
<20	21	14	30
20–25	43	56	116
>25	47	56	217

Is there any association?

In 2×3 or larger tables, it is unsafe to do the calculation if the expected value in any cell is under 1. If this happens, it is best to try to combine adjacent cells. Combination of cells can also be done if the total chi-square is not large but the deviations in adjacent cells go in the same direction. Then it might be appropriate to combine cells to try to obtain a greater deviation, providing that pooling the data from these cells is meaningful. Consider the data in Table 14.12b in which the proportions with recurrent headaches on placebo alone are compared with the other three groups, all of which all received at least 6 mg of sumatriptan, and had more than expected or the same number of recurrent headaches.

Table 14.12b Pooling cells

Headaches	P/P		Rest	Total
No recurrences	85 10.424 9.924 <i>1.321</i>	74.576 98.486	487 −10.424 −9.924 <i>0.198</i>	497.424 98.486 572
Recurrences	15 −10.424 −9.924 <i>3.874</i>	25.424 98.486	180 10.424 9.924 <i>0.581</i>	169.576 98.486 195
Total	100		667	767

$$\chi^2_T = 5.974, 1 \text{ d.f.}, P = 0.0145.$$

We have gained ability to reject the null hypothesis.

Caution is needed when making such combinations. In a study with 20 columns and 6 rows, we could end up making dozens of combinations of cells for the analysis, but how should we interpret a large value of chi-square for any one of them? Just as for multiple *t*-tests (Chapter 24) we might reject the null hypothesis at the 0.05 level when in reality the null hypothesis is true. It is best to make as few comparisons as possible, to combine cells only if combination makes physiological or clinical sense, and preferably to reject the null hypothesis after the combination as a reason to repeat the experiment to test that particular hypothesis.

Fisher's Exact Test

When the total number of counts in a fourfold table <30 , or if $N < 50$ and the smallest expected value is <5 , the chi-square test may be inaccurate, and is better replaced by Fisher's exact test (also called the Fisher-Irwin test). If there is no association between the row and column classifications, the exact probability of obtaining any set of cell counts given the marginal totals can be calculated as:

$$P = \frac{R_1!R_2!C_1!C_2!}{a!b!c!d!N!}$$

where the symbol! indicates the factorial of the number referred to. This is based on the hypergeometric function (Chapter 13).

Table 14.13 presents data reported by Basile et al. (1991) about the activity of benzodiazepine receptors in patients who died from fulminant liver failure and a control group who died from cardiovascular causes.

Table 14.13 Data for Fisher's test
Diazepam equivalents (ng/g tissue)

	Under 100	Over 100	Total
Liver disease	2	9	11
Control	5	3	8
Total	7	12	19

The proportion with low activity in the control group is $5/8 = 0.625$, and the proportion with low activity in those with liver disease is $2/11 = 0.182$. Are these proportions different enough to allow us to reject the null hypothesis?

If the marginal totals are fixed, there are eight different combinations of cell counts possible (Fig. 14.3).

0	11	11	4	7	11
7	1	8	3	5	8
7	12	19	7	12	19
1	10	11	5	6	11
6	2	8	2	6	8
7	12	19	7	12	19
2	9	11	6	5	11
5	3	8	1	7	8
7	12	19	7	12	19
3	8	11	7	4	11
4	4	8	0	8	8
7	12	19	7	12	19

Fig. 14.3 All possible combinations with fixed marginal totals.

From the previous formula, the probabilities of getting each of these combinations of cell counts are (shown in [Table 14.14](#)).

Table 14.14 Probabilities of each combination

$P(\alpha = 0)$	0.000158768
$P(\alpha = 1)$	0.006112566
$P(\alpha = 2)$	0.061125665
$P(\alpha = 3)$	0.229221243
$P(\alpha = 4)$	0.366753989
$P(\alpha = 5)$	0.256727792
$P(\alpha = 6)$	0.073350798
$P(\alpha = 7)$	0.006549178
	0.999999999

These results are plotted in [Fig. 14.4](#), which plots the probabilities on the vertical axis against the value of α shown above, that is, the chances of getting any given number with low receptor activity values in the group with liver disease.

As expected from an exhaustive set of mutually exclusive events, the sum of the probabilities equals 1. From [Table 14.14](#), the probability of getting α equal to 2 if the null hypothesis is true is 0.0611, not strong evidence against the null hypothesis. However, we must also consider the probability of $\alpha = 1$ or 0, which would be more extreme

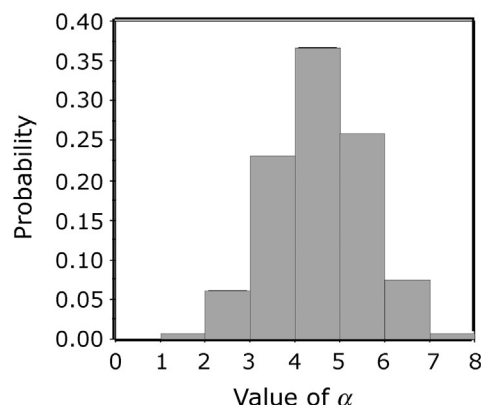


Fig. 14.4 Probabilities for Fisher's exact test.

departures from the hypothesis of equal proportions. These probabilities are 0.0061 and 0.00015, and the three probabilities together sum to 0.0674, equivalent to one tail of the distribution, namely, the probability of getting 2 or less with low values in those with liver disease. From this study, then, we have no compelling reason to reject the null hypothesis. As described here, the test is a one-sided test, which is the more usual requirement. Should a two-sided test be wanted, then according to many statisticians the probability as calculated before should be doubled if the sample sizes in the two groups are similar (Armitage et al., 2002; Everitt, 1992). However, because the distribution is not symmetrical (see Fig. 14.2), others recommend calculating both tails separately (Zar, 2010). This is the procedure adopted by major software programs such as SAS and SPSS. An interactive freeware internet program for doing this and many other aspects of the chi-square test may be found at <http://statpages.org/ctab2x2.html>, <http://www.graphpad.com/quickcalcs/contingency1.cfm>, <http://www.quantpsy.org/fisher/fisher.htm>, <http://www.danielsoper.com/statcalc3/calc.aspx?id=29>, <http://www.quantitativeskills.com/sisa/statistics/fisher.php?n11=2&n12=9&n21=5&n22=3>, or <http://vassarstats.net/> (see frequency data).

Calculating by hand, even with a calculator that gives factorials, is tedious and fraught with errors. Use programs such as those listed before.

Problem 14.4 Examine the data relating gender (M, F) to drinking dietary soft drinks:

	Dietary soda	Nondietary soda
Male	3	9
Female	7	4

Because of the small numbers, perform Fisher's test.

Fisher's exact test (like most tests) has relatively low power when N is small (Everitt, 1992). A substantial difference with the test may be reason to reject the null hypothesis, but with small total numbers failure to reach a low level of α may merely reflect a low power of the test; Table 14.14 is a good example of this low power. If because of low power you decide to increase sample size, calculate the size needed; do not keep on adding more data until a small P value is obtained.

It is possible with the aid of a computer program to do Fisher's exact test for tables larger than 2×2 , for example, by using the interactive freeware found online at: <http://vassarstats.net> (see frequency data) or <http://statpages.org/#CrossTabs>.

Determining Power and Sample Size

To assess the degree of effect, that is, the size of the difference, between observed and expected distributions that is independent of sample size, several indexes have been suggested. An important one is w (Cohen, 1988), defined as:

$$w = \sqrt{\sum_{i=1}^m \frac{(P_{Oij} - P_{Eij})^2}{P_{Eij}}} \text{ where } m \text{ is the number of cells. } w \text{ is the effect size, and } \chi^2 = Nw^2.$$

Therefore $w = \sqrt{\frac{\chi^2}{N}}$.

Here P_{Oij} and P_{Eij} are the observed and expected proportions in any cell, and the fraction is summed for all cells in the table. The expected proportions for any cell are obtained by multiplying the marginal proportions for the row and column which that cell shares. For example, consider Table 14.15, which shows in the upper left-hand corner of each cell the observed proportions (P_{Oij}) of migrainous patients with and without recurrent headaches after treatment with various combinations of placebo and sumatriptan, taken from the raw data in Table 14.12a.

Table 14.15 Sumatriptan data as proportions

	P/P		6mgS/P		6mgS/6mgS		8mgS/P		Total
No recurrence	0.1108	0.0972	0.2529	0.2596	0.2477	0.2547	0.1343	0.1342	0.7457
	0.0136	0.000185	-0.0067	0.000045	-0.007	0.000049	0.0001	0.000	
	0.00190		0.000173		0.000192		0.000		
Recurrence	0.0196	0.0331	0.0952	0.0885	0.0939	0.0868	0.0456	0.0457	0.2543
	-0.0135	0.000182	0.0067	0.000045	0.0071	0.00005	0.0001	0.000	
	0.00550		0.000507		0.000581		0.000		
Total	0.1304		0.3481		0.3416		0.1799		1.0000

Bold type—observed proportions.

The expected proportions (P_{Eij}) are set out in the upper right-hand corners of each cell. The expected proportion (P_{Eij}) of those with no recurrence who had only placebo

(P/P) is calculated as (marginal proportion with $P/P \times$ marginal proportion with no recurrences) $= 0.1304 \times 0.7457 = 0.0972$. This is quicker than calculating the expected numbers and then turning them into proportions. Then to calculate w , sum the values of fractions such as $\frac{(0.1108 - 0.0972)^2}{0.0972}$ for each cell, and take the square root of the sum. This value of w indicates the amount of departure from no association between no recurrences and placebo treatment. If observed and expected values are the same, then $w = 0$. The more w exceeds zero, the less likely is it that the null hypothesis is true.

Another way of calculating w is to note that.

$$\frac{(P_{Oij} - P_{Eij})^2}{P_{Eil}} = \frac{(Oij/N - Eij/N)^2}{Eij/N} = \chi^2/N.$$

Therefore if the values of χ^2 for each cell have already been calculated,

$$w = \sqrt{\frac{\chi^2_T}{N}}; \text{ this involves little extra calculating.}$$

The index w is used in the tables provided by Cohen to evaluate the power of a given chi-square test. To use these tables, a value for α , the Type I error, is selected, and the table is selected for the degrees of freedom involved in the test, the total number of counts, N , and the calculated value of w . For example, in the study presented in Table 14.19a, $w = \sqrt{0.0019 + 0.000173 + 0.000192 + 0.0055 + 0.000507 + 0.000581} = 0.094$. This is almost the same as $\sqrt{\frac{6.834}{767}} = 0.095$, taken from Table 14.11b. With this value, 3 degrees of freedom, and $\alpha = 0.05$, the table gives the power as 0.63. The Type II error (β) is $1 - 0.63 = 0.37$. Clearly the test had low power and the number of subjects was too small to be sure that the degree of difference observed was not due to chance. If we multiply the value of $\frac{(P_{Oi} - P_{Ei})^2}{P_{Ei}}$ in each cell by N , the product is the value of χ^2 for that cell.

Power and sample size calculations are available in many computer programs and online programs such as <https://www.anzmtg.org/stats/PowerCalculator/PowerChiSquare>, and <http://www.real-statistics.com/chi-square-and-f-distributions/power-chi-square-tests/> for 2×2 tables. There is an extensive freeware program called G*Power, at <http://www.pscho.uni-duesseldorf.de/abteilungen/aap/gpower3/download-and-register>. Details of its use have been published and should be consulted. An easy graphically determined power is available at <http://homepage.stat.uiowa.edu/~rlenth/Power/index.html>.

Problem 14.5 Confirm the power calculation for Table 14.12a by using <http://homepage.stat.uiowa.edu/~rlenth/Power/index.html> and use it to determine the number needed for a power of 0.8.

Any 2×2 table can be analyzed by a chi-square test or as proportions, and sample sizes for proportions can be obtained from several online sources ([Chapter 16](#)).

To determine sample size required to give a particular power, consider the same example in which we want to determine what sample size would give a power of 0.8. Look up in a companion set of tables for $\alpha = 0.05$, degrees of freedom = 3, and $w = 0.094$ how many total counts (=patients) are needed for a power of 0.8. Approximately 1090 patients would be needed to achieve the desired level of α , that is another 323 patients would be needed. Of course, if we gathered another 323 patient results, the chi-square test would not allow rejection of the null hypothesis unless the additional patients behaved in the same way as those already studied. If the new patients did not show as big differences as the old patients did, then the total chi-square would not be large, and we would probably conclude that any differences related to the drug dosages were small indeed. Implicit assumptions are that the added patients are similar to those already studied and the different groups are incremented in rough proportion to their present numbers. If any cells have expected numbers below 5, the total sample size needed may be very large.

ADVANCED CONCEPTS

Cochran-Mantel-Haenszel (C-M-H) Test and Confounders

In Simpson's paradox, the paradoxical conclusions were usually due to ignoring a third factor, for example, the relative number of applications per department in the Berkeley study, or the size of the kidney stones. These sometimes disregarded third factors may be confounders, in which all the relationship between X and Y is due to their individual relationship with the third factor S , or may be modifiers, in which the strength of the association between X and Y depends on the level of the S factor. To deal with this problem, the C-M-H test is used. If there are three nominal variables, for example, two 2×2 contingency tables for testing independence and a third nominal variable that indicates repeats, we need to know if it is appropriate to merge the different repeats and obtain a larger and more representative sample. For example, compare two species of snail counted above and below the tide line to determine if there is any relationship between species and site. This is a simple 2×2 table for nominal variables. If we do this in two different months, can we combine the two data sets? There is actually a third nominal variable, the lurking confounder of time, much like the confounder of the size of the kidney stones cited earlier. The repeats may be different times, places, or studies.

To test the independence between cause and effect at two different levels of the possible confounder, there is a method based on the hypergeometric distribution known as the Cochran-Mantel-Haenszel (or sometimes just as the Mantel-Haenszel) test. The null hypothesis is that the deviation $O - E = 0$; because in a 2×2 table the deviations are of the

same magnitude in each cell, calculate only the deviation in any one cell. As shown before, the expected value can be determined as $O_{ij} = \frac{R_i C_j}{N}$. For the upper left cell, the deviation is $O_{11} - \frac{R_1 C_1}{N}$. Do this calculation for each stratum, subtract 0.5 for Yates' continuity correction, and then square the deviations. Pool these squared deviations for each stratum and obtain a weighted mean by dividing the sum of the squared deviations by the sum of their respective variances, calculated as $\text{Variance} = \frac{R_1 C_1 R_2 C_2}{N^2(N-1)}$ for each

stratum. The final formula becomes $\chi^2_{\text{MH}} = \frac{\left\{ \left| \sum \left(O_{ij} - \frac{R_i C_j}{N} \right) - 0.5 \right|^2 \right\}}{\sum \left(\frac{R_1 R_2 C_1 C_2}{N^2(N-1)} \right)}$, sometimes written as

$$\chi^2_{\text{MH}} = \frac{\left\{ \left| \sum \left(a - \frac{(a+b)(a-b)}{N} \right) - 0.5 \right|^2 \right\}}{\sum \frac{(a+b)(a+c)(b+d)(c+d)}{N^2(N-1)}}$$

This value is distributed like χ^2 with 1 degree of freedom.

The combined odds ratio is: $OR_{\text{MH}} = \frac{\sum \left(\frac{ad}{N} \right)}{\sum \left(\frac{bc}{N} \right)}$.

The null hypothesis is that the categories are independent so that the combined odds ratio will be 1. If it is not, it may make sense to calculate a combined odds ratio providing that all the individual odds ratios are similar in magnitude.

As an example, investigators wished to determine the association between thiazide diuretics and hip fractures (LaCroix et al., 1990). They studied subjects in East Boston, Iowa, and New Haven, and data for incidence rates of hip fracture are presented in Table 14.16.

Table 14.16 Example for CM test

Age	Number East Boston			Number Iowa			Number New Haven		
	Men	Women	Total	Men	Women	Total	Men	Women	Total
65–74	11	17	28	7	17	24	8	15	23
≥75	15	42	57	9	49	58	12	40	52
Total	26	59	85	16	66	82	20	55	75
Chi-square	1.48, $P=0.33$			2.04, $P=0.16$			1.11, $P=0.29$		
Odds ratio	1.812			2.24			1.78		

In this subset of their data, although all the odds ratios are >1 , none of the chi-squares are close to $P = 0.05$. The questions are whether gender and age are independent and is it reasonable to combine the data from the groups and obtain a combined odds ratio?

In these tables, there seems to be an association between the age and gender. The combined odds ratio is.

$$OR_{MH} = \frac{\frac{11 \times 42}{85} + \frac{7 \times 49}{82} + \frac{8 \times 40}{75}}{\frac{15 \times 17}{85} + \frac{9 \times 17}{82} + \frac{8 \times 40}{75}} = \frac{13.88}{7.27} = 1.91. \text{ This is the weighted average of}$$

the three odds ratios. The Cochran-Mantel-Haenszel test is then

$$\begin{aligned} \sum O_{11} &= 11 + 7 + 8 = 26 \\ \sum E_{11} &= \frac{26 \times 28}{85} + \frac{16 \times 24}{82} + \frac{20 \times 23}{75} = 19.38. \\ \nu_{EB} &= \frac{26 \times 59 \times 57 \times 28}{85^2 \times 84} = 4.03, \nu_I = \frac{16 \times 66 \times 58 \times 24}{82^2 \times 81} = 2.70, \\ \nu_{NH} &= \frac{20 \times 55 \times 52 \times 23}{75^2 \times 74}, \\ \chi^2_{MH} &= \frac{(|26 - 19.38| - 0.5)^2}{4.03 + 2.70 + 3.16} = 3.79. \text{ With 1 degree of freedom, this indicates that} \end{aligned}$$

$P = 0.0515$, so that we can cautiously reject the null hypothesis that age and gender are independent.

The test can be performed at <http://www.biostathandbook.com/cmh.html>.

We can use more groups. An example of the Berkeley admissions data alluded to above uses some data provided by the Internet Project Simpson's paradox at http://wps.aw.com/wps/media/objects/15/15719/projects/ch2_simpson/index.html. The number of applicants admitted and rejected by four different departments was sorted by gender (Table 14.17a).

Table 14.17a Berkeley admission data by department and gender

	A			B			C			D		
	M	F	Total	M	F	Total	M	F	Total	M	F	Total
Admitted	512	89	601	353	17	370	138	131	269	53	94	147
Rejected	313	19	332	207	8	215	279	244	523	138	299	437
Total	825	108	933	560	25	585	417	375	792	191	393	584

For the data as a whole, ignoring the departments, we have (Table 14.17b).

The χ^2 total is 211.90, $P < 0.00001$, and the odds ratio is 1.94, suggesting that males are admitted disproportionately more than females. As mentioned before, this was the basis for

Table 14.17b Admission data pooled across departments
Stratum

		Stratum		
		M	F	Total
Status	Admitted	1056	331	1387
	Rejected	937	570	1507
Total		1993	901	2894

a complaint to the US Government about discrimination against women in the admissions policy of the Graduate Division. If we examine each of the four departments separately, notice that the four χ^2 total values are, respectively, 16.32, 0.08, 0.22, and 0.81 (all with Yates' correction), with respective odds ratios of 0.35, 0.80, 0.92, and 1.22.

The Mantel-Haenszel test gives

$$\begin{aligned}\chi^2_{MH} &= \frac{\left\{ \left| \left(512 - \frac{601 \times 825}{933} \right) + \left(353 - \frac{370 \times 560}{585} \right) + \left(138 - \frac{269 \times 417}{792} \right) + \left(53 - \frac{147 \times 191}{584} \right) \right| - 0.5 \right\}^2}{\left(\frac{601 \times 332 \times 825 \times 108}{933^2 \times 932} \right) + \left(\frac{370 \times 215 \times 560 \times 25}{595^2 \times 584} \right) + \left(\frac{269 \times 523 \times 417 \times 375}{792^2 \times 791} \right) + \left(\frac{147 \times 437 \times 191 \times 393}{584^2 \times 583} \right)} \\ &= \frac{\{|-19.43 - 1.19 - 3.63 + 4.92| - 0.5\}^2}{21.91 + 5.57 + 44.34 + 24.25} = \frac{\{|-19.33| - 0.5\}^2}{96.07} = \frac{354.57}{96.07} = 3.69\end{aligned}$$

This, with 1 degree of freedom, $P=0.055$. (I would always do this calculation with the online test because of the risk of arithmetic errors.)

We should probably not reject the null hypothesis,

Pooling Data

How can we best combine data from several small 2×2 tables, each with a trend that does not reach statistical significance? First examine the tables to determine if the direction of the trend is similar in all the samples. If the different samples show trends in opposite directions, there is not much sense in combining them. Once we decide to combine them, we have a choice of several methods. The simplest would be to combine all the original data from each table to produce a larger single 2×2 table. This method could produce misleading results if the proportions in the different groups differ markedly from each other (Armitage et al., 2002). Statistical consultation is required.

Testing for Trends in Proportions (Cochran-Armitage Test)

A $2 \times k$ table may have the k groups arranged in some natural order, for examples, time, age groups, or degrees of severity of an illness. A chi-square test examines the null hypothesis that the proportions of the variables in the two rows are the same. However, there might be a trend in the proportions from group 1 to group k . To test this, use an analogy to the testing of linearity with ANOVA (Chapters 25 and 27), by comparing the

total variability of a variate with the amount of variability due to linear regression; if the difference between these is small, there is good evidence for a linear relationship. To do this for the chi-square test, assign a quantitative variable x_i to the k groups. This variable might be ascending integers from 1 to k if, for example, the groups were arranged in decades by age; they might be estimated magnitudes, for example, estimated thyroid weights for different groups of enlarged thyroid glands; or they might merely indicate an order of magnitude, such as -1 for normal tonsils, 0 for moderately enlarged tonsils, and 1 for very enlarged tonsils (Armitage et al., 2002).

Table 14.18 presents the proportion of deaths from coronary heart disease in workers exposed to increasing concentrations of carbon disulfide (Hernberg et al., 1973).

Table 14.18 Carbon disulfide concentration and ischemic deaths

	Concentration score				Total
Heart disease (x_j)	0	1	2	3	
Yes (n_{1i})	3	5	6	5	19 ($R_1 = n_{1\bullet}$)
No	340	157	118	52	667 (R_2)
Total ($n_{1\bullet}$)	343 (C_1)	162 (C_2)	124 (C_3)	57 (C_4)	686 (N)
Proportion %	0.87	3.09	4.83	8.77	

Does the proportion of coronary heart disease tend to increase with carbon disulfide concentration? Inspection shows that the proportions do increase, but could this have occurred by chance? Start by calculating the standard chi-square with 3 degrees of freedom; it is 14.23, and $P = 0.0026$. Then compute a component of chi-square (χ^2_{lin}) that is due to a linear trend. If this is similar to the standard chi-square, then there is probably a trend. This component is

calculated as
$$\frac{N \left[N \sum_{i=1}^k n_{1i}x_i - n_{1\bullet} \sum_{i=1}^k n_{\bullet i}x_i \right]^2}{n_{1\bullet}(N - n_{1\bullet}) \left[N \sum_{i=1}^k n_{\bullet i}x_i^2 - \left(\sum_{i=1}^k n_{\bullet i}x_i \right)^2 \right]} = \frac{N \left[N \sum_{i=1}^k n_{1i}x_i - R_1 \sum_{i=1}^k C_i x_i \right]^2}{R_1 R_2 \left[N \sum_{i=1}^k C_i x_i^2 - \left(\sum_{i=1}^k C_i x_i \right)^2 \right]} \quad (\text{alternative symbols}).$$

From Table 14.18

$$\chi^2_{\text{lin}} = \frac{686[686(0 \times 3 + 1 \times 5 + 2 \times 6 + 3 \times 5) - 19(0 \times 343 + 1 \times 162 + 2 \times 124 + 3 \times 57)]^2}{19 \times 667[686(0^2 \times 343 + 1^2 \times 162 + 2^2 \times 124 + 3^2 \times 57) - (0 \times 343 + 1 \times 162 + 2 \times 124 + 3 \times 57)^2]} = 13.84$$

So large a chi-square strongly favors a linear relationship. The difference between the two chi-squares, which tests departure from a linear trend, is $\chi^2_{\text{T}} - \chi^2_{\text{lin}} = 14.23 - 13.84 = 0.39$ with 2 degrees of freedom, and $P > 0.99$. Therefore conclude that most of the total chi-square is due to the linear trend in the direction of increasing coronary heart disease as the concentration of carbon disulfide increases. The selection of n_1 or n_2 is arbitrary; one could replace n_{1i} and $n_{\bullet i}$ by n_{2i} and $n_{\bullet 2}$ in the previous formula.

Furthermore, the scores assigned are also arbitrary. If the calculation of the linear chi-square is repeated replacing n_{1i} and $n_{\bullet i}$ by n_{2i} and $n_{\bullet 2}$, and assigning scores of -1 , 0 , 1 and 2 , exactly the same result will be obtained. Doing this test by hand is tedious and prone to error, so that it is fortunate that it can be done online at <http://epitools.ausvet.com.au/content.php?page=trend>. (Enter the data into an Excel work sheet—no tabs—and then copy the data to the app.) A similar test, the Mantel-Haenszel extended chi-square test, can be performed online at <http://www.openepi.com/DoseResponse/DoseResponse.htm>.

Sample size can be estimated at <http://resourcetepee.com/free-statistical-calculators/cochran-armitage-sample-size-calculator/>.

Trends can be assessed for tables larger than $2 \times k$; see Armitage et al. (2002).

Problems With $R \times C$ Tables

What do we do when we have rejected the null hypothesis for an $m \times n$ table? Once the total chi-square is big enough to reject the null hypothesis, it would be useful to determine which of the individual χ^2 values appear to be large enough on their own. If the null hypothesis is true, $z = \frac{(O-E)}{\sqrt{E}}$ has an approximately standard normal distribution (as long as E is not too small) that can be read from the z table at http://davidmlane.com/hyperstat/z_table.html, or <http://www.fourmilab.ch/rpkp/experiments/analysis/zCalc.html>. If z calculated in this way is 2.08 , 0.0188 of the area under the standard probability curve is above $z = 2.08$ (one tail). If the null hypothesis is true, the deviation that produced this value of χ^2 is unlikely to have arisen by chance, and we should pay attention to that particular cell in our assessment. The ratio z can be positive or negative. If positive, it indicates a greater than expected frequency, and if negative, a lesser than expected frequency. The square of this z is the same as χ^2 . The z statistic, calculated in this way, is what is used when the alternative hypothesis is one sided. Thus if

$$H_0 : p_1 = p_2 \text{ and } H_A : p_1 < p_2 \text{ or } p_1 > p_2,$$

calculate the square root of chi-square and refer to the z table for the level of α to determine if the null hypothesis can be rejected.

Another approach when chi-square testing with multiple rows and columns is to calculate adjusted standardized residuals as.

Adjusted residual (z_{ij})

$$= \frac{\text{Observed} - \text{Expected}}{\sqrt{\text{Expected} \times \text{Row total proportion} \times \text{Column total proportion}}}.$$

$$\text{These can be computed from } z_{ij} = \frac{O_{ij} - \frac{R_i C_j}{N}}{\sqrt{\frac{R_i C_j}{N} \left(1 - \frac{R_i}{N}\right) \left(1 - \frac{C_j}{N}\right)}}$$

Essentially this formula relates the magnitude of the difference to its standard deviation that is a function of sample size. If the variables in the contingency table are independent, then the terms z_{ij} are approximately normally distributed with mean zero and standard deviation of 1. Any adjusted residuals over 1.96 should occur <5% of the time if the null hypothesis is true, allowing us to decide where there are substantial departures from the null hypothesis of equal proportions in each group.

With many cells, correction for multiplicity can be made by multiplying each P value by $(r-1)(c-1)$ to produce the equivalent of a Bonferroni-type adjustment (Chapter 24). Only if the adjusted P value is below 5% should the null hypothesis be rejected at the 5% level.

Other ways of exploring the data are given by Sharpe (2015).

Multidimensional Tables

The typical $r \times c$ contingency table that assesses independence or association between two categories is computationally simple and relatively easy to interpret. In more complex tables, however, interactions among several categories are common. Their analyses are more complex but can be done with current computer programs. However, their interpretation involves investigating possible interactions among the categories, an issue absent from $r \times c$ tables.

Only the three-dimensional table will be discussed here (Everitt, 1992). Fig. 14.5 shows how a $2 \times 2 \times 2$ table is constructed.

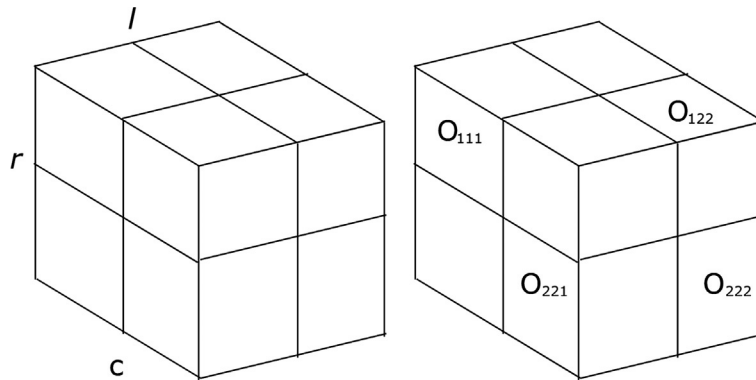


Fig. 14.5 The left panel shows 8 cells specified by rows (r), columns (c), and layers (l). The right-hand panel shows that the cells can be designated in standard format. The observed number in the cell in the first row, first column, and first layer is O_{111} ; in the second row, second column, and first layer is O_{221} ; and in the second row, second column, and second layer is O_{222} and so on. (There may be >2 categories in each row, column, or layer; for example, Rubik's cube is a $4 \times 4 \times 4$ matrix.) To represent this three-dimensional structure in two dimensions either a tree-like table or else an extended two-way table with subgroups can be constructed.

Table 14.19a Three-way table: surgical results

	Dead		Alive		Total
	RACHS 1–3	RACHS 4–6	RACHS 1–3	RACHS 4–6	
Surgeon 1	14	10	397	44	465
Surgeon 2	4	2	342	49	397
Total	18	12	739	93	862

Table 14.19a presents data from a study examining the mortality rates of two cardiac surgeons operating on children with congenital heart disease. Because different forms of congenital heart disease have different operative risks, the patients are classified by a RACHS scale, in which grades 1–3 are relatively simple lesions with a low expected operative mortality, and grades 4–6 are more complex lesions with a higher expected operative mortality.

The table has three categories, arranged in rows (r), columns (c), and layers (l). The observed number in any cell can be represented by O_{ijk} , for $i = 1, 2, \dots, r$; $j = 1, 2, \dots, c$; and $k = 1, 2, \dots, l$. Single variable marginal totals can be obtained by summing all the values for the other two categories:

For example,

$O_{1..}$ = number of patients operated on by Surgeon 1 = $(14 + 10) + (397 + 44) = 465$.

$O_{.1.}$ = Number of deaths = $(14 + 4) + (10 + 2) = 30$.

$O_{..1}$ = number of children with RACHS 1–3 = $(14 + 4) + (397 + 342) = 739$.

Two variable marginal totals can be obtained by summing the O_{ijk} over any single subscript. For example, the number of children who died with Surgeon 1, summed over all RACHS grades, is $(14 + 10) = 24$.

In a three-way table of this type, any of the categories might be dependent or independent. In this table, for example, the outcomes (survival) depend on the severity of the lesions, and possibly depend on the surgeon; it is this last possibility that was being assessed.

Similar to 2×2 tables, calculate the expected frequencies in any cell as

$$E_{ijk} = \frac{n_{i..} \cdot n_{.j.} \cdot n_{..k}}{N^2}.$$

Then calculate the total chi-square as usual by summing the individual chi-squares.

Using the table, the degrees of freedom are $rdl - r - c - 1 + 2 = 2 \times 2 \times 2 - 2 - 2 - 2 + 2 = 4$.

E_{111} is $\frac{465 \times 30 \times 757}{862^2} = 14.212$; the other expected frequencies can be calculated in similar fashion. The value of the total chi-square is 52.34 with 4 degrees of freedom, $P < 0.0001$, so that we would reject the null hypothesis of independence among the three categories.

If we do not reject the null hypothesis, there is no reason to proceed further. If we do accept it, then examine interactions.

Although direct computation of these subgroups is not complicated, it is easier to use log-linear models. These are presented fairly simply by [Everitt \(1992\)](#). Some useful examples are described by Fisher and van Belle, [Chapter 7 \(Fisher and van Belle, 1993\)](#). These models calculate G^2 which is very similar to chi-square.

As an example, consider again the data on surgical mortality ([Table 14.19b](#)). These were analyzed by the Frequency section of <http://vassarstats.net/> which calculates all the interactions.

Table 14.19b Mortality rate (p = percentage) by surgeon

	Surgeon 1				Surgeon 2				
	Dead	Alive	Total	<i>P</i>	Dead	Alive	Total	<i>P</i>	
RACHS 1–3	14	397	411	0.034	RACHS 1–3	4	342	346	0.012
RACHS 4–6	10	44	54	0.23	RACHS 4–6	2	49	51	0.041
Total	24	441	465		Total	6	391	397	

More details can be found in [Table 14.19c](#).

Table 14.19c Partial independence
Surgeon vs RACHS

	1–3	4–6	Total
Surgeon 1	411	54	465
Surgeon 2	346	51	397
Total	757	105	862

Surgeon vs Survival

	Dead	Alive	Total
Surgeon 1	24	441	465
Surgeon 2	6	391	397
Total	30	832	862

RACHS vs survival

	Dead	Alive	Total
1–3	18	739	757
3–6	12	93	105
Total	30	832	862

There are three sets of comparisons: between complexity (A), between survival (B), and between surgeons (C).

When these are analyzed by the log linear method, we get Table 14.20.

Table 14.20 Log linear analysis of data

Source	G^2	Df	P
ABC	26.36	4	<0.0001
AB	15.64	1	<0.0001
AC	0.32	1	0.5716
BC	9.22	1	0.0024
AB(C)	16.82	2	0.0002
AC(B)	1.5	2	0.4724
BC(A)	10.4	2	0.0055

When all three categories are involved (ABC) there is lack of independence ($P < 0.0001$), so we proceed. There are differences between the mortality rates for the two surgeons, (BC $P = 0.0024$) but how do we evaluate these for comparable complexity of lesions?

In Table 14.20, for example, the association between the surgeon and severity is examined, independent of survival. Because AC has a low G^2 , with $P = 0.5716$, we conclude that the choice of surgeon and severity of disease are not related (first panel, perfectly reasonable), and also conclude that survival and complexity are associated, with AB having a high G^2 value, $P \leq 0.0001$ (third panel, also reasonable). In the middle panel, however, the large difference in mortality rates is important and is what we wanted to examine but does not allow for differences in severity of disease. The analysis shows that allowing for complexity of disease, surgeons and mortality are not independent; BC(A) has $G^2 = 10.4$ and $P = 0.0055$. In other words, the choice of surgeon does play a role in the surgical mortalities for matched complexity.

REFERENCES

- Altman, D.G., 1992. Practical Statistics for Medical Research. Chapman and Hall, London, p. 611.
- Armitage, P., Berry, G., Matthews, J.N.S., 2002. Statistical Methods in Medical Research. Blackwell, Oxford.
- Basile, A.S., Hughes, R.D., Harrison, P.M., Murata, Y., Pannell, L., Jones, E.A., Williams, R., Skolnick, P., 1991. Elevated brain concentrations of 1,4-diazepines in fulminant hepatic failure. New Engl. J. Med. 325, 473–478.
- Berkson, J., 1946. Limitations of the fourfold table analysis to hospital data. Biom. Bull. 2, 47–53.
- Bickel, P.J., Hammel, E.A., O'connell, J., W., 1975. Sex Bias in graduate admissions: data from Berkeley. Science 187, 398–404.
- Binder, H.J., Clemett, A.R., Thayer, W.R., Spiro, H.M., 1965. Rarity of hiatus hernia in achalasia. New Engl. J. Med. 272, 680–681.

- Bracey, G., 2003. Those misleading SAT and NAEP trends: Simpson's paradox at work. <https://www.onemeck.org/wp-content/uploads/2015/08/THOSE-MISLEADING-SAT-AND-NAEP-TRENDS.pdf>.
- Charig, C.R., Webb, D.R., Payne, S.R., Wickham, J.E., 1986. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *BMJ (Clin. Res. Ed.)* 292, 879–882.
- Cohen, J., 1988. *Statistical Power Analysis for Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Everitt, B.S., 1992. *The Analysis of Contingency Tables*. Chapman & Hall, London.
- Fisher, L.D., Van Belle, G., 1993. *Biostatistics. A Methodology for the Health Sciences*. John Wiley and Sons, New York.
- Fleiss, J.L., 1981. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, New York.
- Group, T. S. S. I. S., 1991. Treatment of migraine attacks with sumatriptan. *New Engl. J. Med.* 325, 316–321.
- Hernandez-Diaz, S., Schisterman, E.F., Hernan, M.A., 2006. The birth weight "paradox" uncovered? *Am. J. Epidemiol.* 164, 1115–1120.
- Hernberg, S., Nurminen, M., Tolonen, M., 1973. Excess mortality from coronary heart disease in viscose rayon workers exposed to carbon disulfide. *World Environ. Health* 10, 93–99.
- Lacroix, A.Z., Wienpahl, J., White, L.R., Wallace, R.B., Scherr, P.A., George, L.K., Cornoni-Huntley, J., Ostfeld, A.M., 1990. Thiazide diuretic agents and the incidence of hip fracture. *N. Engl. J. Med.* 322, 286–290.
- Mainland, D., 1953. The risk of fallacious conclusions from autopsy data on the incidence of diseases with applications to heart disease. *Am. Heart J.* 45, 644–654.
- Motulsky, H.J., 2015. Common misconceptions about data analysis and statistics. *Br. J. Pharmacol.* 172, 2126–2132.
- Muench, H., 1965. Hiatus hernia in achalasia. *New Engl. J. Med.* 272, 1134.
- Perera, R., 2006. Statistics and death from meningococcal disease in children. *BMJ* 332, 1297–1298.
- Roberts, R.S., Spitzer, W.O., Delmore, T., Sackett, D.L., 1978. An empirical demonstration of Berkson's bias. *J. Chronic Dis.* 31, 119–128.
- Sharpe, D., 2015. Your chi-square test is statistically significant: Now what? *Ptact. Assess. Res. Eval.* 20, 1–10. <http://pareonline.net/getvn.asp?v=20&n=8>.
- Wilcox, A.J., 2006. Invited commentary: The perils of birth weight—a lesson from directed acyclic graphs. *Am. J. Epidemiol.* 164, 1121–1123 (discussion 1124–5).
- Zar, J.H., 2010. *Biostatistical Analysis*. Prentice Hall, Upper Saddle River, NJ.

CHAPTER 15

Categorical and Cross-Classified Data: McNemar's and Bowker's Tests, Kolmogorov-Smirnov Tests, Concordance

PAIRED SAMPLES: McNemar'S TEST

The chi-square tests in [Chapter 14](#) are unpaired tests. It is, however, sometimes appropriate to pair data. [Catalona et al. \(1975\)](#) compared the reactivity to DNCB and PHA in each of 28 patients with genitourinary tract malignancies; reactivity to both agents, one on each arm, was tested in each patient. The data appear in [Table 15.1](#).

Table 15.1 Paired data set

PHA	DNCB		Total
	Reactive	Nonreactive	
Reactive	13	3	16
Nonreactive	4	8	12
Total	17	11	28

This 2×2 table resembles a typical 2×2 chi-square table, but it is not because the results are pairs of results; the units are pairs, not individuals. The table shows that 13 patients reacted to both agents, 3 patients reacted to PHA but not to DNCB, 4 patients reacted to DNCB but not to PHA, and 8 patients did not react to either agent. By chi-square analysis, χ^2_T is 6.39 and $P=0.0115$; the odds ratio is 8.667. No clear interpretation of the chi-square and the odds ratio can be obtained. The odds ratio implies that the proportion of reactors to DNCB is higher in those who are also reactors to PHA, and the chi-square value indicates that we can reject the null hypothesis that this proportion is similar in those who do and who do not react to PHA. However, that was not the scientific question being investigated, which was whether there was a difference in the reactivity of the patients to the two agents, and the chi-square test does not help to evaluate this question. Instead, use McNemar's test. The proportion of PHA reactive patients who also react to DNCB is $\frac{a}{a+b}$, ($=\frac{13}{13+3}$), and the proportion of DNCB reactive patients who react to PHA is $\frac{a}{a+c}$, ($=\frac{13}{13+4}$). The only

way that these two proportions can differ is if b (3) and c (4) are unequal, so it is these two cells that we focus on; those who react to both agents, or who react to neither, do not help to evaluate differences between reactivity to these agents. Therefore focus on the two sets of discordant results: 3 patients who react to PHA but not to DNCB, and 4 patients who react to DNCB but not to PHA. There is no real difference between these, and if formal testing were needed one could use a 2×1 chi-square or the binomial theorem. Although hardly necessary, these calculations can be done online at <http://www.graphpad.com/quickcalcs/McNemar1.cfm>, and <http://vassarstats.net/> (see Frequency data).

Table 15.2 shows data based on a study done of the response of preterm infants with a patent ductus arteriosus to indomethacin, a drug purported to lead to closure of the ductus arteriosus. Because the probability of a persistent ductus arteriosus depends on the degree of prematurity, infants were paired based on their weight at birth, and one member of the pair chosen at random was given indomethacin whereas the other member of the pair was given a placebo.

Table 15.2 Paired patent ductus arteriosus study

Indomethacin	Placebo	
	Closed	Open
Closed	65	27
Open	13	40

Because this is a paired test, compare the 27 pairs of patients whose ductuses closed with indomethacin but not with placebo with the 13 patients whose ductuses closed with placebo but not with indomethacin. The null hypothesis of no difference in responsiveness to the two agents gives us an expected number of $\frac{13+27}{2}=20$, so that

$$\chi^2_T = \frac{(|27-20| - 0.5)^2}{20} + \frac{(|13-20| - 0.5)^2}{20} = 4.22 \text{ which, with one degree of free-}$$

dom, gives $P=0.04$. It is thus reasonable to reject the null hypothesis and to believe that indomethacin is more likely than the placebo to cause closure of the ductus arteriosus. If we had analyzed this table with the chi-square test, then $\chi^2_T=26.954$, and $0.0001 > P$, but we would have difficulty drawing conclusions from the analysis.

(Excluding the Yates' adjustment for discontinuity, this calculation is the same as $\frac{(27-13)^2}{27+13}$, or more generally $\frac{(b-c)^2}{b+c}$, where b and c are the two discordant frequencies. As a rule of thumb, there should be at least 10 discordant pairs for accurate computation of χ^2_T .

It is of interest to analyze this data as an unpaired chi-square test. To extract the correct numbers, remember that the count in each cell indicates the response of two infants. The number of ductuses that closed with indomethacin is 92—the 65 members of the pairs in

which both members closed with indomethacin and placebo and the 27 members of the pairs that closed with indomethacin but not with placebo. The remaining numbers can be worked out similarly, and the results (with Yates' correction) are given in [Table 15.3](#).

Table 15.3 Unpaired patent ductus arteriosus study

	Closed		Open		Total
Indomethacin	92	85	53	60	145
	7		−7		
	6.5	42.25	−6.5	42.25	
	<i>0.50</i>		<i>0.70</i>		
Placebo	78	85	67	60	145
	−7		7		
	−6.5	42.25	6.5	42.25	
	<i>0.50</i>		<i>0.70</i>		
Total		170		120	290

Observed values in bold type. Chi-square in italic type.

$\chi^2_T = 2.40$, 1 d.f. and $P = 0.12$; that is, we cannot reject the null hypothesis. This example confirms that a paired test is more sensitive than an unpaired test.

If the number of discordant observations is small relative to the number of concordant observations, then McNemar's test is still valid but the difference between the groups, although rejecting the null hypothesis, becomes trivial when related to the sample size. Assume, for example, the results shown in [Table 15.4](#).

Table 15.4 Hypothetical paired study

	Group A	
Group B	Success	Failure
Success	4000	27
Failure	13	3000

McNemar's test gives the same conclusion as for the data in [Table 15.2](#). Nevertheless, common sense tells us that either there is a success in both groups or a failure in both groups, and that very seldom will there be discordance. Once again, the distinction between statistical significance and importance needs to be considered. Furthermore, although the concordant cells a and d play no part in the McNemar test per se, they are a factor in estimating the variance of the difference between $\hat{p}_1 = \frac{a+c}{N}$ and $\hat{p}_2 = \frac{a+c}{N}$ (Selvin, 1995).

$$\text{Variance } \hat{p}_1 - \hat{p}_2 = \frac{(a+d)(b+c) + 4bc}{N^2}$$

Thus if a and d are very large, so are the variance and the resulting confidence limits.

BOWKER'S TEST

A similar paired test with 3 or more categories, the Bowker test, can also be done. Consider 3 categories of improved, unchanged, and worse, with the two treatments given to paired patients (Table 15.5).

Table 15.5 Paired test with 3 categories

Treatment B	Treatment A			Total
	Improved	Unchanged	Worse	
Improved	50	18	35	103
Unchanged	7	20	9	36
Worse	8	12	30	50
Total	65	50	74	189

To perform Bowker's test, do three separate McNemar tests: Improved vs Unchanged, Improved vs Worse, and Unchanged vs Worse. Each of these gives a value for χ^2 , calculated as $\frac{(b-c)^2}{b+c}$. These are summed to give the total Bowker χ^2 . This is then evaluated by a chi-square distribution with degrees of freedom $\nu = \binom{k}{2}$, where k is the number of categories (that can be >3). For the example above, the chi-squares for each of the individual McNemar tests are

$$\frac{(18-7)^2}{18+7} = 4.84, \quad \frac{(35-8)^2}{35+8} = 16.95, \quad \text{and} \quad \frac{(9-8)^2}{9+8} = 0.059.$$

The total chi-square is 21.85 with 3 df, and $P=0.00007$, and the universal null hypothesis H_0 : $p_{12}=p_{21}$, $p_{13}=p_{31}$, and $p_{23}=p_{32}$ can be rejected. The investigator then inspects the individual paired tests to decide which of the alternative hypotheses to consider.

TESTING ORDERED CATEGORICAL DATA: KOLMOGOROV-SMIRNOV (K-S) TESTS

If the categories are ordered, the chi-square test may not be the best test to use. Kolmogorov-Smirnov tests were developed to compare ordered distributions. Tests designed to compare an observed sample distribution and a specified theoretical distribution are termed Kolmogorov-Smirnov one-sample tests. Tests designed to compare two (or occasionally more) observed distributions are termed Kolmogorov-Smirnov two-sample tests. These tests do not depend on parameters of the distributions but assume that the distribution of the underlying variable is continuous.

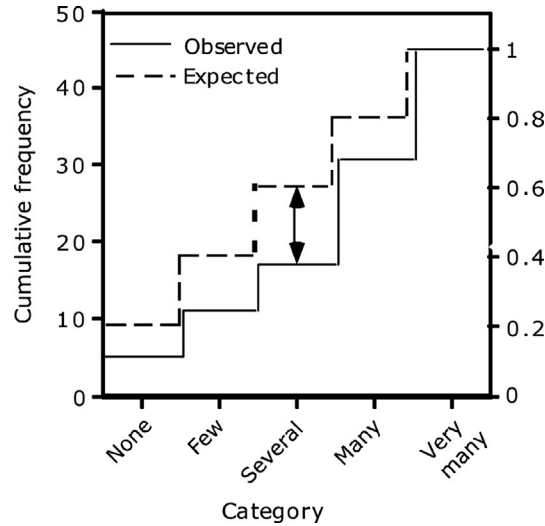


Fig. 15.1 The cumulative frequency for the observed and hypothesized distributions is plotted against the ordered frequencies for the number of ectopic beats (see later). The left-hand vertical scale is in absolute cumulative numbers and the right-hand vertical scale is the relative cumulative frequency. The vertical double arrow indicates the maximal vertical difference.

These tests cumulate the frequencies for each of the distributions, observed or theoretical, convert the cumulative absolute frequencies to cumulative relative frequencies (i.e., as fractions of the total number), plot the relative cumulative frequencies for each distribution on the vertical (Y) axis against the ordered groups on the horizontal (X) axis, and then examine the vertical differences between the relative cumulative frequencies for the two distributions at each ordered group (Fig. 15.1, based on Table 15.6).

If the distributions are consistent with having come from the same population, then the vertical differences are unlikely to be large. The sampling distributions of the maximal differences if the null hypothesis is true have been determined. Therefore if the maximal vertical difference exceeds the $1 - \alpha$ quantile, the null hypothesis can be rejected at the level α .

Kolmogoro-Smirnov One-Sample Test

After surviving a myocardial infarction, patients may be studied for 24h to determine how many ventricular ectopic beats per hour they have to assess their risk of sudden death. The average hourly number of these ectopic beats is assessed as none, few (1–10), several (11–30), many (31–300), and very many (≥ 300). If the numbers in each group are 5, 6, 6, 14, and 14, for a total of 45, is there any evidence that the numbers in each group are or are not randomly distributed? The null hypothesis H_0 is that the numbers should be the same in each group, namely, 9 in each group.

We could treat this as a contingency table and ask if there was any relationship between the frequency category and the number in it by performing a chi-square test (Table 15.6).

Table 15.6 One-sample test

	None	Few	Several	Many	Very many	Total
Observed	5	6	6	14	14	45
Expected	9	9	9	9	9	45
Difference	−4	−3	−3	5	5	
Difference ²	16	9	9	25	25	
χ^2	1.778	1.000	1.000	2.778	2.778	9.334

$$\chi^2_T = 9.334, 4 \text{ d.f.}, P = 0.0533.$$

In this example, it would not make any difference if we interchanged the positions of any of the columns. In the chi-square test, the categories are merely names. However, we know that the categories are ordered, and should ask if we can make use of this information to create a more sensitive test.

One approach is to use the one-sample Kolmogorov-Smirnov test. To do this, H_0 is the same null hypothesis that the observed distribution does not differ substantially from some expected distribution of interest. Then the observed and expected numbers are cumulated (Table 15.7), and their absolute deviations are calculated. Make all the cumulative numbers fractions of the total.

Table 15.7 Layout for Kolmogorov-Smirnov test

	None	Few	Several	Many	Very many	Total
Observed	5	6	6	14	14	45
Cumulative observed (Co)	5	11	17	31	45	
Relative Co	5/45 = 0.111	11/45 = 0.244	17/45 = 0.378	31/45 = 0.6891	1	
Expected	9	9	9	9	9	45
Cumulative expected (Ce)	9	18	27	36	45	
Relative Co	9/45 = 0.200	18/45 = 0.400	27/45 = 0.600	36/45 = 0.800	1	
Relative difference	0.089	0.156	0.222	0.111	0	

If the null hypothesis is true, then none of the relative differences should be very big. The sampling distribution of these differences has been worked out and tables give the critical values of the maximal fractional difference for different total numbers. These are provided at <http://www.real-statistics.com/statistics-tables/kolmogorov-smirnov-table/>. If $N > 35$, the 5% value of α is $\frac{1.36}{\sqrt{N}}$ and the 1% value is $\frac{1.63}{\sqrt{N}}$. The largest relative difference is 0.222. If the biggest difference is greater than the specified value of α , then the null hypothesis is rejected at that value of α . The 0.05 critical value for $N = 45$ is 0.1984, and the 0.02 value is 0.2262, so that $P < 0.05$ that the null hypothesis is true, that is, we can reject the null hypothesis at the 0.05 level. An online program is available at

http://www.physics.csbsju.edu/stats/KS-test.n.plot_form.html (which needs at least 10 data categories), http://www.wessa.net/rwasp_Reddy-Moores%20K-S%20Test.wasp#charts, and <http://www.real-statistics.com/non-parametric-tests/goodness-of-fit-tests/one-sample-kolmogorov-smirnov-test/>.

We would obtain a different maximal fractional difference if we changed the order of the columns (Table 15.8).

Table 15.8 Revised K-S test to change order of data

	None	Many	Few	Several	Very many	Total
Observed	5	14	6	6	14	45
Cumulative observed (Co)	5	19	25	31	45	
Related Co	0.111	0.422	0.556	0.689		
Expected	9	9	9	9	9	45
Cumulative expected (Ce)	9	36	18	27	45	
Related Ce	0.200	0.800	0.400	0.600		
Related difference	0.089	0.378	0.156	0.089	0	

In this instance, the deviation from the null hypothesis is even greater; the critical value for 0.01 is 0.237, so that $P < 0.01$.

As in most hypothesis tests, there are three possibilities if H_0 is rejected. The usual two-sided test is H_0 : Observed distribution = Expected distribution.

H_A : Observed distribution \neq Expected distribution.

However, one-sided tests are also possible. There are two of these:

a. H_0 : Observed distribution = Expected distribution

H_A : Observed distribution $>$ Expected distribution,

and

b. H_0 : Observed distribution = Expected distribution

H_A : Observed distribution $<$ Expected distribution.

For a, calculate the greatest vertical difference by which the observed distribution exceeds the expected distribution; for b, calculate the greatest vertical difference by which the observed distribution falls below the expected distribution. The test is done in the same way, but the Table is examined under one-sided values for α .

Problem 15.1 Test the following distribution for difference from a uniform distribution:

4, 11, 19, 31, 42, 50, 63, 68

Kolmogorov-Smirnov Two-Sample Test

This can be used to compare two (or more) samples to find out if they could have been drawn from a single population. Unlike some other tests, which test differences between

the means or the medians, these tests are sensitive to differences in shapes of the distributions as well.

The assumptions behind the test are that the samples are random samples, that the two samples are mutually independent, that the measurement scale is at least ordinal, and, for exact results, that the random variables should be continuous. If the variables are discrete, the test remains valid but becomes conservative.

Set up the two cumulative frequency distributions, just as in Table 15.9. Because the total numbers (m, n) in each group can be different, the relative cumulative frequency distributions must be calculated. The maximal vertical difference between these distributions is obtained, multiplied by mn ,¹ and the result compared with the quantile values in the appropriate table. Such tables can be found at http://books.google.com/books?id=LlkoZVPzCoC&pg=PA32&source=gbs_toc_r&cad=4#v=onepage&q&f=false, (for equal group numbers) and <http://www.real-statistics.com/statistics-tables/two-sample-kolmogorov-smirnov-table/>, and the test can be performed by <https://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/ks.htm>.

Alternatively the maximal difference allows us to reject the null hypothesis at $\alpha = 0.05$ (two-tailed) if it exceeds $1.36\sqrt{\frac{N_1 + N_2}{N_1 N_2}}$, where N_1 and N_2 are the respective sample sizes (Siegel and Castellan, 1988). As an example, consider a study of the occurrence of nausea after two different chemotherapeutic agents A and B (Table 15.9):

The largest difference is 0.1423. The critical value for $\alpha = 0.05$ is $1.36\sqrt{\frac{N_1 + N_2}{N_1 N_2}} = 1.36\sqrt{\frac{82 + 73}{82 \times 73}} = 0.2188$. Because this exceeds the maximal difference, we should not reject the null hypothesis. This is the same conclusion as obtained from the

Table 15.9 Comparison of degree of nausea after two chemotherapeutic agents

Agent		None	Vague	Slight	Mild	Moderate	Severe	Very severe
A	Observed O_A	9	18	15	10	11	10	9
A	Cumulative O_A	9	27	42	52	63	73	82
A	Relative O_A	0.1098	0.3297	0.5122	0.6341	0.7683	0.8902	1
B	Observed O_B	5	10	12	13	13	10	10
B	Cumulative O_B	5	15	27	40	53	63	73
B	Relative O_B	0.0685	0.2055	0.3699	0.5479	0.7260	0.8630	1
	Relative difference	0.0413	0.1242	0.1423	0.0862	0.0423	0.0272	

¹ If group numbers differ, some tables give critical values for D_{mn} (D is the maximal difference). See <http://sparky.rice.edu/astr360/kstest.pdf>. Alternatively, compute $\chi^2 = 4D_{mn}^2 \frac{mn}{m+n}$, and evaluate this for 2 d.f.

online program <http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/ks.htm>; this allows different group numbers. For $\alpha=0.025$ or 0.01 the previous formula is used with factors 1.48 and 1.63, respectively, in place of 1.36.

If we use the formula in footnote 1 because the numbers are too large for the tables,
$$\chi^2 = 4 \times 0.1423^2 \times \frac{82 \times 73}{82 + 73} = 3.1280.$$
 With 2 df, $P=0.2092$, so that we should not reject the null hypothesis.

Problem 15.2 Use the K-S test to determine if the following two distributions are comparable

8	17	31	47	52	73	89
4	6	19	36	44	68	93

CONCORDANCE (AGREEMENT) BETWEEN OBSERVERS

Two Categories

At times, we want to know how close the agreement (concordance) is between two sets of data; for example, if two radiologists examine the films of an upper gastrointestinal barium study, how often will they agree or disagree about the presence of an ulcer? The results of such studies can be set out in 2×2 tables, but neither chi-square nor McNemar’s tests are appropriate to answer the question. Consider the data set out in Table 15.10.

Here both radiologists diagnose ulcers in 20 patients and agree that there are no ulcers in 85 patients. However, radiologist A diagnoses an ulcer in 9 patients whom radiologist B regards as having no ulcer, whereas radiologist B diagnoses ulcers in 4

Table 15.10 Interobserver agreement
Radiologist A

Radiologist B	Ulcer	No ulcer	Total
Ulcer	20	4	24
No ulcer	9	85	94
Total	29	89	118

patients whom radiologist A considers as being normal. How can we specify the degree of agreement of these two radiologists? The percentage of agreement is.

$$\left(\frac{a+d}{N}\right) \times 100;$$

and for the example in Table 15.10 is $\left(\frac{20+85}{118}\right) \times 100 = 89.0\%$. However, this index does not take account of the degree of agreement expected by chance. We can calculate the expected values in each cell if the null hypothesis is true as we did for the chi-square test: the expected value for a is $\frac{29 \times 24}{118} = 59.90$, and the expected value for d is $\frac{89 \times 94}{118} = 70.9$. Now calculate the percentage of agreement using the expected numbers to get $\left(\frac{5.90+70.9}{118}\right) \times 100 = 65.1\%$. How much better than chance agreement did they do? One way of assessing this is to use the κ (kappa) statistic (Cohen, 1960).

$$k = \frac{p_o - p_e}{1 - p_e},$$

where p_o is the observed proportion of agreements, and p_e is the chance expected proportion of agreements. The agreement expected by chance is obtained by adding the chance-expected agreements in cells a and d —in which both observers agree or disagree with the diagnosis—and dividing by the total number of observations.

If the observed and chance expected agreements are the same, the numerator and κ will be zero. If the observed agreement is perfect and chance agreement is zero, then the numerator and denominator will be 1, and so will κ . If the observed agreement is less than the chance expected agreement, then κ will be negative.

In the example given in Table 15.5, $k = \frac{0.89 - 0.65}{1 - 0.65} = 0.69$. This result can be evaluated in two ways. One is to compute the standard error of κ when the null hypothesis is true, that is, when $\kappa = 0$, as

$$k_{\sigma 0} = \sqrt{\frac{p_e}{N(1 - p_e)}}.$$

In the example, $k_{\sigma 0} = \sqrt{\frac{0.65}{118(1 - 0.65)}} = 0.1255$, and $z = \frac{k}{k_{\sigma 0}} = \frac{0.69}{0.1255} = 5.4980$, $P < 0.00001$

We reject the null hypothesis that the observed agreement is due to chance.

This formula for the standard error is approximate, and a more accurate but complex formula was described by Fleiss (1981). Fortunately, this formula was used in the online

application <https://www.graphpad.com/quickcalcs/kappa1/>. This gave 0.081 for the standard error.

The other way to evaluate κ is to assess qualitatively the degree of agreement as <0 = poor; $0-0.20$ = slight; $0.21-0.40$ = fair; $0.41-0.60$ = moderate; $0.61-0.80$ = substantial; $0.81-1.00$ = almost perfect. Kappa may be calculated online at <https://www.easycalculation.com/statistics/cohens-kappa-index.php>, <http://faculty.vassar.edu/lowry/kappa.html>, and <https://www.graphpad.com/quickcalcs/kappa1.cfm>.

More Than Two Categories

There may be >2 categories to be evaluated. For example, two physicians may be evaluating 4 grades of pneumonia: +, ++, +++, or +++++, based on considering fever, white blood cell count, cognitive state, physical findings in the lungs, and chest X-ray findings (Table 15.11).

Table 15.11 Interrater agreement

Physician B	Physician A				Total
	+	++	+++	++++	
+	72	18	7	1	98
++	22	53	6	1	82
+++	9	7	24	4	44
++++	3	10	7	6	26
Total	106	88	44	12	250

Agreement shown in bold enlarged type.

To avoid the errors of doing these calculations by hand, use interactive online calculators that may include the option of comparing the concordance among more than two observers: <http://graphpad.com/quickcalcs/Kappa2.cfm> and <http://www.singlecaseresearch.org/calculators/pabak-os>.

$$k = \frac{0.62 - 0.32}{1 - 0.32} = 0.44.$$

Weighted Kappa

If the categories are nominal, this form of kappa set out previously has to be used. Sometimes, however, the categories form an ordered array, just as in the example above. Then diagnosing + instead of ++ is not as serious an error as diagnosing + instead of +++ or ++ +. To deal with this, each observed value is weighted in one of two ways. If each category is regarded as one step different from the adjacent one, use unit weights. If the error

is proportionately more serious as the disagreement becomes more marked, then a quadratic weight is given.

Then for linear weights use $\text{Weight} = 1 - \frac{|\text{distance}|}{\text{maximal possible distance}}$, and for quadratic weights use $\text{Weight} = 1 - \frac{|\text{distance}|^2}{\text{maximal possible distance}}$.

Use an online program that provides values for unweighted and both weighted kappa statistics, as well as standard errors and confidence intervals: <http://faculty.vassar.edu/lowry/kappa.html>.

If using weights $p_o = 0.905$ and $p_e = 0.797$, the quadratic weighted kappa is $k = \frac{0.905 - 0.797}{1 - 0.797} = 0.53$. With linear weighting, $\text{kappa} = 0.48$.

Both of these weighted kappa estimates are a little higher than the unweighted kappa.

Problem 15.3 Two subjects A and B are given a series of stimuli and asked to assess the degree of discomfort as

1 (least uncomfortable), 2, 3, and 4 (most uncomfortable). The following matrix shows their ratings. Calculate kappa (weighted and unweighted).

Subject B	Subject A			
Degree	1	2	3	4
1	16	3	1	1
2	3	17	2	2
3	4	6	14	3
4	1	1	5	16

Uebersax (2008) has described some of the issues involved in the rating comparisons. We need to know why they are being done, and how we interpret any differences that are found. If, for example, different radiologists disagree about the presence or absence of an ulcer, is this because they are using different criteria? And if so, can these criteria be clarified?

It is not clear whether calculating kappa gives useful information that is not obvious by inspecting the data. The suggestion by Cichetti and Feinstein (1990) of reporting p_{pos} and p_{neg} (see later) as well makes it easier to determine where agreements and disagreements occur.

Cautionary Tales

Do not accept the kappa statistic unreservedly. Like all omnibus statistics that uses a single number to summarize a batch of data, there are unseen dangers lurking, even with as simple a procedure as kappa.

If a finding is rare (low prevalence), kappa might be very low despite what appears to be good agreement (Viera and Garrett, 2005). As an example, these authors cited a study by Metlay et al. (1997) of pneumonia in which tactile fremitus was diagnosed with 85% agreement but with kappa only 0.01, and provided an explanation based on the rarity of the finding. As a hypothetical example, consider Table 15.12.

There is obviously a high degree of agreement (95%), but $\kappa = 0.26$.

Table 15.12 Problems with kappa
Rater A

Rater B	Rater A	
	Yes	No
Yes	94	3
No	2	1

Cicchetti and Feinstein (1990) and Feinstein and Cicchetti (1990) studied other paradoxes associated with kappa and found that the same high percentage of agreement between 2 observers could produce either a high or a low value of kappa, depending on the distribution of the marginal subtotals (Tables 15.13 and 15.14).

Tables based on examples provided by Feinstein and Cicchetti (1990).

Table 15.13 Variation of kappa
Observer A

Observer B	Observer A		
	Yes	No	Total
Yes	87	7	94
No	6	4	10
Total	93	11	104

Table 15.14 Kappa paradox
Observer A

Observer B	Observer A		
	Yes	No	Total
Yes	72	12	84
No	2	24	26
Total	74	36	110

In both sets of data there is substantial agreement between the two observers, with the observed agreement p_o being 0.88 in panel a and 0.87 in panel b. Nevertheless, the kappa values for each are very different, being 0.31 for panel a and 0.69 for panel b.

Cicchetti and Feinstein (1990) recommended that in addition to kappa the investigator should report the value of p_{pos} and p_{neg} . To calculate p_{pos}

$$p_{\text{pos}} = \frac{\frac{a}{R^+ + C^+}}{\frac{2}{2}} = \frac{2a}{R^+ + C^+}.$$

In this formula use the format for the usual 2×2 table, where a is the number of positive agreements between the two observers, and R^+ and C^+ are the respective row and column subtotals associated with Tables 15.13 and 15.14.

$$p_{\text{pos}} = \frac{2 \times 87}{93 + 94} = 0.93 \text{ for Table 15.13, and } p_{\text{pos}} = \frac{2 \times 72}{74 + 84} = 0.91 \text{ for Table 15.14.}$$

To calculate p_{neg}

$$p_{\text{neg}} = \frac{\frac{d}{R^- + C^-}}{\frac{2}{2}} = \frac{2d}{R^- + C^-}.$$

Here d is the number of negative agreements between the 2 observers, and R^- and C^- are the associated row and column subtotals. Kappa is the weighted sum of the two proportions. For Tables 15.13 and 15.14.

$$p_{\text{neg}} = \frac{2 \times 4}{11 + 10} = 0.38 \text{ for Table 15.13 and } p_{\text{neg}} = \frac{2 \times 24}{36 + 26} = 0.77 \text{ for Table 15.14.}$$

These precautions have been emphasized by others (Byrt et al., 1993). The kappa statistic is affected by bias, the frequency with which raters choose a particular category, because this unbalances the table. The bias index (BI) is $\text{BI} = \frac{b - c}{N}$, and larger values of BI yield a larger kappa. It is also affected by the prevalence index (PI), defined as $\text{PI} = \frac{a - d}{N}$; the higher the PI the more unbalanced the table the lower the value for kappa. Both of these can be dealt with in a single formula (PABAK)

$$\text{PABAK} = \frac{\frac{(a + d)}{2N} - 0.5}{1 - 0.5} = 2p_o - 1$$

For the data in Table 15.11, PABAK is $(2 \times 0.89) - 1 = 0.78$, as compared with the uncorrected kappa of 0.69. For the example in Table 15.12, PABAK is 0.925 compared with $\kappa = 0.26$.

If BI and PI are small, kappa and PABAK are similar. If they are large, the two measures differ, and caution is needed in interpreting the data. Kappa cannot be compared in different studies if BI or PI is large, so that it is best to include calculation of PABAK from <http://www.singlecaseresearch.org/calculators/pabak-os>.

One other difficulty is that although kappa theoretically ranges from +1 to -1, this is true only in the rare instance when all the marginal subtotals are equal. If they are not, the maximal value of kappa may be well below one, thus making it even more difficult to interpret (Flight and Julious, 2015).

REFERENCES

- Byrt, T., Bishop, J., Carlin, J.B., 1993. Bias, prevalence and kappa. *J. Clin. Epidemiol.* 46, 423–429.
- Catalona, W.J., Tarpley, J.L., Potvin, C., Chretien, P.B., 1975. Correlations among cutaneous reactivity to DNCB, PHA-induced lymphocyte blastogenesis and peripheral blood E rosettes. *Clin. Exp. Immunol.* 19, 327–333.
- Cicchetti, D.V., Feinstein, A.R., 1990. High agreement but low kappa: II. Resolving the paradoxes. *J. Clin. Epidemiol.* 43, 551–558.
- Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educ Psychol. Measurement* 20, 37–46.
- Feinstein, A.R., Cicchetti, D.V., 1990. High agreement but low kappa: I. the problems of two paradoxes. *J. Clin. Epidemiol.* 43, 543–549.
- Fleiss, J.L., 1981. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, New York.
- Flight, L., Julious, S.A., 2015. The disagreeable behaviour of the kappa statistic. *Pharm. Stat.* 14, 74–78.
- Metlay, J.P., Kapoor, W.N., Fine, M.J., 1997. Does this patient have community-acquired pneumonia? Diagnosing pneumonia by history and physical examination. *JAMA* 278, 1440–1445.
- Selvin, S. 1995. *Practical Biostatistical Methods.*, Belmont, CA, Wadsworth Publishing Company.
- Siegel, S., Castellan Jr., N.J., 1988. *Nonparametric Statistics for the Behavioral Sciences*, second ed. McGraw-Hill, New York.
- Uebersax, J., 2008. *Statistical Methods for Rater Agreement*. Available:<http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm>.
- Viera, A.J., Garrett, J.M., 2005. Understanding interobserver agreement: the kappa statistic. *Fam. Med.* 37, 360–363.

CHAPTER 16

Binomial and Multinomial Distributions

BASIC CONCEPTS

Introduction

Observations that fall into one of two mutually exclusive categories are dichotomous or binary. Any population with only two categories is a binomial population; newborn infants are male or female or the outcome of a treatment is cure or failure. We can ask certain questions about binomial populations. For example, what is the likelihood that in a family of five children there will be three girls? What is the probability that 5 patients out of 50 will die from meningitis? How likely is it that 7 out of 20 people with a certain type of seizure will develop paralysis?

The allocation to each category must be unambiguous. The number of items (counts) in one group (e.g., r successes) gives information about the proportion of the total number (n) attributed to one of the groups. Define the proportion of the number of one of the groups to the total (r/n) as π for a theoretical population and as p for a sample from that population. Therefore $\pi = \frac{\text{number of newborn girls}}{\text{total number of newborns}}$ in an infinitely large number of births, but the same proportion is p if related to a smaller number of births in a particular hospital. Because this is a dichotomous variable in which the total probability must add to 1, if the probability of one of the two categories is π , then the probability of the other category must be $1-\pi$. In a sample, the probability of one category is p , and the other is $1-p$, also called q . By convention, the probability of one category is a success and of the other category is a failure, without equating the terms “success” and “failure” with the desirability of the outcome. An examination of the dichotomous outcomes of a study is often termed a Bernoulli trial if

1. The number of trials is defined.
2. In each trial, the event of interest either occurs (“Success”) or does not occur (“Failure”).
3. The trials are independent, that is, the outcome of one trial does not affect the outcome of the next trial.
4. The probability of success does not change throughout the trial (Example 16.1).

Example 16.1

The proportion π for male births is approximately 0.5. (0.512 in the United States.) If a family has five children, how often would there be three girls? no girls? four boys? To approach this problem, consider what happens by flipping a fair coin many times, and recording whether it

comes down heads (H) or tails (T). This is a good analogy, because for heads in coin tossing π is exactly 0.5. Table 16.1 gives the possible results for the first few tosses.

Table 16.1 Binomial table

Number of tosses	Possible combinations					
1		1H	1T			
2		1HH	2HT	1TT		
3		1HHH	3HHT	3HTT	1TTT	
4		1HHHH	4HHHT	6HHTT	4HTTT	1TTTT
5	1HHHHH	5HHHHT	10HHHTT	10HHTTT	5HTTTT	1TTTTT

H = head; T = tail. The numbers (coefficients) indicate how many possible combinations there are for each set.

Turning this figure through 90 degrees gives a tree diagram (Fig. 16.1).

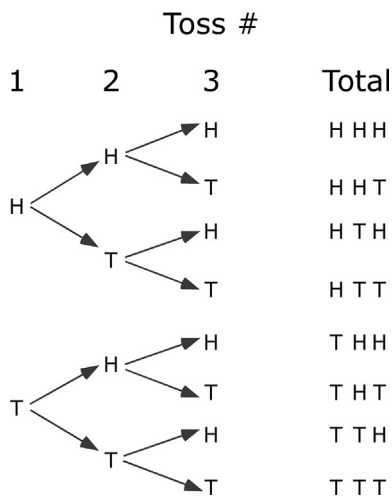


Fig. 16.1 Binomial table as tree diagram.

Consider the numbers of possible combinations of heads and tails for any given number of tosses.

First toss: a head (H) or a tail (T).

Second toss: HH, HT, TH, or TT. The difference between the two middle combinations is the order in which the outcomes occur. If order is immaterial, the outcomes are HH, 2HT, TT.

Third toss: HHH (three heads) once, HHT, HTH, THH (two heads and one tail), HTT, THT, TTH (one head and two tails), or TTT (three tails) once.

Fourth toss: Four heads (HHHH) once; 3 heads and one tail, and there are four ways of doing this (HHHT, HHTH, HTHH, THHH); 2 heads and 2 tails, and there are 6 ways of doing this (HHTT, HTHT, HTTH, TTHH, TTTH, THHT); one head and three tails, with four possible combinations; and 4 tails (TTTT) once.

Fifth toss: HHHHH, with only one such combination possible; four heads and one tail, with 5 possible combinations; three heads and two tails, with 10 possible combinations (work this out for practice); two heads and three tails, also with 10 possible combinations; one head and four tails, with 5 possible combinations; and five tails, with only one such combination.

Isolating the numbers of possible combinations (the coefficients) in the earlier diagram gives (Table 16.2).

Table 16.2 Bernoulli coefficients

<u>Number of tosses</u>	<u>Number of combinations</u>									
1			1		1					
2			1		2		1			
3			1		3		3		1	
4		1		4		6		4		1
5		1		5		10		10		5

Each line begins and ends with a 1, because there can be only one combination with all heads and one with all tails. In each line except the first, the number of combinations is the sum of the two nearest combinations in the preceding line. Thus 5 for five tosses is the sum of 1 and 4 which are above it in line 4, 10 is the sum of the 4 and the 6 above it in line 4, and so on. In each line, the pattern of coefficients is symmetrical because the coefficient for, say, 5–1 heads = 4 heads is 5, and this is the same as the coefficient for 5–4 heads = 1 head. More generally, in n trials, the coefficient for k successes is the same as for $n-k$ successes. This pattern is termed Pascal's triangle, described in a letter to Fermat in 1654, but it was known in China in the 12th century.

These numbers of combinations, also termed coefficients, are one of the two items needed to work out the probabilities. The other is the probability of getting any particular combination. The second item is easy to calculate. Assuming that the results of one toss do not influence any of the other results, then to get any combination with two tosses multiply the probabilities of each result. The probability of getting one head is 0.5.

The probability of getting two heads is $0.5 \times 0.5 = 0.5^2 = 0.25$. The probability of getting three heads is $0.5 \times 0.5 \times 0.5 = 0.5^3 = 0.125$. The probability of getting two heads and a tail is $0.5^2 \times 0.5 = 0.5^3 = 0.125$, and so on for other numbers of tosses and combinations. Now all we need to do is to multiply the probability of getting any particular combination by the number of times that combination occurs, that is, by the coefficient.

For example, for four tosses, we get the following probabilities:

$$P(\text{Four heads}) = 0.5 \times 0.5 \times 0.5 \times 0.5 = 0.5^4 = 0.0625.$$

$P(\text{Three heads and one tail}) = P(3 \text{ heads}) \times P(1 \text{ tail}) \times \text{number of combinations} = 0.5^3 \times 0.5 \times 4 = 0.25$ because there are 4 ways of having three heads and one tail: HHHT, HHTH, HTHH, THHH.

$P(\text{Two heads and two tails}) = 0.5^2 \times 0.5^2 \times 6 = 0.375$, because there are 6 possible combinations: HHTT, HTHT, HTTH, TTHH, TTTH, THHT.

$$P(\text{One head and three tails}) = 0.5^1 \times 0.5^3 \times 4 = 0.25.$$

$$P(\text{Four tails}) = 0.5^4 = 0.0625.$$

Adding up all these probabilities gives $0.0625 + 0.25 + 0.375 + 0.25 + 0.0625 = 1.0000$ (Example 16.2).

Example 16.2

Perform similar calculations when π is not 0.5. Take a 6-sided die, call the number 1 a success, and any of the numbers 2–6 a failure. Then the probability of getting a success is $1/6$, and the probability of getting a failure is $5/6$.

For one toss of the die, the probability of one success = $1/6$, and the probability of one failure = $5/6$.

For two tosses, the probability of two successes is $1/6 \times 1/6 = (1/6)^2 = 1/36$. The probability of one success and one failure is $1/6 \times 5/6 \times 2 = (1/6)^1 \times (5/6)^1 \times 2 = 10/36$. The probability of two failures is $5/6 \times 5/6 = (5/6)^2 = 25/36$.

For three tosses, the probability of one success is $1/6 \times 1/6 \times 1/6 = (1/6)^3 = 1/216$. The probability of two successes and one failure is $1/6 \times 1/6 \times 5/6 \times 3 = (1/6)^2 \times (5/6)^1 \times 3 = 15/216$. The probability of one success and two failures is $1/6 \times 5/6 \times 5/6 \times 3 = (1/6)^1 \times (5/6)^2 \times 3 = 75/216$. The probability of three failures is $5/6 \times 5/6 \times 5/6 = (5/6)^3 = 125/216$. The sum of these probabilities is $(1 + 15 + 75 + 125)/216 = 1$, as expected.

Bernoulli formula

With greater numbers of tosses, the calculations of the probabilities (especially the coefficients) become increasingly tedious. Fortunately, Jakob Bernoulli (1654–1705), one of a large, famous and disputatious family of Swiss physicists and mathematicians, showed that these expressions were the expansion of the algebraic expression $(\pi + [1 - \pi])^n$, where n represents the number of observations. For there to be X successes (with probability π of each success), there must therefore be $n - X$ failures. The probability for any value of r and

n is: $nCr\pi^r[1-\pi]^{n-r}$, where nCr (also written $\binom{n}{r}$) represents the number of possible combinations of r successes and $n-r$ failures (i.e., the binomial coefficient). This is the probability function.

From combinatorial theory (Chapter 12) nCr is: $\frac{n!}{r!(n-r)!}$.

In our example, the probability would be $\frac{n!}{r!(n-r)!}\pi^r[1-\pi]^{n-r}$. This formula has two components. The first, the binomial coefficient, $\frac{n!}{r!(n-r)!}$, gives the number of possible arrangements of r successes and $n-r$ failures as set out in part in Table 16.2. The second factor involving π is the probability for any of the different ways of getting r successes and $n-r$ failures.

The probability of getting 4 successes out of 6 tosses is:

$$\frac{6!}{4!(6-4)!}(1/6)^4(5/6)^{6-4} = \frac{720}{24 \times 2} \times \frac{1}{1296} \times \frac{25}{36} = 0.008038$$

The probability of getting one success out of 6 tosses is:

$$\frac{6!}{1!(6-1)!}(1/6)^1(5/6)^{6-1} = \frac{720}{120} \times \frac{1}{6} \times \frac{3125}{7776} = 0.401878.$$

To determine the probability of getting 7 successes out of 12 total trials in which $p=0.37$, this would be calculated from:

$$\begin{aligned} {}_{12}C^{12}7(0.37)^7(1-0.37)^{12-7} &= \frac{12!}{7!(12-7)!}0.37^7(1-0.37)^{12-7} \\ &= 792 \times 0.0009493 \times 0.09924 = 0.07461. \end{aligned}$$

What is the probability that 5 patients out of 50 with meningitis will die? To answer this question, we need information about the average proportion of these patients who die. Assume that this probability is $p=0.1$. Then the probability that we want is:

$$\begin{aligned} {}_{50}C^{50}5(0.1)^5(1-0.1)^{50-5} &= \frac{50!}{5!(50-5)!}(0.1)^5(1-0.1)^{50-5} \\ &= 2118760 \times 0.000010 \times 0.008727964 = 0.1849. \end{aligned}$$

Free online binomial calculators are available at <http://stattrek.com/onlinecalculator/binomial.aspx>, [<http://vassarstats.net/binomialX.html>](http://vassarstats.net/binomialX.html), <http://binomial.calculatord.com> and <http://nccalculators.com/statistics/binomial-distribution-calculator.htm> and http://www.statstodo.com/BinomialTest_Pgm.php

Problem 16.1 In an influenza epidemic the mortality rate is 0.7%. What is the probability of death in 3 out of the next 50 patients?

Binomial basics

The mean of a binomial distribution is $n\pi$, and its variance is $n\pi(1-\pi)$. Because both π and $1-\pi$ are less than one, their product is even smaller, so that the variance of a binomial distribution is less than the mean. If the mean and variance are calculated as proportions of 1, rather than as numbers based on the sample size, then divide the mean by n and the variance by n^2 to give π and $\frac{n\pi(1-\pi)}{n^2} = \frac{\pi(1-\pi)}{n}$, respectively.

The normal approximation

The binomial distribution is approximately normal if $n\pi$ and $n(1-\pi) > 9$ (Hamilton, 1990).

Fig. 16.2 shows the distribution for different values of n when $p=0.1, 0.3$, or 0.5 .

Fitting a binomial distribution to a set of trial results

With a typical Mendelian recessive gene, for example, for Tay-Sachs disease, the probability of a child having the disease if both parents carry the recessive gene is 0.25. One hundred such families with 5 children are examined, and the children with the disease are identified (Table 16.3). Does the observed distribution fit a binomial distribution?

Calculate the expected probabilities of having 0, 1, 2, 3, 4, or 5 children with the disease.

$$P(r=0) = 1 \times 0.25^0 \times 0.75^5 = 0.237305.$$

$$P(r=1) = 5 \times 0.25^1 \times 0.75^4 = 0.395508.$$

$$P(r=2) = \frac{5!}{2!3!} \times 0.25^2 \times 0.75^3 = 0.263672.$$

$$P(r=3) = \frac{5!}{2!3!} \times 0.25^3 \times 0.75^2 = 0.087891.$$

$$P(r=4) = 5 \times 0.25^4 \times 0.75^1 = 0.014648.$$

$$P(r=5) = 1 \times 0.25^5 \times 0.75^0 = 0.0009766.$$

Table 16.3 The expected values show the theoretical distribution for the recessive population

Number of involved children (r)	Observed f	Expected f	χ^2
0	26	23.73	0.22
1	37	39.55	0.16
2	22	26.37	0.72
3	11	8.79	0.56
4	4	1.46	4.42
5	0	0.10	0.10
Total	100		6.18

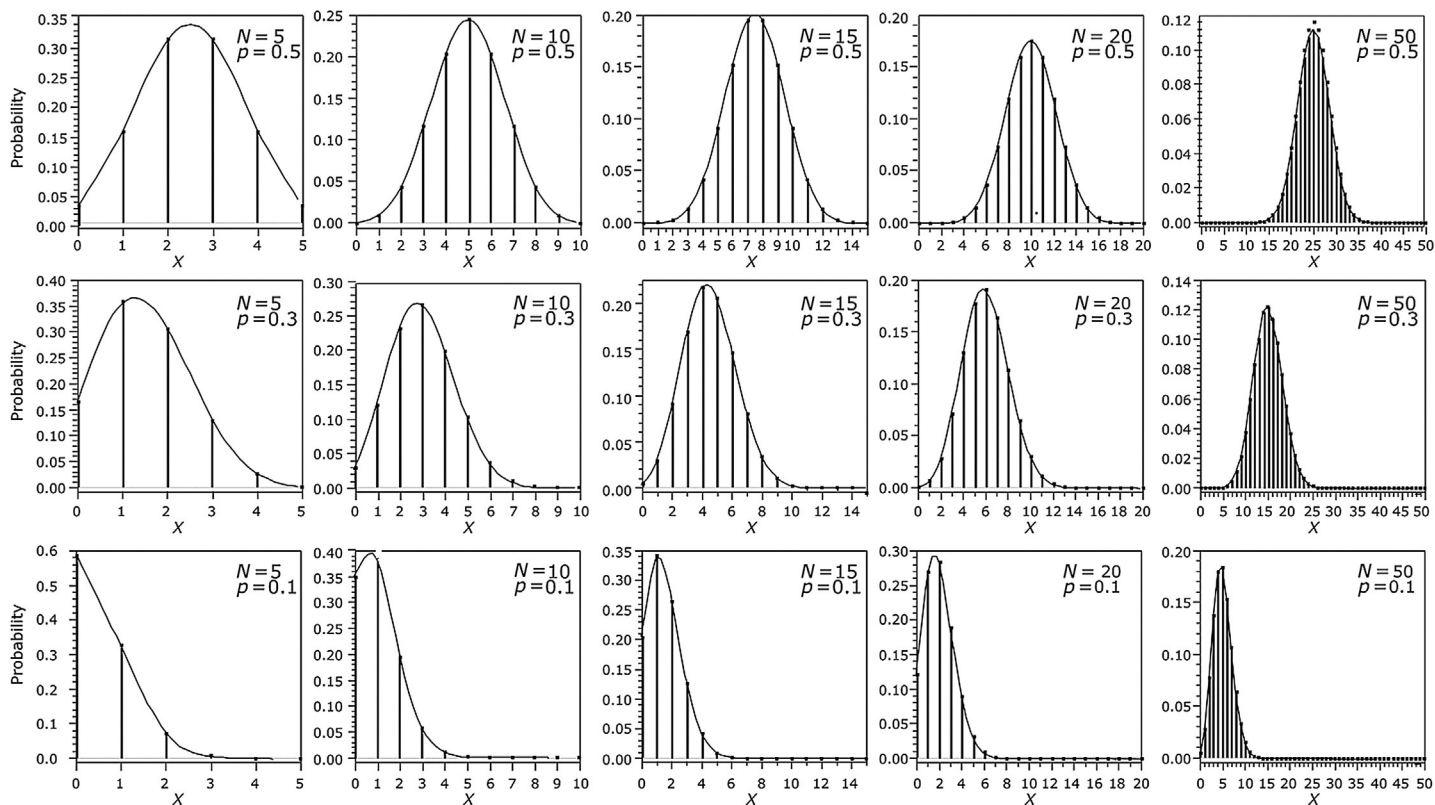


Fig. 16.2 The bigger the product of N and p , the closer the approximation to the normal curve.

These add up to 1.0000006, the difference from 1 being due to rounding off errors. Now multiply each probability by N , the total number, and compare observed and expected numbers by a chi-square test, as in [Table 16.3 \(Chapter 14\)](#).

The observed and expected values are similar. The total chi-square is 6.18 with 4 degrees of freedom, so that $P=0.1861$. There is therefore no convincing evidence against the null hypothesis that these samples are drawn from a population in which the mean probability of the disease occurring is 0.25.

Cumulative binomial probabilities

Some problems call for an estimate of more than one probability. For example, if for a Mendelian recessive characteristic, the risk of a child with the recessive disease is 0.25, how likely is it that there will be families of 5 children with two or more diseased children? This could be calculated by working out the probabilities of having 0 or 1 affected child, adding these together and then subtracting that total from 1. If the sample sizes were larger, however, this could be very time consuming. For example, in samples of 100 people, what is the probability that there will be >20 with type B blood group that occurs in about 14% of the population? Fortunately there are tables in which cumulative probabilities have been calculated and listed for varying values of π ([McGhee, 1985](#); [Kennedy and Neville, 1986](#); [Hahn and Meeker, 1991](#)) ([Example 16.3](#)). The calculations can be done easily on the free interactive websites http://onlinestatbook.com/analysis_lab/binomial_dist.html, <http://www.danielsoper.com/statcalc3/calc.aspx?id=71>

Example 16.3

Using the blood group data before, $n=100$, $p=0.14$. From the calculator, enter 0.14, 20, and 100, and get $P(X>20) \geq 0.0356$ and $P(X \geq 20) = 0.0614$.

Confidence limits

Continuity correction

A binomial series consists of the probabilities of discrete events—3 children, 7 teeth, and so on. Because these events are discrete they are represented by integers. When this probability set is examined as if it were a continuous distribution, an error appears as diagrammed in [Fig. 16.3](#).

For the sample illustrated before, with $n=20$ and $p=0.3$, the probability of $P \geq 10$, that is, the probability of getting 10 or more, can be calculated from the binomial theorem. This probability is 0.0480 (using the online calculator referred to previously). But with the normal approximation and the formula.

$$z = \frac{X_i - p}{\sqrt{Npq}} \text{ (see the following)}$$

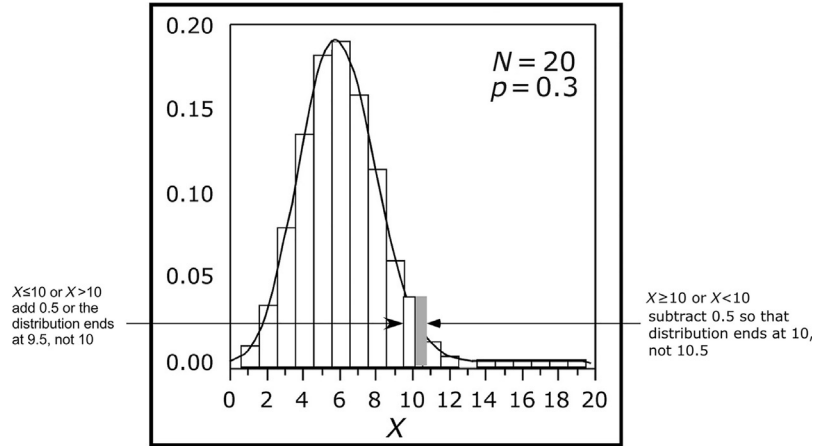


Fig. 16.3 Diagram of continuity correction.

$$z = \frac{10 - 20 \times 0.3}{\sqrt{20 \times 0.3 \times 0.7}} = \frac{4}{2.0494} = 1.95, p = 0.0256.$$

As shown in Fig. 16.3, this value must be too small. The value 10 really extends from 9.5 to 10.5, so that calculating the probability from the continuous curve using only 10 and above ignores the shaded area in the figure. To allow for this, subtract 0.5 from 10 and then

$$z = \frac{9.5 - 20 \times 0.3}{\sqrt{20 \times 0.3 \times 0.7}} = \frac{3.5}{2.0494} = 1.71, P = .0463.$$

This is much closer to the true probability. When N is large, the correction is less important.

The correction is not always by subtracting 0.5. Table 16.4 summarizes the changes. For example, for $X \geq 10$ you have to start at 9.5, because starting at 10.5 does not meet

Table 16.4 Continuity corrections

Discrete	Continuous	Correction
$X = 10$	$9.5 < X < 10.5$	± 0.5
$X > 10$	$X > 10.5$	$+0.5$
$X \leq 10$	$X < 10.5$	$+0.5$
$X < 10$	$X < 9.5$	-0.5
$X \geq 10$	$X > 9.5$	-0.5

the requirements. If you want $X < 10$, again subtract 0.5 or else the distribution will end at 10.5. For $X \leq 10$ add 0.5 or the distribution will end at 9.5. (See Fig. 16.3).

Sometimes the correction is not 0.5. For example, when calculating the confidence limits of a proportion, the numerator includes the number of events as a proportion of n .

The correction has to be on the same scale and becomes $\frac{0.5}{n} = \frac{1}{2n}$.

Using the continuity correction, for a one-tailed test, $z = \frac{(r - np) - 0.5}{\sqrt{npq}}$. (Do not correct if $r - np \leq 0.5$). For two-tailed tests, delete the fractional part of $r - np$ if it lies between 0 and 0.5, and replace it by 0.5 if the fractional part lies between 0.5 and 1.0. Therefore 7.3 becomes 7, 8.73 becomes 8.5, 5.0 becomes 4.5, and 3.5 becomes 3.0.

Confidence limits can be calculated online from <http://statpages.org/confint.html>, <http://www.danielsoper.com/statcalc3/calc.aspx?id=85>, <http://www.graphpad.com/quickcalcs/ConfInterval2.cfm>, and <http://www.biyee.net/data-solution/resources/binomial-confidence-interval-calculator.aspx>. (the last one is only for Windows operating system).

For example, if $N=20$ and $p=0.3$, what are the confidence limits for p ? From the calculator, the 95% confidence limits for p are 0.1189–0.5428, a very wide range because of the small sample size.

In larger samples, the binomial estimate of p is approximately normally distributed about population proportion π with standard deviation $\sqrt{\frac{\pi(1-\pi)}{n}}$. For the true but unknown standard deviation substitute the sample estimate $\sqrt{\frac{pq}{n}}$.

Hence the probability is approximately 0.95 that π lies between the limits $\pi - 1.96$ and $\pi + 1.96\sqrt{\frac{pq}{n}}$. With the correction for continuity, use $p \pm \left(1.96\sqrt{\frac{pq}{n}} + \frac{1}{2n}\right)$ (Example 16.4).

Example 16.4

If the proportion of people with group B blood in a sample of 100 people is 0.14, what are the 95% confidence limits for that proportion?

$p=0.14$, $q=0.86$, $n=100$. The 95% confidence limits are $p \pm \left(1.96\sqrt{\frac{pq}{n}} + \frac{1}{2n}\right) = 0.14 \pm \left[1.96\sqrt{0.14 \times 0.86/100} + 1/200\right] = 0.14 \pm [1.96 \times 0.034699 + 0.005] = 0.14 \pm 0.073 = 0.067$ to 0.213 (The exact interval is 0.0612–0.2072, so that the approximation is close).

Other slightly more complex formulas can be used, with Wilson's method being highly accurate (see later) (Curran-Everett and Benos, 2004).

Problem 16.2

A study (Jousilahti et al., 1998) showed that serum cholesterol in men in 1992 had the following distribution:

<5.0 mmol/L—6%; 5–6.4 mmol/L—37%; 6.5–7.9 mmol/L—41%; >8.0 mmol/L—16%.

Set 95% confidence limits for those with serum cholesterol 5.0–6.4 mmol/L in a sample of 200 men.

Comparing two binomial distributions

This can be done with the chi-square. Tables or graphs for performing these comparisons can be found in [Gardner and Altman \(1995\)](#). It can also be done by turning the mean values into proportions and using the free program <http://vassarstats.net> (see Proportions) that also provides confidence intervals.

Estimating sample size

If two binomial distributions are not substantially different, this may be because they are similar (i.e., that the null hypothesis is true) or else that the null hypothesis may not be true but the numbers in each sample are too small to allow us to reach that conclusion with confidence. If we have the data based on a 2×2 table, we can use the program referred to in the preceding paragraph. Sometimes, however, we have only the two proportions. For example, we know that the proportion of patients who die from treatment A is 0.4, and we want to know how many patients we will need to demonstrate a reduction on treatment B to 0.2,

A simple way is to use the normal approximation with the formula.

$$n_A = \frac{16p_{av}(1-p_{av})}{(p_A - p_B)^2}, \text{ where } p_A \text{ and } p_B \text{ are the two proportions being compared and } p_{av}$$

is their average, and n_A is the number in each group ([Lehr, 1992](#); [van Belle, 2002](#)). For our example

$$n_A = \frac{16 \times 0.4(1 - 0.4)}{(0.4 - 0.2)^2} = 95.$$

Alternatively, to determine sample size, use the free online program <http://www.cct.cuhk.edu.hk/stat/proportion/Casagrande.htm>, <http://statulator.com/SampleSize/ss2P.html>, or <https://www.stat.ubc.ca/~rollin/stats/ssize/b2.html>. These do not give identical answers because they are based on different principles, but they are close enough for practical use. In the order given, these programs give 91, 89, and 82.

Comparing probabilities

We may want to compare the ratio of two of the probabilities p_i/p_j . A $1 - \alpha$ confidence interval can be created from

$$\begin{aligned} \text{Log}_e L &= \log(p_i/p_j) - z(1 - \alpha/2) \left(\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \text{ and} \\ \text{Log}_e U &= \log(p_i/p_j) + z(1 - \alpha/2) \left(\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \end{aligned}$$

In the previous example, consider the ratio of two blood types, A and B ; $pA/pB = 0.43/0.12 = 3.58$. If this ratio is based on the observation of 100 individuals, 43 with A and 12 with B , what is the 95% confidence interval for this ratio?

$$\log L = \log(43/12) - 1.96\sqrt{1/43 + 1/12}, \text{ and}$$

$$\log U = \log(43/12) + 1.96\sqrt{1/43 + 1/12}.$$

$\log L = 1.2763 - 1.96 \times 0.6399 = 0.6364$, and $\log U = 1.9162$. The lower and the upper limits are the antilogarithms of these values, and these are 1.8897–6.7952. These limits would be different for different sample sizes.

ADVANCED OR ALTERNATIVE CONCEPTS

Exact confidence limits

Wilson's method for determining confidence limits (Newcombe, 1998) seems to be the most accurate method available. It involves solving the two formulas for the lower (L) and upper (U) confidence limits of a binomial proportion and includes the continuity correction:

$$L = \frac{2np + z^2 - 1 - z\sqrt{\{z^2 - 2 - 1/n + 4p(nq + 1)\}}}{2(n + z^2)}$$

$$U = \frac{2np + z^2 + 1 + z\sqrt{\{z^2 + 2 - 1/n + 4p(nq + 1)\}}}{2(n + z^2)}$$

For 95% limits, $z = 1.96$, and the equations reduce to:

$$L = \frac{2np + 2.84 - 1.96\sqrt{\{1.84 - 1/n + 4p(nq + 1)\}}}{2(n + 3.84)}$$

$$U = \frac{2np + 4.84 + 1.96\sqrt{\{5.84 - 1/n + 4p(nq + 1)\}}}{2(n + 3.84)}$$

For the example used before, the confidence limits are 0.0829–0.2278, for an interval of 0.145. If p is very low and the lower limit is < 0 , it is set at zero. If p is very high and the upper limit is > 1 , it is set at 1.

The continuity correction is not needed for larger sample sizes, and a simpler formula can then be used:

$$\frac{2np + z^2 \pm z\sqrt{z^2 + 4npq}}{2(n + z^2)}.$$

With the same proportions as used previously, the 95% limits are from 0.0852 to 0.2213, quite close to the more exact formulation. This method seems to be more accurate than other proposed methods. It can be performed online at <http://vassarstats.net/> (see Frequency data).

Problem 16.3 Calculate 95% confidence limits as in this problem, using Wilson's method.

MULTINOMIAL DISTRIBUTION

This is an extension of the binomial distribution to more than two groups. If there are >2 mutually exclusive categories, X_1 is the number of trials producing type 1 outcomes, X_2 is the number of trials producing type 2 outcomes, ..., X_k is the number of trials producing type k outcomes, then the event $(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k)$ follows a multinomial distribution if there are n independent trials, each trial results in one of k mutually exclusive outcomes, the probability of a type i outcome is p_i , and this probability does not change from trial to trial. Then the distribution of the event $(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k)$ is given by

$$P(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1! x_2! \dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$$

where $\sum_{i=1}^k p_i = 1$, and x_1, x_2, \dots, x_k are nonnegative integers satisfying $\sum_{i=1}^k x_i = n$.

The mean of each X_i is given by $\mu_i = np_i$ and its variance is given by $\sigma_i^2 = np_i(1 - p_i)$.

Consider the four main blood group phenotypes. The probabilities of A , B , AB , and O are, respectively, 0.43, 0.12, 0.05, and 0.40. In a study of 17 randomly selected subjects, what is the probability of choosing 5 A s, 2 B s, 1 AB , and 9 O s at random?

By the formula,

$$p(A = 5; B = 2; AB = 1; O = 9) = \frac{17!}{5!2!1!9!} 0.43^5 0.12^2 0.05^1 0.40^9 = 0.01133.$$

In repeated random samples of size 17, the mean number of persons with type A is $17 \times 0.43 = 7.3$, with variance $17 \times 0.43 \times 0.57 = 4.1667$ and standard deviation is about 2.04.

Explanations of the formula with examples are given by Ash (1993), Hogg and Tanis (1977), Murphy (1979), and Ross (1984).

Problem 16.4 Using the serum cholesterol data before, in repeated random samples of 75 subjects, what is the probability of choosing, in order of ascending concentrations, 3, 26, 43, and 9, respectively?

APPENDIX

Calculation of sample size for unequal numbers:

Let sample sizes be n_1 and n_2 and let $k = n_1/n_2$. Then modified sample size $N' = N(1 + k)^2/4k$. That is, calculate N for total sample size, assuming equal sample sizes. Then compute N' . Then the two sample sizes are $N'/(1 + k)$ and $kN'/(1 + k)$.

REFERENCES

- Ash, C., 1993. *The Probability Tutoring Book. An Intuitive Course for Engineers and Scientists*. IEEE Press, New York, p. 470.
- Curran-Everett, D., Benos, D.J., 2004. Guidelines for reporting statistics in journals published by the American Physiological Society. *Am. J. Physiol. Endocrinol. Metab.* 287, E189–E191.
- Gardner, M.J., Altman, D.G., 1995. *Statistics with Confidence—Confidence Intervals and Statistical Guidelines*. British Medical Journal, London.
- Hahn, G.J., Meeker, W.Q., 1991. *Statistical Intervals. A Guide for Practitioners*. John Wiley and Sons, Inc, New York.
- Hamilton, L.C., 1990. *Modern Data Analysis. A First Course in Applied Statistics*. Brooks/Cole Publishing Co, Pacific Grove, CA.
- Hogg, R.V., Tanis, E.A., 1977. *Probability & Statistical Inference*. Macmillan Publishing Company, Inc, New York.
- Jousilahti, P., Vartiainen, E., Pekkanen, J., Tuomilehto, J., Sundvall, J., Puska, P., 1998. Serum cholesterol distribution and coronary heart disease risk: observations and predictions among middle-aged population in eastern Finland. *Circulation* 97, 1087–1094.
- Kennedy, J.B., Neville, A.M., 1986. *Basic Statistical Methods for Engineers and Scientists*. Harper and Row, New York.
- Lehr, R., 1992. Sixteen s-squared over d-squared: a relation for crude sample size estimates. *Stat. Med.* 11, 1099–1102.
- McGhee, J.W., 1985. *Introductory Statistics*. West Publishing Company, St. Paul, MN.
- Murphy, E.A., 1979. *Probability in Medicine*. Johns Hopkins University Press, Baltimore.
- Newcombe, R.G., 1998. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Stat. Med.* 17, 873–890.
- Ross, S., 1984. *A First Course in Probability*. Macmillan Publishing Company, New York.
- Van Belle, G., 2002. *Statistical Rules of Thumb*. Wiley Interscience, New York.

CHAPTER 17

Proportions

INTRODUCTION

Previous chapters dealing with categories sometimes used the proportions within each group. It may be easier to work with proportions rather than absolute numbers.

A proportion is the relationship of a part to the whole and can be given as a fraction of 1 or as a percentage. A ratio is the relationship of one part to another part. If out of 100 patients with a myocardial infarction 30 of them die, the proportion who die is $30/100=0.3$ or 30%, whereas the ratio of those who die to those who survive is $30/70=0.43$ or 43%.

PROPORTIONS AND BINOMIAL THEOREM

If there are 37 deaths in 192 people with a certain disease, the number who survive is $192-37=155$. The proportion of deaths (p) is determined by number with the attribute/total number $=37/192=0.1927$. The proportion that survives is $155/192=0.8073$. This is the value $1-p$, also symbolized by q (Chapter 16). In repeated samples from this population, the value of p would vary; the sample value p is a point estimate of the population value π . The distribution of p is approximately normal as long as np and $nq > 9$. The standard deviation of such a binomial is:

$$\sigma = \sqrt{p(1-p)} = \sqrt{pq}$$

The standard error of the sample proportion p that estimates the population proportion π is equivalent to the standard error of a mean value and is estimated by dividing the standard deviation by \sqrt{n} . Therefore $\sigma = \sqrt{\frac{pq}{n}}$. In the previous example,

$$\sigma_p = \sqrt{\frac{0.1927 \times 0.8073}{192}} = 0.02486.$$

The 95% confidence limits of π are $p \pm 1.96 \times 0.02486 = 0.1440$ to 0.2414 .

Because the formula for the standard error of a proportion comes from the approximation of the discrete binomial distribution to the normal distribution, we need a continuity correction. Test the null hypothesis by

$$z = \frac{|p - \pi| - \frac{1}{2n}}{se_p}.$$

The effect of the correction is to make the difference slightly smaller, and for large sample sizes the correction is unimportant.

CONFIDENCE LIMITS

The observed proportion is a point estimate. To determine its confidence limits, use the Wald method, with the 95% limits determined from $p \pm z_{\alpha/2} \sqrt{\frac{pq}{N}}$. This method yields inaccurate results if N is small or p is close to 0 or 1. A simple way to obtain more accurate limits is to use the Agresti-Coull adjustment by adding 2 successes and 2 failures to the observed counts. Thus $p_a = \frac{X+2}{N+4}$, and the adjusted value of p (p_a) is substituted for the value of p in the Wald formula before.

If $p = \frac{37}{192} = 0.1927$, then the Wald limits are $0.1927 \pm 1.96 \sqrt{\frac{0.1927 \times 0.8073}{192}} = 0.1927 \pm 0.0579 = 0.1348$ to 0.2506 . With the adjustment, the adjusted value of p is $p = \frac{37+2}{192+4} = 0.1990$, and the adjusted Wald limits are

$$0.1990 \pm 1.96 \sqrt{\frac{0.1990 \times 0.8010}{192}} = 0.1990 \pm 0.0565 = 0.1425 \text{ to } 0.2555.$$

Confidence limits may be obtained online from http://www.causascientia.org/math_stat/ProportionCI.html (a Bayesian calculator), <http://www.graphpad.com/quickcalcs/ConfInterval2.cfm>, <http://www.sample-size.net/confidence-interval-proportion/>, and <http://vassarstats.net/prop1.html>.

Problem 17.1 Nineteen out of 113 (16.8%) men have a serum cholesterol < 5.0 mmol/L. What are the 95% and 99% confidence limits for this proportion?

One use for confidence limits is to determine the upper 95% limit for a proportion if zero events occur. If a surgeon operates on 10 patients without a death, what is the upper 95% limit of deaths? It can be determined easily by the rule of 3 (Hanley and Lippman-Hand, 1983; van Belle, 2002). The upper 95% limit for the proportion is determined by $3/N$. The upper 95% mortality proportion for the surgical procedure is $3/10$ or 0.3 (or 30%). If there were no deaths in 50 operations the upper 95% limit would be $3/50 = 0.06$ or 6%. (This assumes that the operations were similar in all respects.)

The corollary to this is to determine how many surgical operations must be observed in order to find at least one death if we know the average mortality. If the mortality of a procedure is 1%, there is a 95% chance of observing 1 death in $3/0.01 = 300$ operations. This calculation does not allow for differences in surgical skill or severity of illness.

SAMPLE AND POPULATION PROPORTIONS

To compare any observed value of p with the population value π , use the normal distribution curve.

$$z = \frac{\pi - p}{\sigma_p}$$

To determine if the sample value of $p = 0.1927$ could have come from a population in which $\pi = 0.3$, calculate $z = \frac{0.3 - 0.1927}{0.02486} = 4.3162$.

$P = 0.000016$ (two-tailed), and we would reject the null hypothesis.

SAMPLE SIZE

To compare two proportions, we need to know how many subjects we will need to minimize type I and type II errors.

The basic formulas are discussed by Fleiss (1981) and Bland (2015). They are relatively complex and approximate, can be written in several ways, and are best replaced by free online calculation at <http://statpages.org/proppowr.html>, <http://www.stat.ubc.ca/~rollin/stats/ssize/b2.html>, <http://www.cct.cuhk.edu.hk/stat/proportion/Casagrande.htm>, <http://www.sample-size.net/confidence-interval-proportion/>, and https://www.statstodo.com/SSiz2Props_Pgm.php.

A simplified approximation was devised by Lehr (1992) who used the equation

$$n = \frac{16pq}{(p_1 - p_2)^2},$$

where p is the average of p_1 and p_2 , and q is 1-average p . This gives the number in each group for a power of 0.8 (Type II error $\beta = 0.2$) and a Type I error $\alpha = 0.05$ (two-tailed). If a power of 0.9 is wanted, the constant is changed from 16 to 21. According to Lehr, this estimate is slightly low for Fisher's exact test and slightly high for a 2×2 chi-square test. As with all sample size calculations, these are estimates, not precise numbers.

As an example, consider how many subjects (equal sized groups) are needed to show a difference between a remission rate of 0.7 for one treatment and 0.6 for another. We wish to set $\alpha = 0.05$ and $\beta = 0.2$, that is, power = 0.8. By Lehr's formula,

$$n = \frac{16 \times 0.65 \times 0.35}{(0.7 - 0.6)^2} = 364 \text{ as the number required in each group. The more formal}$$

calculation online gives 356 in each group, or 376 if the continuity correction is used. For a power of 0.9, Lehr's formula gives 478 in each group, and the online calculation gives 476, or 496 with the continuity correction. The major advantage of using an online calculator is that it allows for unequal group sizes.

In estimating the numbers needed, we need some idea of the effect size that we want. Take 0.50 as the null hypothesis, that is, there is no difference between the two groups. Then either select an effect size based on previous work or else try to decide what minimum effect size to detect. For example, suppose the null hypothesis is that the probability of survival of a disease under standard treatment is 0.50, we would probably not be interested in a treatment that changed the proportion to only 0.49. Cohen (1988) classified effect size of the difference from 0.50 as small (<0.05 , that is, 0.45–0.55), medium (0.15, that is, 0.35–0.65), and large (0.25, i.e., 0.25–0.75). He pointed out that large differences were rare. For example, in presidential elections in the United States there has never been a division as extreme as 65:35, and even a division of 55:45 would be regarded as a landslide victory. Cohen gives tables, and online calculators are available. As always, it takes a huge number of measurements or counts to detect a small difference.

Problem 17.2 In two different populations of men, serum cholesterol concentrations below 5.0 mmol/L occur in 6% and 9%, respectively. What sample size is needed to show that this difference allows us to reject the null hypothesis with $\alpha=0.05$ and $\beta=0.2$? (assume equal sized groups)?

COMPARING PROPORTIONS

To compare two proportions, for example, the survival of patients with a given disease who have two different treatments, let the proportion surviving in group 1 be p_1 and that in group 2 be p_2 . If these two proportions are similar, we conclude that treatment did not affect survival. If they are quite different, then we can ask if the null hypothesis is true. To do this, calculate the standard error of the difference as

$$s_{p_1-p_2} = \sqrt{\left(\frac{p_1 q_1}{n_1}\right) \left(\frac{p_2 q_2}{n_2}\right)}.$$

Then relate the difference p_1-p_2 to the standard error of the difference;

$$z = \frac{p_1 - p_2}{\sqrt{\left(\frac{p_1 q_1}{n_1}\right) \left(\frac{p_2 q_2}{n_2}\right)}},$$

For example, if treatment 1 gives a survival proportion of $37/192 = 0.1927$ and treatment 2 gives a survival of $17/168 = 0.1012$, could that difference have occurred by chance? Calculate z as

$$z = \frac{0.1927 - 0.1012}{\sqrt{\left(\frac{0.1927 \times 0.8073}{192}\right) + \left(\frac{0.1012 \times 0.8988}{168}\right)}} = \frac{0.0915}{\sqrt{0.0008102 + 0.0005414}} = 2.4888$$

Therefore $P = 0.0128$ (two-sided) and we can reasonably reject the null hypothesis and believe that treatment 2 might be better.

The 95% confidence limits for the difference are $0.0915 \pm 1.96 \times 0.03676 = 0.05474$ to 0.1291 . Because this does not include zero, it confirms the results of the z test.

The continuity correction can be done by using

$$Z = \frac{|P_1 - P_2| - \frac{1}{2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}{se_{(p1-p2)}}$$

Sample sizes for adequate power are provided by Cohen (1988).

Proportions can be compared at <http://www.measuringusability.com/ab-calc.php>, https://www.medcalc.org/calc/comparison_of_proportions.php, http://vassarstats.net/prop2_ind.html, or <http://in-silico.net/statistics/ztest/two-proportion> and sample size can be calculated at <http://www.sample-size.net>, and http://www.statstodo.com/SSiz2Props_Pgm.php.

Note: Comparing two proportions is the same as doing a 2×2 chi-square test.

Pooling samples

If there are several small samples from a population, with X_1 successes in N_1 trials, X_2 successes in N_2 trials, up to X_k successes in N_k trials, then an average proportion of successes can be calculated as

$$\bar{P} = \frac{X_1 + X_2 + \dots + X_k}{N_1 + N_2 + \dots + N_k}.$$

REFERENCES

- Bland, M., 2015. *An Introduction to Medical Statistics*. Oxford University Press, Oxford.
- Cohen, J., 1988. *Statistical Power Analysis for Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Fleiss, J.L., 1981. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, New York.
- Hanley, J.A., Lippman-Hand, A., 1983. If nothing goes wrong, is everything alright? *JAMA* 249, 1743–1745.
- Lehr, R., 1992. Sixteen s-squared over d-squared: a relation for crude sample size estimates. *Stat. Med.* 11, 1099–1102.
- van Belle, G., 2002. *Statistical Rules of Thumb*. John Wiley & Sons, Inc., New York.

CHAPTER 18

The Poisson Distribution

INTRODUCTION

The results of a study may be counts of the numbers of discrete events that occur per unit of time, space, or mass.

A. Examples with unit time are:

- a. number of disintegrations per minute of a radioactive isotope,
- b. number of births per day in a busy Metropolitan Hospital,
- c. (a famous historical example) annual number of people kicked to death by mules in different Austrian cavalry corps.

B. Examples with unit space are:

- a. number of flaws per cm length of silk suture material,
- b. number of red cells per hemocytometer field,
- c. number of mutant bacteria per 100 μL of a bacterial suspension,
- d. number of diseased white footed mice (carriers of the agent that causes Lyme disease) per acre of woodland.

C. Examples with unit mass are the number of seeds of poisonous plants per 100 g of grass seeds or the number of pathological *E. coli* in a gram of herbal food supplement.

We expect variation from one unit to the next and also expect some average number of births, bacteria, or poisonous seeds. What form of distribution does this type of variation take?

The mathematical distribution of rare events that occur randomly in time or space is called after a French mathematician, Siméon-Denis Poisson (1781–1840). The term *random* implies that any one section of space or time interval has the same probability as any other of experiencing or not experiencing an event. For the counts to fit a Poisson distribution, they must obey the following assumptions:

1. The probability that a single event occurs in a very tiny time interval or region is proportional to the length of that interval or size of that region. Assume that there is a constant λ , such that the probability of observing one event in a tiny time interval Δt is $\lambda\Delta t$, and the probability of observing no events is $1-\lambda\Delta t$; the probability of observing more than one event is essentially zero, being the probability of a rare event raised to some power.
2. The probability of observing an event in a tiny time interval or region does not change over the whole period of observation. The process shows stationarity.

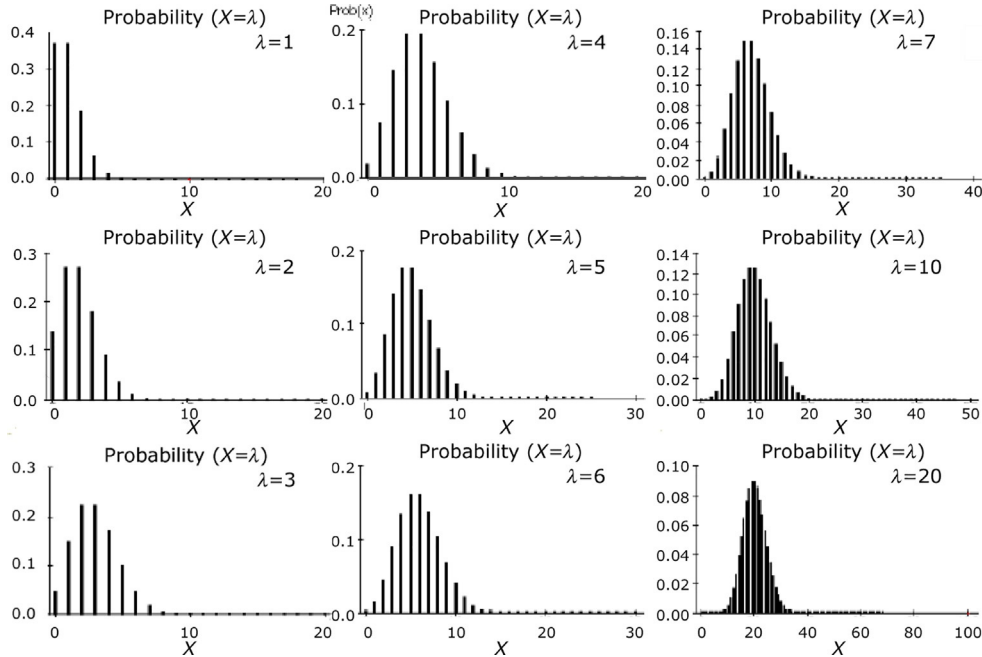


Fig. 18.1 Poisson distribution for $\lambda = 1, 2, 3, 4, 5, 6, 7, 10$, and 20 . After $\lambda = 5$, the distribution is fairly normal.

3. An event that occurs in a time interval or region is independent of events occurring in other time intervals or regions. An event does not influence or is not influenced by any other events.

If the events meet these assumptions, then the probability of k events occurring in a given time period for a Poisson variable with parameter λ is

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}, \text{ where } k = 0, 1, 2, 3, \dots, \text{ and } e \text{ is the base of the natural logarithms,}$$

approximately 2.71828. These are parameters that reflect a population value, hence the Greek symbols. Sample values should be given other symbols, but some authors still use the Greek symbols and rely on the reader to differentiate between population and sample values; others use m , μ , or \bar{X} instead of λ .

The Poisson distribution is asymmetrical, especially for low values of the Poisson parameter λ (Fig. 18.1). As the value for the Poisson variate (or mean count) increases, the distribution becomes more symmetrical and approaches the normal distribution.

RELATIONSHIP TO THE BINOMIAL DISTRIBUTION

The Poisson distribution approximates the binomial distribution closely when n is very large and p is very small. It is the limiting form of the binomial distribution when $n \rightarrow \infty$,

$p \rightarrow 0$, and $np = \mu$ is constant and < 5 . In the binomial distribution, the mean is given by np , and the standard deviation by \sqrt{npq} . If n is large and p is very small, as in the Poisson approximation to the binomial, then the mean is still np , but the standard deviation is now $\sim \sqrt{np}$, because q is almost 1. Consequently, the limiting value of the standard deviation as the binomial distribution approaches the Poisson distribution is the square root of the mean. This mathematical distribution can be applied to a binomial distribution in which the probability of an event, p , is very small and n is large, and also to a rare, random event in which we know the number of events that occur but do not know the number that do not occur. We know how many people in the cavalry corps were kicked to death by mules, but there is no way of knowing how many were not kicked to death by mules. We know how many people were struck by lightning, but not how many were not struck. For the binomial example, we could calculate the probabilities of 0, 1, 2, and so on, events from the binomial distribution, but when p is very small and n is very big, the Poisson calculation is simpler (Example 18.1).

GOODNESS OF FIT TO A POISSON DISTRIBUTION

Example 18.1

Here is how to calculate the Poisson distribution. It shows the process, although in practice we use a computer program. Free online programs are available: <http://stattrek.com/online-calculator/poisson.aspx>, <https://easycalculation.com/statistics/poisson-distribution.php>, <http://www.danielsoper.com/statcalc/calculator.aspx?id=81>, http://statstodo.com/PoissonTest_Pgm.php, and <http://vassarstats.net/poissonfit.html>, this last being the most convenient.

The data recorded by von Bortkiewicz on the chances of a Prussian cavalryman being kicked to death by a mule were taken from 14 cavalry corps over 20 years, for a total of 200 readings (Table 18.1). This is an important example historically, because it was the first time that the Poisson distribution function had been used in practice, and it introduced the Poisson distribution to a wide audience.

Table 18.1 Deaths from mule kicks in the cavalry corps

Number of deaths/ corps/year (x)	Observed frequency (f)	fx	Expected frequency (np)	Chi-square
0	109	0	108.67	0.0010
1	65	65	66.29	0.0251
2	22	44	20.22	0.1567
3	3	9	4.11	0.2998
4	1	4	0.63	0.2173
5	0	0	0.08	0.0078
Total	200	122	200.01	0.7077

There is variation from corps to corps. In 109 corps there were no such deaths, but in one corps in 1 year there were 4 deaths. Was this a random event, could there have been inadequate training in that corps of how to handle mules, or was there one particularly vicious mule in that corps?

First, calculate the average number of deaths per corps per year. (columns 1–3) There were 122 deaths to be averaged over 200 corps-years, for a mean of 0.61 per corps-year. This is the value of \bar{X} , an estimate of λ in the Poisson equation.

The equation tells us that no deaths [$P(X=0)$] will be given by

$$P(X=0) = \frac{e^{-0.61} 0.61^0}{0!} = 0.543351 \text{ (because } 0.61^0 \text{ and } 0! \text{ both equal } 1).$$

By applying the formula, the remaining probabilities can be calculated

$$P(X=1) = \frac{e^{-0.61} 0.61^1}{1!} = 0.331444.$$

$$P(X=2) = 0.101090.$$

$$P(X=3) = \frac{e^{-0.61} 0.61^3}{3!} = 0.020555.$$

$$P(X=4) = \frac{e^{-0.61} 0.61^4}{4!} = 0.003135.$$

$$P(X=5) = \frac{e^{-0.61} 0.61^5}{5!} = 0.000382.$$

$$P(X=6) = \frac{e^{-0.61} 0.61^6}{6!} = 0.000039.$$

The figures in bold type show differences from the formula in the immediately preceding line.

The sum of these 6 probabilities is 0.999996. The remainder is the probability of getting >6 deaths and is very small.

If the data are truly random, then they will fit a Poisson process with the mean value of 0.61. Taking each calculated probability and multiplying it by 200, the total number of observations, gives the fourth column in [Table 18.1](#), a column labeled Expected Frequency. Note how closely the Observed and Expected frequencies agree. To determine objectively how good the fit is, do a chi-square test ([Chapter 14](#)). The total chi-square is very small and indicates a good fit between observed and expected frequencies. The probability of falsely rejecting the null hypothesis (Type I error, or α) can be determined from the chi-square tables, with $k-2$ degrees of freedom, where k is the highest number of events observed. (Two degrees of freedom are lost. One is the usual loss because of the mean, and the other because the mean is used to calculate all the other terms.) In this example, the risk of falsely rejecting the null hypothesis is low.

If the data do not indicate randomness but have excess numbers in some bins and too few in others, they may fit a distribution called a contagious distribution (other terms are clumped, aggregated, overdispersed, or clustered) that has application to epidemics. Detecting and modeling such data will be presented in [Chapter 19](#).

Problem 18.1 [Rutherford and Geiger \(1910\)](#) described the radioactive decay counts of polonium and discovered alpha particles. They observed:

Particles/unit	Number of units
0	57
1	203
2	383
3	525
4	532
5	408
6	273
7	139
8	45
9	27
10	10
11	4
12	0
13	1
14	1
15 or more	0
Total	2,608

Does this fit a Poisson distribution?

Example 18.2

The numbers of bacteria counted in 10 consecutive agar plates were 545, 531, 530, 525, 533, 529, 529, 535, 543, and 544. Because these are counts they should fit a Poisson distribution, and failure to fit this distribution would reflect on the method of preparing and diluting the samples.

The mean of these counts is 534.4. This is the expected number to be obtained, and calculating the chi-square value for each observed number gives (Table 18.2).

Table 18.2 Bacterial counts

Observed(O)	Expected(E)	O-E	χ^2
545	534.4	10.6	0.21
531	534.4	-3.40	0.02
530	534.4	-4.4	0.04
525	534.4	-9.4	0.17
533	534.4	-1.40	0.00
529	534.4	-5.4	0.05
529	534.4	-5.4	0.05
535	534.4	0.60	0.00
543	534.4	8.6	0.14
544	534.4	9.6	0.17
Total			0.85

A total chi-square of 0.85 with 9 degrees of freedom shows that $P > 0.999$, so that there is no reason to reject the null hypothesis that these counts could have come from a Poisson distribution (Example 18.2).

THE RATIO OF THE VARIANCE TO THE MEAN OF A POISSON DISTRIBUTION

If $X_1, X_2, X_3, \dots, X_n$ come from a Poisson distribution with mean λ , then $\frac{\sum_{i=1}^n (X_i - \lambda)^2}{\lambda}$ is approximately distributed as χ^2_n . Remember that the variance of a Poisson variable is the same as the mean (see Appendix).

$$\chi^2 = \frac{\sum_{i=1}^n (X_i - \lambda)^2}{\lambda} = \frac{\sum_{i=1}^n (X_i - \lambda)^2 (n-1)}{\lambda(n-1)} = \frac{s^2(n-1)}{\lambda}$$

If the agreement with the Poisson distribution is perfect, the ratio of variance to mean will be exactly 1, and chi-square will equal $n-1$. If the agreement is not perfect, then the value of chi-square can be referred to a table of chi-square values to assess the probability that the distribution is compatible with a Poisson distribution. If $n > 31$, $\sqrt{2\chi^2}$ is distributed normally about $\sqrt{2\nu - 1}$ with unit variance, or alternatively, the variable $d = \sqrt{2\chi^2} - \sqrt{2\nu - 1}$ is a normal variable with zero mean and unit variance. (ν is the number of degrees of freedom.) Then if the absolute value of $d < 1.96$, the hypothesis that the distribution is consistent with a Poisson distribution cannot be rejected (Examples 18.3 and 18.5).

Example 18.3

The standard deviation of the deaths in the mule example was 0.611. The ratio of the variance to the mean is $0.611/0.61 = 1.0016$, very close to the exact ratio of 1. Alternatively, refer to the chi-square table with $n = 199$. Now $1.0016 \times 199 = 199.318$, and this is close to the 50% value of chi-square.

Example 18.4

In 80 samples of shrimps taken from a river, the mean was 5.3125 and the variance was 13.534 (Elliott, 1983). Then

$$\chi^2 = \frac{s^2(n-1)}{\lambda} = \frac{13.534 \times 79}{5.3125} = 201.2585$$

The normal variable d is

$$d = \sqrt{2\chi^2} - \sqrt{2n-1} = \sqrt{402.5170} - \sqrt{157} = +7.532.$$

Because d is $\gg 1.96$, reject the null hypothesis that the distribution is compatible with a Poisson distribution at the 0.01 level. The high and positive value for d , with the variance much greater than the mean, suggests that this is a contagious distribution. This example is better analyzed as a negative binomial distribution (Chapter 19).

Problem 18.2 Use the polonium data to test the fit to a Poisson distribution by the method of Example 18.4.

Example 18.5

Examine the problem of the successive bacterial counts shown in Table 18.2 by using the ratio of the variance to the mean:

$$\chi^2 = \frac{s^2(n-1)}{\lambda} = \frac{458.4 \times 9}{534.4} = 7.72.$$

For chi-square with 9 degrees of freedom, $P=0.56$, so that there is no reason to reject the null hypothesis. (See online calculators <http://www.fourmilab.ch/rpkp/experiments/analysis/chiCalc.html>. or <http://www.danielsoper.com/statcalc3/calc.aspx?id=12>

SETTING CONFIDENCE LIMITS

Normal Approximation

If \bar{X} is >100 , an approximate $100(1-\alpha)$ confidence interval for λ is $\bar{X} \pm z_{1-\alpha/2} \times \sqrt{\bar{X}}$ (Example 18.6).

Example 18.6

Count a dilute suspension of red blood cells under a microscope. The suspension is pipetted onto a hemocytometer slide and then covered with a coverslip so that the thickness of the suspension is constant. In the first square 400 red blood cells are counted. What are the 95% confidence limits of the number of red blood cells per square?

Because these are counts, and because 400 is a very tiny portion of the millions of red blood cells present, the counts should fit a Poisson distribution. Therefore the standard deviation of the counts is $\sqrt{400} = 20$ and 95% confidence limits for the counts per square are $400 \pm (20 \times z_{0.05}) = 400 \pm (20 \times 1.96) = 400 \pm 39.2 = 360.8$ to 439.2 . The population counts per square range from 361 to 439 with a probability of 95%. However, most red cell counts are reported per mm^3 , so we need to correct for the actual volume examined. The volume that is examined is the area of counting chamber examined multiplied by the depth of the fluid under the coverslip. Assume that this is 10^{-4} mm^3 . Then had the cells been counted over 1 mm^3 , the total number would have been 400×10^4 , and the corresponding 95% confidence limits would therefore be 360.8×10^4 to 439.2×10^4 , or 3.61 to 4.39 million per mm^3 . It is essential to use the actual observed counts to set the confidence limits. If we had taken the calculated number of cells per mm^3 , namely, 4 million, then the standard deviation would have been 2000, and the confidence limits would have been calculated as $4 \times 10^6 \pm 1.96 \times 2000 = 4 \times 10^6 \pm 3920 = 3.996$ to 4.004 million per mm^3 . These limits are too narrow. The precision depends on the actual number counted. Had the number counted been 4 million, then the narrower confidence limits would have been appropriate. These calculations can be done online at <http://www.danielsoper.com/statcalc3/calc.aspx?id=86>, and <http://statpages.org/confint.html>.

Problem 18.3 If the mean number of polonium counts was 425, what are the 95% confidence limits for this mean?

Exact Method

If $\bar{X} < 100$, the approximation ceases to be accurate. Exact confidence limits can be calculated by using the mathematical link between the Poisson and chi-square distributions (Armitage et al., 2002) (Example 18.7).

Example 18.7

Ten deaths are observed from AIDS in 1 week in an urban hospital. What are the exact 95% confidence limits for the number of deaths per 4 weeks?

The exact 95% confidence limits for 10 counts are

$$\lambda_L = 0.5\chi_{20,0.975}^2 \quad \text{and} \quad \lambda_U = 0.5\chi_{22,0.025}^2$$

These limits are thus 4.80 and 18.39 deaths per week, or 19.18 to 73.56 deaths per 4 weeks. These can be determined online at <http://statpages.org/confint.html> or <http://www.danielsoper.com/statcalc3/calc.aspx?id=86>.

With the normal approximation, the limits would have been $10 \pm 1.96\sqrt{10} = 10 \pm 6.20 = 3.80$ to 16.20 per week. These limits, even if incorrect, are not very far from the true limits (Example 18.8).

Example 18.8

A physician observes 2 deaths from a rare form of cancer in 1 year. If deaths from this form of cancer are random, what are the 95% confidence limits of annual deaths? The normal approximation gives $2 \pm 1.96\sqrt{2} = 2 \pm 2.77$, or -0.23 to 4.77 per year. The exact method and the calculators at <http://statpages.org/confint.html> or <http://www.danielsoper.com/statcalc3/calc.aspx?id=86> give limits of 0.24 to 7.22 per year.

If one-sided upper and lower $(100-\alpha)$ limits are wanted, replace $\alpha/2$ by α in the respective equations. The upper 95% limit becomes $\lambda_U = 0.5\chi_{(2\lambda+2),\alpha}^2$ and the lower 95% limit becomes $\lambda_L = 0.5\chi_{2\lambda,1-\alpha}^2$. In the cancer example, the one-sided upper 95% confidence limit would be $\lambda_U = 0.5\chi_{(6),0.05}^2 = 6.30$.

THE SQUARE ROOT TRANSFORMATION

Problem 18.4 What are the 95% confidence limits for a mean count of 4?

We want the variances of different groups to be homogeneous when testing for differences between the means, because if the group variances differ markedly, then comparisons between groups become less efficient (Chapter 25). This could be a problem for Poisson distributions because as their means increase, so do their variances. Therefore it may be useful to use a transformation that keeps the variances stable. If a random variable X has a Poisson distribution with mean λ , then the reexpression \sqrt{X} (or $\sqrt{X+0.5}$ if some counts have zero values) has a more normal distribution with a mean that is a function of λ and a

variance of about $\frac{1}{4}$, as long as λ is >30 (Zar, 2010). The mean and variance are no longer interdependent after the transformation. Freeman and Tukey suggested that for small values of $\lambda < 3$ it is better to use $\sqrt{X_i} + \sqrt{(X_i + 1)}$. This transformation is also used if there are many zero values in a data set that has <15 counts. In addition to stabilizing the variance, the distribution becomes closer to a normal distribution.

CUMULATIVE POISSON PROBABILITIES

We may want the combined probabilities of counts over certain ranges, rather than the absolute probability of each count. For this, use the cumulative Poisson distribution.

Example 18.9

A nursing supervisor has to determine the duty roster for a busy delivery service. If too few nurses are assigned, then patients will not receive good care. If too many are assigned, then several nurses will not have anything to do, and the costs of medical care will be increased. If there were some way to know how many deliveries were likely to occur, then the supervisor could arrange for a certain number of nurses to be on duty and for a certain number to be at home but available in an emergency.

Assume that over a long time there are an average of 16 deliveries per 8-h shift. The delivery rate per shift is likely to be a Poisson variate, so it is possible to calculate the probability of any given number of deliveries per shift. The following calculations are provided to illustrate the process of cumulating probabilities.

It is easy to calculate the probabilities of 0, 1, 2, and so on, events.

$$P(X=0) = e^{-16} = 0.000000113.$$

$$P(X=1) = 0.000000113 \times 16 = 0.000001801.$$

$$P(X=2) = 0.000001801 \times 16/2 = 0.0000144, \text{ etc.}$$

These results are displayed in Fig. 18.2 left panel.

The most likely numbers of deliveries per shift are 15 and 16. However, 16 would not be a safe upper number to pick, because higher numbers of deliveries per shift are fairly frequent.

Now calculate a cumulative probability from these data.

$$P(X=0) = 0.000000113.$$

$$P(X \leq 1) = P(X=0) + P(X=1) = 0.000000113 + 0.000001801 = 0.000001914.$$

$$P(X \leq 2) = P(X \leq 1) + P(X=2) = 0.000001914 + 0.0000144 = 0.000016314, \text{ etc.}$$

The results appear in Fig. 18.2 right panel.

The supervisor can be sure that about 95% of the time, there will not be >22 – 23 deliveries per shift. Therefore with this knowledge and the experience of knowing what happens if there are too few or too many nurses on duty per shift, the supervisor can decide how many nurses to allocate per shift.

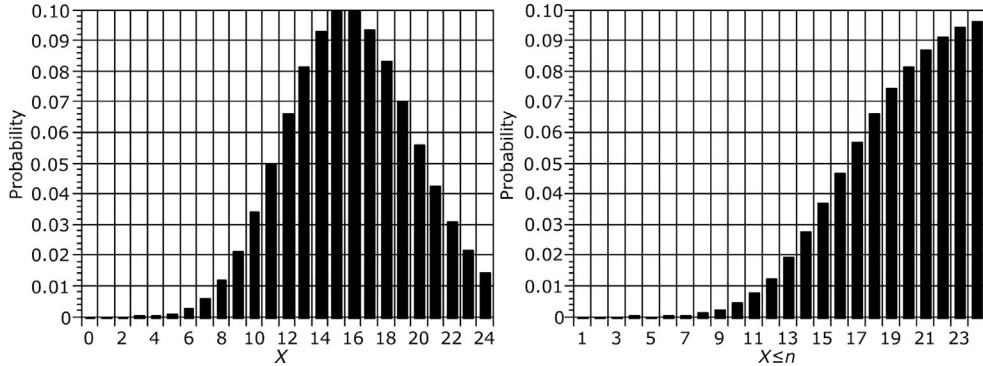


Fig. 18.2 Probabilities of $P(X = n)$ in left panel and $P(X \leq n)$ in right panel for a Poisson with mean number of events 16.

These calculations are very tedious, and it is best to use tables of the cumulative Poisson distribution or an online program such as <http://stattrek.com/Tables/poisson.aspx>, <http://ncalculators.com/statistics/cumulative-poisson-distribution-calculator.htm>, <https://easycalculation.com/statistics/poisson-distribution.php>, or <http://www.daniel-soper.com/statcalc3/calc.aspx?id=86>. In these, the mean number of events (e.g., 16) and some higher number (e.g., 21) are entered, the cumulative probability up to that number is calculated. The second number is reentered until the desired probability is achieved.

The same process can be used to set the probability for a range of values. For example, in [Example 18.9](#), what proportion of the counts lie between 9 and 15? From the calculator the cumulative probability up to 9 deliveries per shift is 0.0433 and up to 15 per shift is 0.4667. The probability of being between these limits is $0.4667 - 0.0433 = 0.4234$.

Problem 18.5 Assume that an emergency room gets an average of 3 patients per hour. What is the 95% probability of seeing 6 patients in the next hour? What is the probability of seeing 32 patients in the next 8 h?

DIFFERENCES BETWEEN MEANS OF POISSON DISTRIBUTIONS

Comparing the mean counts in two Poisson distributions depends in part on the size of the means and the size of the samples from which they come, and on whether the counts are based on the same or different units.

Comparison of Counts Based on Same Units

Assume that the counts in the two groups are both >10 , and that each is based on the same unit of time or space. Then the normal approximation to the Poisson distribution can be used. Let the counts in the two groups be λ_1 and λ_2 . Then.

$$\begin{aligned} z &= \frac{\text{Difference between the counts}}{\text{Square root of the variance of the difference between the counts}} \\ &= \frac{\lambda_1 - \lambda_2}{\sqrt{\lambda_1 + \lambda_2}}. \end{aligned}$$

If the sum of the mean counts in the two groups is >5 but under 20, then use the formula with the continuity correction (Example 18.10):

$$\frac{\lambda_1 - \lambda_2 - 1}{\sqrt{\lambda_1 + \lambda_2}}$$

Example 18.10

If there are 33 counts in the first group and 26 counts in the second group,

$z = \frac{33 - 26}{\sqrt{33 + 26}} = 0.91$, and the null hypothesis of no difference would be accepted with $P = 0.36$ (two sided).

These differences can be tested online at <http://www.quantitativeskills.com/sisa/statistics/t-thlp.htm>.

Comparison of Counts Not Based on Same Units

Often the two counts being compared are not based on the same unit of time or space. For example, we may want to compare the number of abnormal cells in an organ after two methods of treatment, but the number of cells (n, m) in the microscopic fields differs in the two groups. Then we must obtain a proportion that bases the count in each group on a common unit. Let the resulting two rates be μ_1 and μ_2 where $\mu_1 = \lambda_1/n$ and $\mu_2 = \lambda_2/m$. First, calculate a pooled weighted average to estimate the common rate, assuming that the null hypothesis is true. This is

$$\lambda_p = \frac{\lambda_1 n + \lambda_2 m}{n + m} = \frac{\mu_1 \mu_2}{n + m},$$

where λ_p is the pooled rate. Then the standard error of the difference between the two rates is

$$\sqrt{\frac{\lambda_p}{n} + \frac{\lambda_p}{m}}$$

The test of the significance of the difference between the two count rates is thus (Example 18.11)

$$z = \frac{\lambda_1 - \lambda_2}{\sqrt{\frac{\lambda_p}{n} + \frac{\lambda_p}{m}}}$$

Example 18.11

There are 11 abnormal cells out of 1800 cells in treatment 1 and 27 abnormal cells out of 1450 cells in treatment 2. Then the two rates per 1000 cells are $11/1.8 = 6.11$ and $27/1.45 = 18.62$, for a difference between rates of 12.51. The standard error of the difference between these two rates is based on the pooled average rate of $(11 + 27)/(1.8 + 1.45) = 11.69$. The standard error of the difference is thus

$$\sqrt{\frac{11.69}{1.8} + \frac{11.69}{1.45}} = 3.82$$

$$z = \frac{12.51}{3.82} = 3.27, \text{ and the null hypothesis can be rejected with } P < 0.0018 \text{ (two sided).}$$

Another useful method for small mean counts is to calculate the pooled value of λ , as before, and then to calculate z_1 for the smaller value of λ , and z_2 for the larger value of λ , from

$$z_1 = 2(\sqrt{\mu_1 + 1} - \sqrt{\lambda_p \times n}) \quad \text{and} \quad z_2 = 2(\sqrt{\mu_2} - \sqrt{\lambda_p \times m}).$$

Calculate the sum of z_{12} and z_{22} . This is referred to the chi-square table with 1 degree of freedom.

This method can be extended to k groups. The first formula, that is, that for z_1 , is used for all samples where the value of λ is less than λ_p , and the formula for z_2 is used when λ is greater than λ_p . Then testing is done with the formula $\chi^2 = \sum_{i=1}^k z_i^2$ for $k-1$ degrees of freedom (Example 18.12).

Example 18.12

For the data in Example 18.16, we have

$$z_1 = 2(\sqrt{11 + 1} - \sqrt{11.69 \times 1.8}) \quad \text{and} \quad z_2 = 2(\sqrt{27} - \sqrt{11.69 \times 1.45})$$

Therefore $z_1 = -2.2461$ and $z_2 = 2.1581$, and the sum of their squares is 9.7024. From the table for one degree of freedom, $P = 0.0018$, the same as shown before.

If μ_1 and μ_2 are large, but n and m are small, then use the square root transformation. With this transformation, the variance is approximately 0.25. Therefore the variance of the

difference between $\sqrt{\mu_1}$ and $\sqrt{\mu_2} = (\sqrt{\mu_1} + \sqrt{\mu_2}) \approx 0.5$. Under the null hypothesis, the difference d between the means of the square roots has mean zero and variance $\frac{0.25}{n} + \frac{0.25}{m}$. Therefore

$$z = \frac{d}{0.5\sqrt{\frac{1}{n} + \frac{1}{m}}}.$$

Example 18.13

In 11 soil samples from one area there are an average of 34 pathogenic bacteria, and from another area 15 soil samples give an average of 69 pathogenic bacteria. Are these substantially different?

$$d = \sqrt{69} - \sqrt{34} = 2.476.$$

$$z = \frac{2.476}{0.5\sqrt{1/11 + 1/15}} = 12.47, \text{ with } P < 0.001.$$

If we do this calculation by the method set out before in [Example 18.11](#), we get the difference between the mean number of bacteria per sample $= 69/15 - 34/11 = 1.51$. The weighted average of the number of bacteria per sample is $(34 + 69)/(11 + 15) = 3.81$. The standard error of the difference is

$$\sqrt{\frac{3.81}{11} + \frac{3.81}{15}} = 0.7748.$$

Then $z = 3.81/0.7748 = 4.92$, and $P < 0.001$ ([Example 18.13](#)).

Comparing the Ratio of Two Poisson Variates

To compare the ratio of two Poisson variates, use a method recommended by [Armitage et al. \(2002\)](#) ([Example 18.14](#)).

Example 18.14

Let $X_1 = 13$, and $X_2 = 4$. Then $\hat{F} = \frac{13}{4+1} = 2.60$. Entering the F table with degrees of freedom $2(4+1) = 10$, and $2 \times 13 = 26$, we find that $F = 2.59$ with $\alpha(2) = 0.05$. We can reject the two-sided null hypothesis that the two mean values are equal at the 0.05 level of probability.

Using this for the values in example 5.16 of 33 and 26, we have

$$\hat{F} = \frac{33}{26+1} = 1.22.$$

For degrees of freedom 54 and 66, we have $\alpha(2) \approx 0.50$, and the null hypothesis cannot be rejected.

To compare the ratio of two rates derived from Poisson variables, use a similar approach, but multiply the final values for the upper and lower limits by the ratio of the two measurement units (Example 18.15).

Example 18.15

The number of bacteria growing on two culture plates is 13 and 31. Then the 95% confidence limits for the ratio 13/31 are calculated by taking the proportion of bacteria in the first plate as $13/44 = 0.2955$, and from <http://statpages.org/confint.html> the confidence limits for π are 0.1676 and 0.4520. For each of these values we then calculate the 95% confidence limits of $\pi/(1-\pi)$ as $0.1676/(1-0.1676) = 0.2013$, and $0.4520/(1-0.4520) = 0.8248$.

Example 18.16

Assume that the two culture plates before had different areas. The 13 bacteria in plate 1 were based on an area of 6.3 cm^2 , whereas the 31 bacteria in plate 2 were based on an area of 10.9 cm^2 ; the number of bacteria per cm^2 would then be $13/6.3 = 2.0635$ and $31/10.9 = 2.8440$, and their ratio would be 0.7256. Then the lower limit of the ratio per cm^2 would be $0.2013 \times (10.9/6.3) = 0.3483$, and the upper limit would be $0.8248 \times (10.9/6.3) = 1.427$. See [Daly et al. \(1991\)](#).

DETERMINING THE REQUIRED SAMPLE SIZE

There are at least three different ways of determining sample sizes, power, and differences between Poisson means, depending on the underlying assumptions. They give similar results for sample sizes > 100 but differ for smaller numbers.

If two Poisson distributions have means λ_1 and λ_2 , a simple formula to determine the required number of observations to give a difference with $\alpha = 0.05$ is (see van Belle <http://vanbelle.org/chapters/webchapter2.pdf>)

$$N = \frac{4}{(\sqrt{\lambda_1} - \sqrt{\lambda_2})^2}$$

If the means are 30 and 40, the number of events per group needed will be $N = \frac{4}{(\sqrt{30} - \sqrt{40})^2} = 5.57$, or a total of about 12. If the two means are closer together, such

as 7 and 11, the number needed per group is $N = \frac{4}{(\sqrt{7} - \sqrt{11})^2} = 8.89$, or a total of about 18.

These sizes can be also be calculated online from <http://www.quantitativeskills.com/sisa/statistics/t-test.php?mean1=33&mean2=26&N1=000&N2=000&SD1=00.00&SD2=00.00&CI=95&Submit1=Calculate>. For the previous example, this calculator gives 11 in each group.

Sometimes we are interested in calculating sample size for detecting a difference between two Poisson counts when there is a background count rate. Examples might be analyzing two Poisson radiation counts against a background of radiation or evaluating the significance of two sedatives in producing phocomelia (as with thalidomide) when there is already a low incidence of phocomelia in the absence of drugs. Let the background rate be λ^* , and the two experimental rates be λ_1 and λ_2 . Then

$$N = \frac{4}{(\sqrt{\lambda^* + \lambda_1} - \sqrt{\lambda^* + \lambda_2})^2}$$

Taking the example of rates of 7 and 11 used before, but add a background of 3, then we have $N = \frac{4}{(\sqrt{3+7} - \sqrt{3+11})^2} = 11.9$ for a total of about 24.

We may need to determine how large a sample size is needed to determine λ so that the upper or lower confidence bounds do not differ from the sample estimate by more than a specified percentage. These bounds can be calculated or else determined easily from graphs published by [Hahn and Meeker \(1991\)](#) (Example 18.17).

Example 18.17

A problem concerns the accuracy of the microsphere method of measuring regional blood flow. Microspheres are tiny (usually 15 μm in diameter) spheres that when well mixed with blood in the heart are distributed to all the organs and regions within organs in proportion to the flows to those regions and organs. The microspheres are trapped in the organs. At the end of the experiment, the organs are removed, cut up into appropriate pieces, and the microspheres are counted by virtue of radioactivity or contained dye. The question asked is how many microspheres need there be in, say, the left atrial wall for flow to be measured within 10% of the true flow with a probability of 95%. The concept here is that if flows showed stationarity (i.e., did not change from measurement to measurement) and we were to repeat the microsphere injection several times, there would be different numbers of microspheres trapped each time, simply because there would be slight changes in the mixing and distribution of the microspheres with each injection. What critical number of microspheres is needed?

Let the critical average number of microspheres be X . If we require the 95% confidence limits of X to be within 10% of the true flow that would be given by the average number of microspheres for all the injections, then the 95% confidence limits will be $\pm 0.1X$. Because the distribution of the microspheres is a Poisson variate, the 95% confidence limits are given by $1.96\sqrt{X}$. These two numbers are equal. Therefore,

$0.1X = 1.96\sqrt{X}$. Square to remove the square root, so that $0.01X^2 = 3.84X$. Multiply by 100 to remove the decimal point, so that $X^2 = 384X$, and $X = 384$.

If there are 384 microspheres per piece of tissue to be measured, the requirements will be satisfied. [There are, of course, other technical points to be covered before the method can be accurate.]

This same method could be used for other degrees of precision, for example 5%, or for other confidence limits, for example, 99%, by substituting the appropriate figures in the previous equations.

Hahn and Meeker also give a simple computational formula for determining sample size:

$$n = \lambda^* \left[\frac{Z(1 - \alpha/2)}{d} \right]^2,$$

where λ^* is the desired mean rate, and d is the $100(1-\alpha)$ confidence interval of length $\pm d$, usually given as a percentage of λ^* . This approximation works well if $n > 10$ (Example 18.18).

Example 18.18

We wish to determine how many water samples of fixed volume to count to determine if the mean number of pathogenic *E. coli* bacteria is 3, with 95% confidence limits of 20% of the mean value, that is, $3 \times 0.2 = 0.6$. Then

$$n = 3 \left(\frac{1.96}{0.6} \right)^2 = 32$$

APPENDIX

The formula for the variance of a Poisson distribution follows naturally from the equivalent formula for the binomial distribution. In the latter, the mean is $N\pi$ and the variance is $N\pi(1-\pi)$. As π becomes smaller, the expression $(1-\pi)$ approaches 1 and the variance becomes $N\pi$.

REFERENCES

- Armitage, P., Berry, G., Matthews, J.N.S., 2002. *Statistical Methods in Medical Research*. Blackwell, Oxford.
- Daly, L.E., Bourke, G.J., McGilvray, J., 1991. *Interpretation and Uses of Medical Statistics*. Blackwell Scientific Publications, Oxford.
- Elliott, J.M., 1983. *Some Methods for the Statistical Analysis of Samples of Benthic Invertebrates*. Freshwater Biological Association, Ambleside, Cumbria.
- Hahn, G.J., Meeker, W.Q., 1991. *Statistical Intervals. A Guide for Practitioners*. John Wiley and Sons, New York.
- Rutherford, E., Geiger, H., 1910. The probability variations in the distribution of alpha particles. *Phil Mag* 20, 698–704.
- Zar, J.H., 2010. *Biostatistical Analysis*. Prentice Hall, Upper Saddle River, NJ.

CHAPTER 19

Negative Binomial Distribution

INTRODUCTION

This distribution is used to solve two different problems. The first resembles a Bernoulli trial, in which the number of successes in the first n trials has a binomial distribution $(p + q)^n$ with parameters n and p . If we ask instead what random variable r will give the number of trials at which the k -th success is achieved, then we use the negative binomial distribution, because it is derived from the expansion of $(q - p)^{-k}$, where $p = \mu/k$ and μ is the mean number of events.

The parameters of this distribution are the arithmetic mean μ (a measure of location), and k (a measure of dispersion), which is not necessarily an integer. If k is an integer, the distribution is known as the Pascal distribution. The negative binomial distribution is appropriate when:

The experiment consists of x repeated trials.

Each trial can result in just two possible outcomes. One of these outcomes is a success and the other is a failure.

The probability of success, denoted by p , is the same on every trial.

The trials are independent, that is, the outcome of one trial does not affect the outcome on other trials.

The experiment continues until r successes are observed, with r specified in advance.

PROBABILITY OF R SUCCESSES

Suppose that independent trials, each with probability p of being a success, are done until there are r successes. If X is the number of trials required, then

$$P(X = n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}, \text{ where } n = r, r+1, \dots$$

In order for the r -th success to occur in the n -th trial, there must be $r-1$ successes in the first $n-1$ trials, and the n -th trial must be a success. The probability of $r-1$ successes in the first $n-1$ trials is $\binom{n-1}{r-1} p^{r-1} (1-p)^{n-r}$ by the binomial theorem, and the probability of the second event is p . Multiplying these two together gives $\binom{n-1}{r-1} p^r (1-p)^{n-r}$, the required probability.

Example 19.1

What is the probability of getting 3 heads after 8 tosses of a coin? $p = 0.5$, $X = n$ (number of heads) = 3, r = number of tosses. The formula provides the results in Table 19.1.

The probability of getting the third head in the eighth toss is 0.08203125.

Table 19.1 Probability of success

r = number of tosses for $X = 3$	Probability $X = 3$
3	0.125
4	0.1875
5	0.1875
6	0.15625
7	0.1171825
8	0.08203125
9	0.0546875
10	0.003525625 etc.

Example 19.2

Let $P(\text{head} = p = 0.5)$. Then to get 17 heads in 30 tosses, we have $P(16 \text{ heads in } 29 \text{ tosses and heads on } 30\text{th toss}) = \binom{29}{16} p^{16} (1-p)^{13} p = \binom{29}{16} p^{17} (1-p)^{13} = \frac{29!}{16!13!} 0.5^{17} 0.5^{13} = 0.0632$.

Free online calculators at <http://stattrek.com/Tables/NegBinomial.aspx>, <https://www.thecalculator.co/math/Negative-Binomial-Calculator-744.html> and <https://trignosource.com/statistics/negative%20binomial%20distribution.html> solve similar problems easily.

Example 19.3

How might we apply this to a medical problem? An oncologist wants to recruit 6 patients with breast cancer to test a new therapeutic agent to prepare for a large randomized trial. Assume that the probability of a patient agreeing to the trial is $p = 0.25$. What is the probability that $N = 16$ patients will have to be interviewed to obtain 6 ($=r$) consents?

$$\text{Calculate } p = \binom{N-1}{r-1} p^r (1-p)^{N-r} = \frac{15!}{5!10!} 0.25^5 0.75^{10} = 0.0413.$$

This probability may be of little use, and a more important question is the average number of interviews to obtain 6 consents. The mean μ of a negative binomial distribution is $r/p = 6/0.25 = 24$.

Problem 19.1 What is the probability that the oncologist will obtain 8 consents in 30 interviews?

OVERDISPERSED DISTRIBUTION

The negative binomial distribution has a more important use for a contagious or over-dispersed distribution, one with clumps of objects rather than a random distribution. In such a distribution, the variance is much greater than the mean, whereas in a Poisson distribution the variance is approximately equal to the mean, and in a binomial distribution the variance is less than the mean. Unlike the Poisson, the probability of any time or space being occupied by an event is not constant, and the occurrence of an event may affect the occurrence of other events.

An early example of the use of this distribution was provided by Greenwood and Yule in 1930. They examined the numbers of accidents in 414 machinists followed for 3 months (Table 19.2).

Table 19.2 Accidents and machinists

Observed		Expected	
Number of accidents	Number of machinists	Poisson	Negative binomial
0	296	256	299
1	74	122	69
2	26	30	26
3	8	5	11
4	4	1	5
5	4	0	2
6	1	0	1
7	0	0	1
8	1	0	0
Total 200	414		

If the accidents are independent events, then a Poisson distribution would be suitable. As shown, however, the Poisson distribution has a deficit of those with no accidents and an excess of those with one or more accidents. When a negative binomial distribution is fitted, however, the observed and expected numbers match, suggesting that some machinists are accident prone. The negative binomial is preferred to the Poisson distribution when events are more likely to recur in one group than another; for example, some people have more recurrent infections or asthmatic attacks than do others (Glynn and Buring, 1996).

The mean of a negative binomial is kq/p , and variance is kq/p^2 . The variance is also given as $\mu + \frac{\mu^2}{k}$. The expression $1/k$ is a measure of the excess variance due to possible clumping.

USES OF THE NEGATIVE BINOMIAL

It is used to model temporal and geographic variation of parasitic infections of plants, animals, and humans, in all of which zero infestation is frequent but a few have excessive numbers of infestations, for example, [Mwangi et al., 2008](#). It has been used to model accident statistics in many fields—occupational health, automobile accidents ([Ramirez et al., 2009](#)), or falls in the home ([Iinattiniemi et al., 2009](#)). Some have used this distribution to model the sizes of family practices in Canada ([Anderson et al., 1986](#)), the rate of consultations in a practice ([Kilpatrick, 1977](#); [Iinattiniemi et al., 2009](#)), or the number of episodes of psychiatric illness ([Smeeton, 1986](#)). The distribution model has even been extended to evaluating founder germ cell numbers ([Zheng et al., 2005](#)) and vasopressin mRNA distribution in the supraoptic nucleus ([McCabe et al., 1990](#)).

There are several methods for determining k and so being able to test for the fit to a negative binomial distribution. Free online advice for calculating k is given in http://influentialpoints.com/Training/negative_binomial_distribution.htm.

Consultation with a statistician is recommended.

Many negative binomial distributions are monotonic with a huge peak for those with no episodes. This is not a requirement for the negative binomial that could resemble a skewed Gaussian curve ([Mwangi et al., 2008](#)).

Fitting the distribution is not the end of the exercise. Once the value of k is determined, the investigator can then consider why that form of contagious distribution had occurred and postulate mechanisms that might lead to better understanding. One way of thinking about the meaning of k is that it indicates variation among individuals in their intrinsic level of contact that is responsible for departure from randomness.

Other comparable distributions are the zero-inflated Poisson and the zero-inflated Negative Binomial distributions referred to in [Chapter 34](#).

REFERENCES

- Anderson, J.E., Willan, A.R., Gancher, W.A., 1986. The negative binomial model and the denominator problem in a rural family practice. *Fam. Pract.* 3, 174–183.
- Glynn, R.J., Buring, J.E., 1996. Ways of measuring rates of recurrent events. *BMJ (Clin Res ed)* 312, 364–367.
- Iinattiniemi, S., Jokelainen, J., Luukinen, H., 2009. Falls risk among a very old home-dwelling population. *Scand. J. Prim. Health Care* 27, 25–30.
- Kilpatrick, S.J.J., 1977. Consultation frequencies in general practice. *Health Serv. Res.* 12, 284–298.

- McCabe, J.T., Kawata, M., Sano, Y., Pfaff, D.W., Desharnais, R.A., 1990. Quantitative in situ hybridization to measure single-cell changes in vasopressin and oxytocin mRNA levels after osmotic stimulation. *Cell. Mol. Neurobiol.* 10, 59–71.
- Mwangi, T.W., Fegan, G., Williams, T.N., Kinyanjui, S.M., Snow, R.W., Marsh, K., 2008. Evidence for over-dispersion in the distribution of clinical malaria episodes in children. *PLoS One.* 3e2196.
- Ramirez, B.A., Izquierdo, F.A., Fernandez, C.G., Mendez, A.G., 2009. The influence of heavy goods vehicle traffic on accidents on different types of Spanish interurban roads. *Accid Anal Prev* 41, 15–24.
- Smeeton, N.C., 1986. Distribution of episodes of mental illness in general practice: results from the second National Morbidity Survey. *J. Epidemiol. Community Health* 40, 130–133.
- Zheng, C.J., Luebeck, E.G., Byers, B., Moolgavkar, S.H., 2005. On the number of founding germ cells in humans. *Theor Biol Med Model* 2, 32.

SECTION V

Probability in Epidemiology and Medical Diagnosis

CHAPTER 20

Odds Ratio, Relative Risk, Attributable Risk, and Number Needed to Treat

BASIC CONCEPTS

Introduction

The odds ratio in [Chapter 14](#) gave a point estimate of how much a proportion of successes in one group differs from the proportion of successes in another group. By definition, “odds” is the probability of an event occurring, divided by the probability that an event does not occur. The odds ratio (OR) is the odds of an outcome occurring in one group, divided by the odds of an outcome occurring in another group. A comparable ratio is the relative risk ratio (RR), and the following discussion describes the relationships and uses of each of these ratios.

Care is needed to use the correct estimates of disease incidence, because there are many deceptively similar ratios in use. Three common estimates are (i) the incidence rate, which is the number of new incidences of the disease per unit time; (ii) the cumulative incidence which is the proportion of study subjects who develop the outcome of interest at any time during the follow-up period; and (iii) the incidence odds, the ratio of the number of subjects experiencing the outcome to those not experiencing the outcome ([Pearce, 1993](#)). The latter two are not rates and are discussed later. The data must be collected over the same time period. Detailed evaluations of these different estimates are described by [Kleinbaum et al. \(1982\)](#).

Cohort study

Consider a study that follows a large population from time t_0 to time t_n , noting who has a putative risk factor (cause or exposure) at onset and at the end of the period noting how many have response (disease) in those with and without risk factor. This is a prospective cohort study. The data might have been collected previously and we are examining the records, but as long as one moves forward from time t_0 to time t_n it is still a prospective study. People who have the disease at the outset are excluded. Then calculate the relative risk as

$$RR = \frac{\text{Incidence of disease in exposed group}(p_e)}{\text{Incidence of disease in unexposed group}(p_u)}$$

Most often, the data are represented by a 2×2 table (Table 20.1). By convention, exposure is shown as the rows and outcomes (disease, cure) by the columns. Methods for dealing with more complex tables are discussed in epidemiology texts.

Table 20.1 Factors and definitions

	Disease			Odds of disease	Probability of disease
	Yes	No	Total		
Exposure	a	b	R_1	a/b	$a/(a+b) = a/R_1 = p_e$
No exposure	c	d	R_2	c/d	$c/(c+d) = c/R_2 = p_u$
Total	C_1	C_2	N		$C_1/N = p_t$
Odds of exposure	a/c	b/d			

p_e is the probability of disease in the exposed population; p_u is the probability of disease in the unexposed population; p_t is the probability of disease in the whole population whether exposed or not. Another incidence is the probability of exposure in the whole population, R_1/N , denoted by p_{ex} .

Table 20.1 represents a typical data table relating an input (exposure) to an output (disease). In discussing how to make various calculations (see later) I will use the format of Table 20.2 (left side) rather than the format of (right side) that appears in some texts.

Table 20.2 Left side—format used in this chapter; right side—alternative format used in some texts

	Disease	No disease	Disease	No disease
Exposure	a	b	a	c
No exposure	c	d	b	d

Total number N is $a + b + c + d$, and this and cells a and d are unaffected by the different arrangements.

One way to avoid errors due to use of the wrong symbols is to specify the conditional probability in each cell. The probability of disease (D^+) given exposure (E^+) is written as $P(D^+|E^+)$, the probability of no disease in exposed subjects is $P(D^-|E^+)$. These are unambiguous. $P(D^+|E^+) = a/N$, $P(D^-|E^+) = b/N$, (in Table 20.2, left side), and so on. Because N is common to all these proportions, we can work with either absolute numbers or proportions.

As an example, 122,612 normotensive people and 18,310 with systolic hypertension were followed to determine the incidence of cardiovascular (CV) deaths in each group (Table 20.3) (Kelly et al., 2008).

Table 20.3 Data on blood pressure and strokes

	CV death	No CV death	Total	Incidence
Hypertension	2134	16,176	18,310	0.116548
Normotension	3882	118,730	122,612	0.0316608
Total	6016	134,906	140,922	0.0426903

The incidence of CV deaths in hypertensives (p_e) [$P(D^+|E^+)$] is $2134/18310 = 0.116548$, and in normotensives (p_u) [$P(D^+|E^-)$] is $3882/122612 = 0.0316608$. The relative risk RR is therefore $0.116548/0.0316608 = 3.6811$.

This study gives an estimate of the incidence of new CV deaths in each group over the time period. Because the data are set out as a 2×2 contingency table, the odds ratio is $\frac{2134 \times 118730}{16176 \times 3882} = \frac{253369820}{62795232} = 4.0349$. The odds ratio and the relative risk are similar.

For the whole population, the risk p_t is $6016/140,922 = 0.0426903$.

The online calculators at http://www.medcalc.org/calc/relative_risk.php and <http://www.ebm.med.ualberta.ca/TherapyCalc.html> give the relative risk, and <http://statpages.org/ctab2x2.html> gives the odds ratio as well.

Noncohort study

One alternative to a cohort study is a cross-sectional study. This is done at one time on a series of subjects who are not followed. The whole of a specified population is examined at one time, looking for associations between putative causes and outcomes. In another alternative, patients are matched for response, and the risk factor is calculated for each group. This is known as a case-control study. The choice of which type of study to do is often practical. If the response of interest is rare, for example, a congenital disease with an incidence of about $1/10,000$ live births, then it takes a huge number of people to be followed prospectively in a cohort study to obtain enough responses to evaluate. It is less costly and time consuming to select 500 patients with the disease and 500 without the disease, and then look back to determine which group differed in antecedent factors. Because the subjects are selected (based on factor or response) and not chosen at random, no population incidence can be determined. For this reason, we cannot calculate relative risk but must use the odds ratio. An example is shown in Table 20.4 in a cross-sectional study relating exposure to second-hand smoke to ischemic strokes (He et al., 2008).

Table 20.4 Relation of stroke to second-hand smoke exposure

	Stroke	No stroke	Total	Incidence
Exposure	83	394	477	0.174004
No exposure	89	643	732	0.121585
	172	1037	1209	0.142266

The incidence of stroke in those exposed to cigarette smoke [$P(D^+|E^+)$] is $83/477 = 0.174004$, and in those not exposed [$P(D^+|E^-)$] is $89/732 = 0.121585$. Because the data are set out as a 2×2 contingency table, the odds ratio is 1.522.

Odds and risk ratios can be calculated online from <http://vassarstats.net/odds2x2.html>, or <http://statpages.org/ctab2x2.html>. Be careful entering data into each cell.

The “relative risk” from the table is $0.174004/0.121585 = 1.4311$. This is not a true relative risk because each group is selected and may or may not approximate the true relative risk, depending on the relative rarity of the response in the population (p_t). The difference is shown in the artificial example in [Tables 20.5a–d](#) in which the number of patients with strokes is constant but the numbers without strokes increases from a to d.

Table 20.5 Relationship between OR and RR for different prevalences

	a			b			c			d		
	S	No S	T	S	No S	T	S	No S	T	S	No S	T
Smoker	63	50	113	63	542	605	63	1259	1322	63	12,593	12,656
Nonsmoker	33	50	83	33	539	572	33	1260	1293	33	12,732	12,765
Total	96	100	196	96	1081	1177	96	2519	2615	96	25,325	25,421

S, stroke; No S, no stroke; T, total.

These data are analyzed in [Table 20.6](#).

Table 20.6 Analysis of [Table 20.5](#)

	Group			
	a	b	c	d
OR	1.91	1.90	1.91	1.93
“RR”	1.40	1.80	1.87	1.926
p_t	0.49	0.08	0.037	0.0038

OR, odds ratio; RR, approximate relative risk; p_t , population prevalence (total stroke/population).

Relative risk and odds ratio are similar only if the population prevalence p_t is very low because then the data in the “total” and “no stroke” columns are similar. Relative risk (RR) can be written as $\{a/(a+b)\}/\{c/(c+d)\}$. If a and c are very small relative to b and d , this ratio approximates $(a/b)/(c/d)$ which is another way of writing the odds ratio. Therefore although it is relative risk that we want, the odds ratio is a reasonable estimate of relative risk if the response (disease) is relatively rare. It is only in the study of [Table 20.5d](#) that we can even approximate a true population prevalence. These and other problems with the odds ratio are set out clearly by Sackett et al. ([Sackett et al., 1996](#); [Altman, 1998](#)).

[Zhang and Yu \(1998\)](#) developed a formula relating OR to RR, $RR = \frac{OR}{(1 - P_o) + (P_o \times OR)}$, where P_o is the outcome of interest in the nonexposed group (prevalence in control group).

This was rearranged by [Shrier and Steele \(2006\)](#):

$$OR/RR = (OR - 1)P_o + 1.$$

P_o is the prevalence of the disease in the population. If P_o and OR are both small, then $OR/RR \approx 1$ and OR is a good estimate of RR, but if either OR or P_o is large, the close relationship breaks down. They provided appropriate graphs for examining these relationships. Other helpful diagrams are provided by [Davies et al. \(1998\)](#) ([Example 20.1](#)).

Example 20.1

[Plint et al. \(2009\)](#) treated patients who came to the Emergency Room with bronchiolitis and compared the effectiveness of placebo vs nebulized epinephrine + oral dexamethasone as judged by what percentage in each group were admitted to hospital in the next 7 days. They found the following results

	Placebo	Epinephrine + steroid
Readmitted	53	34
Not readmitted	148	165

The odds ratio for readmission while taking epinephrine plus steroids is 0.58. We do not know if this is similar to the relative risk because we do not know the true value of P_o that in this population might be large.

Problem 20.1 A study of the relation between prehypertension (systolic blood pressure 120–139 mmHg or diastolic pressure 80–89 mmHg and presence or absence of diabetes mellitus ([Shrier and Steele, 2006](#)).

	Normal blood pressure	Prehypertension	Total
Diabetes	445	652	1097
No diabetes	794	738	1532
Total	1239	1390	2629

Calculate the odds ratio and “relative risk” Is this a true relative risk?

Cautionary tales

A cross-sectional study allows only legitimate calculation of the odds ratio, but this is frequently used as an estimate of the risk ratio. Failure to appreciate the difference between odds ratio and risk ratio, and when the odds ratio can and cannot be a surrogate for the risk ratio, accounts for many serious errors in the medical literature.

Katz (2006) analyzed a report from Nijsten et al. (2005) who found that 89.3% of National Psoriasis Foundation members reported having heard of treatment with calcipotriene compared with 25.3% of nonmembers. They reported an odds ratio of 24.41 and concluded that “members were more than 20-fold more likely than nonmembers to have heard of calcipotriene.” On the face of it, this result appears to be unlikely. The ratio of members to nonmembers who had heard of the drug was $89.3/25.3 = 3.53$, and this is the relative risk. Where did the error occur? It was due to ignoring the difference between odds ratio and relative risk (Katz, 2006).

A more serious error entered the medical literature when Schulman et al. (1999) reported that both race and gender influenced the way in which physicians managed chest pain, and concluded that blacks and women were 40% less likely than whites and men to be referred for cardiac testing (cardiac catheterization) when they presented with chest pain. These conclusions were given prominence by newspapers and television commentaries. Schulman et al. based this conclusion on finding in their study that blacks had an odds ratio of 0.6 for referral, but in reality, as pointed out by Schwartz et al. (1999) the relative risk was 0.93, a relatively trivial difference compared to whites. Schwartz et al. did not discount the possibility of bias in referrals but did not find good evidence for this in the reported study.

Both the odds ratio and the relative risk are point estimates, and we need their confidence limits. For the odds, the population value is usually written as Ω_i , and the ratio of two odds Ω_i and Ω_j as ω . The corresponding sample values are O_i and O_j , and their ratio is o or OR.

Because the odds ratios do not follow a normal distribution, use the logarithm to base e of the odds ratio OR (Gardner and Altman, 1995)

$$SE \log_e OR = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}.$$

For the data in Table 20.3, the standard error of $\log_n OR$ is

$$SE \log_e OR = \sqrt{\frac{1}{2134} + \frac{1}{3882} + \frac{1}{16176} + \frac{1}{118730}} = 0.02882.$$

To determine the 95% confidence interval, calculate the lower (L) and upper (U) limits in logarithmic units as

$$L = \log_e OR - (Z_{1-\alpha/2} \times SE \log_e OR),$$

and

$$U = \log_e OR + (Z_{1-\alpha/2} \times SE \log_e OR).$$

The only difference between these expressions is the sign between the first and second parts of the right-hand side of the equation

$$L = \log_e 4.035 - (1.96 \times 0.02822) = 1.3950 - 0.05649 = 1.3385,$$

and

$$U = \log_e 4.035 + (1.96 \times 0.028821) = 1.3950 + 0.05649 = 1.4515.$$

These are the 95% confidence limits for the odds ratio in logarithmic units. To recover the actual units, the 95% confidence limits are the antilogarithms $e^{1.3385}$ to $e^{1.4515}$, or 3.8133 to 4.2695. Because this range does not include 1, there is a difference in proportions between the two treatment groups, and we reject the null hypothesis; this is exactly what the chi-square test concluded.

If any counts are zero, it is not possible to calculate an odds or cross-product ratio, and the standard error of OR is undefined. To overcome this problem, add a small number to each count, commonly 0.5, and the odds ratio and its standard error are calculated with these new increased counts. Thus

$$OR = \frac{(a + 0.5) \times (d + 0.5)}{(b + 0.5) \times (c + 0.5)},$$

and

$$SE \log_e OR = \sqrt{\frac{1}{a + 0.5} + \frac{1}{b + 0.5} + \frac{1}{c + 0.5} + \frac{1}{d + 0.5}}.$$

The modification makes little difference to the results when the numbers are large. Online calculations can be performed at <http://vassarstats.net/odds2x2.html>, <http://www.hutchon.net/confidrr.htm>, or <http://statpages.org/ctab2x2.html>.

For the relative risk confidence intervals, use (Gardner and Altman, 1995)

$$SE \log_e RR = \sqrt{\frac{1}{a} - \frac{1}{a + b} + \frac{1}{c} - \frac{1}{c + d}}.$$

(Be careful. The two reciprocals added together involve the bad outcomes in control and exposed groups, and the two reciprocals subtracted are the sums of the exposed and unexposed groups, respectively.)

The $100(1 - \alpha)\%$ confidence interval is given by

$$\log_e RR \pm (z_{1 - \alpha/2} \times SE \log_e RR).$$

From Table 20.3

$$SE \log_e RR = \sqrt{\frac{1}{2134} - \frac{1}{2134 + 16176} + \frac{1}{3882} - \frac{1}{3882 + 118730}} = 0.025757.$$

Because the relative risk for the data in Table 20.3 was 3.6811, the 95% confidence limits (equivalent to $\alpha = 0.05$) are $\ln 3.6811 \pm 1.96 \times 0.025757 = 1.3032 \pm 0.0505 = 1.2697$ to 1.3707 . Taking antilogarithms $e^{1.2697}$ and $e^{1.3707}$ gives the 95% confidence limits as 3.5597 and 3.9381.

These confidence limits can be calculated online at <http://statpages.org/ctab2x2.html>, <http://www.hutchon.net/confidrr.htm>. These limits can be calculated online at <https://mathcracker.com/relative-risk-calculator.php#results>, or <http://vassarstats.net/odds2x2.html>, but these programs give minimally different limits.

To determine the significance of the difference between two different relative risk values, use

$$z = \frac{RR_1 - RR_2}{\sqrt{\frac{RR_1(1 - RR_1)}{n_1} + \frac{RR_2(1 - RR_2)}{n_2}}}$$

Sample size and power

To detect the relative risk of a rare disease in a prospective (cohort) study, the sample size N for each of the exposed and the unexposed groups (R_1 and R_2 as set out in Table 20.1) that gives Type I error $\alpha = 0.05$ and power $1 - \beta = 0.8$ is given by van Belle (2002) as

$$N = \frac{4}{p_u(\sqrt{RR} - 1)^2},$$

where RR is the relative risk, and $p_u = c/R_2$ (the probability of disease in the unexposed population).

The equation allows us to calculate the number of unexposed (and exposed) subjects.

The number of events needed in the unexposed group rises sharply as the relative risk becomes smaller. To determine the number needed in the exposed group, multiply the previous numbers by the relative risk.

The numbers required for a very low population incidence of the disease in the unexposed group are huge (Fig. 20.1).

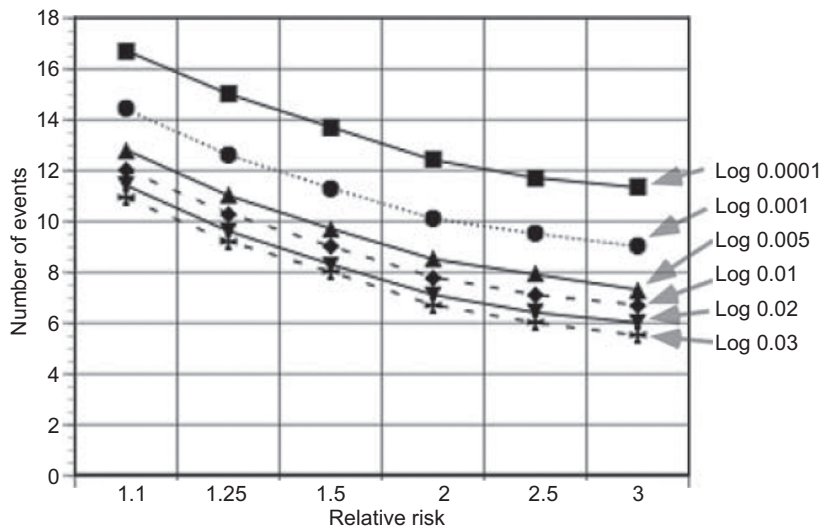


Fig. 20.1 Relative risk vs logarithm of the number needed in exposed group. The values for Y are the population incidences, ranging from $3/100$ to $1/10,000$. A graph for estimating power for a given relative risk and sample size is provided on page 77 in the book by Selvin (1991).

If the incidence in unexposed subjects is 0.0001, the required numbers range from 74,641 for a relative risk of 3 to 16,790,471 for a relative risk of 1.1. This is why the prospective study is seldom used for very rare diseases and is replaced by the cross-sectional study.

Such calculations allow us to decide if the number of subjects required can be recruited for the study in the allotted time or, given the early results, how many more will be needed.

To determine if the results from a completed study had adequate power of 0.8 and $\alpha = 0.05$, use the formula given before. An approximation to these results can be obtained online at <http://www.stat.ubc.ca/~rollin/stats/ssize/caco.html>.

These risk values can also be used to assess the public health benefits of removing a putative factor responsible for diseases. Not all clinical entities have a relative risk. It is not possible for a person to get pneumonic plague unless there is exposure to the plague bacillus, so that although it is possible to calculate the attack rate of the disease as $\frac{\text{number with disease}}{\text{number exposed}}$, there is no way to calculate a relative risk because unexposed people cannot get the disease.

Problem 20.2 Calculate sample size if the risk ratio is 1.5 and p_u is 0.03, if the risk ratio is 2.7 and p_u is 0.3, and if the risk ratio is 2.7 and p_u is 0.8.

When a relative risk can be estimated, it must be interpreted with caution. A relative risk of 2.0 might occur if $p_e = 0.2$ and $p_u = 0.1$, or $p_e = 0.000002$ and $p_u = 0.000001$, where p_e and p_u refer to exposed and unexposed incidence rates, respectively. In the first example disease is potentially preventable in 1 out of 10 people, but in the second example only 1 in 1 million might be helped. The different public health implications of these two different absolute figures are concealed in the single relative risk number. Relative risk should never be referred to without also referring to the absolute risk.

Attributable risk

To derive a more interpretable magnitude, use attributable risk (AR), sometimes referred to as the population attributable risk (PAR), the attributable fraction, the etiologic fraction λ , or the relative risk difference (Sinclair, 2003). The field is clouded by a variety of terms for the same calculation, a variety of calculations for the same term, and a variety of symbols for the same entities. All the different formulas use the same few numbers and so are interrelated. The term attributable risk (AR) is used here to refer to the excess incidence of disease in those exposed to a given (?causal) factor, and AR% is the percentage of the incidence of a disease in the exposed that would be eliminated if exposure were eliminated; it is sometimes expressed as a proportion. The corresponding index

for the whole population is PAR, referring to the excess incidence of disease in the whole population, and PAR% is the percentage of the incidence of a disease in the whole population that would be eliminated if exposure were eliminated; it is sometimes expressed as a proportion.

The attributable fraction indicates how many excess diseased subjects were due to exposure, given that there usually is a baseline incidence of the disease (p_u) in the absence of exposure to a particular factor and the superimposed incidence rate due to exposure (p_e) to that factor. Table 20.1 presents the basic data used in the calculations.

The attributable risk is the excess risk due to the factor as measured by $AR = p_e - p_u$ (Table 20.1). The proportion of the incidence rate due to association with the risk factor is, $(p_e - p_u)/p_e$, either as a fraction (proportional AR) or multiplied by 100% (AR as %).

Applying these concepts to the data in Table 20.3, $p_e = 2134/18310 = 0.116548$, $p_u = 3882/122612 = 0.03166$, and $p_t = 6016/140922 = 0.042693$. Then $RR = 0.116548/0.03166 = 3.6182$, $AR = 0.116548 - 0.03166 = 0.084888$, and AR%, the percentage of diseased subjects that are associated with exposure, is $(0.084888/0.116548) \times 100 = 72.84\%$ (or 0.7284). Exposure is associated with excess disease in about 73/100 patients.

There are other ways of determining AR that may be used if some of the primary data are not available. Because $RR = p_e/p_u$, then $p_e = RR p_u$.

a. Therefore $AR = p_e - p_u = RR p_u - p_u = p_u(RR - 1)$.

Using the previous numbers, $0.03166 \times 2.6182 = 0.0828$, similar to the previous calculation.

b. Dividing the numerator and denominator of the expression for proportional AR by

$$p_u \text{ gives proportional AR} = \frac{\frac{p_e}{p_u} - \frac{p_u}{p_u}}{\frac{p_u}{p_u}} = \frac{RR - 1}{RR}, \text{ as long as } RR > 1.$$

Using the data from Table 20.3, proportional AR = $(3.6182 - 1)/3.6182 = 0.7236$ (or 72.36%), much as before. (Small differences are due to rounding off.)

The incidence rate due to the risk factor among those with the risk factor is

$$p_e \left(\frac{RR - 1}{RR} \right) \text{ as long as } RR > 1.$$

From the data in Table 20.3 this becomes $0.116548 \times 0.7236 = 0.0843$, much as before. These alternative formulations require a cohort study that allows relative risk to be calculated or require the assumption that the odds ratio and the relative risk are similar.

Population attributable risk

The population attributable risk (PAR) as used here is determined by the difference between the incidence in exposed subjects and the incidence in the unexposed

population, that is, $p_e - p_t$, and proportional PAR is $(p_e - p_t)/p_e$. From the data of Table 20.3, $PAR = 0.0427 - 0.0317 = 0.011$, and proportional PAR is $0.011/0.04287 = 0.257$ or 25.8%.

Comparing AR and PAR shows that removing the exposure might reduce the incidence of the disease by 73% in the exposed population and 27% in the whole population. Finally, PAR can be calculated easily from

$$\left(\frac{a}{a+c}\right)\left(\frac{RR-1}{RR}\right).$$

From Table 20.3,

Proportional PAR = $\left(\frac{2134}{2134 + 3882}\right)\left(\frac{3.6811 - 1}{3.6811}\right) = 0.2583$ similar to the previous calculation.

These various risks can be calculated online at <https://www2.ccrb.cuhk.edu.hk/stat/confidence%20interval/CI%20for%20relative%20risk.htm>.

Relative risks below 1; number needed to treat

The relative risk can be <1 if the exposure is protective, for example, with an effective treatment, and then it allows the calculation of how many patients are needed in a clinical trial. For example, Sinclair (2003) asked the question: “Based on existing data in the literature about the possible effect of using steroids to prevent chronic lung disease at 36 weeks in a premature 900 g infant, how many patients will we need to treat to demonstrate one success?”

He used the following empirical data.

Relative risk (RR) = 0.69 (69%);

Risk reduction = $1 - RR = 0.31$ (31%);

Risk difference (RD) = $p_e - p_u = 0.09$ (9%).

This is what others have termed attributable risk and has to be calculated from the primary data.

Then NNT, the number of patients needed to treat in order to prevent chronic lung disease in 1 patient = $1/RD = 1/0.09 \approx 11$ (Laupacis et al., 1988). That is, given the known effect of steroids on these immature lungs, we would have to treat 11 patients in order to see 1 improved patient. This concept is important in terms of cost but even more so in terms of risks. From the database, it is possible to determine how often a given complication would occur. If the NNT for complications is lower than for improvement, the risk might be unacceptable. If it is much larger, then the risk might be tolerable. Sinclair describes how to determine risk-benefit ratios. Online calculations can be performed at <http://graphpad.com/quickcalcs/NNT2.cfm>, <http://araw.mede.uic.edu/cgi-bin/nntcalc.pl>,

<http://www.calctool.org/CALC/prof/medical/NNT>, <https://easycalculation.com/medical/treat-number.php>, <http://statpages.org/ctab2x2.html>.

The NNT concept allows us to attach concrete numbers to concepts of relative risk and attributable risk (or risk difference). For example, the relative risk might be 2 for each of two different diseases. For disease A the attributable risk might be $0.2 - 0.1 = 0.1$, with NNT being $1/0.1 = 10$, whereas for disease B the attributable risk might be $0.002 - 0.001 = 0.001$, so that NNT is $1/0.001 = 1000$. For disease A we need to treat 10 patients for one to benefit, and for disease B we need to treat 1000 for one to benefit. A simple nomogram for determining NNT is available (Chatellier et al., 1996).

If a new treatment turns out to be harmful, then the attributable risk $p_e - p_u$ is negative. Altman (1998) proposed that instead of using the term NNT we should use the terms NNTB, where B=benefit, and NNTH, where H=harm. NNTB is a positive number, and NNTH a negative number. This concept is of value when calculating confidence limits (see later).

When applying the NNT calculation all members of the intended treatment group must be homogeneous with respect to the relative risk. If several strata are used for the trial, then the relative risk should be similar in each stratum if it is to be used to calculate a single number or NNTB. If one person or subgroup has a different baseline risk, say f times as high, then the number to treat based on the rest of the study needs to be divided by f to obtain a realistic NNTB (Cook and Sackett, 1995).

Some investigators have criticized the NNT concept for emphasizing the benefit to the one patient while ignoring the lack of benefit or even harm to the remainder of the group (Bogaty and Brophy, 2005). To stress this point they recommended (perhaps with tongue in cheek) the use of a new index—NTN, the number treated needlessly. Thus if NNT was 250, NTN would be 249, and the investigator must be sure that the benefit to the one patient is sufficiently important to justify treating the remaining patients with no benefit.

As with all summary numbers there are subtleties to consider. If, for example, control (c) and treatment (t) groups are observed then there will be a NNT_c and an NNT_t , and it is the difference between these that demonstrates the value of the treatment (Curiel and Rodriguez-Plaza, 2005). Furthermore, if two groups are to be compared, they need to be followed for similar periods if NNT is to have meaning (Suijsa et al., 2012) (Example 20.2).

Example 20.2

Plint et al. (2009) treated patients who came to the Emergency Room with bronchiolitis and compared the effectiveness of placebo versus nebulized epinephrine with oral dexamethasone as judged by what percentage in each group were admitted to hospital in the next 7 days.

They found the following results:

	Placebo	Epinephrine plus steroid
Readmitted	53	14
Not readmitted	148	155

Calculate the NNT from the formula cited before.
$$NTT = \frac{1}{\frac{53}{201} - \frac{34}{200}} = 10.67$$
 or 11 to

the nearest integer:

An online calculator <http://graphpad.com/quickcalcs/NNT2/> gave 11 as the answer, as well as confidence limits.

Problem 20.3 Assume the Plint data table was as follows.

	Placebo	Epinephrine plus steroid
Readmitted	43	34
Not readmitted	158	165

Calculate the odds and risk ratios with confidence limits, and NNT by using <http://statpages.org/ctab2x2.html> and the formula.

ADVANCED CONCEPTS

Confidence limits for attributable risk

Approximate limits can be set in several ways, depending on whether we examine the difference between proportions or some ratio of proportions.

Differences between proportions

Because attributable risk is the difference between two proportions p_e and p_u , use the formula for the difference between two proportions.

$$(p_e - p_u) \pm 1.96 \sqrt{\frac{p_e(1 - p_e)}{n_1} + \frac{p_u(1 - p_u)}{n_2}},$$

where n_1 is the total number exposed and n_2 is the total number unexposed. For the data of Table 20.3, the confidence limits are

$$\begin{aligned}
 (0.1165 - 0.0317) \pm 1.96 \sqrt{\frac{0.1165(1 - 0.1165)}{18310} + \frac{0.0317(1 - 0.0317)}{122162}} \\
 = 0.0848 \pm 0.004749.
 \end{aligned}$$

This produces 95% confidence limits of 0.0801 to 0.0895. These limits can be calculated online from http://vassarstats.net/prop2_ind.html that gives limits of 0.0802 to 0.0897 with or without a continuity correction; the continuity correction should be used for small sample sizes. Another program <http://in-silico.net/statistics/ztest/two-proportion> gives limits of 0.0792 to 0.0904 and allows for unequal sample variances. The differences between test results depend on the specific method used, with http://vassarstats.net/prop2_ind.html using Wilson's method that has been shown to be more accurate than most.

Similar calculations can be done for PAR, using p_t and p_u as the proportions.

Difference between proportional ratios (Proportional AR and Proportional PAR)

Walter developed a method for confidence limits for the proportional population attributable risk λ_p that depends finding that the distribution of $1 - \lambda_p$ is asymptotically log normal. In the form depicted by [Armitage et al. \(2002\)](#)

$$\begin{aligned}
 \text{SE} \log_e(1 - \lambda) &= \sqrt{\frac{a}{c(a+c)} + \frac{b}{d(b+d)}} \\
 &= \sqrt{\frac{2134}{3882(2134 + 3882)} + \frac{16176}{118730(16176 + 118730)}} = 0.009612.
 \end{aligned}$$

95% limits are $-0.2979 \pm 1.96 \times 0.009612 = -0.3167$ to -0.2791 .

Taking antilogarithms gives 0.7285 to 0.7565.

Subtracting these from 1 gives 95% limits of 0.2435 to 0.2715.

Finally, a simplified method involves substituting the lower and upper 95% confidence limits for $\widehat{\text{RR}}$ in the relevant equation to obtain the corresponding confidence limits for $\widehat{\text{PAR}}$ ([Daly, 1998](#)).

Then the lower confidence limit (L) is

$$\frac{2134}{2134 + 3882} \left(\frac{3.5473 - 1}{3.5473} \right) = 0.2547,$$

Similarly, the upper confidence limit (U) is

$$\frac{2134}{2134 + 3882} \left(\frac{3.8198 - 1}{3.8198} \right) = 0.2619.$$

Calculations involving relative risk must be done cautiously in cross-sectional studies, and designs such as pairing may need special approaches. Statistical consultation is advised.

Confidence limits for NNT

These have conventionally been determined from the reciprocals of the confidence limits for the attributable risk (Cook and Sackett, 1995). If the attributable risk is 0.09 and the 95% confidence limits are 0.06 and 0.14, then the NNT is $1/0.09 = 11$, with confidence limits of $1/0.06 = 17$ and $1/0.14 = 7$. This method, however, works only if the attributable risk is substantially above zero and both confidence limits are positive. Altman (1998) raised a conceptual difficulty if the attributable risk is small because then the confidence limits range from positive to negative. If, as in his example, the attributable risk is 10% with 95% confidence limits of -5% to 25% , then the number to treat is $1/0.1 = 10$, and those limits imply numbers to treat of $-1/0.05 = -20$ to $1/0.25 = 4$. Apart from not making any sense to have a negative number to treat, these numbers do not include the point estimate of NNT. In addition, the limits include the possibility that the attributable risk is zero, which would lead to $\text{NNT} = 1/0 = \infty$. Altman pointed out there are two disjoint sets of data, one from -20 to ∞ and one from ∞ to 4 . In order to resolve the difficulty he developed a diagram plotting the attributable risk reduction from -4 to 20 on a reversed axis on the left and on a double inverted scale on the right, as in Fig. 20.2.

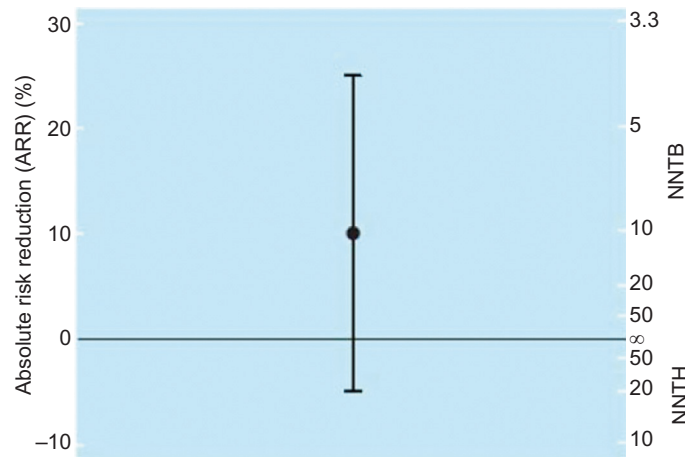


Fig. 20.2 Method of setting confidence limits for NNT.

The confidence limits calculated as suggested before are approximate and usually serve their purpose, but may be greatly in error if samples are small.

REFERENCES

- Altman, D.G., 1998. Confidence intervals for the number needed to treat. *Br. Med. J.* 317, 1309–1312.
- Armitage, P., Berry, G., Matthews, J.N.S., 2002. *Statistical Methods in Medical Research*. Blackwell, Oxford.
- Bogaty, P., Brophy, J., 2005. Numbers needed to treat (needlessly?). *Lancet* 365, 1307–1308.

- Chatellier, G., Zapletal, E., Lemaitre, D., Menard, J., Degoulet, P., 1996. The number needed to treat: a clinically useful nomogram in its proper context. *Br. Med. J.* 312, 426–429.
- Cook, R.J., Sackett, D.L., 1995. The number needed to treat: a clinically useful measure of treatment effect. *Br. Med. J.* 310, 452–454 (correction p. 1056).
- Curiel, R., Rodriguez-Plaza, L., 2005. Basal NNT and interventional NNT. *J. Clin. Epidemiol.* 58, 1074.
- Daly, L.E., 1998. Confidence limits made easy: interval estimation using a substitution method. *Am. J. Epidemiol.* 147, 783–790.
- Davies, H.T., Crombie, I.K., Tavakoli, M., 1998. When can odds ratios mislead? *BMJ (Clin. Res. Ed.)* 316, 989–991.
- Gardner, M.J., Altman, D.G., 1995. *Statistics with Confidence—Confidence Intervals and Statistical Guidelines*. British Medical Journal, London.
- He, Y., Lam, T.H., Jiang, B., Wang, J., Sai, X., Fan, L., Li, X., Qin, Y., Hu, F.B., 2008. Passive smoking and risk of peripheral arterial disease and ischemic stroke in Chinese women who never smoked. *Circulation* 118, 1535–1540.
- Katz, K.A., 2006. The (relative) risks of using odds ratios. *Arch. Dermatol.* 142, 761–764.
- Kelly, T.N., Gu, D., Chen, J., Huang, J.F., Chen, J.C., Duan, X., Wu, X., Yau, C.L., Whelton, P.K., He, J., 2008. Hypertension subtype and risk of cardiovascular disease in Chinese adults. *Circulation* 118, 1558–1566.
- Kleinbaum, D.G., Kupper, L.L., Morgenstern, H., 1982. *Epidemiologic Research. Principles and Quantitative Methods*. Lifetime Learning Publications, Wadsworth, Inc, Belmont, CA.
- Laupacis, A., Sackett, D.L., Roberts, R.S., 1988. An assessment of clinically useful measures of the consequences of treatment. *N. Engl. J. Med.* 318, 1728–1733.
- Nijsten, T., Rolstad, T., Feldman, S.R., Stern, R.S., 2005. Members of the national psoriasis foundation: more extensive disease and better informed about treatment options. *Arch. Dermatol.* 141, 19–26.
- Pearce, N., 1993. What does the odds ratio estimate in a case-control study? *Int. J. Epidemiol.* 22, 1189–1192.
- Plint, A.C., Johnson, D.W., Patel, H., Wiebe, N., Correll, R., Brant, R., Mitton, C., Gouin, S., Bhatt, M., Joubert, G., Black, K.J., Turner, T., Whitehouse, S., Klassen, T.P., Pediatric Emergency Research, C., 2009. Epinephrine and dexamethasone in children with bronchiolitis. *N. Engl. J. Med.* 360, 2079–2089.
- Sackett, D.L., Deeks, J., Altman, D.G., 1996. Down with odds ratios! *Evid. Based Med.* 1, 164–166.
- Schulman, K.A., Berlin, J.A., Harless, W., Kerner, J.F., Sistrunk, S., Gersh, B.J., Dube, R., Taleghani, C.K., Burke, J.E., Williams, S., Eisenberg, J.M., Escarce, J.J., 1999. The effect of race and sex on physicians' recommendations for cardiac catheterization. *N. Engl. J. Med.* 340, 618–626.
- Schwartz, L.M., Woloshin, S., Welch, H.G., 1999. Misunderstandings about the effects of race and sex on physicians' referrals for cardiac catheterization. *N. Engl. J. Med.* 341, 279–283 (discussion 286–7).
- Selvin, S., 1991. *Statistical Analysis of Epidemiologic Data*. Oxford University Press, Oxford.
- Shrier, I., Steele, R., 2006. Understanding the relationship between risks and odds ratios. *Clin. J. Sport Med.* 16, 107–110.
- Sinclair, J.C., 2003. Weighing risks and benefits in treating the individual patient. *Clin. Perinatol.* 30, 251–268.
- Suissa, D., Brassard, P., Smiechowski, B., Suissa, S., 2012. Number needed to treat is incorrect without proper time-related considerations. *J. Clin. Epidemiol.* 65, 42–46.
- van Belle, G., 2002. *Statistical Rules of Thumb*. Wiley Interscience, New York.
- Zhang, J., Yu, K.F., 1998. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA* 280, 1690–1691.

FURTHER READING

- Altman, D.G., Deeks, J.J., Sackett, D.L., 1998. Odds ratios should be avoided when events are common. *BMJ (Clin. Res. Ed.)* 317, 1318.
- Boslaugh, S., McNutt, L.A., 2007. *Encyclopedia of Epidemiology*. Sage Publications, California.
- Feinstein, A.R., 1986. The bias caused by high values of incidence for p_1 in the odds ratio assumption that $1-p_1$ approximately equal to 1. *J. Chronic Dis.* 39, 485–487.
- Schlesselman, J.J., 1982. *Case Control Studies. Design, Conduct, Analysis*. Oxford University Press, Oxford.
- Walter, S.D., 2005. Is NNT now the number needed to traumatize? *J. Clin. Epidemiol.* 58, 1075–1076.

CHAPTER 21

Probability, Bayes Theorem, Medical Diagnostic Evaluation, and Screening

BAYES' THEOREM APPLIED

Almost all of us use a Bayesian approach to medical diagnosis, even if we do not realize it. Consider a 3-day-old patient with a transposition of the great arteries who is referred to a hospital. If the receiving doctor knows nothing more than the age, the list of possible diagnoses is enormous: congenital anomalies of the bowel, urinary tract infections, seizures, aspiration pneumonia, heart disease, and so on. If the patient is cyanotic, then heart and lung diseases go to the head of the list of possible diagnoses. If the baby is breathing normally, then heart disease is more likely than lung disease. If there is no cardiac murmur, then transposition of the great arteries goes to the head of the list. Almost all diagnoses go through this procedure, often so quickly that we do not realize that we are using a sequential logical process. What Bayes' theorem does is to formalize and quantify the process.

Diagnosing a neural tube defect prenatally is a practical example of the use of Bayes' theorem. Some children are born with neural tube defects, that is, the skin and vertebrae have not closed completely over the lower end of the spinal cord. These anomalies are associated with an increased concentration of alpha-fetoprotein in the blood. Therefore a high concentration of the protein suggests that there might be a neural tube defect, but because there are other causes of these high concentrations, they do not diagnose the defect with certainty. The other variable to be considered is that if one child has already been born with a neural tube defect, there is an increased tendency for subsequent children to have the abnormality. In order to give some diagnostic help to the perinatologist in attempting to diagnose neural tube defects in utero, turn to Bayes' theorem. The following primary data are needed:

D^+ is the presence of the disease and T^+ is a positive test, here an increased concentration of alpha-fetoprotein. If the family history suggests that the mother is at risk of having children with neural tube defects, then $P(D^+) = 0.05$. If the mother is not at risk, then $P(D^+) = 0.001$. These are the prior probabilities. Then there are two conditional probabilities that are the same whether the mother is at risk or not. $P(T^+ | D^-) = 0.001$, and $P(T^+ | D^+) = 1$; if the fetus has the disease, the alpha-fetoprotein concentration is always high, and if the fetus is normal, the probability of a false positive test is very low.

The simple form of Bayes' theorem (Chapter 5) can be rewritten in several forms.

Bayes' theorem is $P(D^+ | T^+) = \frac{P(T^+ | D^+)P(D^+)}{P(T^+)}$.

The denominator can be written $P(T^+) = P(T^+ \cap D^+) + P(T^+ \cap D^-)$, where D^- indicates those with no disease.

The two right-hand components can be rewritten:

$$\begin{aligned} P(T^+ \cap D^+) &= P(T^+ | D^+)P(D^+) \\ P(T^+ \cap D^-) &= P(T^+ | D^-)P(D^-) \end{aligned}$$

Now we can rewrite Bayes' theorem as

$$P(D^+ | T^+) = \frac{P(T^+ | D^+)P(D^+)}{P(T^+ | D^+)P(D^+) + P(T^+ | D^-)P(D^-)}$$

Because $P(T^+ | D^-) = 1 - P(T^- | D^-)$ we can also write the theorem as

$$P(D^+ | T^+) = \frac{P(T^+ | D^+)P(D^+)}{P(T^+ | D^+)P(D^+) + [1 - P(T^- | D^-)]}$$

The advantage of this formula is that it can be related to the well-known concepts of sensitivity and specificity (see later).

If the mother is at risk, then Bayes' theorem gives

$$P(D^+ | T^+) = \frac{1 \times 0.05}{(1 \times 0.05) + (0.001 \times [1 - 0.05])} = \frac{0.05}{0.05 + 0.00095} = \frac{0.05}{0.05095} = 0.9831$$

If the mother is not at risk, then the theorem gives

$$P(D^+ | T^+) = \frac{1 \times 0.001}{[(1 \times 0.001) + (0.001 \times \{1 - 0.001\})]} = \frac{0.001}{0.00199} = 0.5002.$$

For the mothers not at risk, the prior (pretest) probability of having a child with a neural tube defect is 0.001. The posterior probability if there is a positive test for an abnormal concentration of alpha-fetoprotein has risen to just over 50%, so that the test has given a great deal of information. If the mother has already had a child with a neural tube defect, then the prior probability of having another child with this defect is 0.05, but a positive test makes it almost certain that the fetus is affected. (In practice, other tests are done to make the diagnosis more certain.)

Cautionary Tale

The failure to apply Bayes' theorem has had devastating consequences in some criminal cases. One striking example was the case of Sally Clark ([Korb, 2012](#)). She had a child who died in infancy from what was diagnosed as SIDS (Sudden Infant Death Syndrome), but when a second infant died from presumed SIDS she was charged with infanticide.

She was convicted of murder largely because of testimony from an expert pediatrician who testified that the chances of two infants in the same family dying from SIDS were one in 73 million, and therefore extremely unlikely.

There were numerous errors in this testimony.

The probability of $1/73,000,000$ was derived from two misconceptions. Information available at that time was that SIDS occurred in $1/1300$ live births, but the expert witness chose to use a probability of $1/8543$ births because of information showing that SIDS was less likely in families who were affluent and nonsmoking. On the other hand, he ignored the fact that both infants were boys, who have a greater risk of dying from SIDS.

If the witness had used the figure of $1/1300$ instead of $1/8543$, then the chances of two successive deaths from SIDS would have been estimated as $1/1,690,000$ instead of $1/72,982,849$.

The second error was assuming that the two consecutive events were independent. It was known that if one child had died from SIDS the risk was 5–10 times higher for the next child. Therefore instead of estimating the probability of two consecutive deaths from SIDS as $1/1300 \times 1/1300 = 1/1,690,000$, the probability should have been 5- to 10-fold greater, namely, $1/169,000$ to $1/338,000$, a far cry from 1 in 73 million.

Later studies showed that the second infant had died from an infection. This information was never produced at the trial.

A more important conceptual error, however, was failing to consider the alternative hypothesis, namely, the probability of two infants being murdered in the same family. By conservative estimates from study groups this probability was about 5–10 times less likely than two consecutive SIDS deaths (Hill, 2004). Therefore in the absence of direct evidence of murder, on probability two consecutive SIDS deaths were about 5–10 times as likely as two consecutive infanticides in the same family. This error is sometimes termed the Prosecutor's Fallacy, namely, the tendency to state that a person is guilty because he or she is associated with some very rare event (Goldacre, 2006; Thompson and Schumann, 1987). Thompson and Schumann (1987) described this fallacy simply:

Suppose you are asked to judge the probability a man is a lawyer based on the fact that he owns a briefcase. Let us assume all lawyers own a briefcase but only one person in ten in the general population owns a briefcase. Following the prosecutor's logic, you would jump to the conclusions that there is a 90% chance the man is a lawyer. But this conclusion is obviously wrong. We know that the number of nonlawyers is many times greater than the number of lawyers. Hence lawyers are probably outnumbered by briefcase owners who are not lawyers (and a given briefcase owner is more likely to be a nonlawyer than a lawyer). To draw conclusions about the probability that the man is a lawyer based on the fact he owns a briefcase, we must consider not just the incidence rate of briefcase ownership, but also the a priori likelihood of being a lawyer. Similarly, to draw conclusions about the probability a criminal suspect is guilty based on evidence of a "match", we must consider not just the percentage of people who would match but also the a priori likelihood that the defendant in question is guilty.

These issues are described clearly in a Wikipedia article (Wikipedia, 2011).

Continued

Cautionary Tale—cont'd

Many pediatricians and statisticians pointed out the fallacies, and eventually after 4 years the sentence was reversed on a second appeal. Unfortunately, the tragedy of her children's deaths and 4 years unjustified imprisonment were too much for Sally Clark, who died a few years later. Failure to consider posterior probabilities and the incorrect prior probabilities caused a grave miscarriage of justice.

SENSITIVITY AND SPECIFICITY

There is another way of looking at diagnostic tests that is closely related to Bayes' theorem. A new test for a given disease should ideally detect all instances of the occurrence of that disease; the test should be *sensitive* to the presence of that disease. Perfect sensitivity of 100% would occur if the test were always positive when the disease was present; sensitivity below 100% would occur if the test were negative in some people who had that disease, that is, if there were false negative tests. Furthermore, ideally the test should be highly *specific* for that disease, and not to be positive if other diseases were present. If this occurred, then the test could be said to be 100% specific for that disease. If, however, the test was positive in some people who had other diseases, then that would be a false positive result, and the test would be <100% specific.

Table 21.1 presents data published by Rubin. Rubin (1992) about evaluating fever in a young child who has no obvious focus of infection. The causes of such a nonspecific symptom range from a self-limiting viral illness that needs no specific treatment to a serious bacterial infection that requires urgent antibiotic treatment. At issue are the facts that it is expensive and frequently unnecessary to perform many laboratory tests; furthermore, decisions about treatment may have to be made before test results return.

Table 21.1 Data from children with occult bacteremia

Neutrophil count (mm ³)	Sensitivity (%)	Specificity (%)
>10,000	100	88
>15,000	100	97
>20,000	40	99

All children with bacteremia have >10,000 neutrophils per mm³, so that for this test the sensitivity is 100%. However, for this cell count the specificity is only 88%; children with fever not due to bacteremia can often have neutrophil counts of this magnitude. If a positive test requires a neutrophil count of over 20,000 per mm³, then the sensitivity is only 40%, that is, only 40% of bacteremic children have such high neutrophil counts. On the other hand, the specificity has risen to 99%, because few children with febrile viral illnesses have such high counts.

Table 21.2 presents results concerning a prostatic acid phosphatase (PAP) test for prostate cancer (Watson and Tang, 1980).

Table 21.2 Prostate cancer data

Test	Disease		Total
	Cancer	Normal	
PAP positive	79	13	92
PAP negative	34	204	238
Total	113	217	330

Assume that the diagnosis of cancer has been established with certainty by another test (“the gold standard”). Then each of the four cells in which the coincidence of a test result and a disease result can be labeled is presented in Table 21.3; some derived ratios are also given.

Table 21.3 Generalized sensitivity and specificity table

Test	Disease		
	Present	Absent	
Positive	True positive TP 79	False positive FP 13	<i>Positive predictive value (PPV)</i> $TP/(TP + FP)$ $79/92 = 0.859$
Negative	False negative FN 34	True negative TN 204	<i>Negative predictive Value (NPV)</i> $TN/(TN + FN)$ $204/238 = 0.86$
	<i>Sensitivity</i> $TP/(TP + FN)$ $79/113 = 0.699$	<i>Specificity</i> $TN/(TN + FP)$ $204/217 = 0.94$	

Basic data in bold type. Derived ratios in italics.

Sensitivity = probability that a test result is **positive** when the disease is present.

Specificity = probability that a test result is **negative** when the disease is not present.

Sensitivity as a percentage is termed the true positive rate, and specificity as a percentage is termed the true negative rate.

The complements of these terms can also be defined:

1–Sensitivity is also termed the **false negative** rate, and is the probability that the test is negative when the disease is present:

$$1 - \text{Sensitivity} = \frac{\text{False negative}}{\text{True positive} + \text{False negative}} = \frac{34}{34 + 79} = 0.301.$$

1-Specificity is also termed the **false positive** rate, and it is the probability that the test is positive when the disease is absent:

$$1\text{-Specificity} = \frac{\text{False positive}}{\text{True negative} + \text{False positive}} = \frac{13}{13 + 204} = 0.060.$$

These numbers are seldom useful on their own in assessing probability of disease which is better done by using the following derived terms:

Positive predictive value = proportion of times that a positive test will detect a diseased person:

Negative predictive value = proportion of times that a negative test will detect a person without that disease:

$$\text{Prevalence} = \frac{\text{True positive}}{\text{Total population}}.$$

This is also sometimes referred to as the pretest probability, that is, the rate in the population before any diagnostic tests have been done. Online calculators for these various ratios are found at <http://statpages.org/ctab2x2.html> and <http://vassarstats.net/clin1.html> (all of which provide confidence limits as well), and <http://www.hutchonnet/Bayes.htm>

Some investigators use the terms predictive value or accuracy (as defined before) as equivalent. Most investigators use the terms predictive *value* for these ratios and reserve the term predictive *accuracy* for the ratio of all the correct tests to all the tests

$$\frac{\text{True positive} + \text{True negative}}{\text{True positive} + \text{True negative} + \text{False positive} + \text{False negative}}.$$

The true prevalence rate cannot be calculated from the data given in Table 21.2, because patients were selected because they did or did not have prostatic cancer and there is no reason to believe that these represent the true ratios in the population.

Because these ratios give point estimates for one particular set of data it is also valuable to calculate the 95% confidence intervals. For sensitivity these are $\text{Sensitivity} \pm 1.96 \sqrt{\frac{\text{Sensitivity}(1-\text{Sensitivity})}{N}}$ where N is the number of people with the disease.

For specificity these limits are $\text{Specificity} \pm 1.96 \sqrt{\frac{\text{Specificity}(1-\text{Specificity})}{N}}$, where N is the number of people without the disease. These confidence limits can be obtained online at <http://vassarstats.net/clin1.html>.

From these results, the PAP test is only 70% sensitive; it misses 30% of those with prostatic cancer, but it is 94% specific because only rarely is it positive in someone who does not have prostatic cancer. The ability of the test to detect the disease, that is, its positive predictive accuracy, has a probability of 86%. (This PAP test is not the same as the Pap test used to diagnose uterine cancer. Furthermore, the PAP test has been replaced by another test, the Prostate Specific Antigen test.)

These probabilities have their exact equivalents in the probability symbols used in Bayes' theorem. The probability that defines sensitivity involves a marginal subtotal,

and so concerns a conditional probability. The probability of having a true positive test is $\frac{79}{330} = 0.239$ and the probability of having prostatic cancer is $\frac{113}{330} = 0.342$. The conditional probability of having a true positive test in someone with prostatic cancer is there-

fore $\frac{\frac{79}{330}}{\frac{113}{330}} = \frac{79}{113} = 0.699$. This ratio, $\frac{79}{113}$, is what we termed sensitivity. Therefore

Sensitivity = $P(T^+ | D^+)$, and by similar reasoning.

Specificity = $P(T^- | D^-)$,

1-Sensitivity = $P(T^- | D^+)$, and

1-Specificity = $P(T^+ | D^-)$.

Based on these identities, rewrite Bayes' theorem as:

$$P(D^+ | T^+) = \frac{P(T^+ | D^+)P(D^+)}{P(T^+)} = \frac{P(T^+ | D^+)P(D^+)}{P(T^+ | D^+)P(D^+) + P(T^+ | D^-)} = \frac{P(T^+ | D^+)P(D^+)}{P(T^+ | D^+)P(D^+) + [(1 - P(T^- | D^-)) + (1 - P(D^+))]}$$

So

$$\text{PPV} = \frac{\text{Sensitivity} \times \text{Prevalence}}{(\text{Sensitivity} \times \text{Prevalence}) + (1 - \text{Specificity})(1 - \text{Prevalence})}.$$

By similar manipulations,

$$P(D^- | T^-) \text{ or NPV} = \frac{\text{Specificity} \times (1 - \text{Prevalence})}{\text{Specificity}(1 - \text{Prevalence}) + \text{Prevalence}(1 - \text{Sensitivity})}.$$

Problem 21.1 Testing for bowel cancer by a fecal blood test gave

		Bowel cancer	
		Present	Absent
Occult blood in stool	Present	25	187
	Absent	14	1947

Calculate sensitivity, specificity, positive and negative predictive values. What do the results tell you?

We are not restricted to evaluating sensitivity and specificity one test at a time, whether done simultaneously or sequentially. For example, consider two independent tests A and B for a disease. Assume that our observed results are presented in [Table 21.4](#).

Table 21.4 Combining test results

Test	Sensitivity (%)	Specificity (%)
A	70	90
B	90	45
A or B	97.5	41
A and B	67.5	93.5

The two most common rules for combined testing are “either positive,” in which the combined test is positive if either test is positive, or “both positive” in which the combined test is positive only if both tests are positive. With the “either positive” rule, the sensitivity of the combined test will be no less than the higher sensitivity of each component test but the specificity will be no greater than the lower specificity of each of component tests. The opposite occurs under the “both positive” rule (Macaskill et al., 2002; Marshall, 1989).

Sensitivity and specificity apply to groups of patients and are seldom useful in confirming or excluding a diagnosis in a single patient. However, a highly sensitive test that is negative argues strongly against a particular diagnosis (Sensitivity Negative rule out or *Snout*), and a highly specific test that is positive strongly supports the diagnosis (Specificity Positive rule in or *Spin*) (Akobeng, 2007a).

An example of a two-stage application of Bayes’ theorem was published by Cook and Puri (2017). Lyme disease is typically confirmed using ELISA, a test that is very sensitive but not very specific. Cook and Puri followed up a positive ELISA test with a Western blot and showed how using Bayes’ theorem demonstrated an improved probability of a positive diagnosis.

Cautionary Tales

Practical Issues of Screening Tests

Whereas we can obtain values for sensitivity and specificity from a table such as Table 18.3a, this does not provide a value for prevalence; that value has to come from other observations. The importance of prevalence is illustrated by an Editorial by Redwood et al. that appeared in *Circulation* (Redwood et al., 1976). They discussed the discrepancy between two sets of studies of the relationship of abnormal ST segments on electrocardiograms taken during an exercise test to the subsequent diagnosis of severe coronary arterial disease. In one set of studies, the positive predictive value of an abnormal ST segment response was high, and in the other it was low. Both sets of studies appeared to have been done carefully. In order to explain this discrepancy, the authors showed how the predictive value of a test depended on the population under study, and in particular on the prevalence of the disease. To illustrate this concept, they assumed that a

given diagnostic test was both 95% sensitive and 95% specific for a particular disease, but that the prevalence of the disease could be either 90% or 2%. The data for each of these are set out in [Tables 21.5 and 21.6](#).

Table 21.5 Disease prevalence 90%

Subjects	Positive test	Negative test	Total
Diseased	855	45	900
Normal	5	95	100
Total	860	140	1000

$$\text{Positive predictive value} = \frac{855}{860} = 0.994 \text{ or } 99.4\%.$$

$$\text{Negative predictive value} = \frac{95}{140} = 0.679 \text{ or } 67.9\%.$$

Table 21.6 Disease prevalence 2%

Subjects	Positive test	Negative test	Total
Diseased	19	1	20
Normal	49	931	980
Total	68	932	1000

$$\text{Positive predictive value} = \frac{19}{68} = 0.279 \text{ or } 27.9\%.$$

$$\text{Negative predictive value} = \frac{931}{932} = 0.999 \text{ or } 99.9\%.$$

High positive predictive value was obtained in studies of groups of males over 50 years old with chest pain, and this is a population with a high prevalence of coronary arterial disease that would lead to a positive electrocardiographic stress test. On the other hand, low positive predictive value was found in studies of asymptomatic subjects, most of whom could be expected to be normal so that the prevalence (prior risk or pretest probability) of coronary arterial disease would be low. Because most people in the latter group would not have coronary arterial disease, even a low false positive rate of 5% produces a large absolute number (49) of false positives compared to the smaller number of true positives (19).

Redwood et al. also showed the relationship of predictive value to disease prevalence for different sensitivities and specificities. [Fig. 21.1](#) shows a modification of their figure.

The curves are numbered from 1 to 9, based initially on decreasing sensitivities and, within a given sensitivity, based on decreasing specificities. Combinations 1 and 5 are almost identical and are not displayed separately. For all the combinations of sensitivity and specificity, the higher the prevalence, the higher the positive predictive value; if

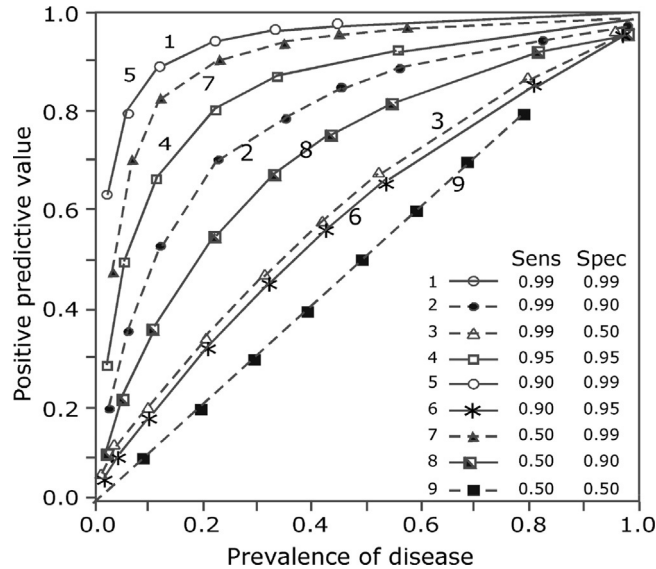


Fig. 21.1 Relationship between positive predictive value, sensitivity (Sens), specificity (Spec), and prevalence.

almost all of a population have a disease, all tests are likely to confirm the presence of that disease. However, in practice, high prevalences are seen only in selected subpopulations known from current medical information to be at high risk; for example, obese, smoking, inactive, stressed diabetic males with chest pain have a high prevalence of coronary arterial disease, and middle-aged women with an extremely high family history of breast cancer are at high risk for breast cancer. Therefore screening whole populations in which no disease has a high prevalence is likely to produce low predictive value with many false positives, thus causing distress to the subjects and substantial costs and even risks in following up the tests. [Watson and Tang \(1980\)](#) using a reasonable figure for the prevalence of prostatic cancer of 35 per 100,000 population, pointed out that if the PAP test was 70% sensitive and 94% specific, based on previously described data, the predictive value of a positive test used as a routine screening examination would be only 0.41%; that is, only 1 out of 244 people with a positive test would have prostatic cancer. They pointed out also that for predictive value to be 50% the prevalence of prostatic cancer would have to be 7894 per 100,000 people, an unrealistic figure.

The best predictive values at any prevalence come with sensitivities and specificities both of 0.99 (curve 1) but decreasing sensitivity to 0.90 makes little difference to positive predictive value (curve 5). In fact, decreasing sensitivity to 0.50 (curve 7) decreases positive predictive value only slightly at any prevalence. On the other hand, decreasing specificity to 0.90 materially lowers the positive predictive value, as can be seen by comparing curves 1 and 2, 5 and 6, and 7 and 8. A specificity of 0.50 produces low positive predictive value, no matter what the sensitivity.

Taking the values for sensitivity, specificity, and prevalence from [Tables 21.5 and 21.6](#) gives:

$$P(D^+ | T^+) = \frac{0.95 \times 0.90}{(0.95 \times 0.90) + (0.05 \times 0.10)} = \frac{0.855}{0.860} = 0.994 \text{ (high prevalence),}$$

and

$$P(D^+ | T^-) = \frac{0.95 \times 0.02}{(0.95 \times 0.02) + (0.05 \times 0.98)} = \frac{0.019}{0.068} = 0.279 \text{ (low prevalence).}$$

Similarly:

$$P(D^- | T^-) = \frac{0.95 \times 0.90}{(0.95 \times 0.90) + (0.90 \times 0.05)} = 0.95 \text{ (high prevalence),}$$

and

$$P(D^- | T^+) = \frac{0.95 \times 0.02}{(0.95 \times 0.02) + (0.02 \times 0.05)} = 0.95 \text{ (low prevalence).}$$

Therefore positive predictive value is merely another term for the conditional probability of a disease, given a positive test. Similarly, negative predictive value (the probability that a negative test means no disease) is another term for the conditional probability of no disease, given a negative test. The posterior probability that is required may be calculated by Bayes' theorem, as before. This form of the expression for predictive value also shows why, in [Fig. 21.1](#), specificity is a more important determinant of positive predictive value than is sensitivity. If sensitivity is below 100% and prevalence is low, there will be only a small number of false negative tests, so that the numerator of the expression will be small. On the other hand, even a small decrease in specificity from 100% with a large number of normal people will give a large number of false positive tests, so that the denominator of the expression will be large and the ratio will be low.

Spectrum Bias or Effect

With a dichotomous test, the sample from which the test data come needs careful assessment. The term "spectrum bias" refers to the effect that a change in the mix of patients may have in the outcomes of a test. [Ransohoff and Feinstein \(1978\)](#) and [Willis \(2008\)](#); specifically, when a diagnostic test has different sensitivities or specificities in patients with different clinical manifestations of disease, it is likely that the more severe the disease, the higher the proportion of positive tests.

As a hypothetical example, consider whether fractional flow reserve (FFR) is useful in deciding whether a coronary angiogram judged to have an intermediate degree of obstruction is the cause of the patient's problems. In their preliminary studies [Pijls et al. \(1995\)](#) established that any value of FFR < 0.75 indicated clinically important obstruction. If we repeated their study, we might find results as presented in [Table 21.7](#).

Table 21.7 Disease severity verified by thallium uptake test

	FFR < 0.75	FFR > 0.75
Severe obstruction	100TP	5FN
Less severe obstruction	15FP	100TN

Sensitivity = $100/106 = 95.2\%$; specificity = $100/115 = 87\%$; and PPV = 87% .

By angiography, some coronary arteries will be so wide that they will not be a cause of cardiac problems, and others so obstructed that there is no doubt about their harmful effect. If our sample included more of these obvious positives and negatives, we might get results as presented in [Table 21.8](#).

Table 21.8 Including many obvious positives and negatives

	FFR < 0.75	FFR > 0.75
Severe obstruction	1000TP	5FN
Less severe obstruction	15FP	1000TN

Sensitivity = $1000/1005 = 99.5\%$; specificity = $1000/1015 = 98.5\%$; and PPV = 99.5% .

Finally, if subjects who are obviously severely affected or with minimal coronary artery disease are excluded, we get the results as presented in [Table 21.9](#)

Table 21.9 Excluding those with severe or minimal coronary artery disease

	FFR < 0.75	FFR > 0.75
Severe obstruction	20TP	5FN
Less severe obstruction	15FP	10TN

Now sensitivity is $20/25 = 80\%$, specificity is $10/25 = 40\%$, and PPV is $20/35 = 57\%$.

Therefore the values for sensitivity, specificity, and PPV depend on the prevalence of the disease and how we draw our samples from the population.

As a real-life example, [Lachs et al. \(1992\)](#) studied the value of the dipstick test for urinary tract infection, and separated the patients into two groups with high or low probabilities of infection, based on signs and symptoms. They found the following results for the whole group ([Table 21.10](#)).

Table 21.10 Testing for urinary tract infection

Clinical diagnosis	Positive test	Negative
Disease present	60TP	84FN
Disease absent	12FP	210TN

The sensitivity was 83% , the specificity was 71% , positive predictive accuracy 42% , negative predictive accuracy 95% . When they separated the two groups they found ([Table 21.11](#)).

Table 21.11 Spectrum bias

Prior probability	Sensitivity (%)	Specificity (%)
High	92	42
Low	56	78

Mulherin and Miller (2002) pointed out that this effect could be minimized if the different strata producing the test were kept separate.

Verification Bias

If a diagnostic test for a given disease is positive, the patients then often have a “gold standard” test for confirmation. These “gold standard” tests are seldom given to those with negative test results. This is a reasonable decision for medical cost containment, but unfortunately it may lead to a large decrease in sensitivity and specificity (Bates et al., 1993). These authors suggest procedures for avoiding this bias.

Likelihood Ratios

A likelihood ratio is the probability of a given test result in those with disease compared to the probability of the same test result for those without the disease.

By definition, a positive likelihood ratio (LR+) is:

$$\frac{\text{True positive rate}}{\text{False positive rate}} = \frac{\frac{\text{TP}}{\text{TP} + \text{FN}}}{\frac{\text{FP}}{\text{FP} + \text{TN}}} = \frac{\text{Sensitivity}}{1 - \text{Specificity}},$$

also symbolized by $\frac{P(T^+|D^+)}{P(T^+|D^-)}$.

The definition of a negative likelihood ratio (LR-) is:

$$\frac{\text{False negative rate}}{\text{True negative rate}} = \frac{\frac{\text{FN}}{\text{TP} + \text{FN}}}{\frac{\text{TN}}{\text{FP} + \text{TN}}} = \frac{1 - \text{Sensitivity}}{\text{Specificity}},$$

also symbolized by $\frac{P(T^-|D^+)}{P(T^-|D^-)}$.

These ratios may not be independent of prevalence, (Brenner and Gefeller, 1997; Willis, 2012) but prevalence is not usually taken into account. Approximate 95% confidence limits for the likelihood ratio of a given test result are: (Simel et al., 1991)

$$\text{LR}^+ = \exp \left(\ln \frac{\text{Sensitivity}}{1 - \text{Specificity}} \pm 1.96 \times \sqrt{\frac{1 - \text{Sensitivity}}{A} + \frac{\text{Specificity}}{B}} \right) \quad \text{and}$$

$$\text{LR}^- = \exp \left(\ln \frac{1 - \text{Sensitivity}}{\text{Specificity}} \pm 1.96 \times \sqrt{\frac{\text{Sensitivity}}{C} + \frac{1 - \text{Specificity}}{D}} \right), \text{ where } A, B,$$

C, D are the cells in the fourfold table.

The ratio can be refined by dividing LR^+ by LR^- to give a diagnostic odds ratio or DOR (Sackett et al., 1985; Lijmer et al., 1999; Glas et al., 2003). This is identical to the odds ratio (cross-product ratio) obtained from a fourfold table. DOR can be calculated in several ways

$$DOR = \frac{LH^+}{LH^-} = \frac{TP/FP}{FN/TN} = \frac{\text{Sensitivity}/(1 - \text{Sensitivity})}{(1 - \text{Specificity})/\text{Specificity}} = \frac{PPV/(1 - PPV)}{(1 - NPV)/NPV}.$$

DOR has the merit of using all the data that help to differentiate true positives from true negatives. The higher the ratio the better the discrimination. The standard error of DOR is

$$SE \log DOR = \sqrt{\frac{1}{TP} + \frac{1}{TN} + \frac{1}{FP} + \frac{1}{FN}}.$$

The approximate 95% confidence interval is obtained from $\ln DOR \pm 1.96 \times SE \log DOR$, and then exponentiating the resultant log confidence interval.

Online calculators include <http://www.hutchon.net/Bayes.htm>, <https://easycalculation.com/statistics/Bayesian-analysis.php>, <http://statpages.org/ctab2x2.html>, <http://araw.mede.uic.edu/cgi-bin/testcalc.pl>, and <http://vassarstats.net/clin2.html>.

Problem 21.2 From the data of Problem 21.1 calculate the positive and negative likelihood ratios and the diagnostic odds ratio.

What does LR^+ indicate? $TP/(TP + FN)$ is the proportion of people with a disease who have a positive test, and $FP/(FP + TN)$ is the proportion of people without the disease who have a positive test. The ratio therefore indicates the likelihood that a person with a positive test has the disease. From the data in Table 21.4, $TP/(TP + FN)$ is $79/113 = 0.6991$, and $FP/(FP + TN)$ is $13/217 = 0.05991$. The ratio of these two is 11.7, showing that a person with a positive test has 11.7 times the risk of having the disease than not having it.

A likelihood ratio >1 indicates an association between a positive test and the disease, whereas a ratio <1 indicates that a test result is not associated with the disease. Ratios over 10 and under 0.1 are strong arguments for or against the diagnosis (Deeks and Altman, 2004). A useful simplification was reported by McGee (2002). Likelihood ratios of 2, 5, and 10 increase the probability of the disease by approximately 15%, 30%, or 45%, respectively, and ratios of $1/2$, $1/5$, and $1/10$ decrease the probability of the disease by approximately 15%, 30%, and 45%, respectively.

Posttest odds can be shown to equal Pretest odds \times LR. The pretest odds are the ratio of the probability of having the disease divided by the probability of not having the disease, and are

$$\frac{P(D^+)}{P(1 - D^+)}.$$

The posttest probability is defined by

$$\frac{\text{Posttest odds}}{1 + \text{Posttest odds}}.$$

The posttest probability can be derived by rearranging the original equation, but to avoid cumbersome calculations use the nomogram developed by Fagan (1975). This nomogram and some of the associated calculations are provided online by <http://araw.mede.uic.edu/cgi-bin/testcalc.pl?DT=0&Dt=0&dT=0&dt=0&2x2=Compute> and <http://www.pmean.org/definitions/fagan.htm>. Results without the nomogram are provided online by <http://www.medcalc.com/bayes.html>. These nomograms can be used with the primary sensitivity and specificity data by referring to articles by Møller-Petersen (1985) and Caraguel and Vanderstichel (2013). The latter authors devised an elegant extension of the Fagan nomogram that allows the physician to start with known sensitivity, specificity, and pretest probability (Fig. 21.2). There is an excellent accompanying website at http://www.adelaide.edu.au/vetsci/research/pub_pop/2step-nomogram/, and a very useful free application for the iPhone, iPad, and iPod touch obtainable from iTunes as DocNomo app at <https://itunes.apple.com/us/app/docnomo/id901279945?mt=8>.

Because a test is usually either positive or negative, their combined probability is 1. In both the sensitivity and the specificity scales, a value of 0.4 for test+ is the same as the value 0.6 for test negative.

To use this nomogram, if the test is positive and has a sensitivity of 0.8 and a specificity of ~ 0.06 , the thin solid line (I) indicates a likelihood ratio (LR) of about 20. If we know the pretest probability (prevalence) is 0.02, then the thick solid line (II) passing from 0.02 through the LR of 20 intercepts the posttest probability at about 0.18, suggesting that the patient does have the disease. (In the total population, most pretest probabilities are low. In selected populations, e.g., middle-aged obese diabetic male smokers with sudden onset chest pain, the pretest probability of coronary artery disease is high.) If the test is negative with a known sensitivity of 0.28 and a specificity of 0.9 (dashed line I), it crosses the LR scale at about 0.22. Given the same prevalence of 0.02, extending this through the LR of 0.22 intersects the posttest probability scale at about 0.005 (thick dashed line II), indicating that the disease is very unlikely to be present.

Problem 21.3 From the data of Problem 21.1, calculate pre- and posttest odds.

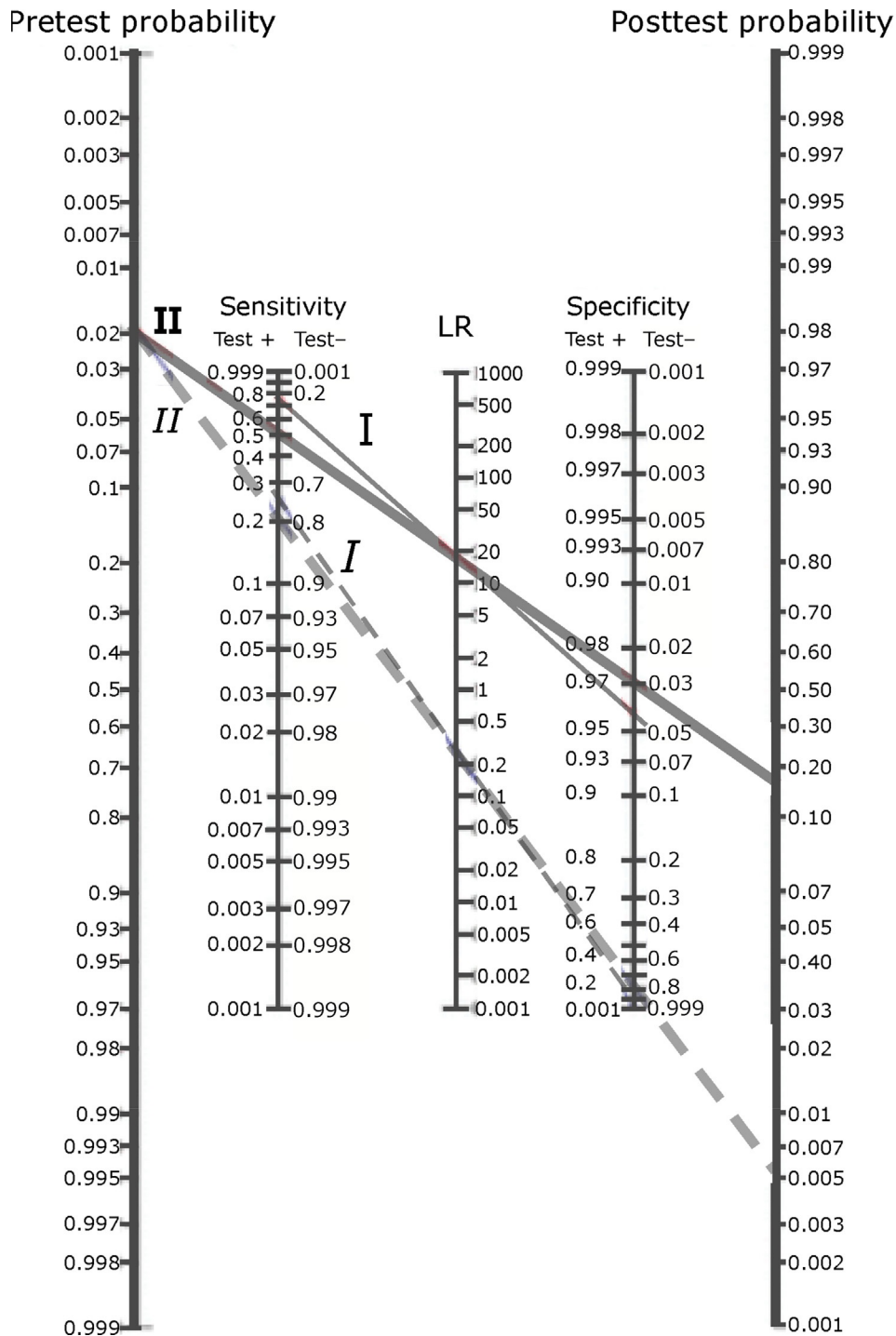


Fig. 21.2 Nomogram for sensitivity, specificity, and likelihood ratios. (Redrawn with permission from Caraguel, C.G., Vanderstichel, R., 2013. The two-step Fagan's nomogram: ad hoc interpretation of a diagnostic test result without calculation. *Evid. Based Med.* 18, 125–128 by permission of the publisher and the authors.)

Problem 21.4 For the fecal blood data, calculate the posttest probability, and comment on the results.

A discussion of likelihood ratios and an online calculator are provided by [Tape, n.d.-a, n.d.-b](#). [Hayden and Brown \(1999\)](#) have emphasized the value of using likelihood ratios rather than sensitivity and specificity because likelihood ratios of necessity lead to a Bayesian approach that takes prior probabilities into consideration. On the other hand, posttest probabilities cannot be calculated if no pretest probability is available.

Likelihood ratios can be applied to a complete data set or to subgroups within it. [Table 21.12](#) presents the sensitivity, specificity, and likelihood ratios (LR^+) for different threshold concentrations of procalcitonin (PCT) in diagnosing septic shock ([Hatherill et al., 1999](#)).

Table 21.12 Effect of using different PCT concentrations to differentiate patients with and without septic shock of bacterial origin

PCT concentration (ng/mL)	Sensitivity	Specificity	LR^+
>2	100	62	2.63
>5	99	78	4.50
>10	88	84	5.50
>20	83	92	10.37

As the critical concentration is increased, sensitivity decreases, but specificity and likelihood ratio increase. If all the data were lumped to regard all patients with an elevated PCT concentration, the averaged LR^+ would be lower than the maximum.

Discussing sensitivity and specificity assumes that when we calculate $P(T^+ | D^+)$ we are certain about the presence of the disease. This may not always be true, and we may end up comparing a new test T^+ with a standard test (“ D^+ ”) that may not always be positive when the disease is present. If, for example, the false positive rate of the standard test is assumed to be zero, but is not, then the false negative rate of the new test is overestimated. Also, if the standard test is assumed to have no false negatives, but this is wrong, then the false positive rate of the new test will be overestimated. Improved results may be obtained by using maximum likelihood methods.

Cutting Points

The data presented in the decision matrix in 21.4a are dichotomous; a test is assumed to be either positive or negative, with no overlap. However, this may not occur with measurements of continuous variables. What is more likely is that the higher (or lower) some test measurement is, the more likely is it that the person has the disease.

Thyroid hormone is essential for development of the nervous system, and children born with thyroid hormone deficiency develop mental retardation unless replacement therapy is started soon after birth. Therefore in many countries it is mandatory to test for thyroid deficiency at birth. A drop of blood from the umbilical cord is placed on filter paper, dried, and then sent to a laboratory for measurement of thyroid hormones. [Morissette and Dussault \(1979\)](#) found that in a large population of infants without hypothyroidism, the distribution of thyroxine (T4) was log normal with a mean of 1.73 ng/spot and a standard deviation of 1.35 ng/spot. In a group of 72 hypothyroid infants, the mean T4 was 0.52 ng/spot with a standard deviation of 0.188 ng/spot, and the distribution was approximately normal (Fig. 21.3).

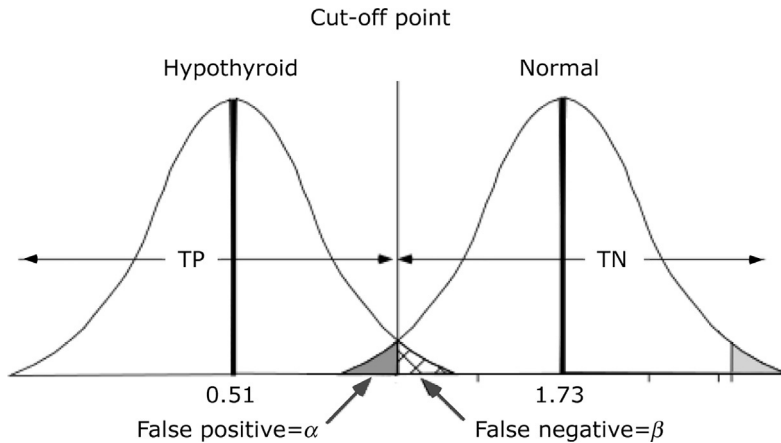


Fig. 21.3 Congenital hypothyroid data.

In this figure, the normal and hypothyroid distributions are drawn overlapping; the means are put in, but the curves are not drawn to scale. A vertical line is drawn where the two lines cross. The shaded area to the left of this vertical line is the proportion α of the normal population ($=TN$) that lies below the cutoff point and would be regarded as abnormal; this shaded area therefore represents false positives. The cross-hatched area to the right of the vertical line indicates the proportion β of the hypothyroid distribution ($=TP$) that lies above the cutoff point and would be regarded as normal; this shaded area therefore represents false negatives.

What happens if the cutting point is moved to the left ([Fig. 21.4](#))?

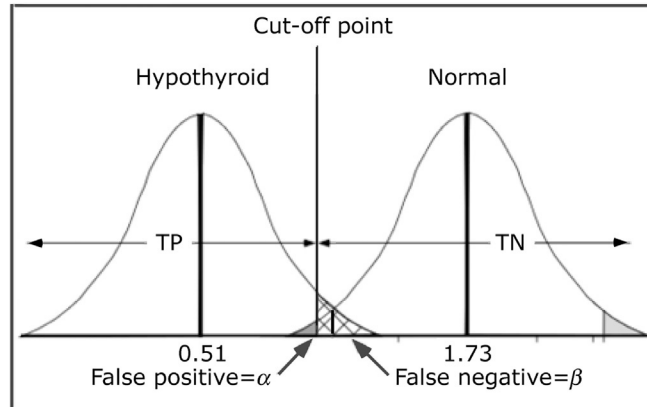


Fig. 21.4 New cutting point.

The proportion of false positives is decreased, but the proportion of false negatives is increased. Conversely, if the cutoff point was moved to the right, the proportion of false negatives is decreased, but the proportion of false positives is increased. From the values for means and standard deviations, Morissette and Dussault computed the proportions of false negatives and false positives for several cutting points and also computed the probability of missing an infant with hypothyroidism by multiplying β , the proportion of false negatives, by the prevalence of the disease in the population. Thus for β equal to 0.016 and prevalence equal to $1/5000$, the probability of missing a patient with disease is $0.016 \times 1/5000 = 0.0000032 = 3.2$ per million infants screened. Changing the cutting point altered these values. Moving the cutting point to the left (more standard deviation units below the normal mean, lower T4 concentrations) decreased the false positives almost 10-fold from the highest to the lowest cutting points, but at the same time increased the false negatives and the chances of missing a hypothyroid infant about 2400-fold. Where the cutting point should be placed is debatable and depends to some extent on the penalty (in human suffering or cost) to be paid on failing to diagnose a disease or in diagnosing too many normal subjects as having disease. Some consequences of varying the cutting point have been discussed (Cheetham, 2011; Krude and Blankenstein, 2011).

One way of examining the relationship of the cutoff point to sensitivity and specificity is to plot both of these against various cutoff points. To illustrate this, Fig. 21.5 (left panel) shows the data in which scores are obtained from the EMPP (Early Motor Pattern Profile) at 6 months of age in an attempt to predict which children will have cerebral palsy (Morgan and Aldag, 1996).

The left-hand panel shows that the sensitivity and specificity curves cross near a cutoff point of 7, where both sensitivity and specificity are high.

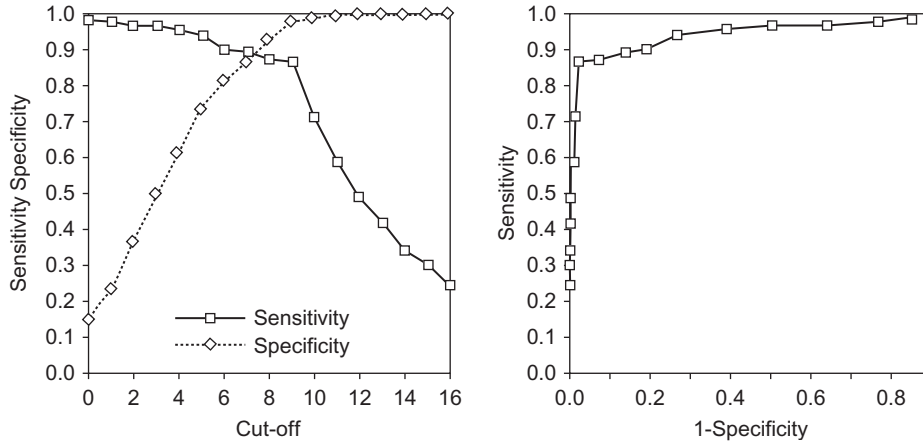


Fig. 21.5 Sensitivity, specificity, and cutoff points. (Based on data in Morgan, A.M., Aldag, J.C., 1996. Early identification of cerebral palsy using a profile of abnormal motor patterns. *Pediatrics* 98, 692–697.)

ROC Curves

One approach to the problems imposed by variations in cutting points is to construct receiver operating characteristic (ROC) curves. A graph is constructed by plotting the false positive fraction (1–specificity) on the horizontal axis against the true positive fraction (sensitivity) on the vertical axis (Fig. 21.6A and B) at each cutting point, each cutting point giving one value for true positive and false positive. To produce such a curve, the sensitivity and specificity are calculated for each of several different cutting points, ideally a large number of cutting points between 0 and 1 with small equal bin sizes. The curve must pass from the lower left-hand corner of the graph if all the tests are negative up to the upper

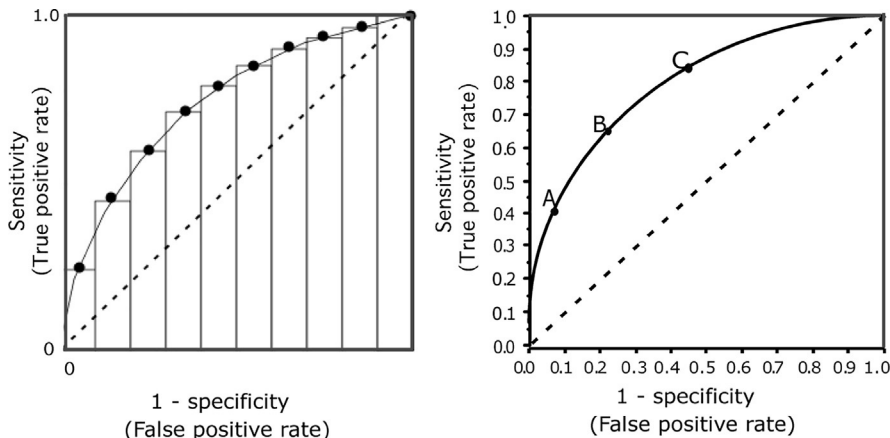


Fig. 21.6 Ideal ROC curves. Sometimes only a few cutting points are used (right panel).

right-hand corner if all the tests are positive. The diagonal dashed line indicates that a positive decision is no more likely if the test is positive than if it is negative and implies that the test is worthless. If the test were a perfect marker for the disease, then the curve would run up the left-hand boundary from 0 to 1, and then run horizontally from the upper left to right hand corners. In reality, tests give a curve above the diagonal line. Any point on the line indicates the ratio of positive to negative tests in patients with the disease (Fig. 21.6).

In the right panel, point A indicates a test with relatively low sensitivity but with very few false positives; many with the disease will be missed but few normal subjects will be falsely diagnosed as having the disease. Point C, on the other hand, indicates that most of the people with the disease will be detected, but at the cost of a large number of false positive tests. Point B gives intermediate values.

The data plotted are from [Morissette and Dussault \(1979\)](#), produces Fig. 21.7.

The false positive increase as sensitivity is increased to the highest level is well shown but changing the cutting point to increase sensitivity from 0.928 to 0.984 incurs little increase in the false positive rate.

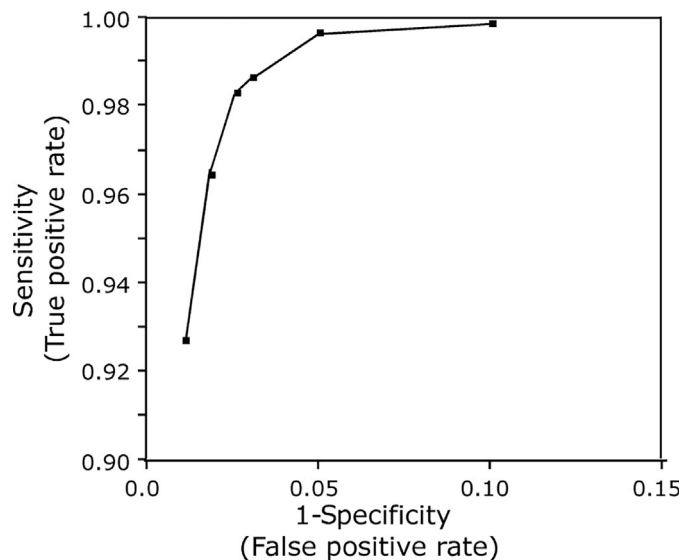


Fig. 21.7 ROC curve for hypothyroidism; only part of the curve is plotted.

Plotting ROC curves for the EMPP data shown in the left panel of Fig. 21.5 gives the curve shown in the right-hand panel of that Figure. As the cutting point is raised from 0 to 7, there is a great increase in sensitivity with very little increase in false positives. At higher cutoff points, there is little further gain in sensitivity but a big increase in false positives.

ROC curves may be constructed to compare the relative merits of different tests in several ways. A single number that summarizes the curve is the area under the curve (AUC).

This area represents the probability that a random person with the disease has a higher measured value than a random person without the disease (Akobeng, 2007b). For the uninformative diagonal line, the probability is 0.5 and thus is of no diagnostic help.

Some authors have recommended a simple scale to indicate the value of the area: (Swets, 1988; Fischer et al., 2003; Tape, n.d.-a, n.d.-b) (Adapted from Tape, n.d.-a, n.d.-b).

Excellent: >90%; Good: 80%–90%; Fair: 70%–80%; Poor: 60%–70%; Bad: 50%–60% but simple inspection is probably adequate for deciding if discrimination is good. The area, however, is only a point estimate, and determining 95% confidence limits as $1.96SE$ is useful. A formula for determining these limits was published by Hanley and McNeil (1982)

$$SE = \sqrt{\frac{A(1-A) + (N_a - 1)(Q_1 - A^2) + (N_n - 1)(Q_2 - A^2)}{N_a N_n}}$$

where A is the area under the curve, N_a and N_n are the number of abnormal and normal results, respectively, $Q_1 = A/(2-A)$, and $Q_2 = 2A^2/(1+A)$. The online program <http://www.anaesthetist.com/mnm/stats/roc/Findex.htm> gives an excellent description of these curves.

If two curves are to be compared, for example, two different tests for the same disease, observation will tell if the two curves are similar or quite different. In a study done on prediction of septic shock in children in which procalcitonin, CRP, and white cell count were compared (Fig. 21.8), based on the study by Hatherill et al. (1999) and used in an explanatory article by Akobeng (2007b).

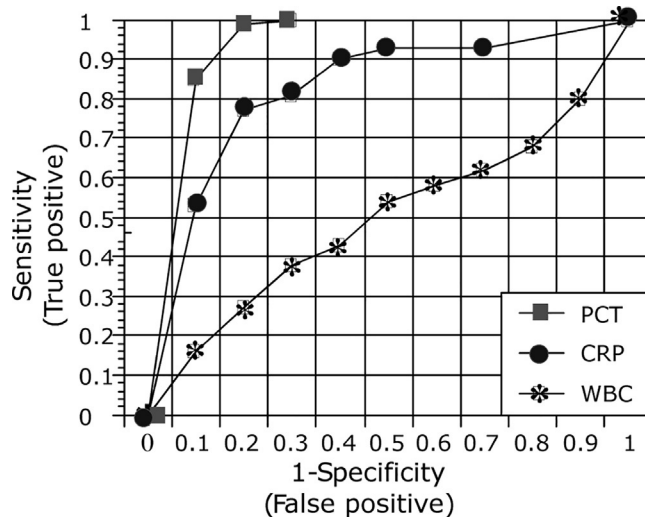


Fig. 21.8 ROC curves comparing procalcitonin (PCT), C-reactive protein (CRP), and white blood cell count (WBC) from the same patients. The white cell count was nondiscriminatory, because at every value there were about equal numbers of true and false positives. The other two curves, however, did show discrimination, with procalcitonin better than CRP.

Single curves can be fitted to a theoretical distribution online at <http://vassarstats.net/index.html>, (see Clinical Research Calculators) and this will also calculate the area under the curve as well as comparing two area under the curve. Unfitted curves cannot be drawn, and curves cannot be superimposed. On the other hand, XY plots can be drawn by several programs (Chapter 27). A more complex program that plots the actual curves and can superimpose two or more is found at <http://www.rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html>.

The power of the test should be determined, preferably in advance. Appropriate numbers can be determined from Table III in the publication by Hanley and McNeil (1982). An online test is available at <http://vassarstats.net/index.html> (see Clinical Research Calculators) for selected data values.

Was the difference between these two curves due to chance? To assess this, perform a z -test

$$z = \frac{A_1 - A_2}{SE_{A_1 - A_2}}$$

where $SE_{A_1 - A_2} = \sqrt{SE^2 A_1 + SE^2 A_2 - 2rSE(A_1)SE(A_2)}$ (Hanley and McNeil, 1982, 1983; Hopley and van Schalkwyk, 2007).

Here r is the correlation coefficient averaged for positive tests and negative tests separately. If the two tests to be compared are done in different groups of patients, the component of the formula involving the correlation coefficient drops out.

The ROC plot has several advantages. It provides an easy visual compilation of all the data, does not depend on the parameters of any distribution, and is independent of the prevalence of the disease. On the other hand, the graph per se does not indicate the patient numbers from which the data were derived, does not show the cutting points used, may be unduly influenced by the initial and final parts of the curve that contain little information, and the decision thresholds used in the calculations are not apparent. Furthermore, although the plots give useful information about individual tests, they do not tell us whether better discrimination could be obtained by combining the results from two or more tests (something that can be done with Bayes' theorem applied sequentially). Finally, the curve does not provide an answer about where the best cutoff point is, because the answer to that question is multifactorial. One recommended method is to examine where the separate curves for sensitivity or specificity versus cutting points cross and set confidence limits around this point (Greiner, 1995, 1996; Greiner et al., 1995, 2000). Another is to calculate the Youden index J , the maximum value of (Sensitivity + Specificity - 1) (Youden, 1950; Fluss et al., 2005).

If it were just a matter of financial cost it would be possible to find an optimal cutoff point by these or similar methods, but some of the human costs cannot be so easily enumerated. It may be more important not to miss early cancer than many other diseases. There is also a difference in cost to a patient versus cost to a community. For example,

now that few people are immunized against smallpox, the failure to diagnose a single patient with smallpox would be catastrophic to the community and would outweigh the costs of dealing with false positives. The subject is discussed by [Zweig and Campbell \(1993\)](#). In this context, it is sometimes appropriate to focus on a specific part of the ROC curve; for example, if two curves cross, or if only cutting points with sensitivities over 90% are essential. Special tests are available in these circumstances.

Another sensible use of ROC curves was provided by [Bhatt et al. \(2009\)](#). They investigated the problem of how to diagnose acute appendicitis in children so as to minimize both false negatives (missed appendicitis) and false positives (unnecessary surgery). They developed a pediatric appendicitis score (PAS) based on clinical and blood findings (no imaging studies) and applied it to 246 patients aged 4–18 years referred for possible acute appendicitis. The score, based on clinical findings and a white blood cell count, ranged from 1 to 10. By relating the scores to what was found at surgery or, if no surgery was done, to the subsequent course, they were able to calculate specificity and sensitivity for each score, and created an ROC curve ([Fig. 21.9](#)).

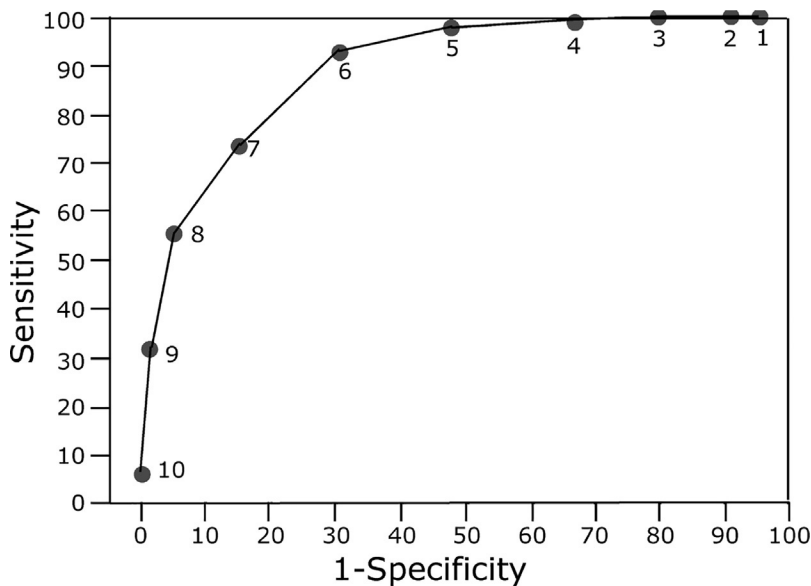


Fig. 21.9 ROC curve based on pediatric appendicitis score (numbers under points on the curve).

No single score was ideal for discriminating between abdominal pain that was or was not due to appendicitis. What they did, however, was to use two threshold scores: a score

of 1–4 allowed them to send home patients without surgery with only 2.4% missed appendicitis, and a score of 8–10 allowed them to operate with a low false positive rate of 8.8%. Patients with scores of 5–7 had an indeterminate diagnosis and required further investigation.

Some Comments on Screening Tests

The aims of a screening test are to detect asymptomatic disease so that it may be treated early and lead to decreased morbidity and mortality.

A screening test must fulfill certain requirements:

1. It should have high sensitivity, so that as many people as possible with the disease can be detected.
2. It should have high specificity, so that the false positive rate is kept low. However, even a small false positive rate when the population incidence is low yields a very large number of people to have follow-up tests.
3. There should be a treatment for the diagnosed disease. Without an effective treatment, the results are useful at best for counseling or research.
4. The results should be cost effective. Although we cannot put a cost-benefit on saving one life, if the costs of the test are so high that they reduce the total pool of money available for other aspects of medical care, then society does not benefit. Certain newborn screening tests for genetic disorders (phenylketonuria and congenital hypothyroidism) have been shown to be cost effective. For many other neonatal screening tests, the cost-benefit ratio is unknown.
5. Screening for cancer in later life has had a checkered course, and this is nowhere better shown than in screening for breast cancer by mammography, ovarian cancer by transvaginal ultrasonography and Cancer antigen-125, or prostatic cancer by blood testing for Prostate Specific Antigen (PSA). These studies have not shown a reduction in deaths. Furthermore, the follow-up testing needed to eliminate false positives may involve potentially harmful radiation, surgery, or invasive biopsies, all of which have substantial cost and may have short- or long-term complications. There is evidence ([Hinkley, 1969](#)) that detecting early lesions in the breast or prostate may not reduce the number of late advanced cancers, so that the screening tests may be detecting small, relatively innocuous lesions. Ductal carcinoma in situ may occur in up to 40% of adult women, and most elderly males have histological evidence of prostatic cancer; both of these changes may be so slowly progressive that they do not account for premature death from cancer. If this concept is true, then these screening tests are not detecting the important lesions and are merely causing anxiety in both patient and doctor ([Wainer, 2011](#)). An excellent personal discussion of the advantages and disadvantages of screening mammography for breast cancer recently appeared in the New York Times magazine ([Orenstein, 2014](#)). Some issues about screening tests have been

highlighted by a recent publication (Wegwarth et al., 2012) in which a survey of primary care physicians showed that relatively few of them could determine if the evidence supporting a hypothetical screening test was relevant or irrelevant. One of the main errors was in using evidence based on survival statistics rather than cancer mortality. Survival statistics, if not obtained by a randomized trial, contain two potentially serious errors: lead-time bias and overdiagnosis bias. The lead-time error is displayed in Fig. 21.10, redrawn from Welch et al. (2007).

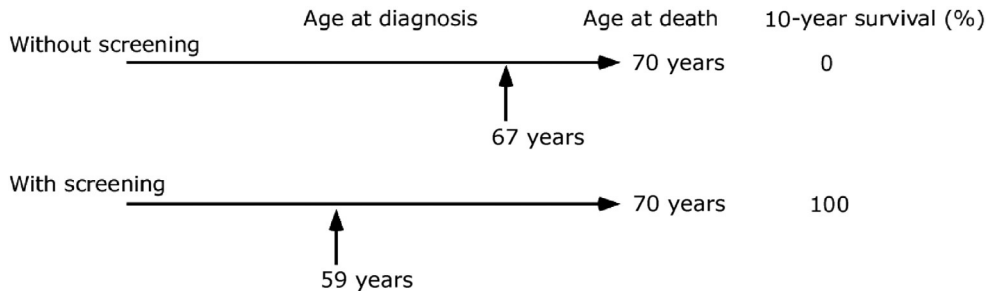


Fig. 21.10 Illustration of lead-time bias. Although cancer death has not been prevented, the 10-year survival estimate shows a great improvement.

The overdiagnosis bias is made when a screening test detects small and often innocuous lesions. As a result, the percent survival will be much increased, even if the number dying from the disease has not changed.

REFERENCES

- Akobeng, A.K., 2007a. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta Paediatr.* 96, 338–341.
- Akobeng, A.K., 2007b. Understanding diagnostic tests 3: receiver operating characteristic curves. *Acta Paediatr.* 96, 644–647.
- Bates, A.S., Margolis, P.A., Evans, A.T., 1993. Verification bias in pediatric studies evaluating diagnostic tests. *J. Pediatr.* 122, 585–590.
- Bhatt, M., Joseph, L., Ducharme, F.M., Dougherty, G., McGillivray, D., 2009. Prospective validation of the pediatric appendicitis score in a Canadian pediatric emergency department. *Acad. Emerg. Med.* 16, 591–596.
- Brenner, H., Gefeller, O., 1997. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat. Med.* 16, 981–991.
- Caraguel, C.G., Vanderstichel, R., 2013. The two-step Fagan's nomogram: ad hoc interpretation of a diagnostic test result without calculation. *Evid Based Med* 18, 125–128.
- Cheetham, T., 2011. Congenital hypothyroidism: managing the hinterland between fact and theory. *Arch. Dis. Child.* 96, 205.
- Cook, M.J., Puri, B.K., 2017. Application of Bayesian decision-making to laboratory testing for Lyme disease and comparison with testing for HIV. *Int J Gen Med* 10, 113–123.
- Deeks, J.J., Altman, D.G., 2004. Diagnostic tests 4: likelihood ratios. *BMJ* 329, 168–169.
- Fagan, T.J., 1975. Letter: nomogram for Bayes theorem. *N. Engl. J. Med.* 293, 257.

- Fischer, J.E., Bachmann, L.M., Jaeschke, R., 2003. A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. *Intensive Care Med.* 29, 1043–1051.
- Fluss, R., Faraggi, D., Reiser, B., 2005. Estimation of the Youden index and its associated cutoff point. *Biom. J.* 47, 458–472.
- Glas, A.S., Lijmer, J.G., Prins, M.H., Bonsel, G.J., Bossuyt, P.M., 2003. The diagnostic odds ratio: a single indicator of test performance. *J. Clin. Epidemiol.* 56, 1129–1135.
- Goldacre, B., 2006. Prosecuting and defending by numbers. *The Guardian*, England. <http://www.guardian.co.uk/science/2006/oct/28/uknews1>.
- Greiner, M., 1995. Two-graph receiver operating characteristic (TG-ROC): a Microsoft-EXCEL template for the selection of cut-off values in diagnostic tests. *J. Immunol. Methods* 185, 145–146.
- Greiner, M., 1996. Two-graph receiver operating characteristic (TG-ROC): update version supports optimisation of cut-off values that minimise overall misclassification costs. *J. Immunol. Methods* 191, 93–94.
- Greiner, M., Sohr, D., Gobel, P., 1995. A modified ROC analysis for the selection of cut-off values and the definition of intermediate results of serodiagnostic tests. *J. Immunol. Methods* 185, 123–132.
- Greiner, M., Pfeiffer, D., Smith, R.D., 2000. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Prev Vet Med* 45, 23–41.
- Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36.
- Hanley, J.A., McNeil, B.J., 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 148, 839–843.
- Hatherill, M., Tibby, S.M., Sykes, K., Turner, C., Murdoch, I.A., 1999. Diagnostic markers of infection: comparison of procalcitonin with C reactive protein and leucocyte count. *Arch. Dis. Child.* 81, 417–421.
- Hayden, S.R., Brown, M.D., 1999. Likelihood ratio: a powerful tool for incorporating the results of a diagnostic test into clinical decisionmaking. *Ann Emerg Med* 33, 575–580.
- Hill, R., 2004. Multiple sudden infant deaths—coincidence or beyond coincidence? *Paediatr Perinatal Epidemiol* 18, 320–326.
- Hinkley, D.V., 1969. Inference about the intersection in two-phase regression. *Biometrika* 56, 495–504.
- Hopley, L., Van Schalkwyk, J., 2007. The magnificent ROC. Available: <http://www.anaesthetist.com/mnm/stats/roc/Findex.htm>.
- Korb, K.B., 2012. Sally Clark is Wrongly Convicted of Murdering Her Children. <http://bayesian-intelligence.com/bwb/2012-03/sally-clark-is-wrongly-convicted-of-murdering-her-children/>.
- Krude, H., Blankenstein, O., 2011. Treating patients not numbers: the benefit and burden of lowering TSH newborn screening cut-offs. *Arch Dis Child* 96, 121–122 [Bates, 1993 #872].
- Lachs, M.S., Nachamkin, I., Edelstein, P.H., Goldman, J., Feinstein, A.R., Schwartz, J.S., 1992. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann. Intern. Med.* 117, 135–140.
- Lijmer, J.G., Mol, B.W., Heisterkamp, S., Bonsel, G.J., Prins, M.H., Van Der Meulen, J.H., Bossuyt, P.M., 1999. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 282, 1061–1066.
- MacAskill, P., Walter, S.D., Irwig, L., Franco, E.L., 2002. Assessing the gain in diagnostic performance when combining two diagnostic tests. *Stat. Med.* 21, 2527–2546.
- Marshall, R.J., 1989. The predictive value of simple rules for combining two diagnostic tests. *Biometrics* 45, 1213–1222.
- McGee, S., 2002. Simplifying likelihood ratios. *J. Gen. Intern. Med.* 17, 646–649.
- Moller-Petersen, J., 1985. Nomogram for predictive values and efficiencies of tests. *Lancet* 1, 348.
- Morgan, A.M., Aldag, J.C., 1996. Early identification of cerebral palsy using a profile of abnormal motor patterns. *Pediatrics* 98, 692–697.
- Morissette, J., Dussault, J.H., 1979. Commentary: the cut-off point for TSH measurement or recalls in a screening program for congenital hypothyroidism using primary T4 screening. *J. Pediatr.* 95, 404–406.
- Mulherin, S.A., Miller, W.C., 2002. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Ann. Intern. Med.* 137, 598–602.
- Orenstein, P., 2014. New York Times Magazine. <https://www.nytimes.com/2013/04/28/magazine/our-feel-good-war-on-breast-cancer.html?pagewanted=all>.

- Pijls, N.H., Van Gelder, B., Van Der Voort, P., Peels, K., Bracke, F.A., Bonnier, H.J., El Gamal, M.I., 1995. Fractional flow reserve. A useful index to evaluate the influence of an epicardial coronary stenosis on myocardial blood flow. *Circulation* 92, 3183–3193.
- Ransohoff, D.F., Feinstein, A.R., 1978. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N. Engl. J. Med.* 299, 926–930.
- Redwood, D.R., Borer, J.S., Epstein, S.E., 1976. Whither the ST segment during exercise. *Circulation* 54, 703–706.
- Rubin, L.G., 1992. Occult bacteremia. *Cur Opin Pediatr* 4, 65–69.
- Sackett, D.L., Haynes, R.B., Tugwell, P., 1985. *Clinical Epidemiology: A Basic Science for Clinical Medicine*. Toronto Little, Brown, Boston.
- Simel, D.L., Samsa, G.P., Matchar, D.B., 1991. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J. Clin. Epidemiol.* 44, 763–770.
- Swets, J.A., 1988. Measuring the accuracy of diagnostic systems. *Science* 240, 1285–1293.
- Tape, G.T., n.d.-a. Interpreting Diagnostic Tests. Available: <http://gim.unmc.edu/dxtests/>.
- Tape, G.T., n.d.-b. Introduction to ROC Curves. Available: <http://gim.unmc.edu/dxtests/ROC1.htm>.
- Thompson, W.C., Schumann, E.L., 1987. Interpretation of statistical evidence in criminal trials. The Prosecutor's fallacy and the defense Attorney's fallacy. *Law Hum. Behav.* 11, 167–187.
- Wainer, H., 2011. How should we screen for breast cancer? Using evidence to make medical decisions. *Significance* 8, 28–30.
- Watson, R.A., Tang, D.B., 1980. The predictive value of prostatic acid phosphatase as a screening test for prostatic cancer. *N. Engl. J. Med.* 303, 497–499.
- Wegwarth, O., Schwartz, L.M., Woloshin, S., Gaissmaier, W., Gigerenzer, G., 2012. Do physicians understand cancer screening statistics? A national survey of primary care physicians in the United States. *Ann. Intern. Med.* 156, 340–349.
- Welch, H.G., Woloshin, S., Schwartz, L.M., Gordis, L., Gotzsche, P.C., Harris, R., Kramer, B.S., Ransohoff, D.F., 2007. Overstating the evidence for lung cancer screening: the international early lung Cancer action program (I-ELCAP) study. *Arch. Intern. Med.* 167, 2289–2295.
- Wikipedia, 2011. Prosecutor's Fallacy. Available: http://en.wikipedia.org/wiki/Prosecutor's_fallacy.
- Willis, B.H., 2008. Spectrum bias—why clinicians need to be cautious when applying diagnostic test studies. *Fam. Pract.* 25, 390–396.
- Willis, B.H., 2012. Empirical evidence that disease prevalence may affect the performance of diagnostic tests with an implicit threshold: a cross-sectional study. *BMJ Open*. 2. e0076 (6 pages).
- Youden, W.J., 1950. Index for rating diagnostic tests. *Cancer* 3, 32–35.
- Zweig, M.H., Campbell, G., 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* 39, 561–577.

SECTION VI

Comparing Means

CHAPTER 22

Comparison of Two Groups: *t*-Tests and Nonparametric Tests

BASIC CONCEPTS

Introduction

One of the most frequent studies is whether the means of two groups are different enough that they are unlikely to have come from the same population. Did the blood sugar decrease more with drug A than drug B? To evaluate this question with continuous data we usually use the *t*-test for comparing the means of two groups. It is the prototype for almost all other statistical inferences and brings into play most of the considerations involved in making these inferences. The *t*-test is a version of the more general analysis of variance restricted to two groups.

There are two types of *t*-tests. In one, data are collected in pairs of subjects and the set of differences between each pair is tested to determine if the variation is consistent with the null hypothesis that the set of differences comes from a population with a mean difference of zero, that is, the two groups are not different. This is termed the paired *t*-test. Its counterpart is where two different groups are compared and the question asked is if the two means could have come from the same or different populations.

Paired *t*-Test

Paired comparisons are frequent. For example, a blood sample is divided in half and placed in two tubes: one is a control and the other has some chemical added, with the question being whether the chemical causes a change in the concentrations of the substance of interest. Blood from several subjects is examined, each time in pairs. Another type of paired experiment might have one pair of rat littermates from several different litters, with one of each pair being given a standard diet and the other member of the pair being given the same diet with a food additive to determine if the additive affects growth. A third experiment might be to study a group of hypertensive people before and after a given dose of a drug to determine if it lowers blood pressure.

In order to determine if the experimental group differs from the control group or if the drug lowers pressure, examine the *differences* between each pair of data values. For example, a study of the biological value of raw (*R*) versus roasted (*P*) peanuts as judged by the weight gain of rat littermates (in grams) produced the data of Table 22.1 (Mitchell, Burroughs and Beadles, 1936).

Table 22.1 Weight gain of paired littermates fed either raw or roasted peanuts in their diet

Raw peanuts <i>R</i>	Roasted peanuts <i>P</i>	Difference <i>D</i>
61	55	6
60	54	6
56	47	9
63	59	4
56	51	5
63	61	2
59	57	2
56	54	2
44	63	−19
61	58	3
		$\Sigma X_i = 20 \overline{X}_D = 2$

For each pair the difference *D* in weight gain is calculated, giving the data in column 3. Now ask: “If in the population there is no average difference between the weight gains on the two diets, how likely is it that in a sample there would be a difference of as much as 2 grams?” That is, $H_0: \mu_D = 0$.

If it is very likely, then we would not consider that roasting affected the nutritional value of peanuts, but if it is an unlikely difference, then we might want to consider that roasting affected their nutritional value. Assess the probability of the null hypothesis by determining how many standard deviations from the mean that difference represents. If the difference is many standard deviations from the mean, then there is reason to reject the null hypothesis. To do the required calculations, calculate the mean and standard deviation of the *differences* $\overline{X}_D = 2$, $\sum (X_i - \overline{X})^2 = 536$. Therefore $s^2 = 59.56$, $s = 7.72$ and $s_{\overline{X}} = 7.72 / \sqrt{10} = 2.44$. Then relate the difference to the standard error to determine the probability of observing that difference if the true population difference is zero.

$$t = \frac{2 - 0}{2.44} = 0.82, P = 0.43. \text{ This does not lead to rejecting the null hypothesis.}$$

The 0.05 value of t for 9 degrees of freedom is 2.262, so that the 95 confidence limits of the mean are $2 \pm 2.262 \times 2.44 = -3.52$ to 7.52. Because zero is included within these limits, the null hypothesis cannot be rejected.

The probabilities for the t values can be obtained from <http://vassarstats.net/tabs.html#t>, <http://in-silico.net/statistics/ttest>, <http://www.statrek.com/online-calculator/t-distribution.aspx>, http://www.wessa.net/rwasp_twosampletests_mean.wasp, <http://www.quantitativeskills.com/sisa/statistics/pairwise.htm>, <http://www.danielsoper.com/statcalc3/calc.aspx?id=98>, and http://www.statstodo.com/PairedDiff_Pgm.phps.

The online programs <http://scistatcalc.blogspot.com/2013/10/paired-students-t-test.html>, <http://www.graphpad.com/quickcalcs/ttest1.cfm>, <http://www.usablestats.com/calcs/2samplet>, <http://www.mathportal.org/calculators/statistics-calculator/t-test-calculator.php>, and <https://mathcracker.com/t-test-for-paired-samples.php> allow you to enter the data, and then perform the test.

Problem 22.1

The following table shows the peak flow rates (L/min) in asthmatic patients before and after exertion.

Subject	Before	After
1	320	297
2	235	200
3	322	220
4	376	334
5	286	210
6	254	255
7	381	338
8	397	341
9	299	227

Did exertion cause a decrease in peak flow rate? Would you reject the null hypothesis?

There are several points to notice about the paired t -test.

1. The numbers must be ratio or interval numbers, and the distribution of the set of differences should be approximately normal. This is not true here, as shown by [Fig. 22.1](#).

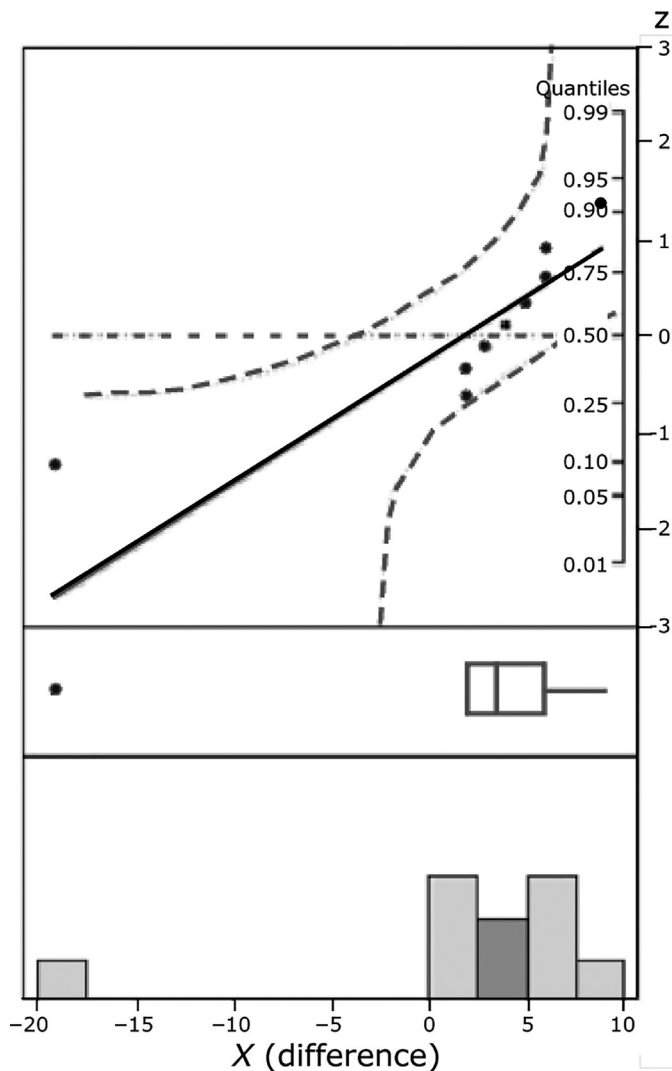


Fig. 22.1 Upper panel: normal quantile plot, with *dashed lines* showing 95% confidence limits. Middle panel: box plot. Lower panel: histogram of differences.

There is one outlier shown in all the panels.

It does not matter if each distribution is not normal; what matters is the distribution of the differences between paired measurements.

2. Pairing must be justified, based on intuitive reasoning or experience. For example, a blood sample divided into two should have the same constituents in each portion; adding something to one of the pair constitutes the difference to be examined. If

theory does not provide the answer, we can turn to experience. Prior work has shown that rat littermates reared on the same diet grow at rates that are more closely matched than are rats from two different litters, or two different species. Patients being tested for airway reactivity tend to have the same reactivity when tested on different days, so that testing a patient before and after administering a potential airway irritant allows pairing. If, however, previous studies have not shown consistency of response to the same stimulus (diet, airway inhalant, etc.) then pairing should not be used. Any investigator using a paired design (shown below to be more efficient than an unpaired design) must justify the use of pairing.

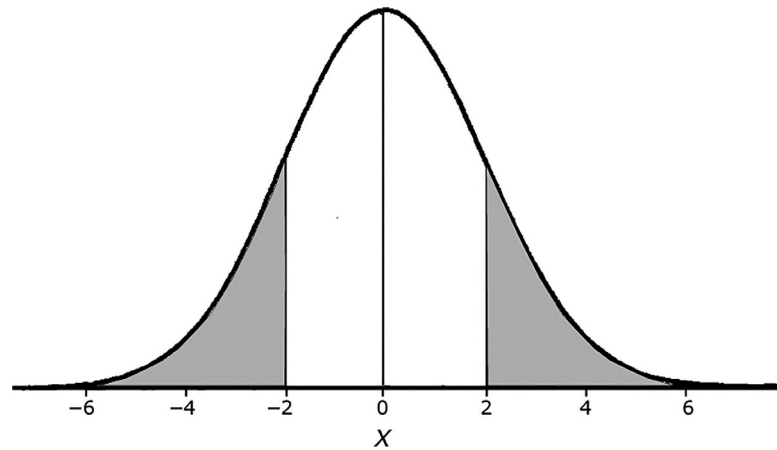


Fig. 22.2 Normal curve with shaded areas indicating proportion under the curve beyond the value of t . That is, assuming normality, if the mean difference in the long run was truly zero, at least 43% of similar samples would have a mean difference > 2 . On this basis we would not reject the null hypothesis.

3. The calculated value for t is 0.82, and the probability of t is 0.43 (Fig. 22.2).
4. t is a ratio of a difference between an observed mean and a population mean (numerator) to a measure of variability (denominator). It indicates the number of standard errors that separate the observed mean difference and the population mean of zero. The numerator is the signal and the denominator is the noise, so that t represents the signal-to-noise ratio. If the difference is small or the variability is big, t will be small, leading to inability to reject the null hypothesis. On the other hand, if the numerator is big relative to the denominator, t will be big, leading to rejection of the null hypothesis.
5. The numerator indicates the absolute difference between two measurements (the effect size), and its importance has to be judged on the physiological or clinical importance of its magnitude. A difference of 2 mg/dL of serum potassium is huge and potentially serious, whereas a difference of 2 mmHg of systolic blood pressure is trivial and not clinically important. Whether a difference is big or small is not a statistical

question but a matter of judgment by the investigator. The numerator reveals the importance of the measured difference.

6. The denominator, the standard deviation of the mean, has variability determined by the variability within the population. For a given numerator to yield a high value of t and lead to rejecting the null hypothesis the denominator should be as small as possible for that set of measurements. This can sometimes be achieved by making the sampled population as homogeneous as possible. For example, there should be less variability of weight gain in the peanut experiment with rats from the same inbred species rather than from different species. At other times minimize variability by avoiding outliers because the standard deviation is not a resistant measurement. Sometimes it is appropriate to transform the data by logarithmic or other transformation to avoid having long tails to the distribution and thus inflating the standard deviation. Another way of minimizing the denominator is to increase the sample size. Because variability is a function of \sqrt{N} , an increase in sample size decreases the standard deviation of the mean.
7. If t is large, so that it is unlikely that the mean difference is zero, then we may wish to reject the null hypothesis. We cannot be certain that the null hypothesis is false; the best we can do is to estimate its probability of being false. If we reject the null hypothesis but are wrong (as shown by future work) we commit a Type I error. It is our choice as to what probability to use to minimize a Type I error. Conventionally the 95% confidence limits are used, giving a 5% chance of making a Type I error, and this is often called “statistical significance.” As discussed in [Chapter 10](#), however, we might prefer not to use the term “significance” and to use $P < 0.001$ or even $P < 0.0001$ before deciding to reject the null hypothesis. (Please reread footnote 2 in [Chapter 10](#).)
 - a. In normal conversation significance implies importance, but that is not true in statistics. Its meaning is confined to the chances of making a Type I error, whether or not the observed difference is physiologically or clinically important. See [Chapter 10](#).
 - b. The 5% (or 0.05) figure for used for rejecting the null hypothesis is arbitrary and probably unsafe to use ([Chapter 10](#)). On the other hand, if an investigator is doing a number of screening tests on different types of peanut preparation, he or she might well use the 10% cut off value to decide which types to study further.
8. If t is large, with a small P value, it argues for a difference between the pairs, but does not prove that the difference observed was due to the experiment. There might have been factors outside our control or knowledge that were the causes. Perhaps the 9/10 rats fed raw peanuts and gained more weight than their littermates were kept warm and slept a lot, whereas their littermates were kept cold and made to be active, so that they burned up more calories. We would see a difference that was *associated* with the type of peanuts, but not due to it.

9. If t is small, first examine the numerator—the difference. If it is small and unimportant, so that we cannot reject the null hypothesis with confidence, we can assume that there are no important differences between the two sets of pairs. If the effect size is potentially important then examine the denominator. Is the large variability due to inhomogeneity that can be reduced? Is it due to a nonnormal distribution that can be normalized? Is it practical (cost, time, manpower) to increase sample size? If none of these remedies is possible, it may be possible to do a nonparametric analysis (see later). Failure to reject the null hypothesis does not mean that the difference was zero, but merely that you have not proved it is not.
10. The one outlier should have been picked up before the analysis was done. Why did 9/10 rats gain more weight on raw peanuts whereas 1 rat not only lost weight, but lost a great deal of weight? Was there an error in weighing or entering the data? Perhaps whoever weighed that rat misread the scale. Did that rat differ in any way? Rats can get pneumonia or tuberculosis, and if that rat was ill it was not validly a member of the group.

The effect of removing the outlying pair from [Table 22.1](#) is shown in [Fig. 22.3](#).

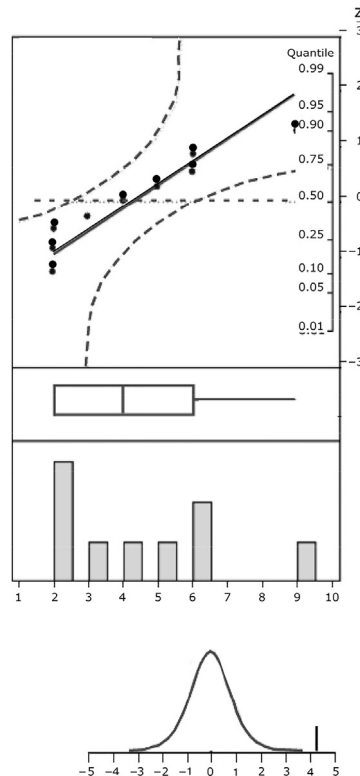


Fig. 22.3 Upper panel: quantile plot. Second panel: box plot. Third panel: histogram. Fourth panel: t -test for the 9 paired observations.

Removing the outlier has changed the mean difference from 2 to 4.33, standard deviation from 7.72 to 2.40, standard deviation of the mean from 2.44 to 0.80, and t from 0.82 to 5.42. That is, the mean difference is now 5.42 standard deviations of the mean from a hypothesized mean of zero, and this would happen with a probability of only 0.00006. The 95% confidence limits become $4.33 \pm 0.80 \times 2.306 = 2.49$ to 6.17. These limits do not include zero, confirming the decision to reject the null hypothesis.

11. If t is large and the P value very low, it does not matter if the distribution was not normal. Uninformed reviewers often make this error.
12. Because the normal curve is symmetrical, the tails of the curve that suggest that the null hypothesis can be rejected contain equal areas. If the null hypothesis is true, there is only a 0.025 (2.5%) probability that the sample mean will lie more than $t_{0.05}\bar{s}_X$ above the mean and another 0.025 probability that it will lie less than $t_{0.05}\bar{s}_X$ below the mean. The two probabilities together add up to 0.05. Whether to use two areas or only one area depends upon the alternative hypothesis H_A . Remember that if the null hypothesis H_0 is rejected, an alternative hypothesis has to be accepted. There are three alternative hypotheses:

$$H_A: \bar{X} \neq \mu;$$

$$H_A: \bar{X} > \mu;$$

$$H_A: \bar{X} < \mu;$$

The first hypothesis states that an excessive deviation from $\mu = 0$ in either direction will lead to rejection of the null hypothesis, the second states that a mean substantially above $\mu = 0$ will lead to rejection of the null hypothesis, and the third states that a mean substantially below $\mu = 0$ will lead to rejection of the null hypothesis. The first is known as the two-tailed test, and the other two each as a one-tailed test. Because the area under the Gaussian curve in a one-tailed test is half that of a two-tailed test, a given value of t will give a probability for a one-tailed test that is half that of a two-tailed test, for example, 0.025 instead of 0.05. It is thus easier to reject the null hypothesis for a one-tailed than for a two-tailed test.

When is a one-tailed test permissible? The answer depends in part on what we are looking for. In the raw versus roasted peanut experiment we might have had no prior guesses as to how the results would turn out, so that a two-tailed test would be appropriate. However, even if we had expected raw peanuts to be better, they could have turned out to be worse. Some proposed treatments are actually harmful, not helpful, and a two-tailed test is appropriate. An example of how predictions can be wrong can be found by examining the CAST trial (CAST, 1989; Ruskin, 1989). The drug flecainide had been shown to be useful in treating and preventing ventricular arrhythmias in experimental animals and in some humans with a normal myocardium and was being used extensively in clinical practice to treat ventricular arrhythmias in patients after myocardial infarction. To investigate further and to

legitimize an accepted practice, the CAST trial randomized patients to a control group or one of several newer antiarrhythmic agents, including flecainide. An interim analysis after about one-third of the patients had been admitted to the study showed to everyone's dismay that four times as many patients had died in the flecainide arm than the control arm of the study. The study was abruptly halted.

One-tailed tests are occasionally used. If a peanut producer is testing the growth potential of a new species of peanut, there is interest in producing it commercially only if it is better than the standard type. Therefore the manufacturer is interested only in a mean growth potential greater than the standard, that is, only in the upper tail. Remember that it is easier to reject the null hypothesis in a one- than a two-tailed test, because to exceed 5% of the area under the curve in one direction takes a smaller deviation from the zero population mean. For example, in a two-tailed test with $N=10$, the value of t that corresponds to 0.025 of the area under the curve for a total of 0.05 for both tails is 2.262, but for 0.05 in one tail is 1.833. See Fig. 22.4.

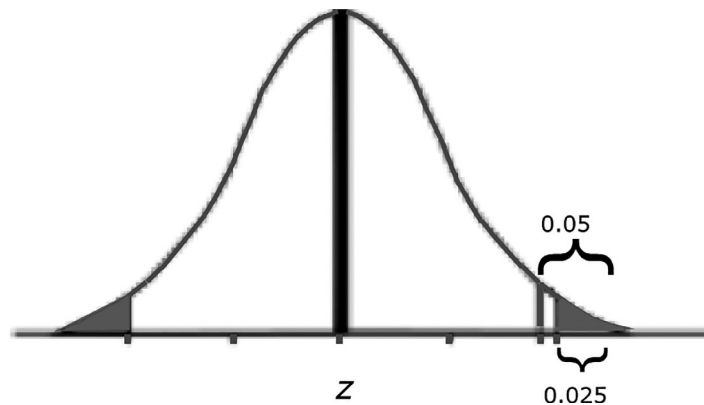


Fig. 22.4 The shaded areas in the two tails make up 0.05 of the total area under the curve, as does the area in the upper tail demarcated by the right of the two lines. The second line is closer to the mean, so that for a given value of α a smaller value of t is needed for a one-tailed test than a two-tailed test.

There is little use for one-tailed tests in Biology or Medicine. We can never be sure that an adverse response might not occur, and so should consider a deviation in either direction. There is no justification for using a one-tailed rather than a two-tailed test simply because it is easier to reject the null hypothesis. It may, however, be used when only one alternative hypothesis is of interest; for example, the new treatment will be used only if it does not cause harm. If it is, the decision to use a one-tailed test must be made before the experiment is done to avoid unconscious bias.

13. Although most two-sided tests are symmetrical, with 2.5% of the area under the curve $>1.96\sigma$ above and below the mean, this is not an absolute requirement. A statistical test provides support for a decision, and decisions have consequences.

Blind adherence to a standard method may not be effective. As discussed in detail by Moyé (2000), there are times when there is more concern for $H_A: \bar{X} < \mu$ than for $H_A: \bar{X} > \mu$. As an example, he discusses testing a new treatment for diabetes mellitus. There is a current standard treatment that is effective in reducing the risk of cardiovascular disease, and its harmful side effects are uncommon and readily recognized and dealt with. A new treatment is proposed and tested. The investigator might be more concerned with an increase in harmful side effects than in an improvement in treatment effect. If in the planning stage the investigator has decided to make the critical value of the Type I error α 0.05, the decision could be made to apportion 0.03 to the tail that indicates harm and 0.02 to the tail that indicates benefit. The decision then is to reject the null hypothesis in favor of harm if $t \leq -1.88$, and in favor of benefit if $t \geq 2.05$. This concept applies to unpaired t -tests and to significance tests in general.

There is no reason not to do the test as usual, and then interpret the observed t value differently for the two alternative hypotheses.

Sample Size for Paired Test

To determine what sample size is needed in any paired experiment, we have to specify the standard deviation of the data, usually taken from similar experiments or a pilot study, the effect size (difference from zero) that is required, and the designated value of α , usually 0.05 or 0.01. Sample size may be calculated online at <http://www.stat.ubc.ca/~rollin/stats/ssize/n2.html>, <http://www.sample-size.net>, https://www.statstodo.com/SSizPairedDiff_Pgm.php, and <http://www.maths.surrey.ac.uk/cgi-bin/stats/sample/twomean.cgi>.

Unpaired t -Test

To compare two unmatched groups, use an unpaired t -test. The numbers in the two groups may be different, but even if they are the same a paired test cannot be done unless pairing is justified.

Return to Table 22.1 and assume that the experiment compared two separate groups of young rats, one group fed with added raw peanuts, the other with added roasted peanuts. No pairing is done. To do the required calculations, first calculate the mean and standard deviation for each group. These values for the means are Raw 57.9, Roasted 55.9g, and for the standard deviations are: Raw 5.59 and Roasted 4.75.

The t ratio then becomes the difference between the means divided by the measure of variability. The variability is, however, derived from two separate variabilities, one for each group. The variability of the difference between two groups is greater than either one alone, because we have to allow for one mean being low and the other high.

To calculate the combined variability, calculate a common or pooled variance s_p^2 that is the weighted average of the two variances (Chapter 3). This can be calculated as

$$s_p^2 = \frac{\sum (X_{1i} - \bar{X}_1)^2 + \sum (X_{2i} - \bar{X}_2)^2}{N_1 + N_2}$$

Then the standard deviation of the difference between two means is

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_p^2}{N_1} + \frac{s_p^2}{N_2}}$$

(This formula implies no correlation between the members of the two groups. If there is correlation, the formula needs to be modified.)

Then

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{s_p^2}{N_1} + \frac{s_p^2}{N_2}\right)}}$$

(Strictly speaking, the numerator should be $|\bar{X}_1 - \bar{X}_2| - \mu$, because the hypothesis is that the difference between the two means is not different from the population mean difference of zero. It is simpler to omit μ from the formula.)

The value obtained for t is tested against $N-2$ degrees of freedom, because two groups are involved.

For the data in Table 22.1 the difference between the means is 2. The argument is that if in the long run there is no difference between the means of the two groups, then the mean difference will be zero. If this null hypothesis is true, how often will a sample difference of 2 arise? The way to determine this is to relate the difference between the means to a measure of variability. If variability is such that a difference of 2 can arise frequently, then we will not reject the null hypothesis. If, however, a difference is unlikely to occur with that population, then we can reject the null hypothesis. For the data in Table 22.1 tested as an unpaired experiment, $t = \frac{2}{2.32} = 0.8621$, degrees of freedom = 18, and $P = 0.40$. About 40% of the time if we drew two samples with $N = 10$ from this population, the means would differ by 2 or more. We therefore do not reject the null hypothesis.

The unpaired t -test can be done online at <http://www.graphpad.com/quickcalcs/ttest1.cfm>, <http://www.usablestats.com/calcs/2samplet>, <https://www.easycalculation.com>.

com/statistics/ttest-calculator.php, <https://www.easycalculation.com/statistics/ttest-calculator.php>, https://www.statstodo.com/MeanDiff_Pgm.php, and <https://www.mathportal.org/calculators/statistics-calculator/t-test-calculator.php>.

Problem 22.2

Reanalyze the peak flow rate data as if they were two different groups of subjects. How do the results differ from a paired test on the same data? Explain why the results are so different.

Requirements for unpaired t -test:

1. First consider if the data meet the requirements for an unpaired t -test. Two of these are similar to the requirements for a paired test, namely, that the numbers are ratio numbers and each distribution is normal. As seen in Fig. 22.5 the second requirement has not been met, but for the moment ignore that.

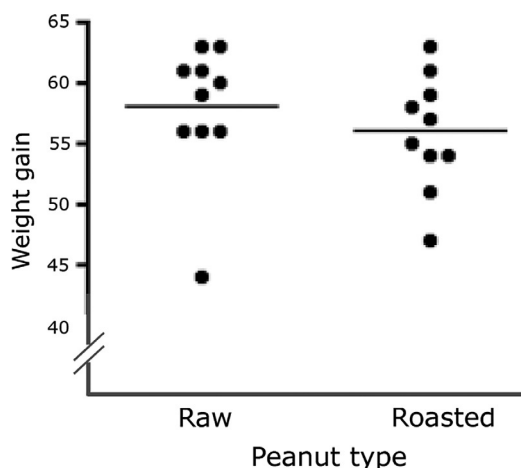


Fig. 22.5 Distribution of data from raw peanuts (left) and roasted peanuts (right).

Another requirement not needed for the paired test is that the variances should be similar in the two data sets. The way of determining this will be described later.

Once again, the 44-g weight gain for the rat fed raw peanuts in the tenth pair is unduly low. Exclude this measurement and repeat the unpaired t -test with 9 in the raw group and 10 in the roasted group. The means become 59.44 g (raw) and 55.90 g (roasted), with respective standard deviations of 2.88 and 4.74, respectively. Now $t=1.939$ with 17 degrees of freedom, and $P=0.0693$. The P value for rejecting the null hypothesis is much bigger than when the paired test was done—0.0693 vs 0.0006—attesting to the greater sensitivity of the paired test *provided it is legitimate to do it*.

2. The remaining considerations about importance versus “significance,” how to interpret the numerator and the denominator, and what to do about a large P value are the same as for the paired test.

We always need to ask if the observed difference is important. This is a value judgment made by workers in that field who ought to know if a difference of 3.54-g weight gain is important. It is possible to quantitate this difference by assessing what contribution the diet makes to total variability. If we take the total variability of the 19 weight gains (excluding the aberrant value), we can ask how much this is reduced if we allow for differences due to the diet. This subject will be taken up more fully in [Chapter 25](#) on ANOVA, but we can estimate the relative reduction, termed ω^2 , as

$$\omega^2 = \frac{t^2 - 1}{t^2 + N_1 + N_2 - 1}.$$

For the unpaired peanut data,

$$\omega^2 = \frac{1.939^2 - 1}{1.939^2 + 10 + 9 - 1} = 0.1268$$

which can be interpreted as showing that about 13% of the variability of weight gains can be accounted for by differences in diet.

Unequal Variances

The logic of the t -test requires a pooled variance from the two groups with comparable variances as a way of obtaining a more representative variance from a larger total sample. If the variances are very different, then their weighted average represents neither sample, with the potential for distorting the value for the standard deviation of the mean that is the denominator for the t -test. If the sample sizes and the variances are very different, the smaller sample has a disproportionate effect on the pooled value because we are dividing by the square root of N . To test the hypothesis of equal variances, divide the larger variance by the smaller variance to obtain the variance ratio termed F ([Chapter 8](#)). This will be discussed fully in the [Chapter 25](#) on Analysis of Variance. Suffice it to state that it is possible to determine whether the observed F ratio is far enough removed from the population ratio of 1 that the hypothesis of equal variances can be rejected.

If the variances are substantially different, use a modified t -test or a nonparametric test. If a nonparametric test is used, it should not be the Mann-Whitney U -test (see later) because if the group variances are very different, Type I error rates may be too high or too low; if too low, the risk of inflating the Type II error is increased ([Ruxton, 2006](#)). Modified t -tests appear in statistics programs as the Welch or the Satterthwaite test, or sometimes the Welch-Satterthwaite test. First determine the ratio d (analogous to t) by

$$d = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}.$$

The value of d is tested by the t distribution with the degrees of freedom being not $N_1 + N_2 - 2$ but rather

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1}\right) + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2}\right)}.$$

This is based on the weighted average of the two standard errors.

Because this will usually not be an integer, the next smallest integer is used. There is no need to compare the two variances before doing the test, and many authorities prefer to perform the test routinely.

As an example, return to the peanut data tested in unpaired groups. The two means are 57.9 and 55.9, and the variances are 31.21 and 22.48, with $N=10$ in each group.

$$\begin{aligned} \text{Then } d \text{ becomes } d &= \frac{57.9 - 55.9}{\sqrt{\left(\frac{31.21}{9} + \frac{22.48}{9}\right)}} = \frac{2}{2.44} = 0.82 (\text{as before}) \\ \nu &= \frac{\left(\frac{31.21}{10} + \frac{22.48}{10}\right)^2}{\frac{\left(\frac{31.21}{10}\right)^2}{9} + \frac{\left(\frac{22.48}{10}\right)^2}{9}} = \frac{28.83}{1.08 + 0.56} = 17.57 \end{aligned}$$

and so $P=0.40$.

The probability of rejecting the null hypothesis of 0.40 is not very different from the value of 0.43 obtained when the differences between the variances were ignored. That is because these two variances are not very different; the one aberrant measurement had more effect on the mean difference than on the variances. Some online tests provide the option for using unequal variances: <http://studentsttest.com/>, <http://www.graphpad.com/quickcalcs/ttest1.cfm>, <http://vassarstats.net/> (see t -tests and procedures), <https://arosh.github.io/ttest/unpaired.html>, and <http://www.quantitativeskills.com/sisa/statistics/t-test.htm>.

In general, the t -test is robust and tolerates moderate departures from the basic requirements. Nonnormality of the distribution is more serious than differences in variances, and it is worse to have the larger variance associated with the smaller group than

with the larger group. Finally, lack of normality or inequality of variances is much worse for small than large samples, and, if practical, samples with over 15 in each group are needed to minimize Type I errors. In addition, nonnormality and unequal variances greatly diminish the power of the t -test.

Sample Size for Unpaired Test

Arguments similar to those used in Chapter 11 apply to the two-sample t -test. Online calculations may be done at <http://www.stat.ubc.ca/~rollin/stats/ssize/n2.html>, <http://www.graphpad.com/quickcalcs/ttest1.cfm>, <http://www.danielsoper.com/statcalc3/calc.aspx?id=47>, https://www.statstodo.com/SSiz2Means_Pgm.php, and www.sample-size.net.

Sample size is discussed clearly at <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3275906/>.

Conclusion

Statistical testing is valuable in emphasizing the variability of the measurements. We can cautiously draw conclusions from the tests as long as we use them in support of reasonable biological hypotheses, allow for the possibility of Type II errors, and make sure that any experiment has sufficient power to allow sensible conclusions to be drawn. This does not mean that we discard unexpected results, but merely that these need stronger confirmation.

Unpaired t -tests are often performed to determine if a specific measurement (e.g., chemical test, echocardiographic data) will distinguish two groups of patients. If the effect size is regarded as important, then it implies that the test is helpful in distinguishing groups and may lead to hypotheses about the mechanisms for the difference. However, that measurement may not be useful in an individual patient if there is overlap of the two data sets (see Fig. 22.5).

Nonparametric or Distribution-Free Tests

Parametric tests such as the t -test lose efficiency, sometimes drastically, when the distributions are severely nonnormal because of skewing, outliers, kurtosis, or grossly unequal variances. They can be replaced by several robust tests that are referred to as distribution-free or nonparametric tests. The two tests to be described later, when applied to normal distributions, are very efficient. The relative efficiency of two tests is determined by the ratio of the sample sizes needed to achieve the same power for a given significance level and a given difference from the null hypothesis (Healy, 1994). When the two distributions are normal, the distribution-free tests are about 95% as efficient as the t -test (Siegal and Castellan, 1988; Conover, 1980). When the distributions are grossly abnormal, then the distribution-free tests have greater efficiency. The main nonparametric test to replace the paired t -test is the Wilcoxon

signed rank test, and the major replacement for the unpaired t -test is the Mann-Whitney U -test (see also Bland, 2015).

The Wilcoxon Signed Rank Test

The paired values for each group are set out, and the difference between each pair is calculated. Then these differences are ranked from the smallest (1) to the biggest (N), ignoring the sign of the difference; **differences of zero are not ranked**. Any tied ranks are averaged. Once the ranking has been done, the negative signs are put back, and the sums of the negative and the positive ranks are calculated.

The theory is that if the paired sets are drawn from the same population, then there will be some small, some medium, and some large positive differences, and approximately the same number of small, medium, and large negative differences. Therefore the sums of the negative and positive ranks should be about the same. If we can calculate the sampling distribution of T , the smaller of these two sums (positive vs negative) for any value of N , then we can determine if one of those sums is so much smaller than the other that the null hypothesis should be rejected. Although the test is part of statistical computer packages, an example to illustrate the principle is presented in Table 22.2.

Table 22.2 Wilcoxon signed rank test used for paired peanut data

Raw peanuts R	Roasted peanuts P	Difference D	Rank	Signed rank
61	55	6	7.5	+7.5
60	54	6	7.5	+7.5
56	47	9	9	+9
63	59	4	5	+5
56	51	5	6	+6
63	61	2	2	+2
59	57	2	2	+2
56	54	2	2	+2
44	63	-19	10	-10
61	58	3	4	+4
		$\Sigma X_i = 20$ $\overline{X_D} = 2$		$\Sigma + = 45$ $\Sigma - = 10$

The three smallest differences are each 2. Because these account for the first 3 ranks, in the fourth column average them $\frac{1 + 2 + 3}{3}$ and assign each a rank of 2. The next value, 3, occupies the next rank, the 4th rank. Similarly, the two 6 differences occupy ranks 7 and 8, but being equal are each assigned a rank of 7.5. The fifth column shows the same ranks, but now the positive and negative ranks are identified. The sums of the positive and negative ranks are different. If the null hypothesis were true, the two sums should be similar.

Calculations or tables show that the probability of such a difference in signed rank sums based on listing all the possible combinations of the signed ranks for the sample size studied is 0.082. This is much smaller than the probability of 0.43 obtained by the paired t -test. If the Wilcoxon test is done for the 9 pairs after excluding the one aberrant pair, the probability from the Wilcoxon test is 0.0039, not as striking as the 0.0006 from the paired t -test but still a good reason to reject the null hypothesis.

The Wilcoxon signed rank test does not give any result if N is ≤ 5 .

If the number of pairs is >10 , T approximates a normal distribution and we do not need special tables to test the null hypothesis. Because the sum of the first N numbers is $\frac{N(N+1)}{2}$, if the sums of the negative and positive ranks were equal, each would be $\frac{N(N+1)}{4}$. Therefore test the difference between the observed and expected value of $T \left(T - \frac{N(N+1)}{4} \right)$ by dividing by the standard deviation of T

$$\sigma_T = \sqrt{\frac{N(N-1)(2N+1)}{24}}$$

Use the z table to test

$$z = \frac{T - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N-1)(2N+1)}{24}}}$$

Pratt (1959) pointed out that ignoring the zeros may produce paradoxical probabilities, and proposed ranking the differences including the zeros, then dropping the zeros when summing the negative and positive ranks, and using the tables of probabilities for the total number of observations, including the zeros.

Online calculations can be done at <http://www.socscistatistics.com/tests/signedranks/Default2.aspx>, <http://www.sdmproject.com/utilities/?show=Wilcoxon>, <http://vassarstats.net/wilcoxon.html>, and https://www.statstodo.com/PairedDiff_Pgm.php.

Problem 22.3

Perform a Wilcoxon test on the data from Problem 22.1.

The Sign Test

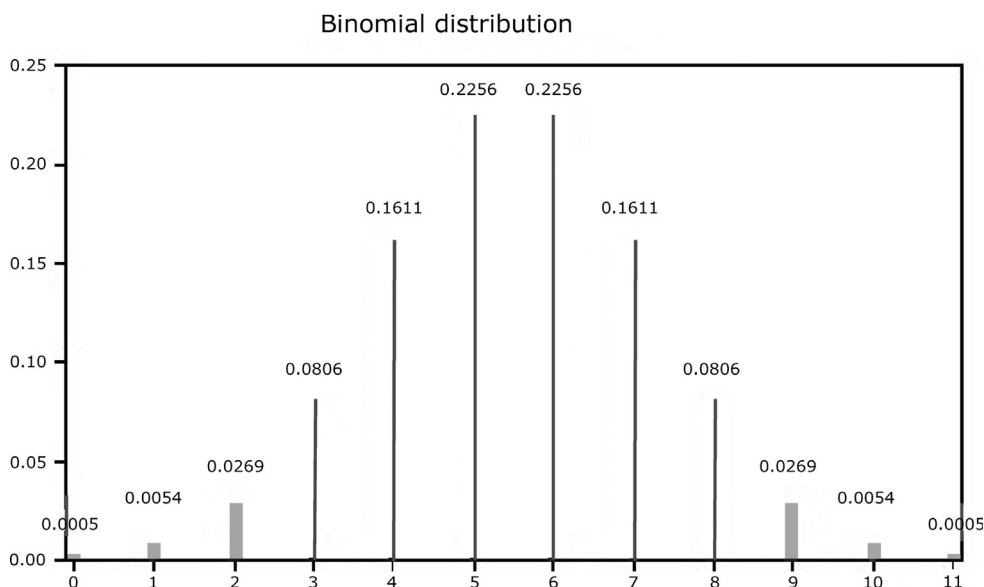
This is a simpler, less powerful version of the Wilcoxon test, used when the data are ordinal or nominal (or categorical). For example, 11 observers rate 2 different bacteriological stains A and B for clarity. Each observer records a preference: if A is better than B the

result is +, and if B is better than A the result is -. On the null hypothesis that there is no difference between the two stains, there should be as many negative results as positive results. If there are more of one sign than the other, then the departure from the null hypothesis can be tested using the binomial distribution for $p=0.5$. Hypothetical data are presented in [Table 22.3](#).

Table 22.3 Sign test

Observer	Result
1	+
2	+
3	-
4	+
5	+
6	+
7	-
8	+
9	+
10	+
11	+

The results show a preference for A in 9 out of 11 trials. On the null hypothesis of no difference between the stains we expect 5 or 6. The question then is to determine if 9 is an unusual event if the null hypothesis is true. To see how this decision is reached, examine [Fig. 22.6](#)

**Fig. 22.6** Binomial distribution for $N=11$, $P=0.50$.

9, 10, or 11 pluses together give a probability of 0.0328, and we would probably reject the null hypothesis. This is the probability of one tail of the distribution. However, a finding of 0, 1, or 2 would also lead to a rejection of the null hypothesis. Because the designation of + or – is arbitrary, finding either 0, 1, 2, 9, 10, or 11+ would occur with a probability of $2 \times 0.0328 = 0.0656$. This is still evidence against the null hypothesis, although not quite as strong.

Using this test in place of the Wilcoxon signed rank test loses the information provided by the size of the differences and so produces a less powerful test. The test can be done easily with online programs <http://www.graphpad.com/quickcalcs/binomial1.cfm>, <http://www.socscistatistics.com/tests/signtest/Default.aspx>, and <https://mathcracker.com/sign-test.php>.

The Mann-Whitney U-Test

In 1945 Wilcoxon developed a ranking test for comparing the positions of two distributions; he called the statistic T that was the sum of the ranks in the smaller group. Two years later, Mann and Whitney extended the theory and they called their statistic U . The two statistics are interconvertible

$$U = N_1 N_2 - \frac{1}{2} N_1 (N_1 + 1) - T.$$

Requirements

1. Each sample is drawn at random from its own population.
2. The values are independent of each other within each sample, and the two samples are independent of each other.
3. The measurement scale is at least ordinal.

The test can be done easily. Consider two groups, A with n_1 members and B with n_2 members, each drawn at random from the same distribution. Because the two sets of measurements come from the same distribution, pool them into a single set and then rank them from the smallest, with a rank of 1, to the largest with a rank of $n_1 + n_2$. Then add up the ranks in each group separately. Intuitively, each group should have similar proportions of low ranks, medium sized ranks, and high ranks, so that if n_1 and n_2 are equal the sums of the ranks in the two groups should be equal or nearly so. If n_2 is twice as big as n_1 , then the sum of the ranks of n_2 should be about twice as large as the sum of ranks from n_1 . The more the sums of ranks in the two groups differ from the expected proportion, the less likely is it that the null hypothesis that they come from the same distribution is true. The possible combinations can be enumerated and the probability of any discrepancy between the sums in the two data sets can be ascertained.

The critical values of rank sums for possible combinations of n_1 and n_2 have been calculated, are given in standard tables, and are available in standard computer programs.

If n_1 and n_2 are >10 use the normal approximation

$$z = \frac{\sum R_m - 0.5 - \frac{n_m(N+1)}{2}}{\sqrt{\frac{mn(N+1)}{12}}}$$

where R_m is the smaller rank total. The 0.5 is the continuity connection.¹ The sum of ranks is Wilcoxon's T .

This test is available in computer programs, but an example will clarify the method (Table 22.4).

Table 22.4 Mann-Whitney test using peanut data

Raw peanuts R	Rank R	Roasted peanuts P	Rank P
61	16	55	6
60	14	54	4.5
56	8	47	2
63	19	59	12.5
56	8	51	3
63	19	61	16
59	12.5	57	10
56	8	54	4.5
44	1	63	19
61	16	58	11
	$\Sigma R = 121.5$		$\Sigma P = 88.5$

Note that some ranks are tied.

Applying these data to the formula, we get

$$z = \frac{88.5 - 0.5 - \frac{10(20+1)}{2}}{\sqrt{\frac{10 \times 10(20+1)}{12}}} = 1.3607.$$

The probability of such a discrepancy is 0.1736 (two-tailed) and suggests that we cannot reject the null hypothesis. For the unpaired t -test the probability of rejecting the null hypothesis was 0.4336. The Mann-Whitney test is closer to rejecting the null hypothesis.

When two or more values are tied, the sum of ranks is modified by averaging the tied ranks. For example, the fourth and fifth measurements are each 54, so allocate them each a rank of 4.5. (If the tied measurements are in the same group, it does not matter if we average their ranks or not, because the sum of ranks 4 and 5 is the same

¹ Many formulas do not apply the continuity correction and so produce slightly different answers.

as the sum of ranks 4.5 and 4.5. If the tied ranks are in different groups, then the ranks must be averaged.) The value of T is usually corrected for ties but the correction factor is usually unimportant; there are several types of correction possible. The whole test can be done online at <http://www.socscistatistics.com/tests/mannwhitney/Default2.aspx>, http://www.wessa.net/rwasp_Reddy-Moores%20Wilcoxon%20Mann-Witney%20Test.wasp, and <http://vassarstats.net> (see Ordinal data), and https://www.statstodo.com/UnpairedDiff_Pgm.php.

Problem 22.4

Perform a Mann-Whitney test on the data from Problem 22.1.

ADVANCED CONCEPTS

Comparing Two Coefficients of Variation

Sometimes we are interested in comparing the coefficients of variation of two groups. This can be done in two ways. If the logarithms of the data are normally distributed, then the ratio

$$F = \frac{s_{\log}^2 X_1}{s_{\log}^2 X_2}$$

can be evaluated from standard F tables. If the data are normally distributed, however, their logarithms will not be normally distributed, so use

$$Z = \frac{CV_1 - CV_2}{\sqrt{\left(\frac{CV_p^2}{N_1 - 1} + \frac{CV_p^2}{N_2 - 1}\right) \left(0.5 + CV_p^2\right)}},$$

where $CV_p = \frac{CV_1(N_1 - 1) + CV_2(N_2 - 1)}{N_1 + N_2 - 1}$ is the weighted mean of the two coefficients of variation CV_i (Zar, 2010).

The Paired t -Test Implies an Additive Model

The paired t -test implies the model.

$$X_{i2} = X_{i1} + \alpha + \varepsilon_i \text{ or } X_{i2} - X_{i1} = \alpha + \varepsilon_i.$$

where X_{i1} and X_{i2} are the two members of each pair, α is the mean difference between them (the effect of the treatment), and ε_i is the error associated with each difference. On the other hand, in any given study the relationship might be multiplicative:

$$X_{i2} = \alpha X_{i1} + \varepsilon_i.$$

In this model, the effect of the treatment is to increase each value for X by a factor α . The difference between these two models is unimportant if all the X_i values are close together but assumes importance if X_i varies widely. For example, [Table 22.5](#) presents data based on hypothetical norepinephrine concentrations (pg/mL) before and after dialysis.

Table 22.5 Additive and multiplicative changes

Initial 1	"Fixed" difference 2	Final 3	Final/ initial ratio	Proportional % difference 4	Final proportional difference 5	Actual difference 6	Final/ initial ratio 7
847	30	817	0.96	10	762	85	0.90
794	28	766	0.96	12	699	95	0.88
439	39	400	0.91	9	399	40	0.91
254	34	220	0.88	9	231	23	0.91
245	27	218	0.89	11	218	27	0.89
174	31	143	0.82	13	151	23	0.87
140	28	112	0.80	8	129	11	0.92
119	32	87	0.73	12	105	14	0.88
81	31	50	0.62	10	73	8	0.90
	Mean 31.11			Mean 10.44		Mean 36.22	Mean 0.896
	sd 3.69					sd 32.03	sd 0.017
	se 1.23					se 10.68	se 0.0056

The initial data are in the first column, the differences are in column 2, and the final values for the additive model are in column 3. The differences between the two are similar for each pair (column 2), with a mean difference of 31.11 and a narrow standard deviation; it is reasonable to reject the null hypothesis that this difference is not substantially different from a mean difference of zero. The ratio of change gets smaller as the initial value decreases. If we postulate a multiplicative model with about a 10% decrease, as shown in columns 4 and 5, then the actual decreases (column 6) vary widely, with a mean of 36.22 but a standard deviation of 32.03, which suggests a skewed distribution as well as a difficulty in rejecting the null hypothesis. On the other hand, taking the ratio of the two gave a mean of 0.896 with a very small standard deviation and standard error, making it easier to reject the null hypothesis that the ratio was 1.

[Motulsky \(2009\)](#) recommended that instead of setting out the data to display proportional differences, as presented in [Table 22.5](#), the two members of the pair should be set out as a ratio of $\frac{\text{treated}}{\text{control}}$. The disadvantage to working with ratios is that they are

asymmetric; below 1 the range can be only from 1 to 0, where above 1 the ratio can in theory be any value >1 . To overcome this, he recommended using the logarithms of the ratios. A zero value means no change, a negative value means a decrease, and a positive value means an increase. If this is done for the initial and final data in Table 22.5, the results in the final column are reproduced, and are interpreted as showing that the ratio is consistently below 1 so that there has been a consistent decrease from initial to final measurements.

Confidence Limits for Medians

Sometimes we may want to determine the confidence limits for the median, or the difference between two medians. Gardner and Altman describe a conservative method for these calculations (Gardner and Altman, 1995).

The $100(1-\alpha)\%$ confidence interval for the population interval requires calculating the lower (R_L) and upper (R_U) ranks, assuming the data are arranged in order from smallest to largest:

$$R_L = \frac{N}{2} - \left(z_{1-\alpha/2} \sqrt{\frac{N}{2}} \right)$$

$$R_U = 1 + \frac{N}{2} + \left(z_{1-\alpha/2} \sqrt{\frac{N}{2}} \right).$$

For the data presented in Table 4.4 there were $N=53$ measurements. Then for 95% confidence limits,

$$R_U = 1 + \frac{53}{2} + \left(z_{1-\alpha/2} \sqrt{\frac{53}{2}} \right) = 27.5 + 1.96 \times 5.1478 = 37.59$$

$$R_L = \frac{53}{2} - \left(z_{1-\alpha/2} \sqrt{\frac{53}{2}} \right) = 26.5 - 1.96 \times 5.1478 = 16.41.$$

Return to the array of measurements and locate the 16th and the 38th ranks as the nearest integers. These are 1.17 and 1.32, which are the required limits of the median. These can be calculated online at http://www.wessa.net/rwasp_bootstrapplot1.wasp.

Calculating the confidence limits of the difference between two medians is not often wanted but can be done by the bootstrap technique (Chapter 37).

Ranking Transforms

If the measurements in the two groups are ranked but then a classical parametric t -test is done on the ranks, a robust test results that can be used for abnormal distributions. This approach is supported by Healy (1994).

The Meaning of the Mann-Whitney Test

An issue that causes confusion is what the Mann-Whitney test indicates. Because the measurements have been turned into ranks, the test cannot allow us to compare the means of the two distributions. Does it compare medians? Consider two groups A and B of equal sample size with the following measurements ([Table 22.6](#)).

Table 22.6 Example for Mann-Whitney test

A: 1.3, 2.7, 4.4		12.7, 15.8, 19.2
B:	5.9, 7.0, 7.2, 8.5, 9.0, 11.4	

The sum of ranks for group A is $1 + 2 + 3 + 10 + 11 + 12 = 39$, and the sum of ranks for B is $4 + 5 + 6 + 7 + 8 + 9 = 39$. The null hypothesis of equality of the sums of ranks of the two groups is obvious, but what is equal? The median of group A is 8.55 and the median of group B is 7.85. These are fairly close to each other. If, however, we make the three largest measurements in group A 127, 158, and 192, the sums of ranks are unaltered, but the median of group A becomes 65.7, much greater than the median of group B. Therefore the test does not compare medians, although with less dispersed distributions it does serve this purpose. What the test actually does is to compare the equality of mean ranks, and thus, by inference, of the distributions. However, like all tests it must not be used without thought, as the previous example shows. Many texts point out that the Mann-Whitney test is most useful when the only difference between the two groups is a measure of location. More formally, that the unspecified distributions of two groups X and Y differ only in location, such that $X = Y + d$, where d is a constant. This may be true for some distributions, but not for others. For example, [Healy \(1994\)](#) has pointed out that many biochemical and endocrine distributions share a common start, for example, a low or zero value, but then the control and experimental groups differ in shape ([Fig. 22.7](#)).

Under these circumstances, the curves differ by more than location.

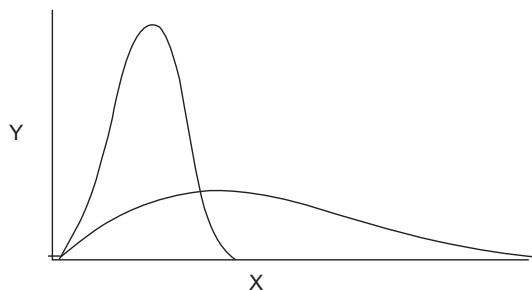


Fig. 22.7 Comparison between two distributions.

Problems

- Often a publication provides the data as the mean (\bar{X}), standard deviation (s), and number of observations (N), but does not give the individual measurements. You may want to perform an unpaired t -test that compares your own data set with the published data, but how can you do this without having all the data? Assume that the published data for cerebral arterial pulsatility in normal neonates has mean 34.3, standard deviation 4.1, and 37 observations. Your own data in neonates with heart disease have values of 51.6, 9.2, and 22, respectively.

To calculate t we need the differences between the two means, which we have, and the standard deviation of that difference which we need to obtain. To derive the required values, you can make use of known relationships:

- $s^2 = \frac{\sum (X_i - \bar{X})^2}{N-1}$ and so $\sum (X_i - \bar{X})^2 = (N-1)s^2$ (columns 3, and 6 below).
- Then the pooled variance can be calculated from equation 19.13, using total in column 7:

$$s_p^2 = \frac{2382.57}{57} = 41.80$$

1	2	3	4	5	6 = 5 ²	7 = 3 × 6
Group	N	N - 1	\bar{X}	s	s ²	(N - 1)s ²
1	37	36	34.3	4.1	16.81	605.16
2	22	21	51.6	9.2	84.64	1777.44
Total	59	57	Difference = 17.3			2382.6

Shaded numbers are computed from the observed data.

Based on these calculations,

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{41.8}{37} + \frac{41.8}{22}} = 1.74,$$

and

$$t = \frac{51.6 - 34.3}{1.74} = 9.94$$

Degrees of freedom 57, $P < 0.0001$.

This simple arithmetic can be automated, and it can be obtained by entering the means, standard deviations, and number of measurements in a computer program. A simple online program can be obtained at <http://graphpad.com/quickcalcs/ttest1.cfm?Format=SD>.

Should you wish to calculate values for $\sum X_i$, multiply the mean by N , and to obtain $\sum X_i^2$ just add $\frac{(\sum X_i)^2}{N}$ to $(N-1)s^2$.

Problem 22.5.

To make sure that you understand the unpaired t -test, try the following problem, and then check the results with the online application.

Group	Number	Mean	Standard Deviation
1	197	14.7	2.9
2	39	19.3	3.3

You should get $t = 6.92$.

REFERENCES

- Bland, M., 2015. An Introduction to Medical Statistics, fourth ed. Oxford University Press, Oxford.
- CAST, 1989. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. The cardiac arrhythmia suppression trial (CAST) investigators. *N. Engl. J. Med.* 321, 406–412.
- Conover, W.J., 1980. Practical Nonparametric Statistics. John Wiley & Sons, New York.
- Gardner, M.J., Altman, D.G., 1995. Statistics with Confidence—Confidence Intervals and Statistical Guidelines. British Medical Journal, London.
- Healy, M.J.R., 1994. Statistics from the inside. 12. Non-normal data. *Arch. Dis. Child.* 70, 158–163.
- Mitchell, H.H., Burroughs, W., Beadles, H.P., 1936. The significance and accuracy of biological values of proteins computed from nitrogen metabolism data. *J Nutrition* 11, 257–274.
- Motulsky, H.J., 2009. Statistical Principles: The Use and Abuse of Logarithmic Axes. <https://s3.amazonaws.com/cdn.graphpad.com/faq/1910/file/1487logaxes.pdf>.
- Moyé, L.A., 2000. Statistical Reasoning in Medicine. The Intuitive P-Value Primer. Springer-Verlag, New York.
- Pratt, J.W., 1959. Remarks on zeros and ties on the Wilcoxon signed rank procedures. *J. Am. Stat. Assoc.* 54, 655–667.
- Ruskin, J.N., 1989. The cardiac arrhythmia suppression trial (CAST). *N. Engl. J. Med.* 321, 386–388.
- Ruxton, G.D., 2006. The unequal variance t -test is an underused alternative to Student's t -test and the Mann–Whitney U test. *Behav. Ecol.* 17, 688–690.
- Siegel, S., Castellan Jr., N.J., 1988. Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill, New York.
- Zar, J.H., 2010. Biostatistical Analysis. Prentice Hall, Upper Saddle River, NJ.

CHAPTER 23

t-Test Variants: Cross-Over Tests, Equivalence Tests

CROSS-OVER TRIALS

One form of paired trial is to give some subjects treatment A, and then after a waiting period give the same subjects treatment B. The null hypothesis is that the two treatments have the same effect, and if we reject the null hypothesis there is some basis for concluding that one treatment is better than the other. An example might be giving a group of hypertensives drug A for a few days, recording the change in blood pressure, and then a week later give the same subjects drug B for a few days to determine which drug caused the greater fall in pressure. For each subject there will be a pressure difference Δ_i . This difference would not be identical in each subject, and for each subject $\Delta_i = \bar{\Delta} + \varepsilon_i$, where ε_i is the individual error term. These error terms have a mean of zero, and their variability allows the calculation of the standard deviation and standard error.

The concern with this design is that the effects of the first treatment might still be present when the second treatment is given. There might be residual blood levels of the drug, some receptors might still be occupied, psychological effects might alter responses, or some long-term physiological changes might have been caused. Some diseases get better or worse with time. If there is any such carry-over effect, then any difference between treatments A and B is a function of a possible real effect of the drugs plus an unknown effect of time, and these cannot be separated. To solve some of these problems, the cross-over design can be used. The following discussion is based on the publication by [Hills and Armitage \(1979\)](#).

Patients are randomized into two similar groups A and B: group A is given treatment X and group B is given treatment Y. One of the treatments can be a placebo. After an appropriate time to allow for washout of the effects of the treatment, the groups are reversed, so that group A gets treatment Y, and group B gets treatment X ([Table 23.1](#)).

The trial is designed to test if the mean values of treatments X and Y are substantially different; any differences due to time are taken into account. Among the assumptions required are similarity between patient groups, attained by randomization, and a response that is on average the same for the two periods on the same treatment; that is, the results of the treatments should not be affected by the order in which they were given ([Hills and Armitage, 1979](#); [Brown, 1980](#); [Jones and Kenward, 2003](#); [Jones, 2008, 2010](#); [Jones and Haughe, 2008](#)). For any given subject in group A, the response in period 1 is Y_1 and can

Table 23.1 Basic 2×2 cross-over trial

Period	Group A	Group B
1	Treatment X Result Washout	Treatment Y Result Washout
2	Treatment Y Result	Treatment X Result

be considered to be the sum of the fixed effect of the treatment T_X and a response that is due to the passage of time ε_{1A} . For that same subject, the response Y_2 in period 2 is $T_Y + \varepsilon_{2A}$. Similarly, a subject in the B group has responses in periods 1 and 2, respectively $T_Y + \varepsilon_{1B}$ and $T_X + \varepsilon_{2B}$ (Table 23.2).

Table 23.2 Individual responses

Period	Group A subject	Group B subject
1	$Y_1 = T_X + \varepsilon_{1A}$	$Y_1 = T_Y + \varepsilon_{1B}$
2	$Y_2 = T_Y + \varepsilon_{2A}$	$Y_2 = T_X + \varepsilon_{2B}$

Based on the assumptions, the values of T_X and T_Y are fixed, but all the other values can change from subject to subject. The effect of treatment (the difference d between X and Y) in a group A subject is determined from $d_A = Y_1 - Y_2 = (T_X - T_Y) + (\varepsilon_{1A} - \varepsilon_{2A})$, and in a group B subject it is $d_B = Y_2 - Y_1 = (T_X - T_Y) - (\varepsilon_{1B} - \varepsilon_{2B})$. If there is no effect of time, then the average values for $\varepsilon_{1A} - \varepsilon_{2A}$ and $\varepsilon_{1B} - \varepsilon_{2B}$ are zero; if there is an effect of time then $\varepsilon_{1A} - \varepsilon_{2A} =$ some mean value δ , with standard error of $\sqrt{N_A}$ for group A and $\sqrt{N_B}$ for group B. Calculate the mean values for each group as \bar{d}_A and \bar{d}_B , and then the average of the difference between these means is $\frac{\bar{d}_A - \bar{d}_B}{2} =$

$$\frac{\left\{ \overline{(T_X - T_Y)_A} + \overline{(\varepsilon_1 - \varepsilon_2)_A} \right\} - \left\{ \overline{(T_X - T_Y)_B} + \overline{(\varepsilon_1 - \varepsilon_2)_B} \right\}}{2} = \frac{\overline{(\varepsilon_1 - \varepsilon_2)_A} + \overline{(\varepsilon_1 - \varepsilon_2)_B}}{2}$$

because the sums of $T_X - T_Y$ for each group cancel out. The standard error of this difference, as in the unpaired t -test, is $\frac{1}{2} \sqrt{\left(\frac{s_p^2}{N_A} + \frac{s_p^2}{N_B} \right)}$, and this can be used to determine if

it is possible to reject the hypothesis that $\bar{d}_A - \bar{d}_B = 0$, that is, that there is no average effect of time. Then test the average effects of the two treatments as

$$\begin{aligned} \frac{\bar{d}_A + \bar{d}_B}{2} &= \frac{\left\{ \overline{(T_X - T_Y)_A} + \overline{(\varepsilon_1 - \varepsilon_2)_A} \right\} + \left\{ \overline{(T_X - T_Y)_B} - \overline{(\varepsilon_1 - \varepsilon_2)_B} \right\}}{2} \\ &= \frac{\overline{(T_X - T_Y)_A} + \overline{(T_X - T_Y)_B}}{2}. \end{aligned}$$

This difference is tested for difference from zero by the same standard error. An alternative set of calculations and a simple explanation are provided by [Wellek and Blettner \(2012\)](#).

Various alternatives have been proposed ([Jones et al., 1996](#)). Some designs include 3 or more periods, for example, group A is given three successive treatments X, Y, Y and group B is given treatments Y, X, X ([Ebbutt, 1984](#); [Laska et al., 1983](#); [Jones and Haughie, 2008](#)). If the second and third identical treatments in each group are similar, it is unlikely that there is a carry-over effect from the first treatment. As an example, [Ramsey et al. \(1993\)](#) studied the effect of aerosolized tobramycin in treating patients with cystic fibrosis who had pneumocystis infection. Group I was given aerosolized tobramycin for 28 days, followed by aerosolized half-normal saline for two 28-day periods. Group II was given aerosolized half-normal saline, followed by two periods of aerosolized tobramycin. The primary outcomes were based on tests of forced vital capacity (FVC), forced expiratory volume (FEV), and forced expiratory flow (FEF). Approximate differences from control values of FEF are presented in [Table 23.3](#).

Table 23.3 Three period cross-over trial

	Period 1	Period 2	Period 3
Group I	Placebo −7	Tobramycin +5	Tobramycin +4
Group II	Tobramycin +8	Placebo +1	Placebo +2

As shown, the duplicate values in periods 2 and 3 are almost identical, suggesting no carry over from period 1 to period 2. In this study, however, there was carry over for FEV.

Cross-over designs can have more groups and can deal with ordinal numbers or binary categories. What are the advantages of the cross-over design? Using each subject as his or her own control minimizes variability as compared with a parallel design with two groups, just as a paired *t*-test has less variability than an unpaired test because it does not have to allow for differences among subjects. The total number of subjects is often considerably less for the cross-over design. This is particularly important when studying treatments for a rare disease. Furthermore, unlike the paired test at two different times, the effect of time can be estimated.

Some key assumptions must be met for the cross-over design to be useful ([Hills and Armitage, 1979](#); [Brown, 1980](#); [Jones, 2008](#)).

(1) The two groups must be equally matched at the onset. (2) The subjects should be in the same clinical state at the beginning of the second period as they were at the beginning of the first period; that is, the first treatment should not leave the subject in a different

state, and the disease process has not changed. (3) The effect of the agent used in the first treatment should not carry over to the beginning the second period; that is, the drug or treatment activity should have a short half-life. (4) The order in which the treatments are given should not affect the results. Therefore cross-over designs are best used for chronic diseases such as chronic obstructive pulmonary disease or rheumatoid arthritis. The design is not restricted to these chronic diseases, though. It has been used to test the ability of acetazolamide to prevent or modify mountain sickness (Greene et al., 1981). It has even been used to study the effect of sumatriptan on acute cluster headaches (Ferrari, 1991). Cross-over designs are often used in equivalence studies.

As an example, treatment with acetazolamide in preventing acute mountain sickness was studied (Greene et al., 1981). Twenty-four amateur mountain climbers were divided at random into two groups. Before climbing Mt. Kilimanjaro (5895 m) one group was given acetazolamide and the other a placebo. After descending, there was a 5-day rest period, and then the treatments were switched when the climbers ascended Mt. Kenya (5186 m). Each climber made daily notes of symptoms, and a scoring system was used; the more symptoms, the higher the score. The results were (Tables 23.4 and 23.5).

Table 23.4 Scores

	Group 1		Group 2			
	Acetazolamide Kilimanjaro (Period 1)	Placebo Mt. Kenya (Period 2)	Period 1–2	Placebo Kilimanjaro (Period 1)	Acetazolamide Mt. Kenya (Period 2)	Period 2–1
	7	0	7	25	–1	–26
	13	7	6	19	5	–14
	3	3	0	17	9	–8
	4	–	–	7	1	–6
	5	–1	6	9	3	–6
	6	–1	7	12	2	–10
	0	0	0	18	2	–16
	1	0	1	12	0	–12
	3	0	3	5	4	–1
	5	2	3	12	–1	–13
	9	9	0	18	–2	–20
	2	2	0	17	8	–9
ΣX	58	21	33	171	30	–141
\bar{X}	4.83	1.91	3	14.25	2.5	–11.75
N	12	11	11	12	12	12
s	3.61	3.30	3	5.74	3.50	6.7
$s\bar{X}$			0.9			1.94

Data adapted from Greene, M.K., Kerr, A.M., McIntosh, I.B., Prescott, R.J. (Eds.), 1981. Acetazolamide in prevention of acute mountain sickness: a double-blind controlled cross-over study. *Br. Med. J. (Clin. Res. Ed)* 283, 811–813.

Table 23.5 Summary of high altitude trial results (see text)

Group	Treatment	Period	Mean score
I	A. Acetazolamide	1	4.83
	B. Placebo	2	1.91
II	A. Acetazolamide	2	2.5
	B. Placebo	1	14.25

The average effect due to time is $\frac{(4.83 - 1.91) - (2.5 - 14.25)}{2} = 7.34$.

The average effect of the drug (difference between scores with acetazolamide and placebo) is

$$\frac{(4.83 - 1.91) + (2.5 - 14.25)}{2} = -4.42.$$

From the data, the standard error was

$$\frac{1}{2} \sqrt{(0.9^2 + 1.94^2)} = 1.07.$$

To test the null hypothesis that time had no effect calculate $t = \frac{7.34}{1.07} = 6.86$. $P < 0.00001$, and we can reject the null hypothesis. This conclusion is reasonable because of the known effect of acclimatization to altitude. The effect of treatment can be tested by $t = \frac{4.42}{1.07} = 4.13$. $P < 0.00001$, also a reason to reject the null hypothesis.

N of 1 Trials

These are variations of the cross-over trials in which treatments are given in random order to a single patient. This is not very different from usual clinical practice in which treatment is changed if it is ineffective or has unacceptable side effects, but it formalizes the approach and uses good statistical methods. If, for example, two medications for back pain are given and symptoms are recorded accurately, it might be possible to show that one treatment is preferable for that particular patient. This avoids the “one size fits all” approach of randomized clinical trials with a gain in efficiency (Lillie et al., 2011).

The ailment treated must be chronic and its symptoms stable. All the issues about carry over pertain. Precautions for randomization and blinding need to be taken. Directions for carrying out these trials are given by Guyatt et al. (1988), Larson (1990), Mahon et al. (1996), and Duan et al. (2013).

EQUIVALENCE AND NONINFERIORITY TESTING

One type of test that compares two samples seems to be the antithesis of a statistical test and that is the equivalence or noninferiority test. Equivalence implies that the new mean is only slightly better or worse than the old mean, whereas noninferiority means that the

new mean is not much worse than the old mean. These tests are aimed at introducing a new treatment that is cheaper, less invasive, has fewer side effects, or has other advantages (Pocock, 2003). A pharmaceutical company might want to establish the merits of a new preparation of a vaccine that can be stored for longer times or may save costs. What is important is to show that the new preparation is not less effective than the old vaccine. If the new item is more effective than the old one that would be advantageous, but all that is required for a license to produce the new vaccine is to show equivalent effectiveness with the previous one. Other examples are comparing two different types of coronary stents (Hofma et al., 2012) or two types of stem cell transplantation (Da Silva et al., 2008). These references describe the methods of testing clearly.

Performing a standard t -test and finding that it does not disprove the null hypothesis is not a substitute for equivalence testing because it may merely reflect a low power. “Absence of evidence is not evidence of absence” (Altman and Bland, 1995). On the other hand, even a trivial difference between two means can lead to rejection of the null hypothesis if the sample size is huge. What is important to consider is the effect size Δ . In noninferiority tests the investigator decides on what size Δ is acceptable. For example, if drug A lowers blood pressure by a mean of 30 mmHg, and drug B lowers blood pressure by a mean of 27 mmHg, then drug B would be regarded as satisfactory. Many regulatory agencies accept a difference of as much as 15% of the mean, that is, if the new treatment is not >15% worse than the old treatment, then noninferiority (accepting the null hypothesis) may be asserted. One Federal standard accepts a 20% difference (Food and Drug Administration, 1977). It would be preferable to have a smaller deviation, for example 5%, but this may demand an impractically large number of subjects.

Often the two one-sided test (TOST) is done to test the joint null hypothesis

$$\begin{aligned} H_{01} : \mu_1 - \mu_2 &\geq \Delta \\ H_{02} : \mu_1 - \mu_2 &\leq -\Delta \end{aligned}$$

Rejection of H_{01} implies that $\mu_1 - \mu_2 \leq \Delta$, and rejecting H_{02} implies that $\mu_1 - \mu_2 \geq -\Delta$. Rejecting both hypotheses implies that the difference lies within the range $+\Delta$ to $-\Delta$ and hence that for practical purposes the two drugs have equivalent effects. Therefore do two

one-sided t -tests $t_1 = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta}{s_{\bar{X}_1 - \bar{X}_2}}$ and $t_2 = \frac{(\bar{X}_1 - \bar{X}_2) - (-\Delta)}{s_{\bar{X}_1 - \bar{X}_2}}$.

If both t -tests are compatible with the null hypothesis, then the observed difference lies within the permissible difference so that the two drugs have equivalent effects.

A variant of this test is to calculate confidence limits for the difference between the two means. If this lies within the limits $\pm\Delta$, which demarcates a zone of scientific or clinical indifference, equivalence is demonstrated. Fig. 23.1, based on a similar figure by Jones et al. (1996), shows the principle.

Lines labeled Equivalent are within this range, so that studies producing these limits are equivalent to existing products. Lines labeled Not equivalent are outside this range so that the two groups are not equivalent. Lines labeled Uncertain are inconclusive and may call for further studies. Two of the lines showing confidence limits that demonstrate

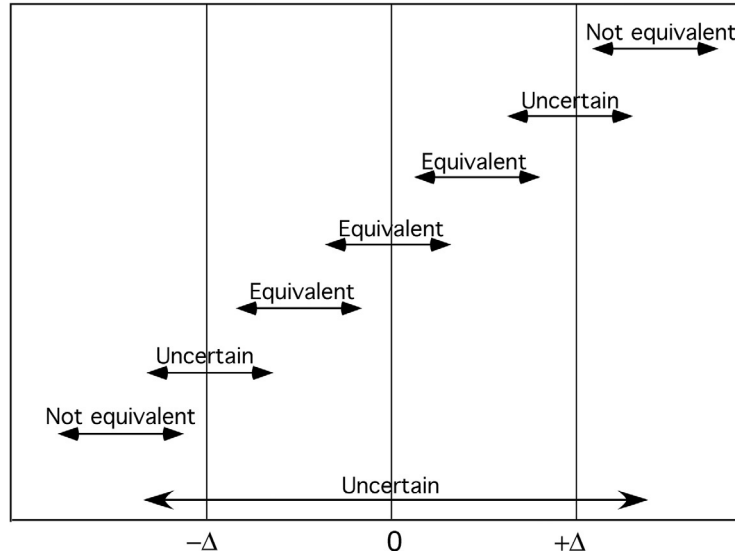


Fig. 23.1 Range from $-\Delta$ to $+\Delta$ within which equivalence is assumed (zone of indifference).

equivalence do not cross zero, so that they argue for rejecting the null hypothesis, but equivalence is still postulated because the observed difference is not meaningful.

The confidence limits are calculated from $(\bar{X}_1 - \bar{X}_2) \pm t_{0.10} s_{\bar{X}_1 - \bar{X}_2}$, and $t_{0.10}$ is chosen so that the chances of rejecting the null hypothesis are 0.05 at each end of the limits.

Details for reporting such trials can be found at <http://www.consort-statement.org/Media/Default/Downloads/Extensions/CONSORT%20Extension%20for%20Non-inferiority%20and%20Equivalence%20Trials.pdf>.

One problem with equivalence testing is that it often requires large numbers of subjects because small differences are being examined. One estimate of numbers required is based on the formula

$$N \geq \frac{(z_\beta + z_\alpha)^2 s^2}{\Delta^2}.$$

Because twice z is squared, four times as many subjects are required as for a simple t or z test. Jacobson and Poland (2005) recommended a modified approach, based on a suggestion by Feinstein. This eliminates consideration of very small differences by setting two thresholds, one for an unimportant difference, designated i , and the other for the threshold of an important difference, Δ . Then it is possible to calculate

$$N \geq \frac{z_\alpha^2 s^2}{(\Delta - i)^2}.$$

What this does is to eliminate the need to consider trivial differences less than i , with consequent reduction in the required numbers.

A simple explanation of this subject can be found at http://graphpad.com/guides/prism/6/statistics/index.htm?stat_testing_for_equivalence_with_c.htm.

REFERENCES

- Altman, D.G., Bland, J.M., 1995. Absence of evidence is not evidence of absence. *BMJ (Clin. Res. Ed.)* 311, 485.
- Brown Jr., B.W., 1980. The crossover experiment for clinical trials. *Biometrics* 36, 69–79.
- Da Silva, G.T., Logan, B.R., Klein, J.P., 2008. Methods for equivalence and noninferiority testing. *Biol. Blood Marrow Transplant.* 15.
- Duan, N., Kravitz, R.L., Schmid, C.H., 2013. Single-patient (n-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research. *J. Clin. Epidemiol.* 66, S21–S28.
- Ebbutt, A.F., 1984. Three-period crossover designs for two treatments. *Biometrics* 40, 219–244.
- Ferrari, M.D., 1991. Treatment of migraine attacks with sumatriptan. The subcutaneous sumatriptan international study group. *New Engl. J. Med.* 325, 316–321.
- Food and Drug Administration, 1977. The Bioavailability Protocol Guideline for ANDA and NDA Submission. Division of Biopharmaceutics, D. M. B. O. D., Food AND Drug Administration. <https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM389370.pdf>.
- Greene, M.K., Kerr, A.M., McIntosh, I.B., Prescott, R.J. (Eds.), 1981. Acetazolamide in prevention of acute mountain sickness: a double-blind controlled cross-over study. *Br. Med. J. (Clin. Res. Ed)* 283, 811–813.
- Guyatt, G., Sackett, D., Adachi, J., Roberts, R., Chong, J., Rosenbloom, D., Keller, J., 1988. A clinician's guide for conducting randomized trials in individual patients. *Can. Med. Assoc. J.* 139, 497–503.
- Hills, M., Armitage, P., 1979. The two-period cross-over clinical trial. *Br. J. Clin. Pharmacol.* 8, 7–20.
- Hofma, S.H., Brouwer, J., Velders, M.A., van THof, A.W., Smits, P.C., Quere, M., De Vries, C.J., Van Boven, A.J., 2012. Second-generation everolimus-eluting stents versus first-generation sirolimus-eluting stents in acute myocardial infarction. 1-year results of the randomized XAMI (XienceV stent vs. cypher stent in primary PCI for acute myocardial infarction) trial. *J. Am. Coll. Cardiol.* 60, 381–387.
- Jacobson, R.M., Poland, G.A., 2005. Studies of equivalence in clinical vaccine research. *Vaccine* 23, 2315–2317.
- Jones, B., 2008. The cross-over trial: a subtle knife. *Significance* 5, 135–137.
- Jones, B., 2010. The waiting game: how long is long enough? *Significance* 2, 40–41.
- Jones, B., Haughie, S., 2008. Cross-over trials in practice: tales of the unexpected. *Significance* 5, 183–184.
- Jones, B., Kenward, M.G., 2003. *Design and Analysis of Cross-Over Trials*. Chapman & Hall/CRC, Boca Raton, FL.
- Jones, B., Jarvis, P., Lewis, J.A., Ebbutt, A.F., 1996. Trials to assess equivalence: the importance of rigorous methods. *BMJ (Clin. Res. Ed.)* 313, 36–39.
- Larson, E.B., 1990. N-of-1 clinical trials. A technique for improving medical therapeutics. *West. J. Med.* 152, 52–56.
- Laska, E., Meisner, M., Kushner, H.B., 1983. Optimal crossover designs in the presence of carryover effects. *Biometrics* 39, 1087–1091.
- Lillie, E.O., Patay, B., Diamant, J., Issell, B., Topol, E.J., Schork, N.J., 2011. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Pers. Med.* 8, 161–173.
- Mahon, J., Laupacis, A., Donner, A., Wood, T., 1996. Randomised study of n of 1 trials versus standard practice. *BMJ (Clin. Res. Ed.)* 312, 1069–1074.
- Pocock, S.J., 2003. The pros and cons of noninferiority trials. *Fundam. Clin. Pharmacol.* 17, 483–490.
- Ramsey, B.W., Dorkin, H.L., Eisenberg, J.D., Gibson, R.L., Harwood, I.R., Kravitz, R.M., Schidlow, D.V., Wilmott, R.W., Astley, S.J., Mcburnie, M.A., et al., 1993. Efficacy of aerosolized tobramycin in patients with cystic fibrosis. *N. Engl. J. Med.* 328, 1740–1746.
- Wellek, S., Blettner, M., 2012. On the proper use of the crossover design in clinical trials: part 18 of a series on evaluation of scientific publications. *Dtsch. Arztebl. Int.* 109, 276–281.

CHAPTER 24

Multiple Comparisons

INTRODUCTION

A population has a mean of 50 units and a standard deviation of 10 units. Draw 10 random samples with $N=25$ from this population. The Central Limit Theorem indicates that the means of the 10 samples are distributed normally about a mean of 50 with a standard error of the mean estimated from $\frac{10}{\sqrt{25}}=2$. Therefore the 95% confidence limits for these means are $50 \pm t_{0.05} \times 2 = 50 \pm 2.262 \times 2 = 45.476$ to 54.524. If the highest and lowest means from these 10 samples are compared by an unpaired t -test, t is 3.199, $P=0.005$.

What is wrong with this scenario? Only about 5% of means of such samples drawn from the same population will be >2 standard errors from the population mean, so that only about 1 time in 20 would we incorrectly reject the null hypothesis. However, by selecting the highest and lowest means from such a series, the difference between them has caused us to reject the null hypothesis. In fact, rejecting the null hypothesis for a mean in the lower tail of the distribution implies rejecting it for comparison between that mean and any mean that is above the population mean.

More realistic scenarios are frequent in the scientific literature. Consider Fig. 24.1 that illustrates figures and comparisons commonly shown in the literature.

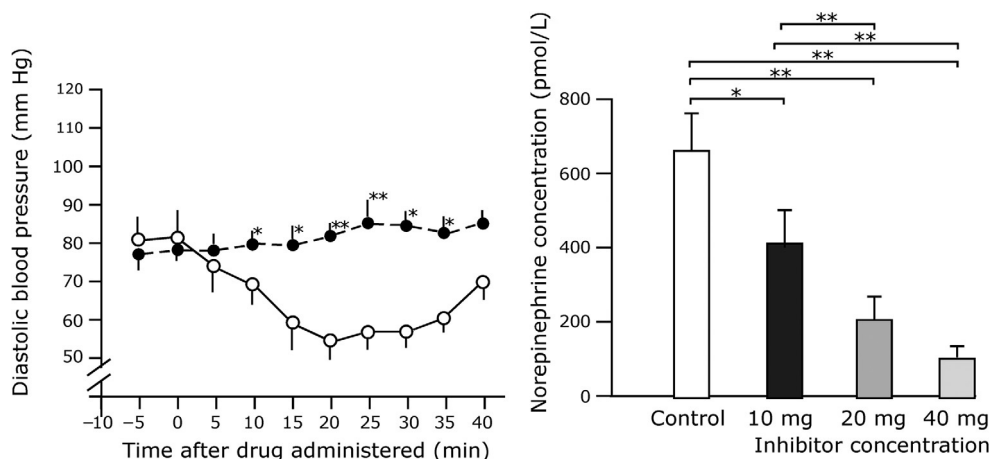


Fig. 24.1 Multiple comparisons. * $P<0.05$; ** $P<0.01$.

The left panel shows 7 sequential time periods with several differences. The right panel shows similar differences for all pairs of comparisons.

Short vertical lines are standard errors.

The questions raised by these multiple comparisons are whether they are legitimate because multiple *t*-tests have been done. The arguments to follow apply equally to other multiple tests, such as chi-square tests. The subject is controversial with some asserting that there is no need to correct for the multiple comparisons and others that correction is always necessary. Most statisticians occupy the middle ground and use correction selectively.

Tukey (1977) provided the following argument. Consider drawing at random 2 samples of the same size from the same population. The logic of the *t*-test tells us that there is a 0.05 probability of rejecting the null hypothesis falsely (α , Type I error), and therefore a $1 - \alpha = 1 - 0.05 = 0.95$ chance of correctly accepting the null hypothesis. If we draw another 2 samples from that population, then by the same argument there is a 0.95 chance of correctly accepting the null hypothesis. What then is the chance of correctly accepting the null hypothesis both times? By the product rule for probabilities, it is $(1 - \alpha)^2 = 0.95 \times 0.95 = 0.95^2 = 0.9025$. Therefore the probability of rejecting the null hypothesis incorrectly is $1 - (1 - \alpha)^2 = 1 - 0.9025 = 0.0975$, even though we use the conventional 0.05 level of α . Draw a third set of samples from that population. Then the probability of correctly accepting the null hypothesis all three times is $(1 - \alpha)^3 = 0.95^3 = 0.8574$. Similarly, drawing 10 pairs of samples, the chances of correctly accepting the null hypothesis in all 10 comparisons is $(1 - \alpha)^{10} = 0.95^{10} = 0.5987$. Therefore the probability of incorrectly rejecting the null hypothesis at least once is $1 - (1 - \alpha)^{10} = 1 - 0.5987 = 0.4013$, even though the 0.05 value for α is what we should get. The value of α has thus been inflated. A similar example illustrated by Bland and Altman (1995) involved 20 comparisons, and then the chances of finding at least one comparison rejecting the null hypothesis at the 0.05 level is 0.64. For 100 comparisons, the probability of incorrectly rejecting the null hypothesis at the 0.05 level is $1 - (1 - \alpha)^{100} = 1 - (1 - 0.05)^{100} = 0.9941$.

A similar inflation of Type I errors occurs when in an experiment the results are divided into very many small groups until a “significant” result appears. This was described by Martin (1984) who termed it “Munchausen’s statistical grid,” the “great virtue of which was that it can resurrect a ‘significant’ result from any foundering therapeutic trial.”

BONFERRONI CORRECTION AND EQUIVALENT TESTS

One way to correct for this inflation factor is attributed to the Italian mathematician Carlo Bonferroni (1892–1960), although the concept had been known for centuries (Bland and Altman, 1995). To keep the value of α at 0.05 for the whole set of comparisons (call this α_f indicating the value of α for the whole family of experiments), solve the equation for the

Table 24.1 Results of Bonferroni correction

k	$\alpha = 0.05$		$\alpha = 0.01$	
	α_c	$\alpha = \alpha/n$	α_c	$\alpha = \alpha/n$
10	0.005116	0.005	0.00100453	0.001
20	0.002561	0.0025	0.00050239	0.0005
50	0.001025	0.001	0.00020099	0.0002
100	0.000513	0.0005	0.0001005	0.0001

value of α_c (c indicating the individual comparison) for each independent comparison: $1 - (1 - \alpha_c)^k = 0.05$ for different values of k , the number of tests conducted.

The expression $\alpha_F = 1 - (1 - \alpha_c)^k$ is sometimes written $\alpha_c = 1 - (1 - \alpha_F)^{1/k}$. In either form it is known as the Dunn-Sidak equation.

Thus if k is 10, then $\alpha_c = 0.005116$. For $k = 20, 50$, and 100, and for $\alpha_c = 0.05$ and 0.01, the values of α_c are given in Table 24.1. These values of α_c are close to what we get by dividing α by k . This is known as the Bonferroni equation, and its results are only very slightly different from the Dunn-Sidak test (Abdi, 2007). Therefore the Bonferroni correction for k t -tests requires that we reject the null hypothesis for each individual comparison only if $P < \alpha/k$. The approximation works because if with k independent t -tests, each with $\alpha = 0.05$ then the probability of getting no differences at this level of α is $(1 - \alpha)^k$. Because α is small, expanding the expression approximates $1 - k\alpha$, because all the higher powers of α are tiny. Then if the null hypothesis is true, for 1 of the k comparisons to have a $P < 0.05$, $k\alpha$ must be < 0.05 , so that α must be $< 0.05/k$ (Bland and Altman, 1995). Reminder: $P < 0.05$ may be an unsafe level to use.

Table 24.1 shows that with 100 comparisons the corrected value of α is so small that it may be difficult to achieve. By trying to avoid Type I errors, the Type II error becomes larger, with the risk of failing to reject the null hypothesis falsely too many times. There is thus loss of statistical power. More efficient procedures were described by Holm and by Hochberg and Benjamini. For the Holm test, rank the P values from the k comparisons from smallest to biggest. Then test the smallest against $0.05/k$, the number of comparisons. If that does not allow rejection of the null hypothesis, accept the null hypothesis for all the comparisons. If we reject the null hypothesis, test the second smallest P value against $0.05/(k - 1)$, and so on. Assume that the P values are 0.005, 0.020, 0.026, and 0.09. Then test 0.005 against $0.05/4 = 0.0125$. Because it is smaller, reject the null hypothesis for that comparison. Then test the second smallest P value, 0.020, against $0.05/3 = 0.017$. Because it is bigger, accept the null hypothesis for this and all remaining comparisons. An online calculator using Excel is at http://www.researchgate.net/publication/236969037_Holm-Bonferroni_Sequential_Correction_An_EXCEL_Calculator; alternatively use <http://www.quantitativeskills.com/sisa/calculations/bonfer.htm>

For the Hochberg test, rank the P values from largest (p_k) to smallest (p_1). If $p_k < \alpha$, reject $H_0(i)$ for $i = 1 \dots k$ for all comparisons. If $p_1 > \alpha$, determine if $p_k - 1 < \alpha/2$. If it is, reject $H_0(i)$ for $i = 1 \dots k - 1$, etc. Assume that these values are 0.09, 0.026, 0.020, and 0.005. Because the first P value exceeds 0.05, do not reject the null hypothesis. Then compare the next P value with $0.05/2 = 0.025$. Because $P = 0.026$ exceeds this, do not reject the null hypothesis for this comparison. For the third comparison, the critical P value is $0.05/3 = 0.017$. Because $P = 0.020$ exceeds this critical value, do not reject the null hypothesis for the third comparison. The fourth P value is compared with the critical value of $0.05/4 = 0.0125$. Because $P = 0.005$ is less than the critical value, reject the null hypothesis for the fourth comparison. If the second comparison had had a P value of 0.021, then this would have been less than the critical value of 0.025, and we would have rejected the null hypothesis for this and *all subsequent comparisons*. Both of these tests keep the error rate for the whole set of tests at 0.05, but greatly reduce the Type II error. The Hochberg test is more powerful than the Holm test, and both improve considerably on the original Bonferroni correction. An online but complex program for this test is at <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3263024/>. The procedure is discussed clearly with examples at <http://onbiostatistics.blogspot.com/2009/08/hochberg-procedure-for-adjustment-for.html>.

Variations of these methods are used when the various end points are correlated with each other; for example, different symptoms in a disease are often correlated (Wright, 1992; Yao and Wei, 1996). Several statistical programs, including R, implement these tests.

The Bonferroni correction is often misused. Glantz (1980) and Wallenstein et al. (1980) wrote editorials to address some common statistical errors in the journals *Circulation* and *Circulation Research*, one of the errors being the use of multiple t -tests. They recommended the Bonferroni correction. Similar conclusions were reached by Pocock et al. (1987) who examined the reports of 45 comparative trials published in the *British Medical Journal*, the *Lancet*, and *New England Journal of Medicine*. Unfortunately, the Bonferroni correction is needed for only some multiple t -tests, and there is a tendency to apply it to all such tests, whether or not they require the correction (Kusuoka and Hoffman, 2002).

Error Rates

Some statisticians reject the Bonferroni adjustment (Rothman, 1990) and others restrict its use to certain types of experiments (Perneger, 1998). An important distinction is between *experiment-wise* and *comparison-wise* error rates. A simple example to explain the difference was published by Carter (2010). If we want to make bicycles with only a 5% chance of being defective, the defective (comparison-wise) error rate for each component (frame, chain, handlebars, etc.) must be well below 5% if the 5% defective (experiment-wise) error rate for the whole bicycle is to be met.

As an example, Creasy et al. (1972) embolized the placenta in pregnant sheep with microspheres and compared the runted and control lambs. They measured body and organ weights and organ blood flows, as well as hematocrit, blood glucose concentrations, and

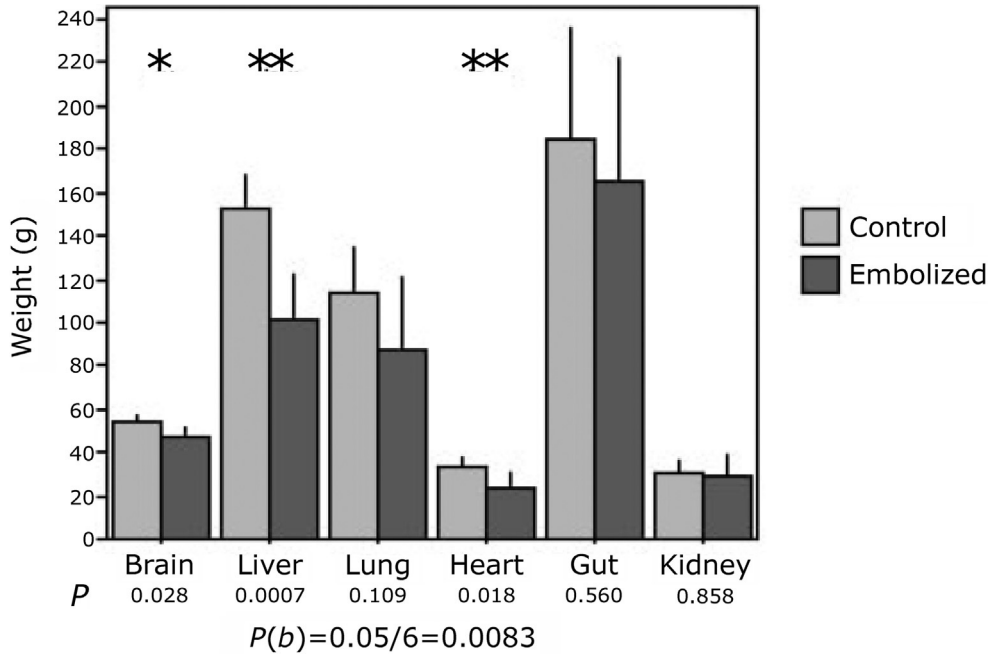


Fig. 24.2 Comparison between control and runted lambs. * $P < 0.05$; ** $P < 0.01$. $P(b)$ Bonferroni adjustment. Vertical lines are standard errors.

arterial pH and oxygen and carbon dioxide tensions. In all, 14 t -test comparisons were made. Fig. 24.2 shows some of their data. Was it correct to do so many t -tests, and did they have to use the Bonferroni correction?

There are two main types of error rates (O'Neill and Wetherill, 1971; O'Brien and Shampo, 1988a; Abdi, 2007). One is the *comparison-wise* error rate, defined as:

$$\frac{\text{Number of comparisons leading to erroneous rejection of the null hypothesis}}{\text{Total number of comparisons}}.$$

The second is the *experiment-wise* (or *family-wise*) error rate, defined as:

$$\frac{\text{Number of families with one or more erroneous rejection of the null hypothesis}}{\text{Total number of families}}.$$

Most statisticians agree that the Bonferroni adjustment should be used if the universal null hypothesis of no differences in any variable is used. For example, doing a battery of biochemical tests on a presumed normal subject. In the Creasy example above, each comparison-wise error rate is 5% and would not change if another six variables had been measured. The family-wise error rate decides whether we accept the hypothesis that runting does or does not change any of the variables and needs protection by the Bonferroni or other methods.

There are many scenarios on the theme of simultaneous multiple comparisons. For example, making consecutive measurements over time such as Fig. 24.1 (left panel) (O'Brien and Shampo, 1988b) using multiple statistical tests to examine possible heterogeneity of response (O'Brien and Shampo, 1988c) or combining the results of several different end points to provide a global measure of superiority of one treatment over another (O'Brien and Shampo, 1988d).

The controversies about doing multiple t -tests extend to more complex analyses and will be discussed in detail in Chapter 25 on Analysis of Variance.

Extreme Multiplicity and False Discovery Rates

Multiplicity problems become extreme when sets of hundreds or thousands of data points are examined, most notably in examining microarrays used to evaluate gene or protein expression or voxels used in imaging. In these studies, the test object (e.g., blood for genes and proteins, organs such as brain or heart for imaging) is divided into thousands of samples, each compared with control normal values. These normal values have their own variability, and some threshold is needed to determine if any one locus differs between test and control.

The comparison-wise error rate (CWER) with $\alpha = 0.05$ examines each comparison separately. For any single comparison, there is a 5% chance of falsely rejecting the null hypothesis with $P = 0.05$, so that standard t -tests on 10,000 spots in a microarray chip provide ~ 500 false positive “significant” differences even if the test and control material were identical. Family-wise error rates (FWER) reduce this potential error and keep the total error rate for the whole array below 5%. An example of such a test is the Bonferroni Inequality in which the null hypothesis for each spot is rejected only with a value of α/k , where k is the number of spots. With $k = 10,000$ the power of the test is very low, and the Type I error is controlled at a given value for α at the cost of an inflated Type II error. Consequently, many differences that might be important would not be detected.

In 1995 Benjamini and Hochberg proposed the concept of the false discovery rate (FDR) to deal with these problems. Results of testing a large number of samples are given in Table 24.2.

The comparison-wise error rate is stipulated to be $E(V/m)$, where E indicates the expectation of the ratio in the long run. Testing each hypothesis at level α guarantees that $E(V/m) \leq \alpha$. Setting a value for α sets a limit for the value of V/m . Testing each hypothesis at level α/m gives the family-wise error rate that keeps the Type I error for the whole data set $\leq \alpha$. The false discovery error rate is represented by the random variable $Q = V/(V + S)$ or V/R and indicates the proportion of rejected null hypotheses that are erroneously rejected. If all the null hypotheses are true, then Q is zero, and the false discovery and family-wise error rates are the same. If not, then the false discovery rate is much lower.

Table 24.2 Type I and Type II errors applied to microarrays

	Do not reject H_0 (Non-DE)	Reject H_0 (DE)	Total
H_0 true (Non-DE)	U = true negative	V = false positive (Type I error)	m_0
H_0 false (H_A true; DE)	T = false negative (Type II error)	S = true positive	$m - m_0$
	$m - R$	R	m

DE, differential expression (difference between test and control value); m , total number of hypotheses tested; m_0 , number of true null hypotheses; $m - m_0$, number of true alternative hypotheses. m is known in advance and R is an observed random variable, but none of the other values are known. S, T, U, and V are rates.

Given that the FDR decreases both Type I and Type II errors, how is it possible to calculate V/R ? One method has been to plot the histogram of the relative frequency with which each p value (for the individual t -tests) occurs (Fig. 24.3; Karp and Lilley, 2007; Storey and Tibshirani, 2003).

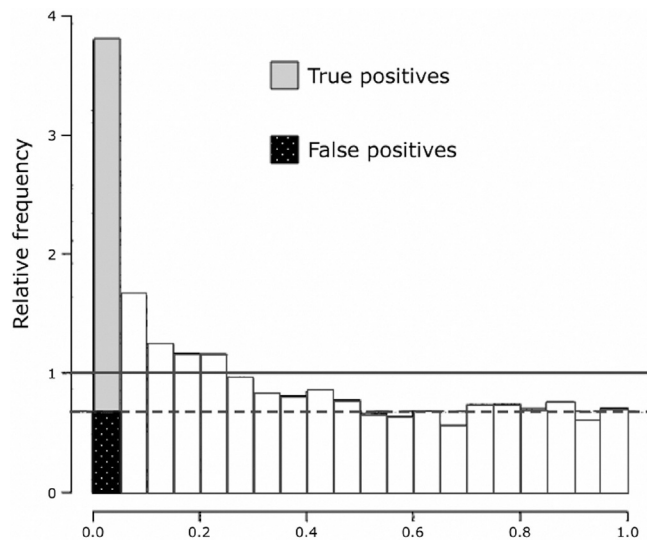


Fig. 24.3 Diagram to illustrate one approach to calculating FDR. If there were no differences between test and control materials, then the relative frequency of each p value (the number of times it occurred relative to what is expected from a normal distribution) would be constant, as indicated by the horizontal solid line. On the reasonable assumption that high p values indicate that the null hypothesis is likely and that there is no difference between the control and the test material, the relatively uniform part of the histogram with $p > 0.5$ provides an estimate of the true background level, as shown by the dashed line. Portions of the columns above this dashed line indicate the true positives (shaded area for the lowest p values) as compared to the baseline false positives shown in dark area. (Based on published figures by Karp, N.A., Lilley, K.S., 2007. *Design and analysis issues in quantitative proteomics studies. Proteomics*, 7(Suppl 1), 42–50; Storey, J.D., Tibshirani, R., 2003. *Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. U. S. A.* 100, 9440–5.)

A simple technique is to set a value Q for the FDR. (Q is the expected proportion of false positives as determined from Fig. 24.3) Assume we choose $Q=0.2$. List the P values from smallest (rank 1) to largest (rank $N=m$) and calculate $(i/m)Q$. Thus for the 4th ranked P value out of 17, $(i/m)Q=4/17 \times 0.2=0.0471$. If the P value is <0.0471 , reject the null hypothesis; if it is greater, it is unsafe to reject the null hypothesis. As an example, consider the data in Table 24.3.

Table 24.3 Calculation of FDR

Rank	P value	$(i/m)Q$	Comment
1	0.010	0.0118	Reject H_0
2	0.022	0.0235	Reject H_0
3	0.029	0.0353	Reject H_0
4	0.043	0.0471	Reject H_0
5	0.11	0.0588	Do not reject H_0
6	0.16	0.0706	Do not reject H_0
7	0.27	0.0824	Do not reject H_0
8	0.31	0.0941	Do not reject H_0
9	0.38	0.1059	Do not reject H_0
10	0.49	0.1176	Do not reject H_0
11	0.55	0.1294	Do not reject H_0
12	0.61	0.1412	Do not reject H_0
13	0.63	0.1529	Do not reject H_0
14	0.69	0.1647	Do not reject H_0
15	0.75	0.1765	Do not reject H_0
16	0.83	0.1882	Do not reject H_0
17	0.98	0.2000	Do not reject H_0

Excel spread sheets for performing this test are available at www.biostathandbook.com/benjaminihochberg.xls.

The subject is discussed simply at <http://www.biostathandbook.com/multiplecomparisons.html>.

There are numerous approaches to computing FDR, and numerous complications such as the fact that often several genes or proteins rise or fall together, so that consultation with an expert in the field is essential. Different methods of setting the threshold for declaring a difference that rejects the null hypothesis give different results. There is a trade-off between sensitivity and specificity and the results should not be accepted blindly without realizing their limitations.

GROUP SEQUENTIAL BOUNDARIES

It is common in clinical trials to plan a study involving many subjects for a long time, but to check the results at intervals to determine if the study should be ended prematurely

because of unexpectedly bad or good results in one group. It would be unethical to continue the study when one group receives a less good treatment. In this context, the multiple comparisons problem is invoked (McPherson, 1974). The approach depends in part on how the trial is likely to proceed. In some trials, many subjects are readily available and unambiguous results are obtained early for each patient. The trial supervisors then estimate how many are needed in each group, what the primary outcome will be, and how long the trial is expected to continue. At the other extreme is a trial when patient accrual is slow and irregular, the time to completion of the trial is uncertain, and the outcome may not be known for several years. These two extremes need different analyses. Furthermore, because most clinical trials test treatments that will produce only modest improvements, the patient numbers required can be very large (Mehta et al., 2009).

One of the most often used methods for interim checks was devised by O'Brien and Fleming. The number of interim tests is defined in advance. Because earlier interim tests involve smaller numbers of patients they have large standard errors and confidence limits and demand a higher level of significance before the trial should be stopped. With each succeeding interim examination, the criterion for accepting the null hypothesis becomes less strict, until for the final test at the end of the trial the conventional predetermined value of α is achieved. O'Brien and Fleming calculated the critical values of α for a predetermined per experiment error rate of 0.05 for 2, 3, 4, or 5 interim tests (Table 24.4, column 2).

Table 24.4 column 2 shows critical P values for a final Type I error rate of 0.05, based on Tables from Pocock (2006) and O'Brien and Shampo. The final test on the completed trial has a critical value close to the designated 0.05. If the final error rate is to be 0.01, then the O'Brien-Fleming method requires critical P values of 0.0000001, 0.00001, 0.001, and 0.004 for the first to fourth interim analyses, respectively. Column 3 shows alternative critical values developed by Haybittle and Peto et al. (1976). These authors use a constant boundary for the interim analyses but retain the α value of 0.05 if early termination does not occur (Table 24.2).

Finally, one other type of test has been proposed to deal with the problem that preliminary data may suggest the need to change the times at which interim analyses are made, or due to slow recruitment the trial has to be extended so that more interim analyses may be needed. The adaptive procedures used were developed by Lan and DeMets and are variants of the O'Brien and Fleming method. They proposed an "alpha spending function" that controlled how much of the false positive error can be used at each interim analysis as a function of the proportion of total information available for the whole test. This proportion is usually based on the fraction of total patients enrolled or the proportion of expected events that had occurred. Some specialized computer programs such as PASS or Cytel will perform the calculations.

A simple rough test was advocated by Pocock (2006) when examining interim results. With the reasonable assumptions that the two groups being compared had approximately

Table 24.4 Probabilities for interim tests

Test number	Critical <i>P</i> value	
	O'Brien-Fleming	Haybittle-Peto
<i>One interim test</i>		
1	0.005	0.001
2	0.049	0.050
<i>Two interim tests</i>		
1	0.0006	0.001
2	0.0151	0.001
3	0.0471	0.0495
<i>Three interim tests</i>		
1	0.00005	0.001
2	0.004	0.001
3	0.018	0.001
4	0.042	0.0492
<i>Four interim tests</i>		
1	0.000005	0.001
2	0.0013	0.001
3	0.009	0.001
4	0.023	0.001
5	0.042	0.0489

equal numbers and that the incidence of events was small and therefore fitted a Poisson distribution, the ratio $z = \frac{a-b}{\sqrt{a+b}}$, where a and b are the numbers of events in the two groups, has an approximately normal distribution. If z is 1.96, then $P=0.05$. The P values of 0.01, 0.001, and 0.0001 are equivalent to z values of 2.68, 3.29, and 3.89, respectively. This method does not eliminate the need for more accurate interim analyses but serves as a check on the arithmetic and provides a more easily understandable figure to evaluate. The reasons for early termination, however, are not affected by this calculation.

The issue of prematurely stopping a clinical trial is complex, and data monitoring committees must take more into account than a P value that refers to the primary end point. Complications need to be taken into account, so does the possibility of late occurring results, and the importance of that particular trial. Pocock discussed these issues in detail (Pocock and White, 1999; Pocock, 2006). He pointed out that most reported trials that were terminated early for substantial benefit were based on limited data and showed unrealistically and unexpectedly large treatment effects. In fact, in some trials that

were continued after apparent interim significance had been reached, subsequent results indicated a less marked difference between the treatments, and Pocock termed this effect “regression to the truth” (Pocock and White, 1999).

The decision to correct for multiplicity is often complicated. The problem is intensified in clinical trials where multiple end points may be involved.

An error due to multiple comparisons occurs also when in an experiment a low P value is not achieved, and the investigators then add more subjects, retest, and continue adding subjects until a satisfactorily low P value is obtained. This illicit process, termed P-hacking (Simmons et al., 2011), cannot be condoned. An adequate sample size should be determined before beginning the experiment. A similar error occurs when comparing two different treatments that are “not significantly different” and the large samples are then subdivided into many smaller samples, each of which is tested (Lee et al., 1980).

SEQUENTIAL ANALYSIS

The ultimate form of interim analysis is sequential analysis, performed after each pair of data points is accrued, whether the data are measured values or preferences (Armitage, 1975). As an example, for comparing preferences (e.g., A is better than B or worse than B) a grid is prepared (Fig. 24.4). To construct the figure, the α and β errors are designated, usually $\alpha=0.05$ and $\beta=0.1-0.2$, and the magnitude of the expected difference θ is selected (often 0.85). Based on these three numbers, the appropriate tables are consulted to determine how to draw the boundary lines.

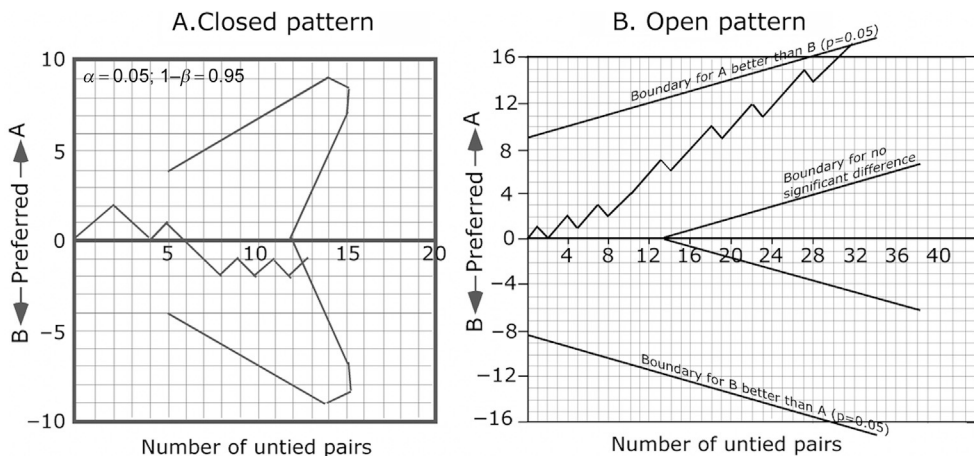


Fig. 24.4 Closed pattern with treatment A being no better than treatment B, and open pattern with A being better than B.

Left panel: For the first pair of preferences (two subjects, two forearms, two cough remedies, etc.) if A is better than B then a diagonal line is drawn upwards in the first square. This is true of the second preference set, so another diagonal line is drawn upwards. If the next preference is for B to be better than A, the next diagonal line passes downwards, as in the third box. Eventually one of three patterns will occur. The jagged preference line crosses the upper boundary, indicating that A is better than B with the stated α and β levels; or it crosses the lower boundary which means that B is better than A at the stated α and β levels; or it crosses the middle boundary which means that the trial has failed to show a difference of the expected magnitude.

The figure shows a closed pattern because the middle boundaries are set. It is also possible to have an open design (right panel) for which the tables provide pairs of parallel lines that serve the same function as the boundaries above. This design has some advantages except that with unbounded parallel lines it is possible for the observed data preference line to meander forever without crossing a boundary. A good example of sequential analysis with an open design was published by [Lewis et al. \(1983\)](#) on the use of aspirin in angina pectoris.

Cautionary Tale

It is important to graph correctly. One study ([Hellier, 1963](#)) published results comparing the effects of trimeprazine versus amylobarbitone in controlling pruritus. A preference for trimeprazine over amylobarbitone produced a line going up at 45 degrees, and a preference for amylobarbitone over trimeprazine produced a line going down at 45 degrees. This graph also showed horizontal lines that indicated no preference. These should not be part of the figure construction, and drawing horizontal portions has the effect of decreasing the mean angle and making the preference line cross the null boundary, giving a false sense of no difference between the treatments. It is, of course, possible for the vast majority of comparisons to result in ties. If that happened, the correct conclusion would be that for most subjects there was no difference between the two treatments, but that for the few who expressed a difference A was preferred more often than B.

Sequential analysis can also be done for measured values. For example, if paired differences are examined, then for each pair calculate a ratio

$$z = \frac{(\sum d)^2}{\sum d^2}, \text{ where } d \text{ is the difference (A-B) between the pairs and may be positive or negative.}$$

$\sum d$ is the result obtained by accumulating successive values for d . If the null hypothesis is true, then $\sum d$ will remain small, and become a smaller and smaller fraction of the expression. If there is a meaningful difference, z will increase and eventually reach and cross the critical boundary. The data are plotted on a figure ([Fig. 24.5](#)) in which the boundaries are set by tables available in Armitage's book ([Armitage, 1975](#)).

Several variations of these designs are possible ([Armitage, 1975](#)).

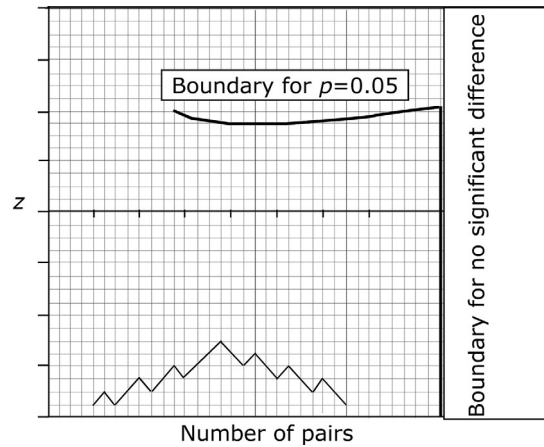


Fig. 24.5 Sequential trial of paired measurements that do not allow us to reject the null hypothesis.

Sequential analyses may minimize the number of patients involved in a trial (Armitage, 1975). All the precautions needed in thinking about early termination of a grouped sequential clinical trial discussed before apply here too.

ADAPTIVE METHODS

Randomized sampling is the hallmark of good clinical trials (Chapter 38) but apart from practical problems of designing an impeccable trial, such trials have a major ethical issue, namely, is it ethical to give one group of patients what will turn out to be an inferior treatment (Royall et al., 1991)? This cannot be known beforehand because that is why the trial is being done, but is there a way to shorten the time of use of the less effective treatment? One method described before is to examine the results at different times after starting the trial, but this cannot be done too often. A second method, sequential sampling, on the average decreases the needed sample size, but is sometimes difficult to implement. An interesting variant is an adaptive strategy termed randomized play-the-winner that can be used if there are binary outcomes detectable soon after treatment has started. As described by Rosenberger (1999) for two treatments A and B an urn (actual or theoretical) is set up to contain α_A balls of type A and α_B balls of type B. For the null hypothesis that A and B are equally good, $\alpha_A = \alpha_B$. After the first few patients are selected at random, the adaptive strategy begins. Any success with A or failure with B causes another type A ball to be added to the urn; conversely, any success with B or failure with A causes another type B ball to be added to the urn. In this way, if treatment A tends to produce better results, the number of A type balls grows faster than the number of B type balls, and the next patient has a greater chance of receiving the apparently better

treatment, and only a minimum number of patients will have received inferior treatment. The theory and practice of this technique have been described (Rosenberger, 1999; Yao and Wei, 1996).

Cautionary Tale

In the first reported clinical trial with this method (Bartlett et al., 1985) the value of ECMO (extracorporeal membrane oxygenation) was assessed in moribund infants with respiratory failure and no response to optimal therapy; their mortality risk was estimated to be 80%–100%. The first patient was randomized to receive ECMO and survived. The next patient was assigned to no ECMO and died. The next patient was assigned to ECMO and survived, and all subsequent patients were assigned to ECMO and survived. Eventually there were 11 survivors of ECMO and 1 death of a patient who did not receive ECMO. Unfortunately, there were problems in that trial, not the least of which were that only one subject did not receive ECMO, and that none of the infants not in the trial who did not receive ECMO died (Paneth and Wallenstein, 1985; Ware and Epstein, 1985). (Subsequently, more extensive trials confirmed the value of ECMO (UK Collaborative ECMO Trial Group, 1996).)

Care is needed when selecting adaptive designs, and power is difficult to determine. Nevertheless, sometimes such a design might give a clear answer using the minimum number of subjects. In a conventional trial of AZT in preventing the vertical transmission of HIV from mother to infant, 239 women received AZT and 238 received placebo; 60 infants in the placebo group had HIV as against 20 in the treated group (Connor et al., 1994). An analysis of this study using an adaptive design suggested that a similar result could have been attained with only 7 failures in the placebo group, and thus would have involved a much smaller trial (Yao and Wei, 1996).

Adaptive methods have recently been recommended as a more effective way to conduct clinical trials (Berry et al., 2015; Meurer et al., 2012). A readable review of various adaptive designs is provided by Chow and Chang (2008).

REFERENCES

- Abdi, H., 2007. The Bonferroni and Šidák Corrections for Multiple Comparisons. Sage, Thousand Oaks, CA. Available: <http://wwwpub.utdallas.edu/~herve/Abdi-Bonferroni2007-pretty.pdf>.
- Armitage, P., 1975. Sequential Medical Trials. Wiley & Sons, New York.
- Bartlett, R.H., Roloff, D.W., Cornell, R.G., Andrews, A.F., Dillon, P.W., Zwischenberger, J.B., 1985. Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study. *Pediatrics* 76, 479–487.
- Berry, S.M., Connor, J.T., Lewis, R.J., 2015. The platform trial: an efficient strategy for evaluating multiple treatments. *JAMA* 313, 1619–1620.
- Bland, J.M., Altman, D.G., 1995. Multiple significance tests: the Bonferroni method. *BMJ* 310, 170.
- Carter, R.E., 2010. A simple illustration for the need of multiple comparison procedures. *Teach. Signif.* 32, 90–91.
- Chow, S.C., Chang, M., 2008. Adaptive design methods in clinical trials—a review. *Orphanet J. Rare Dis.* 3, 11.

- Connor, E.M., Sperling, R.S., Gelber, R., Kiselev, P., Scott, G., O'sullivan, M.J., Vandyke, R., Bey, M., Shearer, W., Jacobson, R.L., et al., 1994. Reduction of maternal-infant transmission of human immunodeficiency virus type 1 with zidovudine treatment. *Pediatric AIDS Clinical Trials Group Protocol 076 Study Group. New Engl. J. Med.* 331, 1173–1180.
- Creasy, R.K., Barrett, C.T., De Swiet, M., Kahanpaa, K.V., Rudolph, A.M., 1972. Experimental intrauterine growth retardation in the sheep. *Am. J. Obstet. Gynecol.* 112, 566–573.
- Glantz, S.A., 1980. Biostatistics: how to detect, correct, and prevent errors in the medical literature. *Circulation* 61, 1–7.
- Hellier, F.F., 1963. A comparative trial of trimeprazine and amylobarbitone in pruritus. *Lancet* 1, 471–472.
- Karp, N.A., Lilley, K.S., 2007. Design and analysis issues in quantitative proteomics studies. *Proteomics* 7 (Suppl. 1), 42–50.
- Kusuoka, H., Hoffman, J.I., 2002. Advice on statistical analysis for circulation research. *Circ. Res.* 91, 662–671.
- Lee, K.L., Mcneer, J.F., Starmer, C.F., Harris, P.J., Rosati, R.A., 1980. Clinical judgment and statistics. Lessons from a simulated randomized trial in coronary artery disease. *Circulation* 61, 508–515.
- Lewis, H.D.J., Davis, J.W., Archibald, D.G., Steinke, W.E., Smitherman, T.C., Doherty III, J.E., Schnaper, H.W., Lewinter, M.M., Linares, E., Pouget, J.M., Sabharwal, S.C., Chesler, E., Demots, H., 1983. Protective effects of aspirin against acute myocardial infarction and death in men with unstable angina. Results of a Veterans Administration Cooperative Study. *New Engl. J. Med.* 309, 396–403.
- Martin, G., 1984. Munchausen's statistical grid, which makes all trials significant. *Lancet* 2, 1457.
- McPherson, K., 1974. Statistics: the problem of examining accumulating data more than once. *New Engl. J. Med.* 290, 501–502.
- Mehta, C., Gao, P., Bhatt, D.L., Harrington, R.A., Skerjanec, S., Ware, J.H., 2009. Optimizing trial design: sequential, adaptive, and enrichment strategies. *Circulation* 119, 597–605.
- Meurer, W.J., Lewis, R.J., Tagle, D., Feters, M.D., Legocki, L., Berry, S., Connor, J., Durkalski, V., Elm, J., Zhao, W., Frederiksen, S., Silbergleit, R., Palesch, Y., Berry, D.A., Barsan, W.G., 2012. An overview of the adaptive designs accelerating promising trials into treatments (ADAPT-IT) project. *Ann. Emerg. Med.* 60, 451–457.
- O'Brien, P.C., Shampo, M.A., 1988a. Statistical considerations for performing multiple tests in a single experiment. 1. Introduction. *Mayo Clin. Proc.* 63, 813–815.
- O'Brien, P.C., Shampo, M.A., 1988b. Statistical considerations for performing multiple tests in a single experiment. 3. Repeated measures over time. *Mayo Clin. Proc.* 63, 918–920.
- O'Brien, P.C., Shampo, M.A., 1988c. Statistical considerations for performing multiple tests in a single experiment. 4. Performing multiple statistical tests on the same data. *Mayo Clin. Proc.* 63, 1043–1045.
- O'Brien, P.C., Shampo, M.A., 1988d. Statistical considerations for performing multiple tests in a single experiment. 5. Comparing two therapies with respect to several endpoints. *Mayo Clin. Proc.* 63, 1140–1143.
- O'Neill, R., Wetherill, G.B., 1971. The present state of multiple comparison methods. *J. R. Stat. Soc. Ser. B* 33, 218–241.
- Paneth, N., Wallenstein, S., 1985. Extracorporeal membrane oxygenation and the play the winner rule. *Pediatrics* 76, 622–623.
- Perneger, T.V., 1998. What's wrong with Bonferroni adjustments. *BMJ* 316, 1236–1238.
- Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J., Smith, P.G., 1976. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br. J. Cancer* 34, 585–612.
- Pocock, S.J., 2006. Current controversies in data monitoring for clinical trials. *Clin. Trials* 3, 513–521.
- Pocock, S., White, I., 1999. Trials stopped early: too good to be true? *Lancet* 353, 943–944.
- Pocock, S.J., Hughes, M.D., Lee, R.J., 1987. Statistical problems in the reporting of clinical trials. *New Engl. J. Med.* 317, 426–432.
- Rosenberger, W.F., 1999. Randomized play-the-winner clinical trials: review and recommendations. *Control. Clin. Trials* 20, 328–342.
- Rothman, K.J., 1990. No adjustments are needed for multiple comparisons. *Epidemiology* 1, 43–46.

- Royall, R.M., Bartlett, R.H., Cornell, R.G., Byar, D.P., DuPont, W.D., Levine, R.J., Lindley, F., Simes, R.J., Zelen, M., 1991. Ethics and statistics in randomized clinical trials. *Stat. Sci.* 6, 52–88.
- Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22, 1359–1366.
- Storey, J.D., Tibshirani, R., 2003. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* 100, 9440–9445.
- Tukey, J.W., 1977. Some thoughts on clinical trials, especially problems of multiplicity. *Science* 198, 679–684.
- UK Collaborative ECMO Trail Group, 1996. UK collaborative randomised trial of neonatal extracorporeal membrane oxygenation. *Lancet* 348, 75–82.
- Wallenstein, S., Zucker, C.L., Fleiss, J.L., 1980. Some statistical methods useful in circulation research. *Circ. Res.* 47, 1–9.
- Ware, J.H., Epstein, M.F., 1985. Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study. *Pediatrics* 76, 849–851.
- Wright, S.P., 1992. Adjusted P-values for simultaneous inference. *Biometrics* 48, 1005–1013.
- Yao, Q., Wei, L.J., 1996. Play the winner for phase II/III clinical trials. *Stat. Med.* 15, 2413–2423 (discussion 2455–8).

CHAPTER 25

Analysis of Variance. I. One-Way

BASIC CONCEPTS

Basic Test

The most effective way of comparing the means of several groups is analysis of variance (ANOVA). The one-way ANOVA is simple conceptually and computationally but has powerful extensions that allow more complex analyses. When comparing the means of several groups, several replicate determinations are made for each independent variable, and measurements within each data set will vary. The independent variables are termed factors. For example, different drugs that might lower blood pressure are tested, using 10 different subjects for each drug. The drugs are factors. In some experiments, different concentrations of each drug might be used; for example, a low dose and a high dose. These subdivisions of the factors are termed levels. The set of levels and factors is termed a treatment.

One requirement for the unpaired t -test was that the variances in the two groups should be homogeneous, and this applies if >2 groups are compared. Take twelve random measurements from a population and calculate the sum of squares of deviations from the mean. This term is shortened to “sum of squares” and abbreviated to SS, with a subscript to indicate the group. Here the subscript is T to indicate the total group. Calculate SS_T (right-hand panel, Fig. 25.1). Dividing SS_T by degrees of freedom $N_T - 1$ gives the variance. In ANOVA the variance is usually called the mean square, which is what it is.

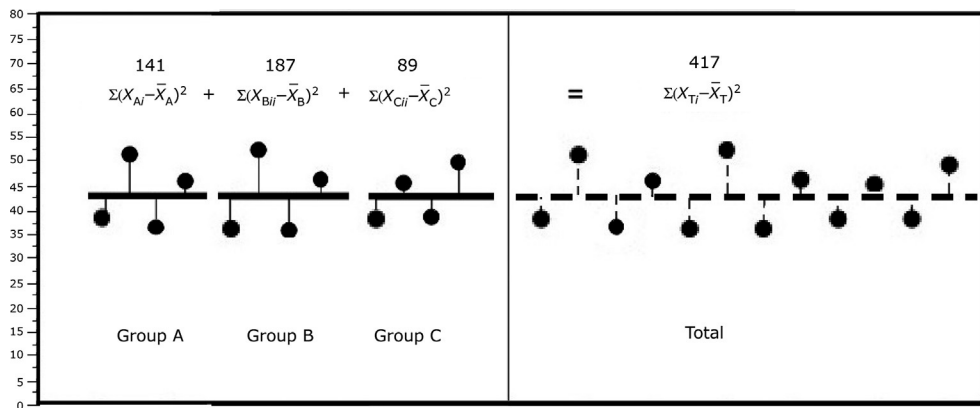


Fig. 25.1 Diagram to illustrate basis of ANOVA. See text.

Assume that there are 3 data sets A, B, and C with 4 measurements in each, each set taken at random from the same population as used for the total group of 12. As shown in Fig. 25.1 the subgroup means might be equal—an unlikely occurrence. If that occurred, then there would be no difference by adding up the sums of squared deviations from the means in sets of 4 for the three subgroups separately, or by adding up all 12 squared deviations as in the total group on the right. The two sums of squares are equal.

It is more likely that there will be three different means: \bar{X}_A , \bar{X}_B , and \bar{X}_C , as in Fig. 25.2; the numbers for group A have been increased by 9 and the numbers for group C have been decreased by 5. The variability of points about the means has not been changed.

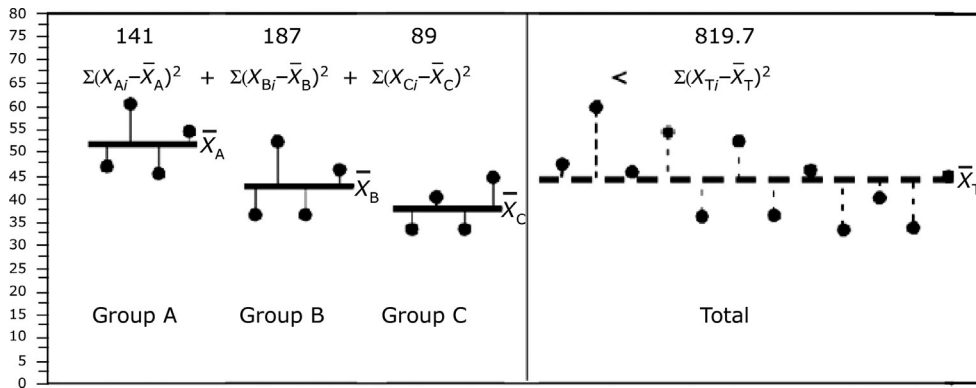


Fig. 25.2 ANOVA with slightly different means. See text.

The thin solid vertical lines show the deviations of each point from its own mean. Squaring these deviations and adding them up gives the respective sums of squares for each group shown in the formula above each data set: these are SS_A (141), SS_B (187), and SS_C (89). When these are added together, their total remains 417 because the variability about each mean has not changed. This sum is, however, much smaller than the sum of squares for the total group because now many of the deviations from the mean of the total group have increased, as shown by the dashed vertical lines in the right-hand panel. This is due to the Principle of Least Squares (Chapter 4); the sum of squared deviations is always least about the mean of its own data set. As a result, if we add the three subgroup Sums of Squares, that total will be less than the Sum of Squares for the total group.

This increased total sum of squares therefore reflects differences among the means. If the means are close to each other the difference will be small, and if the means differ markedly the difference will be large.

Dividing each subgroup SS by its own degrees of freedom (here $4-1=3$) gives the mean square (or variance) for each group, and we can average these three mean squares to obtain another estimate of the population mean square. (Although the example has equal numbers in each subgroup, this is not a requirement for ANOVA. If the numbers in each subgroup were different, derive a weighted mean square by adding up all the SS and dividing by the total degrees of freedom, namely, $N_T - k$ where k is the number of subgroups.)

There is a third way to estimate the population variance. The means of the three subgroups are averaged to give an overall mean \bar{X}_T and we can calculate the sums of squares of deviations from these means (SS_B) as $\sum (\bar{X}_i - \bar{X}_T)^2$. The subscript B indicates that the SS is derived from differences between the means. Divide this sum by $k - 1$ degrees of freedom ($3 - 1 = 2$ because there are 3 subgroups) to obtain the variance of the mean, and the square root of this is the standard deviation of the mean. But the standard deviation of the mean can also be estimated from $s_{\bar{X}} = \frac{s}{\sqrt{N}}$, where N is the sample size. Square this expression to get $s_{\bar{X}}^2 = \frac{s^2}{N}$. Therefore multiplying the variance of the mean $s_{\bar{X}}^2$ by N , the subgroup sample size, (here =4), gives another estimate of the population variance.

There are thus three ways of estimating the population mean square or variance: one from the total group (MS_T), one by averaging the subgroup variances (MS_W), and one from the means (MS_B). In many texts, the SS within groups is referred to as SS_E , where E stands for error, or SS_{res} , where res stands for residual, or $SS_{\underline{W}}$ where \underline{W} stands for Within.

ANOVA calculations are best done by computer, with free online programs available at http://www.physics.csbsju.edu/stats/anova_NGROUP_NMAX_form.html, <http://vassarstats.net/anova1u.html>, <http://turner.faculty.swau.edu/mathematics/math241/materials/anova/aentry.php>, all of these allow entry of all the data points. <http://www.socscistatistics.com/tests/anova/Default2.aspx> allows you to cut and paste. Other programs such as <http://danielsoper.com/statcalc3/calc.aspx?id=43> and <http://statpages.org/anova1sm.html> allow entry of sample size, mean, and standard deviation of groups when the primary measurements are unavailable. The discussion to follow shows what the computations do and helps to clarify the method.

Table 25.1, from an experiment by Richardson et al. (1951) on the requirement of weanling pigs for vitamin B₁₂, presents the average daily weight gain (in lbs) of 3 animals in each of four litters, each litter receiving a different amount of vitamin B₁₂.

Table 25.1 Vitamin B₁₂ data

	Litter A	Litter B	Litter C	Litter D	Total
Daily weight gain (lbs)	1.30	1.26	1.29	1.38	
	1.19	1.21	1.23	1.27	
	1.08	1.19	1.23	1.22	
ΣX	3.57	3.66	3.75	3.87	14.85
N	3	3	3	3	12
\bar{X}	1.19	1.22	1.25	1.29	1.2375
$\Sigma (X_i - \bar{X})^2$	0.0242	0.0026	0.0024	0.0134	0.059025
MS	0.0121	0.0013	0.0012	0.0067	0.005366
$(\Sigma X_i)^2$	12.7449	13.3956	14.0625	14.9769	220.5225
$\frac{(\Sigma X_i)^2}{N}$	4.2483	4.4652	4.6875	4.9923	18.3769

The last two rows are inserted to show an alternative calculation for the between-group sum of squares, discussed later.

The table shows for each litter in columns 2–5 the sums ΣX , the number N , and the mean value \bar{X} . The last column shows the sum of weights, number of pigs, and mean value for the total set of 12 pigs. The next row shows the sum of squares for each litter separately ($\sum (X_i - \bar{X})^2$), and for the total group; the four litter sums of squares do not add up to the total sum of squares, for the reason shown in Fig. 25.2. Divide the sums of squares for each litter and for the total group by the corresponding degrees of freedom ($N - 1$) to obtain the mean square MS in the next line. These values do vary, and whether this is sampling variation or not will be described later. Finally, estimate the population mean square from the group means:

$$s_{\bar{X}}^2 = \frac{(1.19 - 1.2375)^2 + (1.22 - 1.2375)^2 + (1.25 - 1.2375)^2 + (1.29 - 1.2375)^2}{3} \\ = 0.001825$$

Therefore estimate s^2 as $s_{\bar{X}}^2 N = 0.001825 \times 3 = 0.005475$.

Summary of results (Table 25.2).

Table 25.2 Basic ANOVA

Source of variation	SS	Df	MS	F	P
Total (SS_T)	0.059025	11	0.005366		
Within (SS_W)	0.0426	8	0.005325		
Between (SS_B)	0.016425	3	0.005475	1.028	0.4109

The value of SS_B calculated directly is 0.016425, but this is also the same as $SS_T - SS_W$: 0.059025–0.0426. Similarly, the degrees of freedom for SS_B is 4–1=3, but this is also the same as total degrees of freedom minus within degrees of freedom = 11–8=3. These are identities. We have partitioned the total SS and total Df into two components: one due to the within-group SS and Df, and one due to the between-group SS and Df. Divide each SS by its own Df, to obtain the within-group and between-group mean squares. (These are **not** additive). To see how partitioning SS_T works, note that the deviation of each value from the mean of the whole set (\bar{X}^T) is $(X_i - \bar{X}_T)$. This can be separated into two components:

$$(X_i - \bar{X}_T) = (X_i - \bar{X}_i) + (\bar{X}_i - \bar{X}_T).$$

The first part on the right is the difference between an individual measurement and the mean of its own subgroup, the second is the difference between the subgroup mean and the total mean. Square both sides and sum them over all observations and all subgroups (hence the double summation signs) and we get.

$\sum_{i=1}^k \sum_{j=1}^n (X_i - \bar{X}_T)^2 = \sum_{i=1}^k \sum_{j=1}^n (X_i - \bar{X}_i)^2 + \sum_{i=1}^k \sum_{j=1}^n (\bar{X}_i - \bar{X}_T)^2$ because all the cross-products are zero. The first component is SS_T , the second is SS_W , and the third is SS_B .

What would happen if the means were very different? (Fig. 25.3). Adding 9 to each member of group A and subtracting 5 from each member of group C in Fig. 25.2 gives.

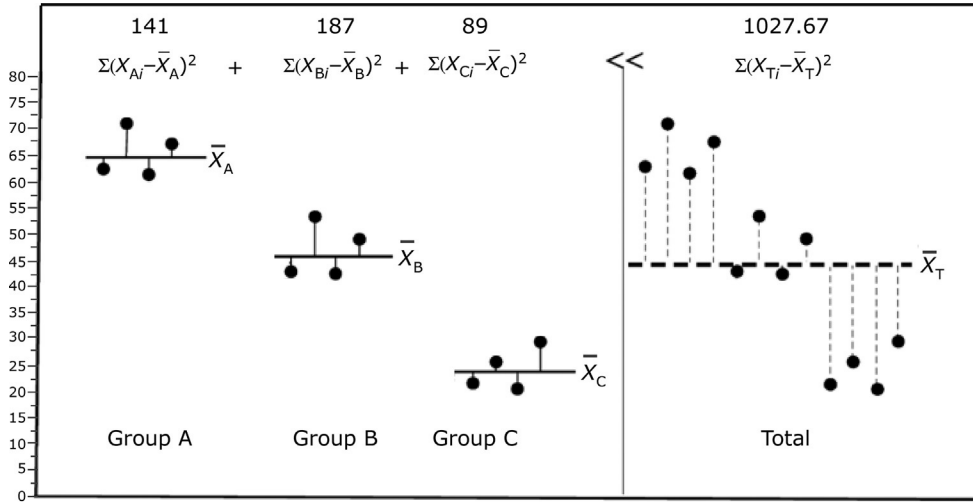


Fig. 25.3 Diagram to show effect of large differences among means. Note the large increase in $\alpha_i (= \bar{X}_i - \bar{X}_T$, where \bar{X}_i is the mean of subgroup i and \bar{X}_T is the mean of the whole population) when compared with Fig. 25.2.

In this figure, the within-group sums of squares has not changed, because each SS is calculated from its own mean. As shown by the dashed vertical lines in the right panel, however, the SS_T has become very large because the mean of the total set is now very different from the individual group means. The MS estimated from the SS_B would be much greater than the MS estimated from SS_W , and we might consider rejecting the null hypothesis that all the group means come from the same population. Because the SS_W is the smallest estimate of variation available, being based on deviations of points from their own means, any marked increase from this value indicates a component due to differences among the means. The between-groups mean square is an unbiased estimate of $\sigma^2 +$ component due to differences among means. This is formally

$$\sigma^2 + \frac{n \sum_{i=1}^k (\mu_i - \bar{\mu})^2}{k-1},$$

where σ^2 is the population variance, k is the number of groups, μ_i is any subgroup mean, $\bar{\mu}$ is the mean of all the measurements, and n is the number in each group. If the subgroup means are very close to each other and the total mean, then the numerator of the second

part of the expression is very small, and SS_B is only slightly larger than SS_W . If differences among the means are large, then $SS_B \gg SS_W$, and F , the ratio of MS_B to MS_W , will be large enough to reject the null hypothesis.

The between-group SS can be obtained directly by calculating

$$\frac{(\sum X_1)^2}{N_1} + \frac{(\sum X_2)^2}{N_2} + \dots + \frac{(\sum X_k)^2}{N_k} - \frac{(\sum X_T)^2}{N_T} \text{ (see last two rows of Table 25.1).}$$

That is, square each subgroup sum of X and divide by the number in that subgroup, add all these subgroup values together, and subtract the square of the total measurements divided by the total number of items. In more complex forms of ANOVA it is always possible to calculate the between-group sum of squares in this way but may be difficult to calculate the within-group sum of squares directly.

How do we determine a probability of accepting or rejecting the null hypothesis? R.A. Fisher determined the distribution of the ratio MS_B/MS_W , and Snedecor subsequently termed it F . Just as for the t -test, ANOVA estimates the probability of obtaining any given F ratio if the null hypothesis is true. This ratio depends on two different degrees of freedom, one for the between-group SS and one for the within-group SS. In this example, $F=1.028$ with Df 3,8, and $P=0.43$ (Chapter 8). The probabilities of given F ratios can be determined online at <http://stattrek.com/online-calculator/f-distribution.aspx>, <http://www.danielsoper.com/statcalc3/calc.aspx?id=7>, <https://easycalculation.com/statistics/f-test-p-value.php>, and <http://www.appliedregression.com/statistical-calculators/fdistribution>, vassarstats.net (see Distributions).

Some programs do not include the row for total values, inasmuch as these are used only to obtain the other two rows.

The term “error” for the within group refers to the minimal natural variation of the data, independent of any differences among the means. Other programs may use the term “treatment” for between groups, because what distinguishes the groups is often the application of different treatments or levels; in the pig data, the four groups represent four different doses of vitamin B₁₂. This type of ANOVA is known as a one-way or one-factor ANOVA, because there is one independent variable being studied. In the previous example, the factor was Vitamin B₁₂ added to the diet in one of four levels.

Problem 25.1

Here is a simple set for practice, based on the blood pressure response (area under the curve) after drinking 480 mL water rapidly in four groups of subjects: MSA—multiple system atrophy (Shy-Drager syndrome), PFA—pure autonomic failure (Bradley-Eggleston syndrome), older controls, and younger controls (Jordan et al., 2000).

MSA	PFA	Older controls	Younger controls
3750	6330	1209	−40
4879	4234	786	−60
3145	3346	826	20
2681	2762	524	20
2580	1875	544	121
2278	1774	444	−504
1935	1854	363	
1774	1552	423	
1572	1431	605	
1230	1310	544	
1350	1169	181	
1512	1048	60	
1310	1230	161	
1190	826	0	
907	604		
665	665		
544	−262		
423			
927			
1048			
645			
464			
262			
161			
−40			
20			
80			

Perform an ANOVA, and think carefully about the results.

Requirements for Test

What are the requirements for a one-way ANOVA? These are similar to those for the unpaired *t*-test that is equivalent to an ANOVA with two groups. (For two groups, the critical 0.05 value of *F* is 3.84, and this is the square of 1.96, the critical value of *t* for large numbers.)

- 1. The numbers must be ratio or interval numbers.
- 2. The distributions should be approximately normal.
- 3. The variances should be homogeneous.
- 4. There is no need for the numbers of measurements in the groups to be the same.
- 5. The measurements must be independent of each other.

An example with a bigger *F* value is presented in [Table 25.3](#) and [Fig. 25.4](#):

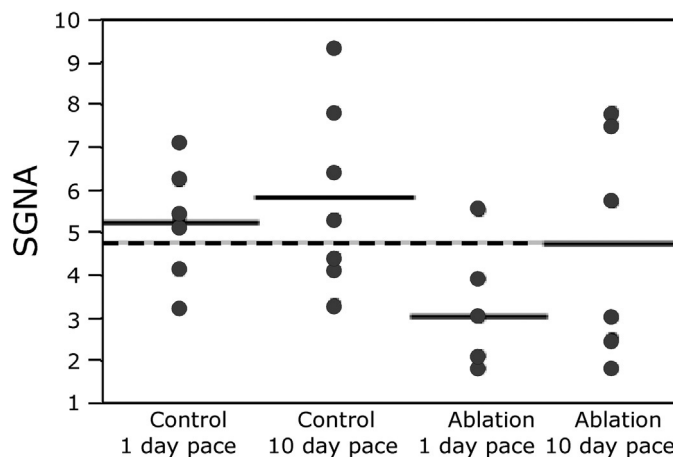


Fig. 25.4 Stellate ganglion nerve activity. *Dashed line is mean of all data; solid lines are group means.*

In this study by [Tan et al. \(2008\)](#) nerve activity in the stellate ganglion (SGNA) was measured after pacing the left atrium in control dogs for 1 day (group A) and 10 days (group B) and 1 day (group C) and 10 days (group D) after cryoablation of sympathetic nerves. Analyze this first as if it were an unplanned experiment. (The term “unplanned” implies that the authors had no specific hypotheses about which groups would differ from the others; this is sometimes referred to as a fishing expedition. The alternative is a planned experiment in which even before the study is performed the authors intend to compare selected groups. Tan et al. actually performed a planned experiment, and the unplanned analysis is used here merely for illustration.)

The basic ANOVA is presented in [Table 25.3](#).

Table 25.3 ANOVA table for stellate ganglion study

Source of variation	SS	Df	MS	F
Total (SS_T)	110.4250	25		
Within (SS_W)	83.1652	22	3.7802	
Between (SS_B)	27.2598	3	9.0866	2.4037
				$P=0.0948$

There were 26 total measurements, so that the total degrees of freedom are $26-1=25$. There were 4 groups so that the between-group degrees of freedom were $4-1=3$. The F ratio is not quite at the 0.05 level. Is this because we have not met the requirements? The data appear to be fairly normally distributed, and the numbers are too small to be worth further normality testing. However, we can try to assess the homogeneity of variances.

Homogeneity of Variance

This is a contentious subject that needs discussion because readers will see these tests mentioned frequently.

How do we tell if the variances are homogeneous, and what difference does it make? (Homogeneous variances are termed “homoscedastic” and heterogeneous variances are termed “heteroscedastic.”) Recall that ANOVA compares the average subgroup variance with the total variance, the difference being a function of the differences among the means. If the subgroup variances are very different themselves, they are no longer good estimates of the population variance. More practically, one or a few large variances among several subgroups inflate the value of SS_W and make it more difficult to achieve a critical F ratio. Various tests are used to assess homogeneity, but they are all imperfect.

In the unpaired t -test, homogeneity of variances is tested by dividing the larger by the smaller variance to obtain an F ratio and determining the probability of that ratio for 1 and $N - 2$ degrees of freedom. In ANOVA, it is more complicated. In particular, the frequent combination of big differences in variance, sample size, and departure from normality makes it difficult to develop a single effective statistical test for homogeneity of variance that minimizes Type I and II errors. The literature has references to tests by Levene, Bartlett, Brown and Forsythe, Hartley, O’Brien, Cochran, and several others, attesting to the unsatisfactory nature of these tests (Zhang, 1998).

There are two main types of tests. One uses the subgroup variances in several ways; calculating $F_{\max} = \frac{s_{\max}^2}{s_{\min}^2}$ (Hartley), $C = \frac{s_{\max}^2}{\sum_{i=1}^N s_i^2}$ (Cochran), or $z = \frac{\ln s_{\max}^2 - \ln s_{\min}^2}{\sqrt{\frac{N}{2}}}$,

where N is the subgroup size, or an average size if subgroup sizes are not very different or Bartlett’s χ^2 test that involves logarithms of all the subgroup variances, (Snedecor and Cochran, 1989, pp. 251–2) and which can be implemented online at <http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/BartletTest.htm>. Some of these tests require special tables (see, e.g., <http://www.statisticshowto.com/fmax-hartleys-test/> or <http://webpace.ship.edu/pgmarr/Geo441/Tables/Hartley%27s%20Fmax%20Table.pdf> for Hartley’s F_{\max} test). All are inefficient when the distributions are markedly asymmetrical.

The second type of test examines deviations from the mean. These are typified by Levene’s test, in which the absolute deviation from the mean is obtained from each subgroup, and an ANOVA is done on these deviations. If one or more subgroups have abnormally large variances their mean deviations will also be large, and a large F value will be obtained. The test can be done at http://www.wessa.net/rwasp_One%20Factor%20ANOVA.wasp. A second form of Levene’s test is to use the squared deviations, although this exaggerates the effect of outliers. More robust versions of Levene’s test are performed by taking deviations either from the median of each subgroup or from a 10% trimmed mean. O’Brien’s test (Abdi, 2007) may reduce the Type I and II errors.

Unfortunately, all these tests become unreliable when the distributions are very abnormal, and most of them do not give any information about how many variance differences there are.

In the example of nerve activity used before, JMP gives the results of several of these tests of homogeneity of variance (Table 25.4).

Table 25.4 Variance homogeneity tests

Test	F ratio	Df	Prob > F
O'Brien	2.5555	3	0.0814
Brown-Forsythe	2.1793	3	0.0602
Levene	3.4274	3	0.0348
Bartlett	1.1484	3	0.3279

In addition, Hartley's maximal to minimal variance ratio was $7.3940/1.7036 = 4.3402$, and this was much above the 0.05 value of 3.27. Note the discrepancies; some tests (e.g., Bartlett's) give no indication of heterogeneity and others (Brown-Forsythe, Levene, and Hartley) suggesting the strong possibility of heterogeneity.

The consensus among statisticians is that ANOVA is robust despite considerable heterogeneity of variances as long as the groups have similar sizes. If group sizes are very unequal, then if the larger variances are associated with the larger samples, the Type I error is $< \alpha$, and if they are associated with the smaller sizes the Type I error is $> \alpha$, sometimes by a large amount. The more important question is what should be done if the variances are inhomogeneous. Possibilities are as follows:

1. Carry out the ANOVA, and if we reject the null hypothesis under these circumstances the rejection is more compelling than if the variances had been homogeneous.
2. If there appear to be samples differing in size, variance, and with severe abnormality (especially with long tails) Lix and Keselman (1998) found by simulation studies that using a trimmed mean and Winsorized variances almost always controlled the Type I error within narrow limits. Given that Type I errors can be as high as 50% under some circumstances, their approach should be considered whenever there is severe abnormality of distributions.
3. Transform the data by, for example, logarithmic or square root transformation (the latter most useful for counts) and then perform an ANOVA. Although the transformed means are not the same as the original means, the ANOVA still tests for differences in location. Methods of finding the correct transformation are available (Emerson and Stoto, 1983; Sokal and Rohlf, 1995).
4. Use a test such as the Welch or the Games and Howell test that allows for inhomogeneity of variance (see later). In the example used before, the F ratio was higher with

the Welch than the regular ANOVA (3.34 vs 2.43 respectively) with a reduction in the value of α from 0.0925 to 0.0574.

5. Use a nonparametric or distribution-free test. The test most often used in the Kruskal-Wallis test that is an extension of the Mann-Whitney test to more than two groups.

Kruskal-Wallis Test

All the data are pooled and ranked from smallest (1) to largest (N), then the sums of ranks in each subgroup are added up, and the probability is calculated. The statistic H is

$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1), \text{ or } H = \frac{12}{N(N+1)} \sum n_i \bar{r}_i^2 - 3(N+1)$$

where N is the total number, n_i is the number in the i -th group, and R_i is the total sum of ranks in the i -th group; in the second equation $\bar{r}_i^2 = \frac{\sum R_i^2}{n_i^2}$. Either equation can be used.

The value of H is tested against the chi-square distribution for $k - 1$ degrees of freedom, where k is the number of groups. If there are tied ranks a correction is used but makes very little difference.

Example: the stellate nerve ganglion data from Fig. 25.3 are ranked (Table 25.5).

Table 25.5 Kruskal-Wallis test

A	B	C	D
7.19 (5)	9.38 (1)	5.47 (9)	7.89 (2)
6.25 (7)	7.81 (3)	3.91 (17)	7.66 (4)
5.39 (10)	6.48 (6)	2.97 (20.5)	5.86 (8)
5.23 (13)	5.31 (11.5)	1.72 (25.5)	2.97 (20.5)
5.31 (11.5)	4.53 (14)	1.80 (24)	2.50 (22)
4.14 (15.5)	4.14 (15.5)	2.11 (23)	1.72 (25.5)
3.20 (19)	3.28 (18)		
$\Sigma R_i = 81$	69	119	82
$\bar{r}_i = 11.37$	9.86	19.83	13.67

The ranks are in parentheses.

$$H = \frac{12}{26 \times 27} \left(\frac{81^2}{7} + \frac{69^2}{7} + \frac{119^2}{6} + \frac{82^2}{6} \right) - 3 \times 27 = 6.1498, \text{ with 3 degrees of freedom, so that } P = 0.1045. \text{ This was no better than the original ANOVA.}$$

Both tests suggest differences among the group means, but with small sample sizes the power of the ANOVA was low (0.53). Standard computer programs perform the analysis, and online calculators are available: <http://vassarstats.net> (see Ordinal), <http://astatsa.com/KruskalWallisTest/>, <http://www.mathcracker.com/kruskal-wallis.php>, <http://www.socscistatistics.com/tests/kruskal/Default.aspx>, and at http://statstodo.com/UnpairedDiff_Pgm.php.

The Kruskal-Wallis test is also used when data sets are composed of ordinal values.

If the Kruskal-Wallis test shows significance, which means are different? Of the several recommended tests, Dunn's is the most useful because it allows for different sample sizes (Dunn, 1964). To perform the test, calculate the statistic Q as

$$Q = \frac{r_i - r_j}{\sqrt{\frac{N(N+1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}}$$

where r_i and r_j are the average ranks for the two groups being compared, with n_i and n_j their respective sample sizes, and N the total sample size.

The critical values for Q depend on the number of groups. It is a Bonferroni type correction in which a critical value to hold the family wide comparison rate at 0.05 is a z value equivalent to $\frac{0.05}{\frac{k(k-1)}{2}}$. The denominator is the number of possible comparisons

that can be made with k groups. The critical value can be determined online at <http://www.graphpad.com/faq/viewfaq.cfm?faq=1156>

The Kruskal-Wallis test with only two groups is equivalent to the Mann-Whitney U -test.

Problem 25.2

Savman et al. (1998) studied newborns who were normal or else had grades I–III of hypoxic-ischemic brain injury (HIE). They measured interleukin-6 (IK-6) concentrations in the cerebrospinal fluid and found the following results:

Raw data for sample			
A	B	C	D
0	26	0	28
0	27	33	79
0	56	42	261
0	247	171	292
0		236	538
35		247	551
43		549	624
		574	640

IL-6 concentrations in pg/mL. A is the control group, and B, C, and D are grades I, II and III, respectively, of HIE.

The differences in sample size and ranges are large enough to suggest that ANOVA might not be the best choice to analyze these data. Use the Kruskal-Wallis test instead. Then compare the mean ranks of groups A and C by Dunn's test.

Conover and Iman developed a rank test by performing ANOVA on the ranks. The rank statistic F_R is evaluated with degrees of freedom $k - 1$ and $N - k$. F_R and H are equivalent. Unlike a classical ANOVA that requires normality of distributions (but certainly tolerates quite wide departures from normality) these rank tests do not require any specific form of distribution but do require that each group have a similar distribution.

No one method for dealing with heteroscedasticity is ideal (Grissom, 2000), and caution is needed in interpreting the results of ANOVA when it is marked.

Independence of Observations

An assumption is that when successive measurements of a variate are made the errors are independent of each other, without any correlation due to space or time. When sequential errors are examined they should be random, not a series of positive errors and other series with negative errors. Because uniformity and independence of errors is an essential part of ANOVA, failure of independence makes the test less sensitive.

Several ways of testing for independence of errors are mentioned in Chapter 31. Another simple one was developed in 1941 by von Neumann et al., who computed the square of each difference $d^2 = (X_{i+1} - X_i)^2$ and added all these differences up to get $\sum d^2$. (The X_i values indicate each successive value.) Divide this sum by $\sum (X_i - \bar{X})^2$. If the measurements are independent the ratio $\eta = \frac{\sum d^2}{\sum (X_i - \bar{X})^2}$ should equal 2. Values other than near 2 can be assessed from tables (Bartels, 1982) or, if $N > 25$, from the normal approximation

$$t = \frac{\left|1 - \frac{\eta}{2}\right|}{\sqrt{\frac{N-2}{N^2-1}}}.$$

Lack of independence suggests some bias in making the measurements and indicates the need to redesign the experiment. Values $\ll 2$ suggest some correlation between the measurements. The test can be performed online at <http://www.wessa.net/slr.wasp>.

Independence may be more important than homogeneity of variance or normality when testing hypotheses (van Belle, 2002). Correlation between successive measurements increases the true over the calculated standard error and increases the chances of a Type I error, sometimes considerably.

Effect Size

In the t -test the effect size is evaluated not only by the absolute difference between the means, but also by the relative difference, namely, absolute difference/standard deviation. Similarly, in ANOVA we need a relative difference. Several statistics have been recommended. One of them is the coefficient of multiple determination, R^2

$$R^2 = \frac{SS_{\text{Between}}}{SS_{\text{Total}}}.$$

This indicates what proportion of the total sum of squares the between-group sum of squares makes. To correct for a slight overestimation an adjusted formula is usually used:

$$R^2_{\text{adj}} = 1 - \left(\frac{N-1}{N-k} \right) (1 - R^2).$$

For the example in [Table 25.2](#), this gives $R^2 = \frac{0.016425}{0.059025} = 0.2783$, so that

$$R^2_{\text{adj}} = 1 - \frac{11}{8} (1 - 0.2783) = 0.0077.$$

This result suggests that differences between the group means contribute little to the total variability and is consistent with the failure to reject the null hypothesis. It shows too the importance of the adjustment.

An alternative approach based on the contribution of the treatment variability to the total variability (effect size) ([Cohen, 1992](#)) is

$$\omega^2_B = \frac{\sigma^2_B}{\sigma^2_B + \sigma^2_w},$$

where $\sigma^2_B = \frac{DF_B(MS_B - MS_w)}{kn}$, k is the number of groups and n the number per group (assumed equal). If they are unequal, an average of N can be substituted. Another way of calculating ω^2_B is $\omega^2_B = \frac{(k-1)(F-1)}{(k-1)(F-1) + kn}$.

Sample Size

There are different formulas when all the subsample sizes are the same ($n_i = n_j$) and if they differ ($n_i \neq n_j$), and if all the means but one are similar versus all the means appear to be different versus the highest and lowest means are different but the rest are bunched in the middle versus the means are fairly evenly scattered from highest to lowest. The basic principle is to calculate a standardized difference, sometimes termed ϕ , the noncentrality parameter, as

$$\phi = \frac{\delta}{\sigma} \sqrt{\frac{n}{2k}},$$

where δ is the difference between the one mean and the average of the others (the effect size), σ is the common within group standard deviation, n is the common sample size, and k is the number of groups. Alternatively, if sample sizes are the same and the means are all different, then use $\phi = \sqrt{\frac{n \sum (X_i - \mu)^2}{k\sigma^2}}$, where \bar{X} refers to each individual expected mean and μ is the mean of the whole set. These two formulas are described by [Glantz](#)

(2005) who gives tables that are entered with the estimates of φ and the degrees of freedom for the between-group sum of squares and the within-group sum of squares.

Cohen (1988) rearranges the formula so that $f = \frac{\phi}{\sqrt{n}}$. The calculation of f depends on the distributions of the means and whether sample sizes are or are not the same. He provides extensive Tables.

A small effect is $\omega^2_B \leq 0.01$, a medium effect is $0.01 \leq \omega^2_B \leq 0.15$, and a large effect is $\omega^2_B > 0.15$. Based on estimated effect sizes, formulas are available for estimating sample size. The programs can be used also to calculate the power of a completed ANOVA. Into the program put values for alpha, total sample size N , the number of groups k , and the calculated value of Cohen's f (see before).

Calculations can be done for simple analyses online by interpolation in the simple online program, <http://www.divms.uiowa.edu/~rlenth/Power/>, <https://anzmtg.org/stats/PowerCalculator/PowerANOVA>, and with the program G*Power found at <http://www.psych.uni-duesseldorf.de/abteilungen/aap/gpower3/>. Because of all the variables concerned, any results obtained should be regarded as conservative.

Problem 25.3

Take the data from Table 25.1 and calculate the sample size needed for a power of 0.8 and $\alpha = 0.05$.

ADVANCED CONCEPTS

Multiple Comparisons

This section is necessary to any study of ANOVA but should be omitted at a first reading until the reader is satisfied about how to perform an ANOVA. There are numerous tests available, no one test is ideal, and there is no consistency in their use. Standard statistics programs implement most of these tests, but readers need to know what the tests do and how reliable they are. I will describe the preferred ones at the end of this section, but because the biomedical community still uses the older tests, readers need to be aware of what they are and how they perform in making statistical inferences. Inferences based on these tests are even less secure if sample sizes differ greatly, variances are very heterogeneous, or distributions are grossly abnormal.

Although ANOVA is recommended for comparing the means of several groups, if the null hypothesis of equality of means is rejected the problem is to decide how many subgroups there are. If the F test allows us to reject the null hypothesis for 6 groups, are there 6 different groups, or does 1 group differ from the other 5, or are there 3 sets of two groups, and so on? The approach to this subject depends on what type of experiment is being done. There are two main types of experiments—*planned* and *unplanned*. The planned experiment

is one in which a specific hypothesis is being tested. For example, if we prepare a batch of dough to make doughnuts, we might hypothesize that more unsaturated than saturated fat is taken up by the dough during the frying process. It is convenient to make one large batch of dough and test many different types of fats at the same time in one experiment. In the analysis, however, the primary interest is in comparing saturated with unsaturated fat uptake. The unplanned experiment is one with no specific hypothesis and is referred to as data dredging, data mining, or fishing. The experiments are performed, the means of each group are determined, and then all the means are inspected to find out if any large differences exist. If they do, they enter a second more focused round of experiments. The planned study is termed an *a priori* study, one in which the important comparisons to be made are designated ahead of time. The unplanned study is termed an *a posteriori* experiment in which, based on the mean values observed, we try to decide which groups are different and worth further study. The latter approach is less efficient.

Evaluating unplanned comparisons in an *a posteriori* study involves striking a balance between keeping the Type I error low without inflating the Type II error. In 2000 Curran-Everett (2000) reported that in 1997 about 40% of the articles published in the American Journal of Physiology used multiple comparison procedures. The tests most often used were the Student-Newman-Keuls, Fisher's LSD, and the Bonferroni tests, but other tests such as Duncan, Dunnett, Scheffé, Tukey, and unnamed procedures formed 45% of the tests used. Tests vary with philosophy and may differ if there are variations in sample size, variances, or shape of the distribution. The tests may be classified as single step (Tukey, Fisher, Scheffé) and multiple step (Student-Newman-Keuls, Duncan, Dunnett, Holm, Hochberg, Ryan). The single step tests use a single factor in a formula to calculate an "allowance"; if the difference between the means exceeds that allowance then the means are regarded as different at a selected level of α , the Type I error. The multiple step methods change the value of the factor depending on the number of means being examined; they start by comparing the two means that show the greatest difference, then select the next smaller difference, and so on.

Studentized Range Test

Multiple comparisons for ANOVA are usually performed with a modified *t*-test based on the studentized range that allows for several groups. The studentized range, attributed to Gosset, is symbolized by *q* or *Q*:

$$q_{ar} = \frac{\bar{X}_{\max} - \bar{X}_{\min}}{\sqrt{\frac{MS_W}{n}}} \left(\text{or } \frac{X_{\max} - X_{\min}}{s} \right), \text{ where the independent } X \text{ variates come from a}$$

normal distribution, \bar{X}_{\min} and \bar{X}_{\max} are the smallest and largest of the means, *n* is the number in each subgroup (assumed equal), *s* is the standard deviation, and MS_W is the within group, residual, or error mean square. This formula is very similar to the

formula for the unpaired t -test that is $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{2MS_W}{n}}}$. The only difference is $\sqrt{2}$ in the denominator. In both tests, the difference between the means (absolute effect size) is compared to a measure of variability that is standardized to 1-unit standard deviation.

The distribution of q is available from Tables or statistics programs. This distribution for mean values depends on r , the number of means (or groups), k ; the within-group degrees of freedom, ν ; and the desired value of α . The value of q is shown for selected values of k and ν for $\alpha = 0.05$ (Table 25.6).

Table 25.6 0.05 one-sided values of the studentized range value q for selected numbers of groups and degrees of freedom

Within-group degrees of freedom			
$r = \text{No. of means}$	10	20	∞
2	3.151	2.950	2.772
5	4.654	4.232	3.858
10	5.599	5.008	4.474

More complete tables or calculations can be found online at, <http://academic.udayton.edu/gregelvers/psy216/tables/qtab.htm>.

These critical values can be interpreted as follows. If there are 10 independent means from a normal distribution, and 20 Df, then the ratio q as defined before exceeds 5.008 only 5% of the time.

The pooled value of the variance is calculated from MS_W . The use of q as related to the number of groups is like the Bonferroni adjustment that requires a bigger value of t for any critical probability value.

There are two ways of utilizing the q -test. One is just like the t -test. The ratio of the difference between the means to the estimate of variability is calculated and compared with the value of q for $\alpha = 0.05$, 0.01, or some smaller value. If the ratio exceeds the critical

value of q , $\frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{MS_W}{n}}} \geq q_{0.05}$, the null hypothesis may be rejected at the given probability. The second way of performing the test is to multiply both sides of the equation by the estimate of variability $\sqrt{\frac{MS_W}{n}}$ to give $q_r \sqrt{\frac{MS_W}{n}} \leq \bar{X}_i - \bar{X}_j$. The difference between the means is known as the minimum significant difference (MSD) or the allowance. Therefore if the MSD is greater than the critical value of q multiplied by the estimate of variability, the null hypothesis may be rejected at that level of α .

Single Step Tests

Fisher introduced the least significant difference (LSD) method in which each pair of means is compared by simple t -test, the only difference being that the pooled variance is derived from SS_W . Because the LSD test is performed only after a large F ratio allows rejection of the universal null hypothesis, this test is sometimes called the protected LSD test. The number of groups does not affect the results, and the critical value is the same as that used for the ANOVA. It is symbolized by $LSD = t_\alpha \sqrt{\frac{2MS_W}{n}}$, where n is the number per subgroup (assumed equal) or if unequal, the harmonic mean of n_i and n_j , the two groups being compared. Any difference $> LSD$ allows rejection of the null hypothesis at the specified value of α . This test loses power and is not recommended if there are > 3 groups.

Similarly, the number of groups does not change the results in the honest significant difference method (HSD), introduced by Tukey: $HSD = q_\alpha \sqrt{\frac{MS_W}{n_h}}$, where i and j are the two groups being compared and n_h is the harmonic mean of n_i and n_j . That is, the HSD is the studentized t -test for the two groups being compared. (The factor of 2 that appears in the LSD formula is included in the value of q .) This test can be done online at <http://web.mst.edu/~psyworld/virtualstat/tukeys/tukeycalc.html>, http://astatsa.com/OneWay_Anova_with_TukeyHSD/_result/, and <http://vassarstats.net/anova1u.html>. The test, like the Bonferroni adjustment, keeps the experiment-wise error rate at the specified level, for example, 0.05, but at the cost of loss of power. If the sample sizes of the two groups compared are so different that averaging them is of concern, then the Tukey-Kramer test is used. In this variation, the standard error of the difference is calculated as $\sqrt{\left[\frac{s^2}{2} \left(\frac{1}{n_i} + \frac{1}{n_k} \right) \right]}$.

If the variances are very inhomogeneous, then the Welch or Games and Howell test can be used. The Games and Howell criterion is $\left(\sqrt{\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}} \right) \left(\frac{q_{\alpha k \nu_{ij}}}{\sqrt{2}} \right)$, where

$$\nu_{ij} = \frac{\left(\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j} \right)^2}{\frac{\left(\frac{s_i^2}{n_i} \right)^2}{n_i - 1} + \frac{\left(\frac{s_j^2}{n_j} \right)^2}{n_j - 1}}$$

ilar to the Welch-Satterthwaite formulas.

Looking at the data for the stellate ganglion activity example (Fig. 25.3), the means of groups A, B, and D seem to be similar and higher than the mean for group C. The differences are given in Table 25.7. The 0.05 critical value of t was 2.074.

Table 25.7 Differences among means by Tukey test

Level	Minus level	Difference	Lower 95% CL	Upper 95% CL
B	C	2.85	−0.21	5.91
A	C	2.25	−0.81	5.31
D	C	1.77	−1.40	4.94
B	D	1.08	−1.98	4.14
B	A	0.60	−2.34	3.54
A	D	0.48	−2.58	3.54

With this more conservative test, none of the differences allowed rejection of the null hypothesis. (The power of the test was low.)

The Scheffé test determines the critical value of the minimal studentized difference (MSD) not from the studentized range but from the value of $\sqrt{(k-1)F_{\alpha, (k-1)(N-k)}}$. For $\alpha=0.05$, $k=4$, and $N=25$, and F , the MSD is 3.005. This MSD is kept constant for all the comparisons and is not adjusted for different numbers of groups in that particular experiment. In the stellate ganglion example, the MSD exceeds the largest difference between the means, so that none of them allow rejection of the null hypothesis. These are the same conclusions as reached by the Tukey test. The Scheffé test is excellent for complex comparisons (e.g., the average of means of groups 1 and 2 vs the average of means of groups 3, 4, 5) but is less powerful than the other tests for paired comparisons. The Scheffé test can be performed online at http://www.statstodo.com/UnpairedDiff_Pgm.php, http://www.statstodo.com/UnpairedDiff_Pgm.php, and <https://www.easycalculation.com/statistics/one-way-anova-scheffe-method.php>.

Multiple Step Tests

The loss of power was partly addressed by the Newman-Keuls method, also known as the Student-Newman-Keuls (SNK) method. Rank the means in order of size and compare the largest and smallest means of the k groups. If these do not allow rejection of the null hypothesis by the studentized range test (HSD test), then any lesser differences are ignored. If, however, the two means with the greatest difference exceed the allowance, then compare two means that differ by a range of $k-1$. For example, if the means are 10, 11, 14, 18, 25, each with $n=5$, and $s=30$, then compare 10 and 25 over a range of 5 groups. With 20 Df for the within-group mean square the value of q is 6.12, and the 0.05 level is 4.23, so we might reject the null hypothesis. Then compare two pairs of means over the range of 4 groups: 11 with 25, and 10 with 18. Each of these is compared with the critical value that is now 3.96. These critical values are based on tables that relate degrees of freedom to the size of the range. As the range gets bigger, so does the critical value, ensuring that the Type I error is not exceeded.

For the stellate ganglion example, the means were in order 5.847 (B), 5.376 (D), 5.244 (A), and 2.997 (C). Test first B vs C, with difference 2.850. The studentized range table for degrees of freedom 4, 21 gives a critical MSD of 3.944. Because this exceeds the observed difference, we should not reject the null hypothesis and there is no need to test smaller differences. The SNK test has relatively low power and is no longer favored.

A similar test by Duncan uses different critical values, and according to some authorities gives too many Type I errors.

To show an example where several pair-wise differences are shown, take the stellate ganglion data of Fig. 25.3 and add a hypothetical fifth group E (Fig. 25.5).

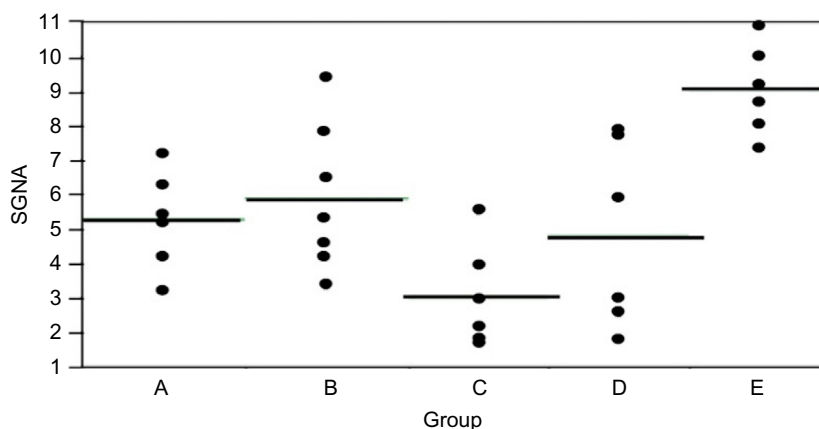


Fig. 25.5 Stellate ganglion data with one added group.

The ANOVA is presented in Table 25.8.

Table 25.8 ANOVA with added group

Source of variation	SS	Df	MS	F
Total (SS_T)	211.0440	31		
Within (SS_W)	94.8825	27	3.5412	
Between (SS_B)	116.1615	4	29.0404	8.2638
				$P=0.0002$

Now the F ratio is very high. The results of the LSD and HSD tests are usually set out as tables with mean differences, upper and lower confidence limits of the differences, and sometimes the P values for the comparisons (Table 25.9).

Table 25.9 LSD and HSD tests

Level	–Level	Difference	Lower 95% CL	Upper 95% CL	P
E	C	6.02	3.80 2.86	8.24 9.18	<0.0001
E	D	4.25	2.0 1.09	6.47 1.41	0.0005
E	A	3.77	1.63 0.73	5.91 8.82	0.0012
E	B	3.17	1.03 0.12	5.31 6.22	0.0052
B	C	2.85	0.71 –0.20	4.99 5.90	0.0109
A	C	2.25	0.11 –0.80	4.39 5.29	0.0402
D	C	1.77	–0.45 –1.39	3.99 4.93	0.1146
B	D	1.08	–1.06 –1.97	3.22 4.13	0.3094
B	A	0.50	–1.45 –2.32	2.66 3.53	0.5524
A	D	0.48	–1.66 –2.57	2.62 3.52	0.6506

The HSD test limits are shown in bold type under the LSD tests. For each comparison, the confidence limits (CL) are wider for HSD than LSD.

For the LSD test group E is different from all the other groups (as intended by the artificial construction) but now both groups A and B differ from group C. Adding new data has altered some of the values used in the comparison. For the HSD test group E differs from all the other groups, and no other comparisons allow rejection of the null hypothesis.

What does the SNK test indicate? The means in order of size are 9.017 (E), 5.847 (B), 5.244 (A), 4.767 (D), and 2.997 (C). Compare first groups E and C, having the greatest difference and a span of 5. If that allows rejection of the null hypothesis (as it will from the Tukey test), then compare E with D, the second smallest, and then B, the second largest, with C, the smallest; each of these comparisons has a span of 4. Then compare means with a span of 3: E vs A, B vs D, and A vs C. The values of $q_{0.05}$ vary with the span and the within-group degrees of freedom, which here are 27. For a span of 5 $q = 4.134$, for a span of 4 it is 3.873, and for a span of 3 it is 3.509. Calculate the standard error if $n_i = n_j$ as $\sqrt{\frac{MS_w}{n_1}}$, and if they are not equal as $\sqrt{\frac{MS_w}{2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$. With these considerations, set up [Table 25.10](#).

It makes no difference if we compare the MSD with the difference between the means or, as in some texts, compare the critical value of q with the difference divided by the standard error.

Table 25.10 Student-Newman-Keuls results

Comparison	$\bar{X}_i - \bar{X}$	SE	k	$q_{0.05, k, 27}$	MSD	H_0
E vs C	6.02	0.7653	5	4.134	3.164	R
E vs D	4.25	0.7653	4	3.873	2.964	R
B vs C	2.85	0.7375	4	3.873	2.856	DNR
E vs A	3.773	0.7375	3	3.509	2.588	R
B vs D	1.08	0.7375	3	3.509	2.588	DNR
A vs C	2.247	0.7375	3	3.509	2.588	DNR
E vs B	3.17	0.7375	2	2.903	2.141	R
B vs A	0.603	0.7085	2	2.903	2.057	DNR
A vs D	0.477	0.7375	2	2.903	2.141	DNR
D vs C	1.77	0.7653	2	2.903	2.222	DNR

DNR, do not reject H_0 at $\alpha=0.05$; R, reject H_0 at $\alpha=0.05$; MSD, $SE \times q_{0.05, k, 27}$.

Therefore with this test reject the null hypothesis for comparisons E vs C, E vs D, E vs B, and E vs A. None of the other differences allow rejection of the null hypothesis.

The differences among the various multiple comparison methods are shown in Fig. 25.6.

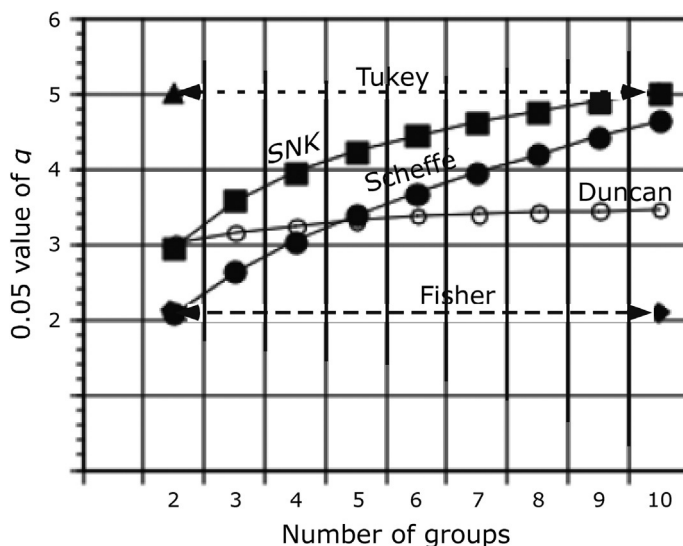


Fig. 25.6 Relation of 0.05 value of q to the number of groups for five multiple comparison tests when the error degrees of freedom are 20.

The Tukey HSD and the Fisher LSD tests are fixed range tests. (So is the Scheffé test, but its factor changes with the total number of groups tested.) The Tukey test is the most conservative and holds the experiment-wise error rate constant. The Fisher LSD test is the most liberal of all these tests and is intended to keep the comparison-wise error rate constant. However, if there are >3 groups, the LSD test sometimes allows the

Type I error rate to rise well above 0.05. To reduce this risk, Hayter used the studentized range q value for $k - 1$ means or all paired comparisons. Because q is bigger than t , the Type I error rate is lower. The Student-Newman-Keuls (SNK) and Scheffé tests start at small values for q and then increase toward the Tukey value as the number of groups increases. They both hold the experiment-wise error rate constant. Duncan's test also has an increased value of q at higher group numbers but the increase is small and regarded by many statisticians as too liberal, that is giving too many Type I errors.

Recommendations

The best advice was given by Dallal (2008). If a difference between means allows rejection of the null hypothesis by Tukey's HSD test, then it probably is correct. If it does not allow rejection of the null hypothesis by Fisher's LSD test, then it does not warrant further investigation. Anything between the two limits should be investigated further. This is one of the penalties paid for doing unplanned tests.

One possibility to be considered is the Hochberg test (Chapter 24) that has good power. An even better test is the extension of the Ryan test. The SNK test is more powerful than the Tukey test, but does not keep the family-wise error rate low as the number of groups increases. To overcome this, Ryan introduced a way to keep the error rate low and constant by entering the SNK tables with $\alpha_r = \frac{\alpha_{0.05}}{k/r}$, where k is the number of groups and r is the span of the means being compared. If there are 6 groups and we are comparing the third and fifth means, $k=6$ and $r=2$. This basic method was modified to provide the Ryan-Einot-Gabriel-Welsch (R-E-G-W) test in which $\alpha_r = 1 - (1 - \alpha)^{r/k}$. Note the similarity between these corrections and the Bonferroni and Dunn-Sidak approaches. The test is best done by a computer program.

To show the difference between the SNK and the R-E-G-W tests, Table 25.10 is repeated as Table 25.11 with the addition of the Ryan factors:

Table 25.11 Comparison of SNK and Ryan factors (highlighted)

Comparison	$\bar{X}_i - \bar{X}_j$	SE	k	$q_{0.05, k, 27}$	MSD	Ryan q'	MSD'	H_0
E vs C	6.02	0.7653	5	4.134	3.164	4.13	3.16	R
E vs D	4.25	0.7653	4	3.873	2.964	3.87	2.96	R
B vs C	2.85	0.7375	4	3.873	2.856	3.87	2.85	DNR
E vs A	3.773	0.7375	3	3.509	2.588	3.828	2.82	R
B vs D	1.08	0.7375	3	3.509	2.588	3.828	2.82	DNR
A vs C	2.247	0.7375	3	3.509	2.588	3.828	2.82	DNR
E vs B	3.17	0.7375	2	2.903	2.141	3.487	2.57	R
B vs A	0.603	0.7085	2	2.903	2.057	3.487	2.47	DNR
A vs D	0.477	0.7375	2	2.903	2.141	3.487	2.57	DNR
D vs C	1.77	0.7653	2	2.903	2.222	3.487	2.67	DNR

Column headed $q_{0.05, k, 27}$ shows the factors from the SNK table; Column headed Ryan q' gives $q_{0.05}$ values modified for total number of groups and is a better way of keeping the family-wise error rate near 0.05. $MSD = SE \times q_{0.05, k, 27}$; $MSD' = SE \times q'$. The Ryan factors increase the size of the MSD when subgroups closer together are compared. H_0 , null hypothesis; R, reject; DNR, do not reject.

Linear Combinations

By definition, a linear combination L is

$$L = \lambda_1 \bar{X}_1 + \lambda_2 \bar{X}_2 + \dots \lambda_n \bar{X}_n,$$

where the λ_s are fixed numbers. The linear combination is called a comparison or contrast if $\sum \lambda_i = 0$. L has a standard error $\sqrt{\sum \lambda_i^2 \left(\frac{s_w^2}{n}\right)}$.

Here n is the number of observations per group. To see how this information can be used, consider the unpaired t -test that examines the difference between \bar{X}_i and \bar{X}_j . This compares one whole part of \bar{X}_i with one whole part of \bar{X}_j for the null hypothesis $\bar{X}_i - \bar{X}_j = 0$, so that the coefficients are $+1$ and -1 . Because $1 - 1 = 0$, this is a contrast in which the standard error is $\sqrt{2\left(\frac{s_w^2}{n}\right)}$, and this is the same as the denominator in the standard t -test (if $n_1 = n_2$).

The advantage of this concept is that it allows us to explore contrasts among more than two groups. For example, in an experiment on yields of different wheat varieties (discussed in detail in [Chapter 26](#)), assume that there were three varieties of winter wheat (W) with mean yields of 22.5, 18.83, and 20.62 bushels/acre, and that the summer wheat (S) yielded 32.83 bushels/acre; each had 6 replications. How different is the summer yield from the average of the three winter yields? The null hypothesis is

$$\bar{S} - \frac{\bar{W}_1 + \bar{W}_2 + \bar{W}_3}{3} = 0.$$

From the ANOVA ([Chapter 26](#)) the within group or residual mean square was 76.225, so that its square root was 8.73; there were 20 degrees of freedom. Then the standard error of the linear contrast is

$$\sqrt{1^2 + \left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^2 \left(\frac{8.73}{6}\right)} = 1.68.$$

The observed difference was $32.83 - \frac{22.5 + 18.83 + 20.62}{3} = 12.18$.

Therefore $t = \frac{12.18 - 0}{1.68} = 7.25$ $P = 0.000001$ (two-tailed), and we can reject the null hypothesis.

Why not use the Bonferroni adjustment to make these multiple comparisons? The maximal Bonferroni adjustment for k groups involves $\frac{k(k-1)}{2}$ comparisons if all pairwise comparisons are to be tested, even if after inspecting the data fewer comparisons are actually performed, and even if initially a subset of comparisons was planned but more comparisons were done after the results were obtained. This makes the Bonferroni adjustment less powerful than the alternative methods used before.

Planned Experiments

In planned experiments, certain comparisons are designated before the experiment is done. The investigator tests certain specific hypotheses. For these a priori hypotheses, and only for these, once an F ratio above the critical ratio has been found, t -tests may be done. Returning to the stellate ganglion ablation experiment of Tan et al. (2008), the investigators a priori wanted to compare stellate ganglion activity on day 1 in dogs with and without ablation, on day 10 in dogs with and without ablation, and on day 1 vs day 10 in dogs after ablation. By t -test these differences had respective probabilities of 0.0157, 0.4527, and 0.1998.

Why not designate every possible pair-wise combination a priori, and do t -tests on all the pairs of means, thus avoiding all the problems discussed before for unplanned experiments? Firstly, such a suggestion is illegal because it negates all the arguments that multiple comparisons inflate the Type I error. Secondly, comparing every mean against every other mean reuses information. If we compare A with B and A with C, we have used the information in all three groups, and then comparing B with C is repetitious. How then do we know how many t -tests may legitimately be done?

Formally, if we want to know if two comparisons are orthogonal (independent), we consider the coefficients of each comparison. In an example about the growth of sugar beets planted at four different combinations of times and ways (Snedecor and Cochran, 1989, p. 229) there were two sets of comparisons. The first one compared the average of combinations 2 and 3 with the average of combinations 1 and 4:

$\frac{1}{2}X_2 + \frac{1}{2}X_3 - \frac{1}{2}X_1 - \frac{1}{2}X_4$ that provided the coefficients for calculating L . The second comparison was the average of combinations 2 and 3 against combination 4 alone:

$\frac{1}{2}X_2 + \frac{1}{2}X_3 - 1X_4 + 0X_1$, in which the yield with no fertilizer (X_1) had a zero coefficient because it was not part of the comparison. Now multiply the corresponding coefficients, and add up the results:

$$\left(\frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{1}{2}\right) - \left(\frac{1}{2} \times 1\right) - \left(\frac{1}{2} \times 0\right) = 0.$$

Because these products sum to zero, the two comparisons are orthogonal and both can safely be tested. If an experiment includes several groups, there may be many ways in which an orthogonal set can be constructed. The experimenter must decide which set to choose.

Often an experiment is done in which several groups are each compared with a control. For example, an investigator may wish to find out if BNP is elevated in a number of different forms of heart disease. Each subgroup—rheumatic, several forms of congenital heart disease, ischemic, cardiomyopathic, and so on—has BNP measured and compared with the normal control; the a priori plan is to make certain selected comparisons. After a

critical F ratio is reached for the whole test, each group is compared with the control by Dunnett's test. This is essentially a form of t or q test in which the critical values are determined by a special set of numbers that are in between t and q . For example, with 5 groups and 20 degrees of freedom within groups, the 0.05 value is 4.232 for q , 2.65 for Dunnett's test, and 2.086 for t . By restricting the comparisons to a specific subset, Dunnett was able to increase the sensitivity of the test without paying the penalty incurred by making an all pairs comparison. Critical values for Dunnett's test may be found online at http://www.stat.ufl.edu/~winner/tables/dunnett_1side.pDf, and http://davidmlane.com/hyperstat/table_Dunnett.html.

In any form of a priori test, including Dunnett's test, the investigator may be struck by some unexpected differences between the means of two groups that were not involved in the a priori planning. For example, in comparing each subgroup with a control group, the original purpose of the study, the investigator may note a large difference between the mean values for BNP in, for example, patients with a large ventricular septal defect and with a large atrial septal defect. These two means can validly be compared by one of the post hoc tests mentioned before for a posteriori experiments. The same applies to those components of an a priori experiment that are not orthogonal and therefore cannot be compared validly by t -test.

In planning a study with one control and several treatment groups, the control sample must be large enough to reflect the control population because that control sample is to be used in all the subsequent comparisons, and if the control group is unrepresentative, all comparisons with the treatment groups will be in jeopardy. Given enough time and money, the control group can be made very large, but often the total number of subjects in the experiment is limited. Assume that there are N total subjects to be divided into k treatment groups, each replicated r times, and one control group replicated ar times: $N = ar + kr$. Then the lowest residual error is produced if $a = \sqrt{k}$ so that the number of control replicates is \sqrt{kr} . As an example, with $N = 30$ subjects, and $k =$ four treatment groups and a control, then $\sqrt{k} = 2$, and we need twice as many replicates in the control as in each treatment group. Therefore we need 10 control replicates and 5 replicates in each treatment group.

As shown in some of the figures before, it is possible (and desirable) to calculate confidence limits for the differences between pairs of means. This allows more careful interpretation of the data.

Computation of power is necessary when planning an ANOVA and can also be done post hoc when the results are available. This computation is more complicated than when comparing two groups and is best done by computer programs.

REFERENCES

- Abdi, H., 2007. O'Brien test for homogeneity of variance. In: Salkind, N. (Ed.), Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA. <http://www.utdallas.edu/~herve/Abdi-OBrien2007-pretty.pdf>.

- Bartels, R., 1982. The rank version of von Neumann's ratio test for randomness. *J Amer Stat Assoc* 77, 40–46.
- Cohen, J., 1988. *Statistical Power Analysis for Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Cohen, J., 1992. A power primer. *Psychol. Bull.* 112, 155–159.
- Curran-Everett, D., 2000. Multiple comparisons: philosophies and illustrations. *Am J Physiol Regul Integr Comp Physiol* 279, R1–R8.
- Dallal, G.E., 2008. Multiple Comparison Procedures. Available: <http://www.jerrydallal.com/LHSP/mc.htm>.
- Dunn, O.J., 1964. Multiple comparisons using rank sums. *Technometrics* 6, 241–252.
- Emerson, J.D., Stoto, M.A., 1983. Transforming data. In: Hoaglin, D.C., Mosteller, F., Tukey, J.W. (Eds.), *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons, New York.
- Glantz, S.A., 2005. *Primer of Biostatistics*. McGraw-Hill, New York.
- Grissom, R.J., 2000. Heterogeneity of variance in clinical data. *J. Consul. Clin. Psychol.* 68, 155–165.
- Jordan, J., Shannon, J.R., Black, B.K., et al., 2000. The pressor response to water drinking in humans : a sympathetic reflex? *Circulation* 101, 504–509.
- Lix, L.M., Keselman, H.J., 1998. To trim or not to trim: tests of location equality under heterocedasticity and nonnormality. *Education Psychol Measurement* 58, 409–429.
- Richardson, D., Catron, D.V., Underkofler, L.A., et al., 1951. Vitamin B₁₂ requirement of male weanling pigs. *J. Nutr.* 44, 371–381.
- Savman, et al., 1998. Cytokine response in cerebrospinal fluid after birth asphyxia. *Pediatr Res* 43 (6), 746–751.
- Snedecor, G.W., Cochran, W.G., 1989. *Statistical Methods*. Iowa State University Press, Ames, IA.
- Sokal, R.R., Rohlf, F.J., 1995. *Biometry*. In: *The Principles and Practice of Statistics in Biological Research*. W.H. Freeman and Company, New York.
- Tan, A.Y., Zhou, S., Ogawa, M., et al., 2008. Neural mechanisms of paroxysmal atrial fibrillation and paroxysmal atrial tachycardia in ambulatory canines. *Circulation* 118, 916–925.
- Van Belle, G., 2002. *Statistical Rules of Thumb*. Wiley Interscience, New York.
- Zhang, S., 1998. Fourteen homogeneity of variance tests: when and how to use them. In: *Annual Meeting of the American Educational Research Association*. San Diego, CA, April 13–17, 1998.

CHAPTER 26

Analysis of Variance. II. More Complex Forms

BASIC CONCEPTS

Introduction

The one factor ANOVA type of analysis is termed model I, or fixed effects analysis, because the different treatments have potentially different effects. The study of the amount of 4 different levels of vitamin B12 on piglet growth tested the hypothesis that one or more of these levels has a greater effect than the others. Each measurement is made on a randomly drawn piglet, and the measurements are independent of each other. Increasing the sample size in each group assumes that the treatment effects remain the same with the means remaining similar in each group but with larger samples giving smaller standard errors and hence making it easier to detect differences among the means. Model II is discussed later in this chapter.

The one-way ANOVA of [Chapter 25](#) is analogous to the unpaired *t*-test. What, then, is the ANOVA analog of the paired *t*-test? In its simplest form it is the two-way (or two factor) analysis of variance, two-way because two different categories are tested simultaneously. Two designs are possible. In one, the treatment factors may be categorical (e.g., different medications or sites), or else ratio or interval numbers (e.g., dosages). Each group in the second factor is tested against each treatment factor, but a different set of subjects is used for each group \times treatment combination (see [Table 26.1](#)). These are termed between-subject designs. In the other, the treatment factor is usually time (but could be a variable such as dosage or type of drug), the *same* subject being measured at designated time intervals; these are termed within-subject designs. The term “repeated measures” is usually applied to the latter design. The difference is that in the second design, the same subject is measured several times, whereas in the first design different samples from the given subject or group are analyzed.

The standard two-factor analysis is the randomized complete block design, based on the agricultural studies of R.A. Fisher. Soils in which crops are grown often have different growth potentials that may affect the conclusions of experiments on, for example, different fertilizers. There are two ways of dealing with this problem. One is to randomize the plots of ground so that each treatment (fertilizer) is given to soils with many different growth potentials. This is the standard one-factor ANOVA. The other way is to select blocks of ground with similar growth potential (based on prior knowledge) and allocate

the treatments at random to smaller plots within the block. This is equivalent to pairing in the t -test and has the same effect of reducing variability. The gain in efficiency can be large (Glantz and Slinker, 2001, pp. 409–10). The term blocking is used not in the sense of obstructing but in the sense of creating homogeneous sets (blocks).

Little new is needed for such a test. Graybill used data from a wheat varietal test performed by the Oklahoma Agricultural Experimental Station (Graybill, 1954). Some of the data are used here. Four varieties of wheat are grown in six different locations (Table 26.1).

Table 26.1 Experiment with different locations (factor 1; treatment) and different wheats (factor 2; subjects)

Wheat	Location						ΣX_r	N	\bar{X}
	1	2	3	4	5	6			
A	44	40	18	20	55	20	197	6	32.83
B	24	22	14	19	34	22	135	6	22.5
C	19	17	17	18	19	23	113	6	18.83
D	19	24	16	18	25	22	124	6	20.67
ΣX_c	106	103	65	75	133	87	569	24	23.7083
N	4	4	4	4	4	4	24		
\bar{X}	26.5	25.75	16.25	18.75	33.25	21.75			

Subscript r = row; subscript c = column.

Our object is to answer the following questions:

1. Does the mean yield (bushels/acre) differ for different types of wheat?
2. Does the mean yield differ for different locations?
3. Do the two factors interact, that is, is the yield of a particular wheat increased for one location but decreased for another?

Although these calculations are done best with statistical programs, they are described here to show what the programs are doing.

Online programs for performing two-way ANOVA may be found at <http://vassarstats.net/> (see ANOVA) <http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/ANOVATwo.htm>, <http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/ANOVA2Rep.htm>, (with replications), http://www.wessa.net/rwasp_Two%20Factor%20ANOVA.wasp, but these are of limited use.

Two-Way ANOVA

Table 26.1 presents the data.

For the moment, ignore factor 2 (wheats), and perform a one-way ANOVA on the locations (Fig. 26.1).

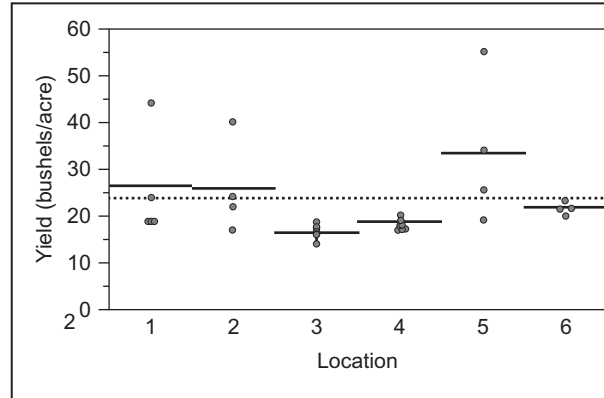


Fig. 26.1 ANOVA for locations. *Short horizontal bars are the means. Dotted line is the mean of all the measurements.*

The ANOVA table is presented in [Table 26.2](#).

Table 26.2 ANOVA table for locations

Source of variation	SS	Df	MS	F
Total (SS_T)	2230.9583	23		
Within (SS_W)	1482.75	18	82.375	
Between (SS_{BL}) (location)	748.2083	5	149.642	1.8166 $P=0.1603$

The ANOVA confirms that there were some differences among the means with $P > 0.05$. The residual error SS or SS_W was 1482.75.

Because we have more information, we know that this SS_W consists of two portions, one due to differences between subjects (wheats) and the true residual error. We compute the between-subject SS in the usual way and obtain [Table 26.3](#).

Table 26.3 Partitioning within group SS_W

Source of variation	SS	Df	MS	F
Within groups (SS_W)	1482.75	18		
Between (SS_{BV}) (variety)	706.4583	3	235.486	4.65
Residual variability (SS_{res})	776.2917	15	51.7528	

We can now put together the final ANOVA, taking into account both factors ([Table 26.4](#)).

We have partitioned the total SS into a component due to differences between locations (SS_{BL}), a component due to differences between varieties (SS_{BD}), and the residual SS_W . Note how removing components related to varieties and locations has reduced the within-group sum of squares.

Table 26.4 Final two-way ANOVA

Source of variation	SS	Df	MS	F
Total (SS_T)	2230.9583	23		
Between (SS_{BL}) (location)	748.2083	5	149.642	2.8914 $P=0.0505$
Between (SS_{BV}) (variety)	706.4583	3	235.486	4.5502 $P=0.0185$
Within (SS_W)	776.2917	15	51.7528	

The model for a two-way or two-factor ANOVA is that the value of any measurement X_{ij} depends only on the mean value of X_i , the mean value of X_j (i.e., on the means of each factor) and the residual error ε_{ij} :

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij},$$

where $\alpha_i = \bar{X}_i - \mu$ and $\beta_j = \bar{X}_j - \mu$.

The residuals are assumed to be normally distributed, and the two factors to be independent. In other words, if wheat variety 1 has a large yield in location 2, all the other varieties should produce a large yield at location 2. A test of this assumption, known as additivity, is given later.

The conclusion is that we might reject the null hypothesis for differences between varieties, but differences between location are borderline. This procedure involves performing a one-way ANOVA, but then taking the residual variation and separating it into two portions: the variation associated with differences between subjects and the “pure” residual variation. The procedure is diagrammed in Fig. 26.2.

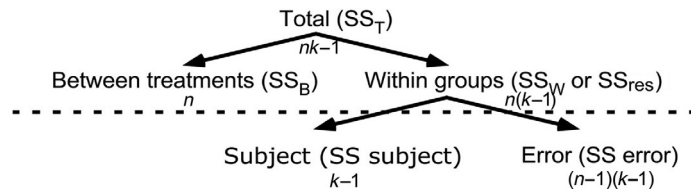


Fig. 26.2 Typical two-way ANOVA. The portion above the *dashed line* is a one-way ANOVA. The degrees of freedom for each portion are shown just below the respective heading. n —number per group (assumed equal); k —number of groups.

If different subjects are used for each treatment, this is a two-way ANOVA. If the same subjects are used for each treatment, this is a repeated measures ANOVA that needs added calculations. More complex designs can be dealt by suitable partitioning.

The two-way ANOVA has considerable value. Not only does it reduce the residual (nonspecific) variability, thereby making the F -test more sensitive, but it reduces the amount of work involved. If we just tested 4 varieties of wheat, each with 6 replicates,

that comes to 24 plots. If we studied just locations, each with 4 replicates, that comes to another 24 plots. Combining the two into one analysis halves the work and makes the comparisons more sensitive. Furthermore, if animals or patients are used, a one-way ANOVA is an expensive and unethical use of more animals or patients than needed for the experiment. Note, too, that by reducing the residual error, the difference between locations now has a smaller P value.

The two-way ANOVA has similar requirements to the paired t -test.

1. Ratio or interval numbers
2. Normal distributions (hard to tell if sample sizes are small)
3. Homogeneous variances (hard to tell if sample sizes are small)
4. Justification for pairing.

This last point leads us to another requirement, namely, additivity, the cardinal feature of the model described before.

Additivity

To test additivity, calculate the value of ε_{ij} for each value of X_{ij} (Tukey, 1977).

The model is $X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$, where α_i is the difference between the grand mean μ and the mean of the first factor and β_j is the difference between the grand mean μ and the mean of the second factor. With perfect additivity, all the residuals will be zero; in practice, additivity is confirmed if all the residuals ε_{ij} are small.

Return to the original data (Table 26.1), and subtract the row mean value (last column) from each of the six measurements in that row for each location separately (Table 26.5). Then calculate the mean deviations for each location as shown in the last row.

Table 26.5 Additivity stage 1

Location (last row)							
Wheat	1	2	3	4	5	6	\bar{X}
A	$44 - 32.83 = 11.17$	7.17	-14.83	-12.83	22.17	-12.83	32.83
B	$24 - 22.5 = 1.5$	-0.5	-8.50	-3.50	11.50	-0.17	22.50
C	0.17	-1.83	-1.83	-0.83	0.17	4.17	18.83
D	-1.67	3.33	-4.67	-2.67	4.33	1.33	20.67
\bar{X}	2.79	2.04	-7.46	-4.96	9.54	-1.87	23.7083

This procedure removes the mean variability for wheats from each location measurement, leaving the effect of each given location on this residual. Subtract the means of these new values (last row in Table 26.5) for each variety (as shown in the table) from the residual values (Table 26.6).

Table 26.6 Additivity stage 2

	Variety A	Variety B	Variety C	Variety D
Location 1	$11.17 - 2.79 = 8.38$	$-0.68 - 2.04 = -2.72$	-0.12	-4.81
Location 2	$-1.5 - 2.70 = -4.29$	-0.11	-8.83	0.94
Location 3	-7.375	-2.47	15.59	2.44
Location 4	-7.875	0.03	-0.93	2.00
Location 5	12.625	0.53	0.38	-5.56
Location 6	-10.875	0.03	16.00	2.94

Final table of residuals.

Ideally, there should be as many positive as negative differences; in the table there are 10 negatives and 14 positives. Furthermore, most difference should be small, and this is not true here.

Additivity can be checked formally by Tukey's test that is implemented by some commercial programs. It is tedious but not difficult to do by hand, and it can be done with the free program BrightStat at <https://secure.brightstat.com/index.php?p=8&e=1>. This program confirms nonadditivity, with $P < 0.0001$.

What if the additivity assumption is unlikely? Check the residuals to determine where the discrepancies might be (outliers), and also consider a transformation, possibly logarithmic or square root. One general type of transformation is the Anscombe-Tukey transformation, a power function $X' = X^{\text{power}}$ that frequently produces acceptable additivity without having to eliminate outliers. For this example, a power of -2.9 restores additivity.

Sometimes we may want to make serial measurements on one of the factors. It is common in research to take several subjects and then measure a given variable, for example, blood pressure, at different times after each subject is given a particular drug. Subject 1 is given amlodipine, at another time atenolol, at another time losartan, and so on. The other subjects are also given each of these agents. The benefit to such a design is that each subject is his or her own control, thereby making this equivalent to a paired t -test with the gain in sensitivity that a paired test has over an unpaired test. The subjects are the blocks, and we know that different subjects react differently to these agents. In addition, for a 4×6 two-factor design, with 4 drugs we need only 6 patients and not 24 patients. If such a design is used, care must be taken to avoid a carry-over effect of a previous treatment or any effect related to a given order of treatments. Apart from pharmacological carry over, order effects can occur for psychological reasons, so that treatments should therefore be given in random order to different patients. The procedures are described online at <https://statistics.laerd.com/statistical-guides/repeated-measures-anova-statistical-guide.php>. These calculations for repeated measures samples can be done online at <http://vassarstats.net> (see ANOVA), and for 3 groups of up to 40 in each group at <http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/ANOVADep.htm>.

To summarize, instead of comparing treatments with a one-way ANOVA, allocate subjects (patients, animals) at random to subgroups (or blocks) that are expected to be homogeneous. For example, several litters of animals may be chosen, and within each litter 4 animals are selected at random to receive one of the treatments. The calculations show differences between treatments, differences between litters (blocks), and residual variability. This makes testing the differences among treatments more sensitive. The differences among blocks may or may not have meaning. Often, as in the example of different litters, the differences are random variables with a mean of zero and a given mean square (see following for model II). On the other hand, it would be possible to do an experiment in which each block represented a different strain or breed, and then differences among blocks may be of interest.

Multiple Factors

More than two factors can be analyzed by using Latin Squares; the name was based on an early mathematical puzzle. The method was introduced into statistical analysis for use in agricultural experiments. To allow for possible underlying differences in soil fertility, the plot of land is divided into smaller blocks to which seeds and treatments (fertilizers) are allocated at random, but in such a way that each treatment occurs once in each row and each column, and each seed occurs once in each row and each column. There are two blocking factors. This principle can be applied to biological experiments as well ([Finney, 1955](#)).

It is possible to do an ANOVA with 4 factors, by allocating the fourth factor (symbolized by Greek letters) once to each of the other 3 factors. This is known as a Graeco-Latin square and is analyzed in similar fashion. Latin squares must have the same number of cells for each factor. The requirements for ratio numbers, normality, and homogeneity of variance apply equally to these more complex ANOVA.

Friedman Test

If there is concern about the normality or variability of the distributions, or if ordinal numbers are involved, then a straightforward two-way ANOVA might yield misleading results. Then a distribution-free test, the Friedman test (developed by Milton Friedman, the economist), can be used. The data table is set up with the columns representing the j factors and the k rows representing the subjects. The measurements for each subject are then ranked 1, 2, 3, and so on, and the mean rank for each factor is calculated. Then a modified ANOVA is done to calculate the SS between ranks. The theoretical mean rank with a random distribution is $(j+1)/2$; for 3 factors this is 2.0. Then for each factor calculate k (observed mean rank-theoretical mean rank)², and the sum of these is the SS between ranks. This is then referred to the chi-square table with $j-1$ degrees of freedom.

An alternative formulation involves summing the ranks in each column (S_i) and the Friedman statistic is calculated as

$$T = \frac{12 \sum S_i^2}{jk(j+1)} - 3k(j+1).$$

This is then assigned a probability from a Table or program. With 3 treatments and >7 subjects, or 4 treatments and >4 subjects, then the distribution of T is similar to that of chi-square for 2 or 3 degrees of freedom, respectively. Online calculations for 3 or 4 factors (columns) can be found at <http://vassarstats.net/> (see Ordinal data) or at http://www.statstodo.com/FriedmanTwoWayAnalysisOfVariance_Pgm.php.

Once the null hypothesis is rejected for the whole test, which groups are different? There are nonparametric equivalents of the Student-Newman-Keuls (SNK) and Dunnett tests. The SNK equivalent for all the pairwise comparisons is

$$q = \frac{R_i - R_j}{\sqrt{\frac{kN(k+1)}{12}}}$$

where k is the number of groups, N is the total number, and R_i and R_j are the rank sums for the groups being compared.

For Dunnett's test, replace R_i by R_{control} in the previous formula, and divide by 6 instead of 12. The critical values are presented online at http://davidmlane.com/hyperstat/table_Dunnett.html.

Cochrane's Q-Test

This is a nonparametric test (Cochran and Bliss, 1970) to answer the question about whether two or more treatments are equally effective when the data are dichotomous (Binary: yes, no) in a two-way randomized block design. It is equivalent to the Friedman test with dichotomous variables. The results are set out as in Table 26.7.

H_0 : all treatments are equally effective.

H_A : There is a difference between some of the treatments

Table 26.7 Cochran's Q-test

Block	Treatment 1	Treatment 2	...	Treatment m	Sum
Block 1	X_{11}	X_{12}	...	X_{1m}	$\sum_{i=1}^m X_{1i}$
Block 2	X_{21}	X_{22}	...	X_{2m}	$\sum_{i=1}^m X_{2i}$
...
Block n	X_{n1}	X_{n2}	...	X_{nm}	$\sum_{i=1}^m X_{ni}$
Sum	$\sum_{i=1}^n X_{i1}$	$\sum_{i=1}^n X_{i2}$...	$\sum_{i=1}^n X_{im}$	N

m , number of treatments; n , number of blocks; N , grand total; X_{ij} is the column total for the j -th treatment; $X_{i\bullet}$ is the row total for the i -th block.

$$Q = m(m-1) \frac{\sum_{j=1}^m \left(X_{\bullet j} - \frac{N}{m} \right)^2}{\sum_{i=1}^n X_{i\bullet} (m - X_{i\bullet})}.$$

The critical value of Q is

$$Q > \chi^2_{1-\alpha, m-1}$$

If H_0 is rejected, then pairwise comparisons of treatment can be done.

As an example, consider testing three different medications (treatments) in six different patients, and recording the results as improved (0) and unimproved (1) (Table 26.8).

Table 26.8 Example of Cochran's Q -test

Block	Treatment 1	Treatment 2	Treatment 3	Sum
Patient 1	1	0	0	1
Patient 2	0	0	1	1
Patient 3	0	0	0	0
Patient 4	1	0	1	1
Patient 5	0	1	1	2
Patient 6	0	0	1	1
Sum	2	1	4	7

Then

$$\begin{aligned}
 Q &= 3(3-1) \frac{\left(2 - \frac{7}{3}\right)^2 + \left(1 - \frac{7}{3}\right)^2 + \left(4 - \frac{7}{3}\right)^2}{1(3-1) + 1(3-1) + 0(3-0) + 2(3-2) + 2(3-2) + 1(3-1)} \\
 &= \frac{6 \times (0.11 + 1.78 + 2.78)}{2 + 2 + 0 + 2 + 2 + 2} = \frac{28.02}{10} = 2.80
 \end{aligned}$$

With $3-1=2$ degrees of freedom, this chi-square of 2.80 has a probability of 0.25, so that there is no evidence provided against the null hypothesis that the treatments are similar.

Because all the results are either 1 or 0, the arithmetic is trivial. An online calculator is <http://scitstatcalc.blogspot.com/2013/12/cochrans-q-test-calculator.html#>.

Interaction

In the study such as the wheat study there is a possibility that one wheat might grow more in one location than another, so that there would not have been additivity. This form of dependence is termed interaction. To illustrate the effect of interaction, Fig. 26.3 shows results of a hypothetical experiment to determine the effects of two drugs (A and B) for lowering blood pressure on 20 patients in each of three different racial groups (a, b, and c).

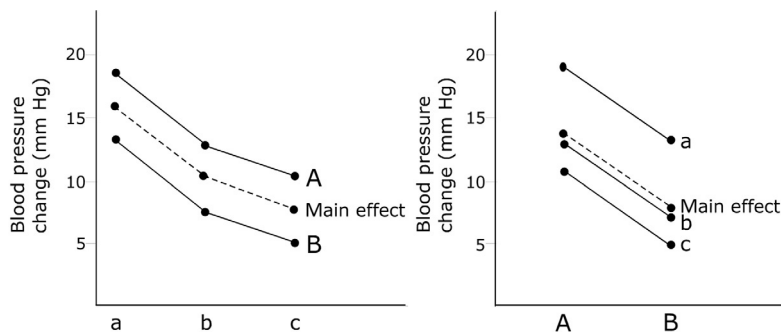


Fig. 26.3 3×2 ANOVA with no interaction.

The left-hand panel shows that group a has a larger change in blood pressure than does group b and this in turn exceeds the change in group c. The average change of blood pressure for the two drugs is shown by the dashed line labeled “main effect.” In group a the average decrease in pressure is about 16 mmHg, in group b about 11 mmHg, and in group c about 8 mmHg. The right-hand panel shows a larger change in blood pressure for drug A than for drug B. The main effect that averages out differences due to the racial group is about 14 mmHg for drug A and about 8 mmHg for drug B. There is no interaction between drugs and racial groups.

Fig. 26.4, however, shows interaction between drugs and racial groups.

In the left panel, groups a and b have the same relative response to drugs, with $A < B$, but now group c has the effect of drug $A > B$. In the right-hand panel, drug A has its order of effects $a > c > b$, whereas the order for drug B is $a > b > c$. It makes no sense to examine main effects because these are not the same across all classes. Therefore when a two-way ANOVA *with replications* is done to test interaction, first examine if you can reject the null hypothesis for the interaction. If you cannot reject it, it is reasonable to proceed to

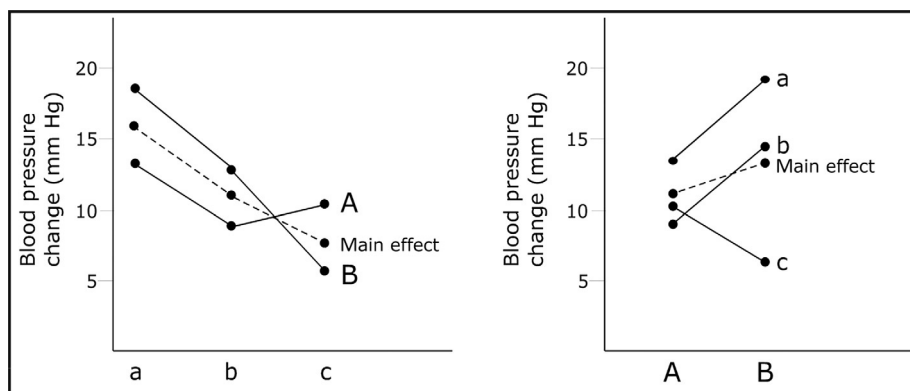


Fig. 26.4 3×2 ANOVA with interaction.

examine the main effects to find out if $A > B$ or if $a > b > c$. On the other hand, if there is substantial interaction, testing main effects is unrealistic, because drug A is better than B in some racial groups but not others.

How can we test for interaction? As an example, an experiment is done in which subjects classified as having high or low anxiety were divided into two groups, one untreated and one given a serotonin reuptake inhibitor, and their norepinephrine concentrations were measured after being given a problem-solving test (Table 26.9).

Table 26.9 Effects of treatment

Agent	High anxiety	Low anxiety
Untreated	180	122
	190	87
	140	65
	163	104
	122	63
	187	55
ΣX	982	496
N	6	6
\overline{X}_i	163.67	82.67
Treated	63	107
	47	96
	24	97
	44	98
	27	63
	13	115
ΣX	218	576
N	6	6
\overline{X}_i	36.33	96

Inspection of the means suggests that whereas treatment decreases mean catecholamine concentrations in high anxiety patients (163.67 vs 36.33), it has little effect in low anxiety patients (82.67 vs 96).

Begin the analysis as a one-way ANOVA on the smallest subgroups: high anxiety untreated, high anxiety treated, low anxiety untreated, and low anxiety treated (Fig. 26.5).

There are large differences, with high anxiety untreated patients having the highest catecholamine concentrations.

Now combine the groups into larger sets (treated and untreated, or high and low anxiety) and do a one-way ANOVA for each of the combined groups. The results are shown in Fig. 26.6.

(When measurements are superimposed, the points on the graph may be staggered out of alignment; this is termed “jittering.”)

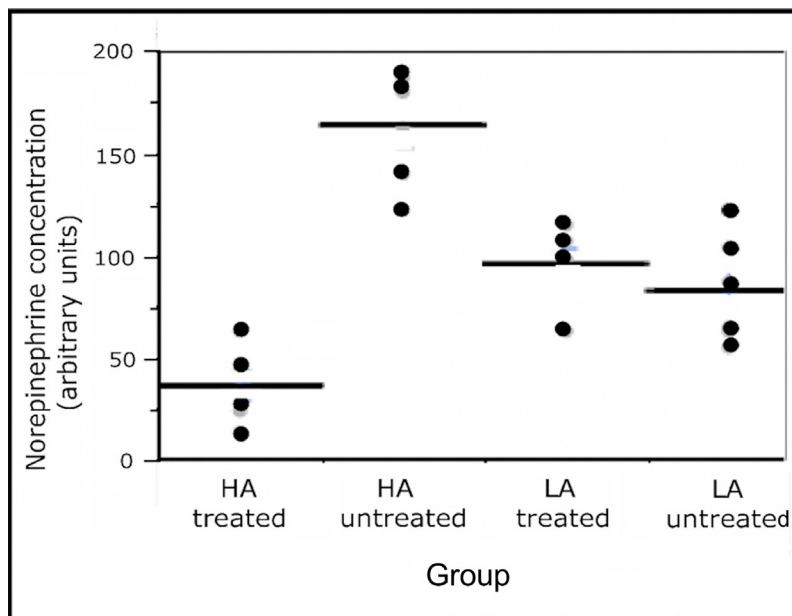


Fig. 26.5 ANOVA for all anxiety and treatment data. HA—high anxiety; LA—low anxiety. Horizontal bars are means.

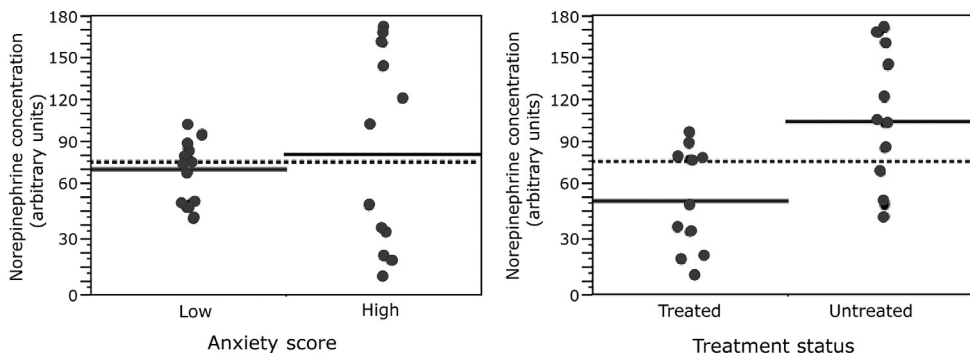


Fig. 26.6 ANOVA for high vs low anxiety, and treated vs untreated. Dotted line—mean of all measurements; solid horizontal lines—group means.

Assemble these results into the standard ANOVA format (Table 26.10).

Tables 26.11 and 26.12 give the ANOVA calculations for anxiety score and treatment status. They are used not for themselves but for the next stage of the calculation.

The four groups, however, are not merely 4 different groups with one factor, but that there are actually two factors—anxiety level and treatment. Therefore partition the between-group sum of squares (Table 26.13).

Table 26.10 ANOVA table for all 4 groups

Source of variation	SS	Df	MS	F	P
Total	60,387.33	23			
Between groups	49,857.33	3	16,619.10	31.56	<0.0001
Within (residual)	10,530.00	20	526.5		

Table 26.11 ANOVA for anxiety score

Source of variation	SS	Df	MS	F	P
Total	60,387.33	23			
Between groups	682.67	1	682.67	0.2515	0.6210
Within (residual)	59,074.67	22	2713.85		

Table 26.12 ANOVA for treatment status

Source of variation	SS	Df	MS	F	P
Total	60,387.33	23			
Between groups	19,494.00	1	19,494.00	10.4875	0.0038
Within (residual)	40,893.33	22	1858.8		

Table 26.13 Partitioned ANOVA

Source of variation	SS	Df	MS	F	P
Between groups	49,857.33	3	16,619.10		
Between anxiety level	682.67	1	682.67		
Between treatments	19,494.00	1	19,494.00		
Sum	20,176.67	2			
Between groups—sum	29,680.66	1	29,680.66		

From the between-group SS of 49,857.33 with 3 degrees of freedom, 682.67 SS is between anxiety levels with 1 Df and 19,494.00 SS between treatments with 1 Df; these are referred to as main effects. There are 29,680.66 SS and 1 Df unaccounted for (bottom row), and these are termed the interaction SS and Df. The full ANOVA table is presented in [Table 26.14](#). Each of the between subgroups is tested against the error term.

Therefore the experiment has shown a large difference due to treatments, enough to reject the null hypothesis, but little difference related to anxiety levels (treatment and no

Table 26.14 Full ANOVA

Source of variation	SS	Df	MS	F	P
Total	60,387.332	23			
Between groups	49,857.33	3	16,619.10		
Between anxiety levels	682.67	1	682.67	1.30	0.2677
Between treatments	19,494.00	1	19,494.00	37.03	0.00001
Interaction	29,680.66	1	29,680.66	5637.35	<0.00001
Within groups (residual)	10,530.00	20	526.5		

treatment combined). There is a large interaction effect with treatment having a large effect in high anxiety patients but none in the low anxiety patients, so that the null hypothesis can be rejected.

As long as there are replications, interaction effects can be assessed, no matter how many groups there are.

Interaction can be tested by online programs at <http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/ANOVA2Rep.htm> (up to 4×6 blocks and 4 replications) and <http://vassarstats.net/anova2u.html>, http://www.wessa.net/rwasp_Two%20Factor%20ANOVA.wasp#output.

An alternative method for assessing interaction is by using multiple regression techniques (Chapter 30).

Problem 26.1

Because a two-way ANOVA with replication is an efficient design, this problem involves testing for interaction.

	Low	Medium	High
Male	4	7	10
	5	9	12
	6	8	11
	5	12	9
Female	6	13	12
	6	25	13
	4	12	10
	4	12	13

Weight gain (g) after eating breakfast with high, medium, or low carbohydrate content. Did the carbohydrate content affect weight gain, and did males and females respond in the same way?

ADVANCED AND ALTERNATIVE CONCEPTS

Missing Data or Unequal Cell Sizes

More complex forms of ANOVA have symmetry. All Latin square designs have equal numbers for each factor. A two-way ANOVA need not have the same number of factors for factor 1 as for factor 2, but all the cells should contain data. In replicated experiments, too, ideally there should be equal numbers of replications in each cell for maximal efficiency. This may not always be possible. For example, if factor 1 involves 3 diseases, one might be rare so that finding enough measurements is difficult or impossible. It is still possible to carry out the ANOVA if there is equal representation within rows and proportional representation within columns, for example, row 1 has 3, 3, 3, and 3 measurements (one in each column). Row 2 has 4, 4, 4, and 4, and row 3 has 2, 2, 2, and 2 measurements. Alternatively, there can be proportional representation within rows and columns, for example, 3, 6, 9, and 6 in row 1; 4, 8, 12, and 8 in row 2; and 2, 4, 6, and 4 in row 3. The ANOVA is carried out with slight modifications (see [Zar, 2010, p. 247](#)).

In practice, though, sometimes data are lost: an animal may die, a test tube is broken, a patient decides to leave the trial. There will thus be missing data and the symmetry of cells is lost. To deal with this, the missing values can be estimated in various ways that to some extent depend upon why the data are missing.

Missing data may be classified as:

1. Missing completely at random (MCAR). The missing data are not related to the subject or the subject's other data. For example, a test tube is dropped and its contents lost.
2. Missing at random (MAR). Here the missing data are related to other data in the subject's data set, but do not depend on the value of the missing data. For example, an examination score is missing because the person was habitually late and paid little attention to instructions.
3. Missing not at random (MNAR). In the previous example, if the person did not attend the examination because she knew she would fail, the absent datum reflects the putative missing value.

Calculations can be done by simple or multiple imputation. For simple imputation regression is performed on all the completed sets relating the variable in the field with the missing datum to the other variables. With this information the missing data are predicted for each person's data set. This can be done with standard regression methods. For multiple imputation, similar regression relationships are calculated, and a random error value is assigned to the new calculated datum. The process is repeated, and new random errors are added to the parameters. After many repeated calculations (20 – 100) a consistent set of values is produced. This method can be used for MCAR and MAR but requires special software.

Statistical consultation is required. An understandable explanation is given by Baraldi and Enders (2010) and another by Howell can be found at https://www.uvm.edu/~dhowell/StatPages/Missing_Data/Missing.html, and at.

http://faculty.smu.edu/kyler/courses/7312/presentations/2013/parker/Handling_Missing_Data.pdf.

Nested Designs

In the two-way ANOVA described before, every member of factor 1 is associated with every member of factor 2; each type of wheat is represented for each location, and each location is represented for each type of wheat. The factors are said to be crossed. By contrast, the nested design does not have every member of factor 1 associated with every member of factor 2, as shown in Fig. 26.7.

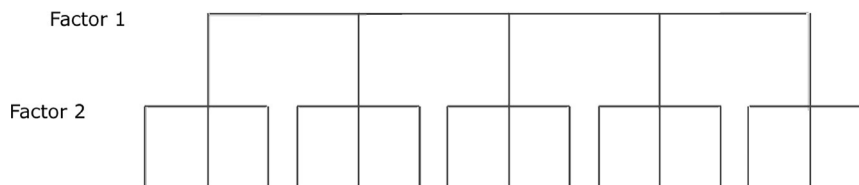


Fig. 26.7 Nested design. Here there are 5 members of factor 1, each associated with 3 members of factor 2. However, the first member of factor 2, group 1 is not related to the first member of factor 2, group 2, and so on.

For example, how similar are different laboratories in measuring blood LDL cholesterol in a standard blood sample? To analyze this, the blood sample is split into 4 equal portions, and one portion is given to each of 4 laboratories in a city. In each laboratory three technicians measure cholesterol independently, each with replicate measurements (Table 26.15).

Table 26.15 Nested data

Laboratory	A				B			C			D		Total
Technician	a	b	c	d	e	f	g	h	i	j	k	l	
Replication	108	109	107	113	110	112	108	107	105	114	113	117	
	106	110	108	112	112	113	108	108	105	115	115	113	
	106	109	107	111	111	114	110	111	108	116	117	114	
ΣX (tech)	320	328	322	336	333	337	326	343	318	345	345	344	3997
ΣX (lab)	970				1006			987			1034		3997

The technicians are not crossed with laboratories. This would happen only if each technician made measurements in each laboratory. However, the first technician in laboratory A is not the same as the first technician in laboratory B, and so on. The main question is whether laboratories get the same results. They should, but it is possible that

differences in reagents used, type of measuring equipment, and conditions such as temperature may cause differences. Another question is how much technicians differ from each other or, to put it another way, is there more variability between technicians than can be accounted for by variability of the method of measurement. It is of no value only to compare technicians with each other, because then differences among laboratories might be responsible.

Looking at the 12 sums (ΣX) for each technician, there are certainly differences, sometimes marked; for example, 320 vs 345. On the other hand, the sums for the 4 laboratories ΣX (lab) have equally great variations. How much of the variation among technicians is due to the differences among laboratories?

Begin by considering the 12 smallest groups, that is, 3 replications in each of 12 technicians, and do a one-way ANOVA (Fig. 26.8).

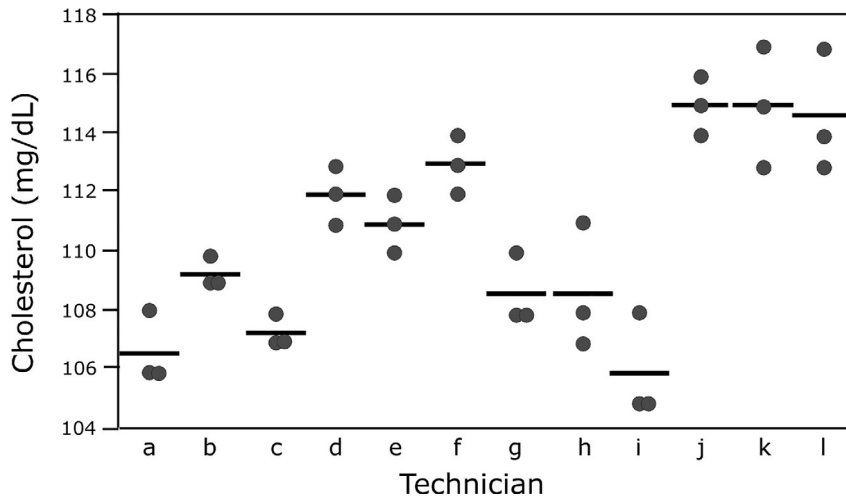


Fig. 26.8 ANOVA on the most basic groups, the different technicians. Horizontal lines indicate means.

The ANOVA is presented in Table 26.16.

Table 26.16 ANOVA for difference among technicians

Source of variation	SS	Df	MS	F	P
Total	404.5556	35			
Between technicians	358.5556	11	32.5960	17.0066	<0.0001
Within (residual)	46.0000	24	1.9167		

There are substantial differences among technicians that allow rejection of the null hypothesis, but how much of the difference is due to the laboratories? To evaluate this, pool the data from each laboratory, and do another one-way ANOVA (Fig. 26.9).

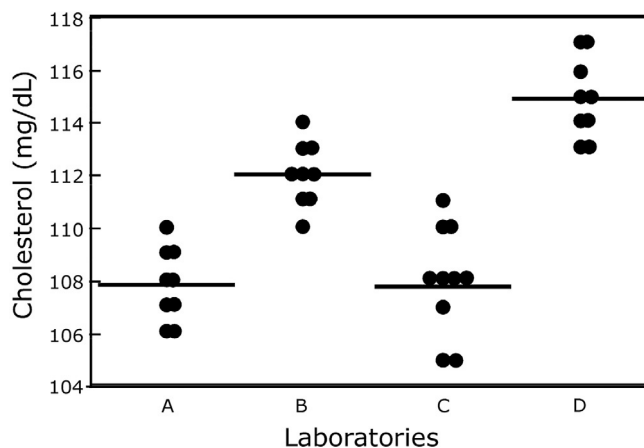


Fig. 26.9 ANOVA for differences among laboratories.

The analysis is presented in Table 26.17.

Table 26.17 ANOVA for differences between laboratories

Source of variation	SS	Df	MS	F	P
Total	404.5556	35			
Between laboratories	326.5556	3	108.852	44.6572	<0.0001
Within groups (residual)	78	32	2.437		

We can reject the null hypothesis for laboratories too. The interaction among laboratories and technicians is presented in Table 26.18.

Table 26.18 ANOVA between technicians and between laboratories

Source of variation	SS	Df	MS	F	P
Total	404.55	35			
Between technicians	358.55	11	32.60		
Between laboratories	326.55	3	108.85		
Difference between laboratories and technicians	(32)	(8)			
Within groups (residual)	46	24	1.92		

Underneath the line for the between technician data in Table 26.18 there is a row for the between laboratory data. The SS_B for technicians is greater than the SS_B for laboratories by $358.55 - 326.55 = 32$, and the degrees of freedom for technicians is greater than that for laboratories by $11 - 3 = 8$ (second last row). These discrepancies are due to the variability of technicians within laboratories (Table 26.19).

Table 26.19 Complete nested design ANOVA

Source of variation	SS	Df	MS	F	P
Total	404.55	35			
Between technicians	358.55	11	32.60		
Between laboratories	326.55	3	108.85	27.21	0.0002
Between technicians in laboratories	32	8	4.00	2.08	0.079
Within groups (residual)	46	24	1.92		

We have partitioned the differences among technicians (part of which is due to differences among laboratories) into a component due to differences among laboratories and a component due to differences between technicians within laboratories. The testing of the null hypothesis in a nested design differs from the usual two-way or three-way ANOVA in which each between-group MS is compared to the within-group MS. In the nested design start by comparing the between technicians in laboratories MS to the within-group MS, there are no excessive differences over and above the variability of replicates (although perhaps a larger sample with more replicates might disclose such a difference). Then to test the difference among laboratories, divide the between-laboratory MS by the MS between technicians in laboratories, and not by the residual MS. The reason for this is that we are no longer estimating the population variability due to replication, but instead want to know if the difference among laboratories is greater than expected from the combined variability of replicates and differences among technicians. Therefore divide 108.85 by 4 to get an F ratio of 27.21 and assess this with degrees of freedom in numerator and denominator, respectively, of 3 and 4.

To understand the technicians within laboratories component, alter the data to make each technician within a laboratory have similar results (Table 26.20).

Table 26.20 ANOVA with altered data

Laboratory	A				B			C			D			Total
Technician	a	b	c	d	e	f	g	h	i	j	k	l		
Replication	108	107	106	113	113	111	108	107	108	114	114	117		
	106	107	108	112	113	113	108	108	110	115	117	113		
	106	106	106	111	110	112	110	111	108	116	114	115		
ΣX (tech)	320	320	320	336	336	336	326	326	326	345	345	345		3978
ΣX (lab)	960				1008			978			1035			3978

The final ANOVA Table is presented in Table 26.21.

The contribution of differences among technicians within laboratories is now very small and almost all the variability is associated with the laboratories.

Table 26.21 Revised ANOVA

Source of variation	SS	Df	MS	F	P
Total	388.97	35			
Between technicians	348.31	11	31.66		0.0007
Between laboratories	322.50	3	107.50	33.49	<0.0001
Between technicians in laboratories	25.81	8	3.21	1.54	0.18
Within groups (residual)	6.48	32	2.08		

Transformations

Although ANOVA is usually performed on ratio numbers, it can be used for counts with some modifications (Snedecor and Cochran, 1989; Zar, 2010).

Counts often come from Poisson distributions. If there is a large variation in counts from cell to cell, then there will be a correspondingly large variation in cell variances because in a Poisson distribution the variance is similar to the mean. The effect of this is to make ANOVA less sensitive in detecting differences because the residual variance is inflated. One way of dealing with this problem is to use the square root transformation that stabilizes the variance and makes the distribution of data more normal. If the counts are very small, for example, most <10 , then $\sqrt{X} + \sqrt{X+1}$ is more effective.

If the standard deviation is proportional to the mean, so that the coefficient of variation is constant, then a logarithmic transformation not only stabilizes the variance but also restores additivity.

Model II ANOVA

Model I ANOVA is the fixed effects model. The factors represent groups that might have different means, such as different fats, or different weight gains on different diets. Even without hypotheses about the outcomes, in the wheat experiment there were selected four *specific* wheat cultivars to be tested. Each member of the group has a mean value (that may or may not be different from the others), and taking more samples makes the estimates of the means more accurate. Model II ANOVA is the random effects model, and it is designed to reveal different components of variability without placing any importance on mean values. For example, how much do different technicians vary in making measurements? The technicians are not specified but are merely random representatives of the class of technicians. After all, one particular technician might be used in a particular experiment but never used again.

The ANOVA is performed as usual, but the analysis of the results is altered. To illustrate this, consider measuring a rare growth factor in 3 random parts of the liver taken from four randomly selected baboons ($a=4$). The hypothetical data are given in Table 26.22.

Table 26.22 Growth factor concentration study

Baboon	Growth factor (pmoles/kg wet weight)			ΣX	\bar{X}_i
1	2.82	2.38	3.18	8.38	2.79
2	2.33	2.30	2.88	7.51	2.50
3	1.67	1.77	2.44	5.88	1.96
4	1.68	1.53	2.06	5.27	1.76

The ANOVA results are presented in [Table 26.23](#).

Table 26.23 ANOVA for growth factor data

Source	Df	SS	MS	F
Baboons	3	2.06	0.69	5.31 $P=0.0263$
Error	8	1.03	0.13	
Total	11	3.09		

The differences among baboons make it fairly safe to reject the null hypothesis, but their mean values are of no interest. Those baboons might never be used again. Furthermore, liver samples from another baboon have no more reason to provide information about baboon 1 than about baboon 4. What the measurements provide are some estimates of error related to different animals and to different parts of the liver. The error mean square S_W^2 is 0.13 and is an estimate of the population variability σ^2 of the estimates of growth factor concentrations of baboon livers. The between-group mean square S_B^2 is composed of the population variance σ^2 plus a component due to differences among the baboons $n\sigma_A^2$ ([Snedecor and Cochran, 1989](#)).

Thus $\sigma_B^2 = \sigma^2 + n\sigma_A^2$ and so $\sigma_A^2 = \frac{\sigma_B^2 - \sigma^2}{n}$.

An unbiased estimate of σ_A^2 is

$$s_A^2 = \frac{s_B^2 - s_W^2}{3} = \frac{0.686822 - 0.129275}{3} = 0.185849.$$

There are now two variances, one associated with replicate determinations of growth factor in baboon livers and one associated with differences among baboons.

What is the interest in knowing these two components of variance? If there is only enough money and reagents to do only 30 measurements of some complex growth factor in the liver, should we take one baboon and make 30 measurements on its liver? That is inefficient because baboons will differ in how much hepatic growth factor they have. Should we take one piece of liver from each of 30 baboons? That is also inefficient. Baboons are expensive, there are ethical questions about their use, and besides growth factor may vary in different parts of the liver. If we knew how much variability was associated with differences in baboons and how much with differences in parts of the liver, and the relative costs of baboons and growth factor measurements we could work out an optimal strategy ([Snedecor and Cochran, 1989](#)).

Some experimental designs include both fixed and random effects and are sometimes referred to as model III ANOVA. In general, biologists and medical investigators use model I most of the time, and more complex designs should be done only with statistical consultation.

More About Repeated Measures Designs

The examples discussed before all had a single observation of the dependent variable for each subject and each treatment and are often referred to as *between-subject* designs. It is, however, common to use a number of subjects and then make a series of measurements

on each of them; this is referred to as a *within-subjects* design. The different measurements might be different treatments or different times.

These results cannot be analyzed by the usual ANOVA, because the results within each subject are likely to be correlated. If the repeated measurements have a correlation ρ , the correlation coefficient (Chapter 29) then the variance of the treatment mean is not σ^2/n but is $\sigma^2 = \frac{[1 + (n-1)\rho]}{n}$, and the estimate of the mean square is not σ^2 but is $\sigma^2(1 - \rho)$. This distorts the true variance of the treatment mean. Furthermore, the treatments are fixed effects, that is, there are possible differences between treatment effects, something the experiment is designed to assess. On the other hand, if the subjects are randomly selected, a different set of subjects might have different responses. What is important is that each subject might have different mean responses, averaged across all the treatments. We need a method of removing the intrasubject variability that is not of direct interest to us and has the result of making the experiment less sensitive. The subject is complex (Glantz and Slinker, 2001; Maxwell and Delaney, 1990).

Problem 26.2.

Datar et al. (personal communication) studied the effect of acetylcholine in relaxing rings of the thoracic duct in control lambs and lambs in which an aorto-pulmonary shunt had been created in utero. The following table gives the data for one concentration of acetylcholine on four replicate rings from each of 6 control lambs:

	Animal	Response e-7
1	3516	0.732
2	3516	0.668
3	3516	0.741
4	3516	0.624
5	3517	0.681
6	3517	0.203
7	3517	0.102
8	3517	0.894
9	3518	0.192
10	3518	0.273
11	3518	0.649
12	3518	0.65
13	3549	0.316
14	3549	0.598
15	3549	0.709
16	3549	0.942
17	3569	0.734
18	3569	0.735
19	3569	0.506
20	3569	0.681
21	3570	0.388
22	3570	0.438
23	3570	0.418
24	3570	0.554

How much of the variability is due to differences among the lambs?

BIBLIOGRAPHY

- Baraldi, A.N., Enders, C.K., 2010. An introduction to modern missing data analyses. *J. Sch. Psychol.* 48, 5–37.
- Cochran, W.G., Bliss, C.F., 1970. In: McArthur, J.W., Colton, T. (Eds.), *Statistics in Endocrinology*. The MIT Press, Cambridge, MA.
- Finney, D.J., 1955. *Experimental Design and its Statistical Basis*. Cambridge University press, London.
- Glantz, S.A., Slinker, B.K., 2001. *Primer of Applied Regression and Analysis of Variance*. McGraw-Hill, Inc, New York.
- Graybill, F.A., 1954. Heterogeneity in a randomized block design. *Biometrics* 10, 516–520.
- Maxwell, S.E., Delaney, H.D., 1990. *Designing Experiments and Analyzing Data. A Model Comparison Perspective*. Wadsworth Publishing Company, Belmont, CA.
- Snedecor, G.W., Cochran, W.G., 1989. *Statistical Methods*. Iowa State University Press, Ames, IA.
- Tukey, J.W., 1977. *Exploratory Data Analysis*. Addison-Wesley Publishing Co, Menlo Park, CA.
- Zar, J.H., 2010. *Biostatistical Analysis*. Prentice Hall, Upper Saddle River, NJ.

SECTION VII

Regression and Correlation

CHAPTER 27

Linear Regression

BASIC CONCEPTS

Introduction

Previous chapters concerned single variables. Now we consider problems that involve relationships between two (X , Y) or more (X , Y , Z ...) variables, such as height and weight, or age, blood pressure, and blood cholesterol. It is everyone's experience that as a child gets older it gets bigger, that as people become more obese their fasting blood glucose levels rise, or that as incomes decrease more people become homeless. There are two ways of examining such relationships. One is to ask how closely a change in one variable is associated with a change in the other variable; this concerns the subject of correlation. The other is to ask how much on the average one variable Y changes as the other variable X changes; this concerns the subject of regression. The two subjects are closely related.

The term regression was introduced by Francis Galton who studied the heights of 930 adult children in relation to the heights of 205 of their respective parents. He observed that tall parents usually had tall children and that short parents usually had short children, but that tall parents usually had children a little shorter than they were, whereas short parents had children a little taller than they were. He entitled his article "Regression towards Mediocrity in Hereditary Stature," and the term "regression" has been used to denote the relationship between two or more variables ever since that time.

There is no *necessary* implication of causality in an X - Y relationship. The increase in age of a child is a factor in its increase in height. An increase in time of a car traveling at constant speed is a direct causative factor in the distance it goes. On the other hand, some variates are associated indirectly; this is referred to as confounding and is a common trap in scientific inference. One study showed a linear relationship between the annual birth rate in Oldenburg in Germany and the number of storks observed annually in the city (Box et al., 1978). A high birth rate caused the population to increase and so more houses were needed, and with more nesting sites more storks came to the city. The storks were definitely not the cause of the increased numbers of babies! Another example of association without causality is when X is a measurement of a process (e.g., cardiac output) measured one way and Y is another way of measuring the process. As cardiac output changes both X and Y will change with it, but neither one causes the other.

If X does cause Y , in whole or in part, then X is called an independent, predictor, explanatory variable (or covariate) and Y is the dependent, response, or explained

variable, but that classification can be made only from other considerations. With this convention, the X values are **always** measured along the horizontal axis (or abscissa) and the Y values **always** along the vertical axis (or ordinate).

The data for regression analysis may be obtained either by bivariate random sampling, the investigator selecting subjects at random from a population, and then measuring X and Y , or by selected sampling in which the investigator specifies the values of X (e.g., ages 1–10 in yearly intervals) and then draws a random sample of each of these samples and measures Y . In general, bivariate sampling is more applicable to observational studies and selected sampling to experimental studies where X (e.g., age, temperature) may be specified in advance.

Exploratory Data Analysis

Just as exploratory data analysis should be done for univariate measurements before launching into calculations and judgments, so should it be done for bivariate analysis. First plot the X and Y data pairs on a *scattergram* in which paired XY values are put into a graph. Each pair of associated values, for example, X_i and Y_i , produces one point on the plot, as indicated by the dashed lines in Fig. 27.1.

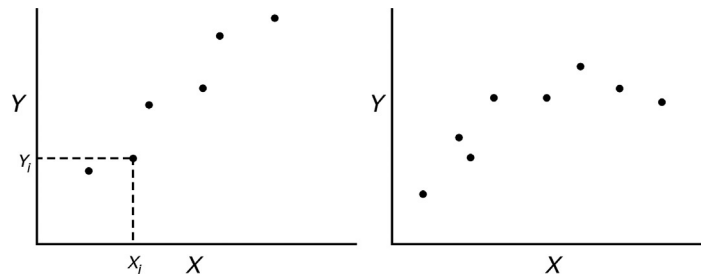


Fig. 27.1 Two scattergrams. Left panel: roughly linear relationship between X and Y . Right panel: roughly curvilinear relationship between X and Y . *Dashed lines* indicate the two coordinates for the point X_1Y_1 .

Sometimes there is no regularity, and XY points wander up and down, as in the daily Dow-Jones index. Because it is easier to demonstrate the principles of regression analysis with a simple linear relationship than with any other form, the rest of this chapter will be restricted to linear regression with two variables.

Transforming Curves

Although there are statistical procedures for dealing with curvilinear regression (Chapter 30), often it is simpler to transform the data so that linearity is achieved. A line is the expression of a relationship $Y = c + bX^p$ where c is the intercept on the Y -axis,

b is a coefficient, and p is the power to which X is raised (Mosteller and Tukey, 1977). If the power of X is 1, the line is straight and b is its slope, if power is >1 then the curve is concave upward, and if power is <1 the curve is concave downward (Fig. 27.2).

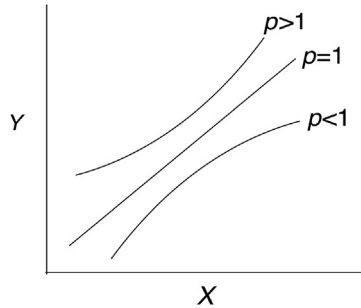


Fig. 27.2 Power curves.

If the points show decreasing curvature (concave down), taking the square root or the logarithm of X , which shortens the X -axis, or squaring or cubing Y , which lengthens the Y -axis, tends to straighten the points. If the points display an increasing curvature (concave up), then squaring or cubing X , or taking the square root or logarithm of Y may straighten out the points. Tukey (1977) introduced a range of possible transformations termed the ladder of powers or ladder of reexpression based on the relationship $Y = c + X^\lambda$ (Table 27.1).

Table 27.1 Ladder of powers

Power λ	Transformation	Name
3	X^3	Cube
2	X^2	Square
1	X	Raw
$\frac{1}{2}$	\sqrt{X}	Square root
0	Log X	Logarithm, usually to base 10
$-1/2$	$-1/\sqrt{X}$	Reciprocal root ^a
-1	$-1/X$	Reciprocal
-2	$-1/X^2$	Reciprocal square

Because anything to the power 0 is 1, logarithms are used.

^aSome publications list the reciprocals as positive, but this reverses the order of X . The reciprocals of $X = 2, 5$, and 10 are $0.5, 0.2$, and 0.1 . The negative sign preserves the order, so that $X = 2, 5, 10$ become $-0.5, -0.2$, and -0.1 .

Higher, lower, or intermediate powers are rarely needed.

Sometimes the underlying process is known to be exponential or logarithmic, and then appropriate procedures can be used. Depending on the approximate curvature of

the points, [Mosteller and Tukey \(1977\)](#) recommended general methods of reexpression (transformation) that they termed the bulging rule ([Fig. 27.3](#)).

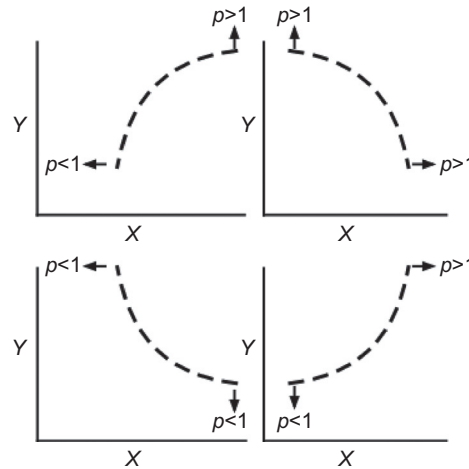


Fig. 27.3 Direction of reexpression needed to straighten out the *dashed line*. Horizontal arrows indicate direction of needed change in X values, vertical arrows indicate direction of needed change in Y values. $p > 1$ indicates increasing power above 1, $p < 1$ indicates decreasing power below 1. Any one or a combination of both indicated transformations tends to straighten the line. (Based on diagram by Mosteller, F., Tukey, J.W., 1977. *Data Analysis and Regression. A Second Course in Statistics*, Addison-Wesley, Reading, CA.)

Whether to reexpress X , Y or both depends upon the effect that reexpression has on the variability of Y . [Mosteller and Tukey \(1977\)](#) described simple methods for deciding what was the most appropriate power. In brief, the slope from a low value of X to a value near the middle of the X array is compared to the slope from that middle value to a high value of X . If the ratio of these slopes is near 1, then the line is approximately straight. Any nonlinear model that can be transformed into a linear model is called “intrinsically linear,” and transformation is likely to provide more normally distributed residuals with constant variance than fitting a more complex nonlinear model ([Montgomery and Peck, 1982](#)). This produces an improvement that shows up as smaller standard errors and therefore more precise prediction of the model parameters.

[Fig. 27.4](#) gives examples of several transformations of data to achieve linearity.

Upper four panels: Upper left panel shows original data with curved relationship between diameter (X) and pressure (Y). The next 3 panels show reexpression of the Y variable: natural logarithm vs diameter, square root vs diameter, and reciprocal vs diameter. Only the reciprocal of pressure shows a good linear relationship.

Lower four panels: First two panels show transformation of the X variate—square and then cube vs pressure. Neither straightens the set of points. Next two panels show reexpression of both X and Y variates, plotting the square and then the cube of X against the

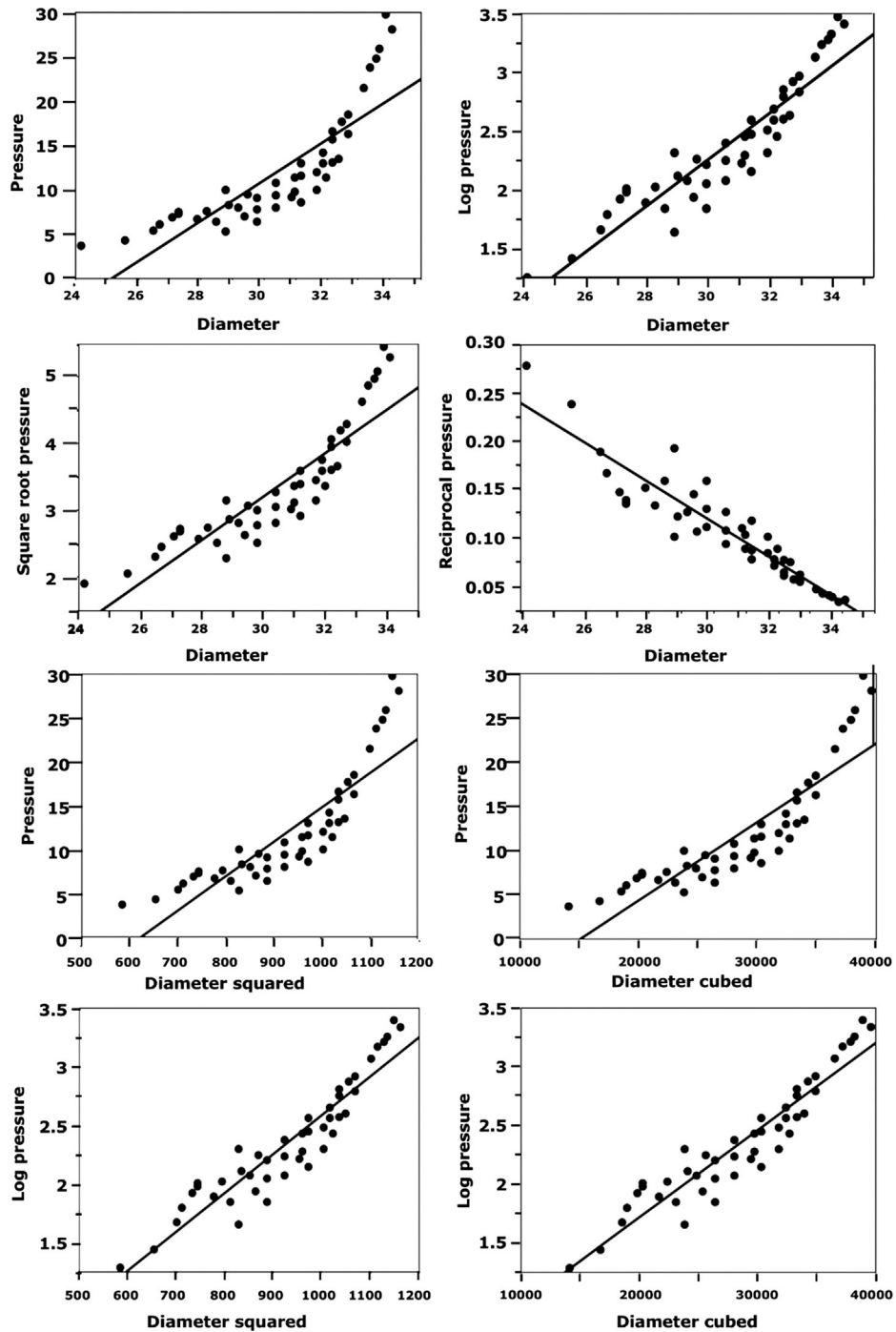


Fig. 27.4 Data taken from [Senzaki et al. \(2000\)](#) on pressure-diameter relations of failing dog hearts.

natural logarithm of Y . The cube of diameter against the logarithm of pressure gives good linearity.

These transformations are merely made to achieve linearity, and more than one may be suitable. Whether the transformation has an underlying mechanistic interpretation cannot be judged from the information presented.

We summarize X - Y relations by the mean slope of the relationship (does Y increase a lot or a little for a unit increase in X ?) and a measure of variability called the standard deviation from regression.

Model I Linear Regression Basics

The requirements for performing a simple linear (model I) regression are as follows:

1. Y must be linearly related to X
2. The numbers must be ratio or interval numbers.
3. The X values should be measured with no or little error.
4. The distribution of Y values at any X value is normal. That is, if there are several measurements of reaction products at each of several specified temperatures of 35°C, 36°C, 37°C, 38°C, and 39°C, then at each temperature (X value) the amount of reaction product (Y value) would be normally distributed.
5. The Y values are independent of one another. (See later and [Chapter 31](#).)
6. At each value of X , the variance of the Y values is constant (homoscedasticity).

Some of these requirements are indicated in [Fig. 27.5](#).

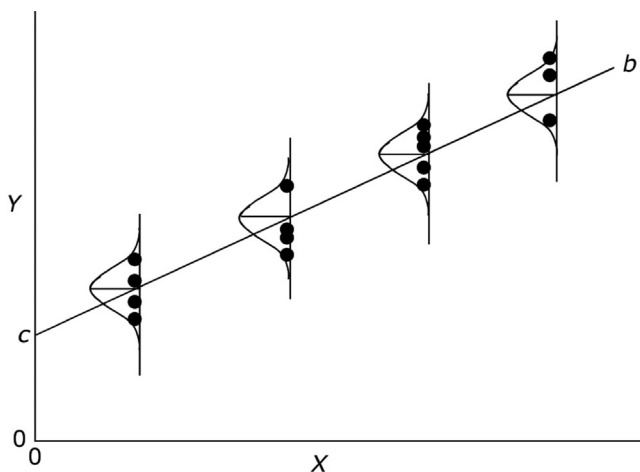


Fig. 27.5 Basic linear regression concepts. The line does not necessarily pass through Y zero. The sets of values of Y at each X value are approximately normal, have similar variances, and the short solid horizontal lines indicate the mean of each Y set. If the relationship of Y to X is linear, then the means of each set of Y values lie on a straight line with a slope indicated by b and an intercept on the Y -axis of c .

The requirement that X is measured with little error is poorly defined. Measurements of temperature, height, age, and so on are never exact, but are usually reproducible with very small standard deviations. This requirement is usually not tested, but if there is reason to be concerned alternative tests can be done (see [Chapter 28](#)).

The distribution of Y at any X can be seen clearly only when several values of Y are measured at each value of X . It is more usual for there to be one Y value with each value, as shown in [Fig. 27.1](#), but the scatter of points above and below the line gives a rough estimate of normality. Fortunately testing of normality is rarely required, and regression analysis is robust in this respect.

Homogeneity of variance is often absent ([Fig. 27.6](#)).

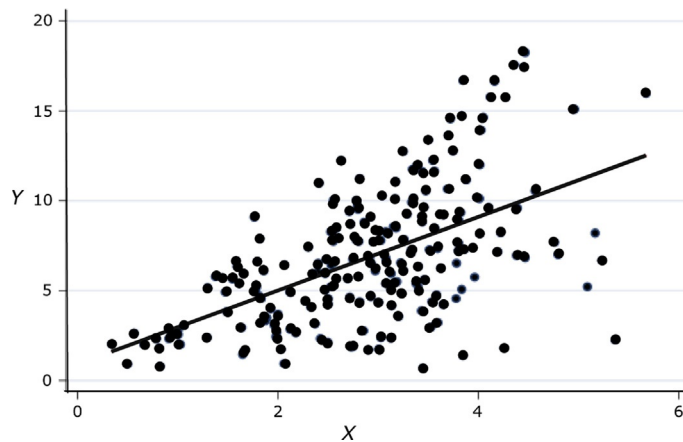


Fig. 27.6 The increasing variability of Y as X increases is obvious.

The increase in variability is expected, because large measurements usually vary more than small measurements do. Remember from [Chapter 4](#) that the coefficient of variation is often constant for a particular set of measurements. This means that if the ratio of standard deviation to mean is constant, then as the mean of X increases the standard deviation of Y increases proportionally. If the inhomogeneity of variance (heteroscedasticity) is not marked by eye the standard regression methods are adequate. If the change in variance is very marked, then although the estimates of slope and intercept are unbiased, the variability is excessive and reduces the efficiency of the test. Methods of testing heteroscedasticity and transforming the Y values are discussed later, (Transforming the X variate only changes its placement on the horizontal axis and does not alter the variability of the Y variate.)

Whether the line passes through zero, that is, the intercept on the Y -axis is zero, does not affect the ability to perform classical regression analysis. However, if the intercept truly is zero, there are sometimes easier ways of doing the calculations (see [Chapter 28](#)).

Linear regression, if the requirements are met, answers the questions:

How much does Y increase for a unit increase in X ? In other words, what is the slope of the line? and

What is the variability of the Y values at each X value?

If all the points fall exactly on a line there is no difficulty in defining the line, but what if they do not fall on a line, as in Fig. 27.1? Then use what is termed “the line of best fit” that can have several definitions. Most often the line of best fit used is one that minimizes the vertical deviations of the Y values from the line (Fig. 27.7).

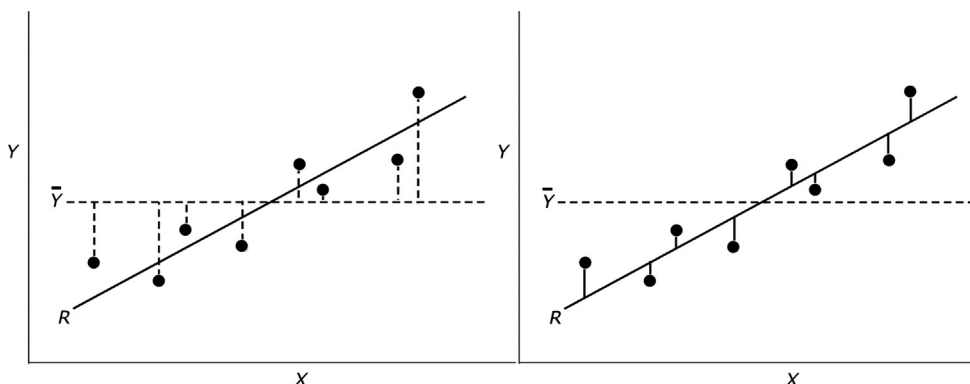


Fig. 27.7 Principle of least squares in linear regression. \bar{Y} is the mean of all the Y values (*horizontal dashed line*), and the regression line is shown as a *solid diagonal line*. The *vertical dashed lines* in the left panel are deviations from the mean, and the *solid vertical lines* in the right panel are deviations from the regression line.

The principle states that the line of best fit minimizes the sum of the squared vertical deviations of the points from the regression line. As shown in Fig. 27.7, the deviations are usually less from the regression line than from the mean. It is essential that in the scattergram the X (independent) values are plotted on the horizontal axis and the Y (dependent) values are plotted on the vertical axis so that the vertical deviations of Y from the regression line will be minimized. If the X and Y axes are interchanged, it is the deviations of X from the regression line that are minimized, giving a different line of best fit and a different interpretation.

How do we calculate the slope of the line of best fit? Although this is done easily in computer programs, readers should understand how this is done, because the elements of the resulting formulas appear frequently, and because understanding the process leads to greater insights. Online calculations can be done at <http://vassarstats.net/> (see Correlation and Regression); XY data can be entered either one by one in a table or by copying data from a spreadsheet. Other programs online are <https://www.easycalculation.com/statistics/regression.php>, http://www.statstodo.com/CorReg_Pgm.php, and

http://www.xycoon.com/simple_linear_regression.htm (which also plot the scatter-gram) and <http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Regression.htm> (which give some additional statistics), <http://www.meracalculator.com/math/regression.php>, <http://www.alcula.com/calculators/statistics/linear-regression/>, <http://www.quantitativeskills.com/sisa/dataprocs/rdata.htm>, <http://www.xuru.org/rt/PowR.asp>, and <https://nccalculators.com/statistics/linear-regression-calculator.htm>.

Fig. 27.8 shows the rationale of the calculations.

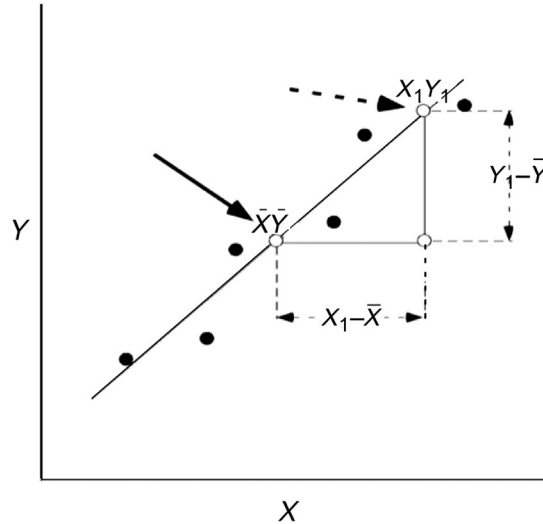


Fig. 27.8 Basis of slope calculation. *Solid circles* are data points. *Open circles* are hypothetical points, two on the regression line, with their XY coordinates.

One of the advantages of the least squares line of best fit is that the regression line passes through a point that is the mean of X and the mean of Y , indicated by the solid arrow. Consider any point on the regression line X_iY_i indicated by the dashed arrow (see figure). Construct a right-angled triangle from these two points, as shown; designate the right angle by the point $X_1\bar{Y}$. By trigonometry, the slope of the line, b , is the tangent. Therefore

$$b = \frac{Y_i - \bar{Y}}{X_i - \bar{X}}.$$

Multiply both sides by the denominator:

$$b(X_i - \bar{X}) = Y_i - \bar{Y},$$

$$\therefore b(X_i - \bar{X}) + \bar{Y} = Y_i, \text{ and by rearrangement.}$$

$$Y_i = bX_i - b\bar{X} + \bar{Y} = bX_i + (\bar{Y} - b\bar{X}) = bX_i + c, \text{ where } c = (\bar{Y} - b\bar{X}).$$

Because this is not any Y_i , but the hypothetical Y_i that lies on the regression line, characterize it as \hat{Y}_i (termed Y hat) where the hat indicates a theoretical point. c , the intercept on the Y -axis, is composed of two portions: the mean of Y and b times the mean of X . Depending on the values for \bar{Y} and \bar{X} , c may be positive or negative.

The remaining step is to find out how to compute b , the slope.

As in univariate statistics, compute $\sum X_i$, $\sum Y_i$, $\sum (X_i^2)$, $\sum (Y_i^2)$ and then for X and Y separately derive the sum of squared deviations from the mean.

$$\sum (X_i - \bar{X})^2 = \sum (X_i^2) - \frac{(\sum X_i)^2}{N} \text{ and } \sum (Y_i - \bar{Y})^2 = \sum (Y_i^2) - \frac{(\sum Y_i)^2}{N},$$

as described in [Chapter 4](#). All that remains is to compute the sum of the product deviations from the mean, also known as the covariance. To understand covariance, examine [Table 27.2](#).

Table 27.2 Product deviation from the mean

X	$X_i - \bar{X}$	Y	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	XY
7	-2	13	-4	$-2 \times -4 = 8$	91
6	-1	11	-2	$-1 \times -2 = 2$	66
5	0	9	0	0	45
4	1	7	2	$1 \times 2 = 2$	28
3	2	5	4	$2 \times 4 = 8$	15
$\Sigma X = 25$		$\Sigma Y = 45$		$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 20$	$\Sigma XY = 245$
$\bar{X} = 5$		$\bar{Y} = 9$			

For each X_i calculate its deviation from mean X , and for each Y_i calculate its deviation from mean Y . Then multiply each deviation from the mean of X by the corresponding deviation from the mean of Y and add up all the products; this sum is the covariance.

To avoid this lengthy procedure with large samples and noninteger deviations from the mean, use the identity

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_i Y_i - \frac{\sum X_i \sum Y_i}{N}.$$

For the previous example, this would be.

$$245 - \frac{25 \times 45}{5} = 20, \text{ as shown before.}$$

If X and Y are related ([Table 27.1](#)), then the product deviations from the mean tend to be all positive (or all negative, if the slope is downward) and they will sum to some reasonable value. If X and Y are not associated, then the product deviations from the mean will not be all positive or all negative, and their sum will be small, as presented in [Table 27.3](#).

Table 27.3 Rearrangement to show lack of association between X and Y

X	$X_i - \bar{X}$	Y	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$
8	3	13	4	$3 \times 4 = 12$
1	-4	11	2	$-4 \times 2 = -8$
6	1	9	0	0
1	-4	7	-2	$-4 \times -2 = 8$
9	4	5	-4	$4 \times -4 = -16$
$\Sigma X = 25$		$\Sigma Y = 45$		$\Sigma (X_i - \bar{X})(Y_i - \bar{Y}) = -4$
$\bar{X} = 5$		$\bar{Y} = 9$		

A large product deviation from the mean indicates an association between X and Y , and a small product deviation from the mean indicates little or no such association. Is the product deviation from the mean is large or small? Relate it to the variability of the X array as assessed by the sum of squared deviations:

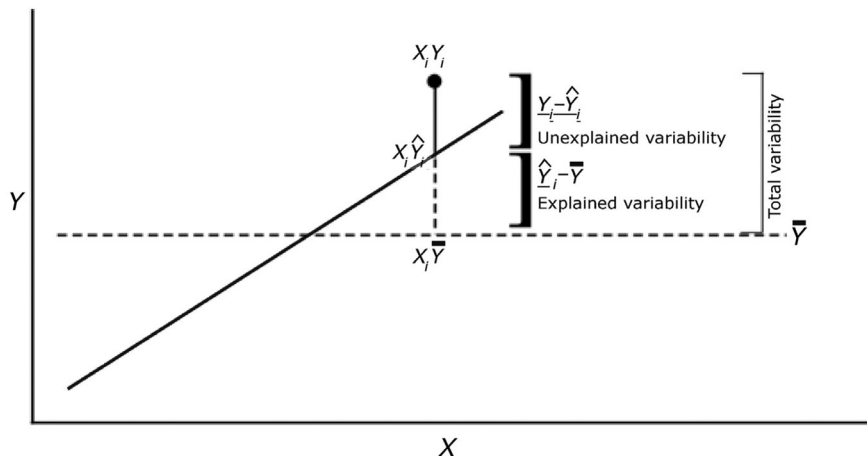
$$b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}.$$

With this calculation, the line of best fit has the equation:

$$\hat{Y}_i = bX_i + (\bar{Y} - b\bar{X}).$$

To draw the line by hand, select two values of X_i , calculate the values of the corresponding values of \hat{Y} , and join them by a straight line. The slope of a relationship depends on the units used; the slope of a height-weight relationship is not the same if these are measured in inches and pounds or in centimeters and kilograms.

Next calculate the variability of the points about the regression line (Fig. 27.9).

**Fig. 27.9** Diagram to show the components of Y .

For any data point X_iY_i , the deviation of Y_i from the mean of Y can be divided into the difference between Y_i and the corresponding point on the regression line, $(Y_i - \hat{Y}_i)$, and the difference between the point on the regression line and the mean of Y , $(\hat{Y}_i - \bar{Y})$. Then the sum of the squared deviations of all the Y values from their mean is

$$\begin{aligned}\sum (Y_i - \bar{Y})^2 &= \sum [(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})]^2 \\ &= \sum (Y_i - \hat{Y}_i)^2 + 2\sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + \sum (\hat{Y}_i - \bar{Y})^2.\end{aligned}$$

But $2\sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0$ because it involves the sum of deviations from the mean, so that

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2.$$

Therefore the sum of squared deviations from the mean of the Y values can be partitioned into a component due to regression $\sum (\hat{Y}_i - \bar{Y})^2$ and the residual variation from the regression line $\sum (Y_i - \hat{Y}_i)^2$. These are the counterparts, respectively, of the between-group sum of squares and the within-group sum of squares in ANOVA (Table 27.4).

Table 27.4 ANOVA format for linear regression

Source	SS	Df	MS	F
Total	$\sum (Y_i - \bar{Y})^2$	$N - 1$		
Due to regression	$\sum (\hat{Y}_i - \bar{Y})^2$	1	$\sum (\hat{Y}_i - \bar{Y})^2$	
Residual	$\sum (Y_i - \bar{Y})^2 - \sum (\hat{Y}_i - \bar{Y})^2$	$N - 2$	$\frac{\sum (Y_i - \bar{Y})^2 - \sum (\hat{Y}_i - \bar{Y})^2}{N - 2}$	

Information about the association of Y with X has reduced the total variability of Y by an amount related to the relationship between X and Y .

For the total variation calculate

$$\sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - \frac{\sum (Y_i)^2}{N}.$$

For the component of variation due to the regression line calculate

$$\sum (\hat{Y}_i - \bar{Y})^2 = \frac{\sum \left(X_i Y_i - \frac{\sum X_i \sum Y_i}{N} \right)^2}{\sum (X_i - \bar{X})^2}.$$

Many of these additional statistics can be calculated online at <http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Regression.htm> and <http://www.wessa.net/slr.wasp>.

As an example Harker and Slichter (1972) studied the relationship of platelet count to bleeding time (Table 27.5).

Table 27.5 Bleeding time and platelet count for each patient

	Bleeding time (min)	Platelet count $\times 10,000/\mu\text{L}$		Bleeding time (min)	Platelet count $\times 10,000/\mu\text{L}$
1	7.1	10.5	27	17	6
2	5.1	10.3	28	17	4.9
3	4.6	10.1	29	16	4.7
4	4	9.9	30	15.5	4.5
5	5.6	9.9	31	15.9	4.3
6	4.1	9.5	32	17	4.2
7	5.1	9.2	33	17.9	4.3
8	6.1	9.3	34	18	4
9	6.6	9.2	35	19.1	4.3
10	6.5	8.9	36	19	4
11	7	8.6	37	19	3.7
12	8.5	9	38	20.1	3.5
13	9.6	8.6	39	23	3.4
14	9.1	7.8	40	23.1	2.3
15	7.7	7.5	41	24.1	1.8
16	10	6.9	42	25	2.3
17	10.1	6.5	43	25	2.1
18	11.1	6.9	44	25	1.9
19	12	7.2	45	26.1	1.7
20	13	7.3	46	25.4	1.5
21	14.1	6.5	47	28	2
22	14	5.8	48	28.9	1.2
23	14.6	5.3	49	30	0.8
24	14	5.2	50	20.9	4.2
25	15.9	5.3	51	18.9	2.9
26	16.9	5.6			

The first thing to do is to plot these data on an X-Y plot (Fig. 27.10). Because bleeding time depends on platelet count, the platelet count is plotted on the X-axis.

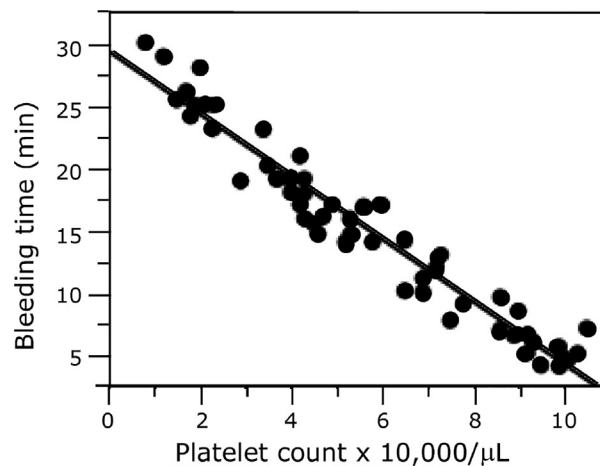


Fig. 27.10 Plot of bleeding time versus platelet count.

The figure shows linearity and homogeneity of variance. The line slopes downwards because as platelet count decreases the bleeding time increases.

Calculate the basic features of regression (Table 27.6).

Table 27.6 Basic regression results

ANOVA				
Source	Df	SS	MS	F
Regression	1	2557.5092	2557.51	825.9038
Error	49	151.7343	3.10	
Total	50	2709.3425		
Parameter Estimates				
Term	Estimate	Standard error	t	P
Intercept	29.4097	0.5512	53.36	<0.0001
Platelet count $\times 10,000/\mu\text{L}$	-2.5151	0.0875	-28.74	<0.0001

Bleeding time (min) = $29.4097 - 2.5151 \times \text{platelet count} \times 10,000/\mu\text{L}$.

Summary of fit

R^2	0.9440
R^2 adjusted	0.9429
Root mean square error	1.7597
No. of observations	51

This gives the line of best fit: the intercept on the Y-axis at $X=0$ is 29.41 and the slope is -2.52. The ANOVA shows that the total variability of bleeding time (Y) has been reduced from 2709.24 to 151.73 (residual, error, or within group SS, also termed s_{res}^2) by virtue of the regression (model) accounting for 2557.51. The F ratio that, as usual divides MS_B by MS_W , is $2557.51/3.1 = 825.90$ and is very large. In a univariate distribution the sum of squared deviations from the mean divided by degrees of freedom was termed the variance s^2 , and the comparable value in regression analysis is the within-group error, or residual sum of squares, divided by degrees of freedom to give the MS_W , sometimes symbolized by $s_{Y.X}^2$ or as $s_{y|x}^2$ to show that the variability of Y is dependent on X. Its square root $s_{y|x}$ is the standard deviation from regression.

What does the F ratio indicate? It indicates that the null hypothesis of zero slope can be rejected confidently. Imagine, for example, a population of unrelated X-Y points with no association so that the true population slope $\beta = 0$. Regression analysis on a sample from that population would very likely get a nonzero slope. To find out if the null hypothesis ($H_0: \beta = 0$) is true, test the difference between the observed slope b and zero against a measure of variability. As seen from Table 27.5, the less association there is between X and Y, the smaller the model component of regression and the larger the residual SS will be.

Problem 27.1 Nagy et al. (2014) compared the pulmonary capillary wedge (PCWP) and left atrial (LAP) pressures in patients with mitral stenosis. The following table presents a subset of their results.

x	y	x	y
2.098	3.980	15.108	16.913
6.055	6.964	15.228	17.985
6.115	7.959	16.067	15.995
7.134	8.036	16.067	15.077
9.113	9.031	16.127	13.852
10.132	10.026	17.146	15.000
10.132	11.020	17.086	15.995
11.091	12.015	17.086	16.990
11.151	13.010	17.026	18.214
11.271	14.005	18.106	21.964
13.129	13.852	18.106	21.046
13.129	15.230	18.106	19.974
14.029	15.995	18.106	18.903
14.029	15.077	18.165	16.990
14.149	13.929	18.165	15.995
15.048	13.010	19.065	15.918
15.108	14.923	19.125	16.990
15.168	16.148		

Draw an XY plot and calculate the linear regression for these data.

Additional Calculations

The basic calculations are used to compute other useful statistics, such as the confidence intervals of the slope and the intercept, of the estimated value of Y_i , and of the standard deviation from regression.

To test the null hypothesis that the slope is zero, calculate the variance of the slope as

$$s_b^2 = \frac{s_{\text{res}}^2}{\sum (X_i - \bar{X})^2}, \text{ and then do a } t\text{-test}$$

$$t_{0.05} = \frac{b - \mu}{s_b}. \text{ For the null hypothesis, } \mu = 0.$$

For the example used before in Table 27.6, $s_{\text{res}}^2 = 3.1$, $\sum (X_i - \bar{X})^2 = 404.29$ [calculated from the platelet counts above], so that $s_b = \sqrt{\frac{3.1}{404.29}} = 0.088$ and

$t_{0.05, N-2} = \frac{2.52 - 0}{0.088} = 28.64$. Reject the hypothesis that the slope is zero, exactly as presented in Table 27.6. From this it is simple to calculate the 95% confidence limits as $-2.52 \pm t_{0.05, n-2} \times 0.088 = -2.52 \pm 2.004 \times 0.088 = -2.70$ to -2.34 .

Occasionally the population mean is not zero; for example, if theoretically there should be a 1:1 ratio with a slope of 1. Then the t -test is done with $\mu = 1$.

To determine the 95% confidence interval of the predicted value of Y_i at X_i , first determine the variance of Y_i as $s_{\text{res}}^2 \left(\frac{1}{N} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right)$.

To calculate the 95% confidence limits for the intercept, use the same equation but substitute zero for X_i . For the data of Table 27.4, this becomes $\text{Variance} = 3.1 \left(\frac{1}{52} + \frac{5.633^2}{404.29} \right) = 0.30$, so that the standard deviation is $\sqrt{0.30} = 0.55$. Then the 95% confidence limits of c are $29.41 \pm 2.004 \times 0.55 = 28.31$ to 30.51 .

Many of these additional statistics can be calculated online at <http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Regression.htm>, http://www.xycoon.com/simple_linear_regression.htm, and <http://www.wessa.net/slr.wasp>.

Residuals

Before going further, check linearity and homogeneity of variance. One simple way is to plot the residuals, the values of $Y_i - \hat{Y}_i$ against the corresponding value of X_i (or \hat{Y}_i which has an exact correspondence to X_i) (Fig. 27.11). If there is reason to suspect an influence of time, the residuals can be plotted against the order or the time at which they were acquired.

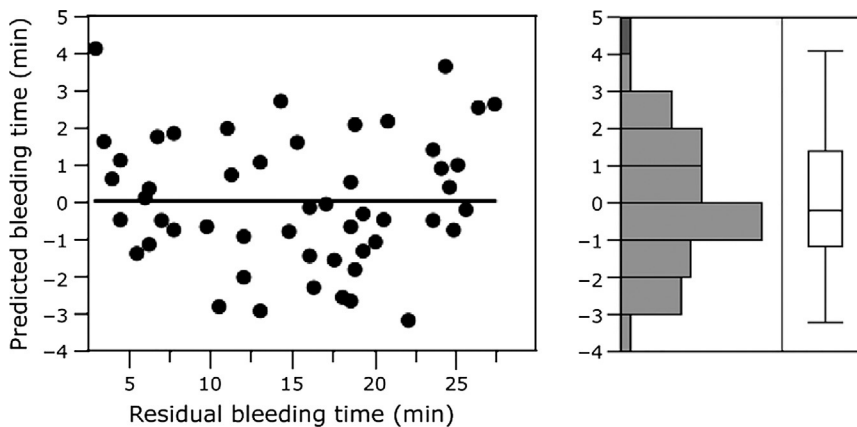


Fig. 27.11 Residual plot. This shows a satisfactory plot with the residuals scattered equally above and below the line. Residual variability does not change materially as Y changes.

Although free online programs for plotting residuals directly are not available, the online programs http://vassarstats.net/corr_stats.html and <http://www.xuru.org/rt/LR.asp#> calculate the residuals (termed “error” in the latter), and these can then be

plotted against the original X or Y values. Another useful assessment of residuals is to plot them on normal probability paper to find out if they form a straight line (as they should) (Fig. 27.12) or if there are distortions that might indicate some abnormality of their distribution.

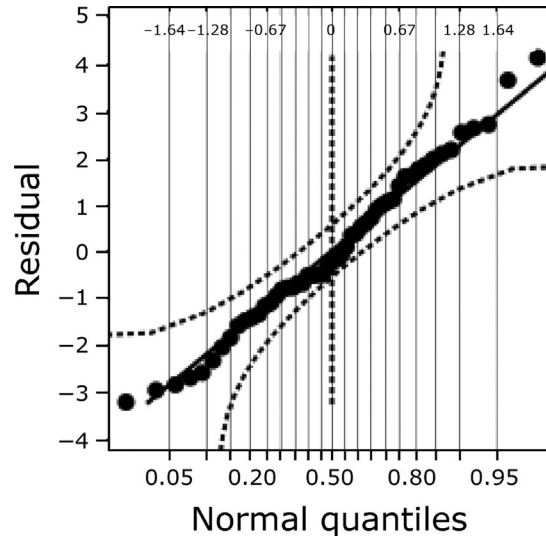


Fig. 27.12 Normal quantile plot of residuals.

Fig. 27.13 shows data from Jamal et al. (2001) who were comparing how changes in dP/dt max affected systolic strain (%).

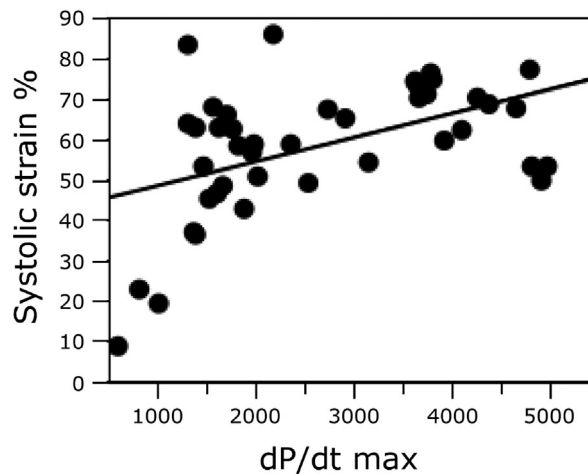


Fig. 27.13 Data from Jamal et al. of systolic strain % vs dP/dt max. The line of best fit by linear regression (which they did **not** use) fits the points poorly, as confirmed by the residual plot (Fig. 27.14).

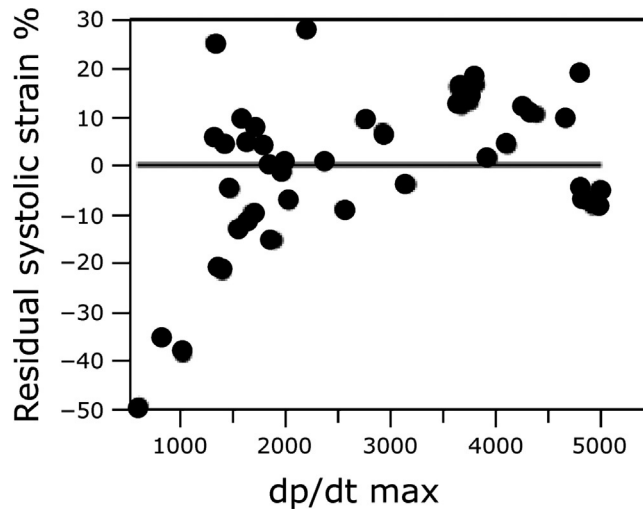


Fig. 27.14 Residual plot of linear regression. At each end of the X values the residuals are below the line, and in the center most are above it. This is typical of a curvilinear association. When the data were correctly fitted by a second-order polynomial (as they did correctly), the residuals show more acceptable behavior (Figs. 27.15 and 27.16). Redrawn from Wolters Kluwer.

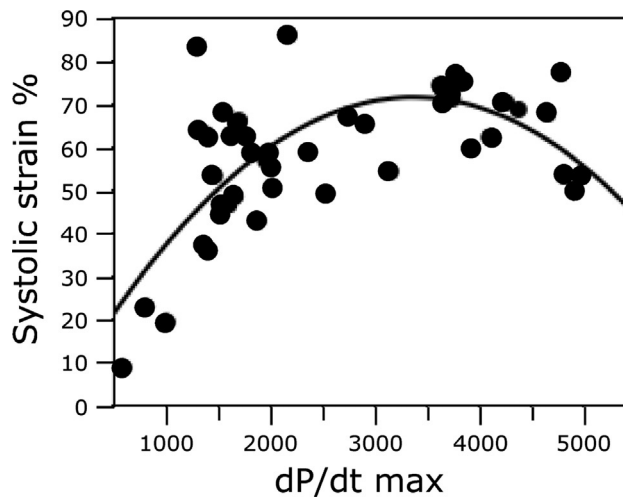


Fig. 27.15 Quadratic fit. (Data from Jamal, F., Strotmann, J., Weidemann, F., Kukulski, T., D'hooge, J., Bijmens, B., Van De Werf, F., De Scheerder, I., Sutherland, G. R. 2001. Noninvasive quantification of the contractile reserve of stunned myocardium by ultrasonic strain rate and strain. *Circulation* 104, 1059–65.)

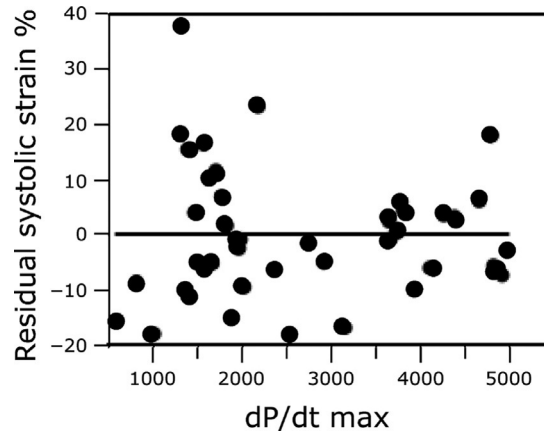


Fig. 27.16 Residuals from quadratic fit. The residuals are normally distributed.

Another example of the use of residual plots is shown in [Fig. 27.17](#) in which the residuals from the reexpressed curves from [Fig. 27.4](#) are shown.

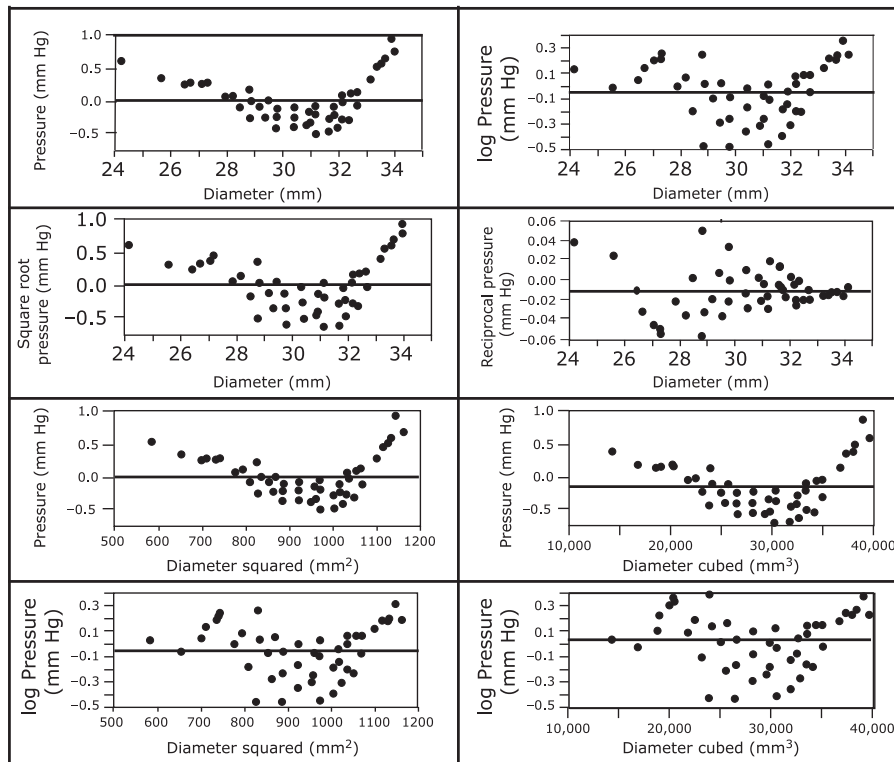


Fig. 27.17 Residual plots from a number of reexpressed curves. The best results with fairly even dispersion of the residuals occur with diameter vs reciprocal of pressure and diameter cubed vs logarithm of pressure, confirming the impression of linearity gained from [Fig. 27.4](#).

Problem 27.2 Draw the residual plot for the data from Problem 27.1.

The absolute residual variability varies with the units used and the size of the deviations. To allow for this, standardized residuals are calculated by dividing each absolute residual by the standard deviation from regression. This yields the residuals in standard deviation units, which are dimensionless and scattered above and below a zero line, just as in a univariate distribution the X variate is transformed into standard deviation units above and below zero mean by the z transformation. Then about 95% of the standardized residuals should lie within 2 standard deviations of the zero line.

The effects of a marked increase in variability as X increases can be seen well in a residual plot based on the artificial data of Fig. 27.6 (Fig. 27.18).

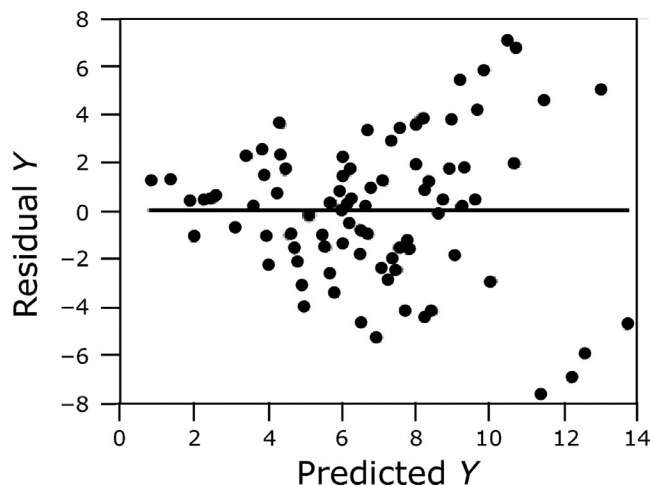


Fig. 27.18 Residual plot to show variability increasing greatly as X increases.

Two other residual plots need to be considered. Fig. 27.19 shows residuals that are not independent of each other:

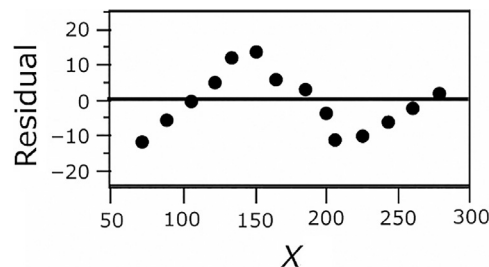


Fig. 27.19 Positive autocorrelation of residuals.

The plot shows a tendency for one positive or negative value to be associated with another positive or negative value, respectively, suggesting that some factor, frequently time, is influencing these residuals. There are fewer changes in direction than would be expected by chance. Negative autocorrelation is shown if positive and negative residuals alternate, again something unexpected by chance (Fig. 27.20). Time-dependent residuals may be tested specifically by the Wald-Wolfowitz runs test or the Durbin-Watson test (Chapter 31).

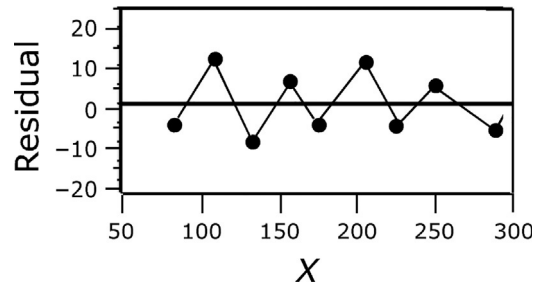


Fig. 27.20 Negative autocorrelation.

Methods of evaluating autocorrelation are described in Chapter 31.

When the results of a linear regression are presented, in addition to the X-Y plot give the equation to the line of best fit, and **always** give the standard deviation from regression. This is the square root of the MS_W (residual or error mean square). In Table 27.6 the MS_W was 5.856, so that the standard deviation from regression was 2.42 (In the computer printout, this value was termed “root mean square error”). In other words, the “average” deviation of data points from the regression line was 2.42. This is the equivalent of the standard deviation of data from a mean in a univariate distribution.

In Chapter 25 ANOVA was used for different dietary additions of vitamin B₁₂ to determine the effect in growing pigs (Richardson et al., 1951). The different groups were labeled A, B, C, and D. If these were different vitamins, then an ANOVA would be the right form of analysis. In fact, the columns represented increasing amounts added to the diet (in µg/lb. ration): 0, 5, 10, and 20 µg/lb. This means that there is the possibility of a regression relationship between amount of additive and weight gain, and this is shown in Fig. 27.21.

There is a rough linear relationship between the amount of Vitamin B₁₂ added and the weight gain with a low slope, with $P = 0.0802$. The 95% confidence limits for the slope are $0.0049714 \pm 2.228 \times 0.002554$, or from -0.0007189 to 0.01066 . In the direct ANOVA of Chapter 25 the standard deviation of weight was 0.0730, and the F ratio was 1.08, with a P value of 0.4109. Introducing the regression relationship has reduced the standard deviation of Y to 0.065 when X has been taken into account. If the effect of vitamin B₁₂ additive might be important, it would be worth repeating the experiment with more animals.

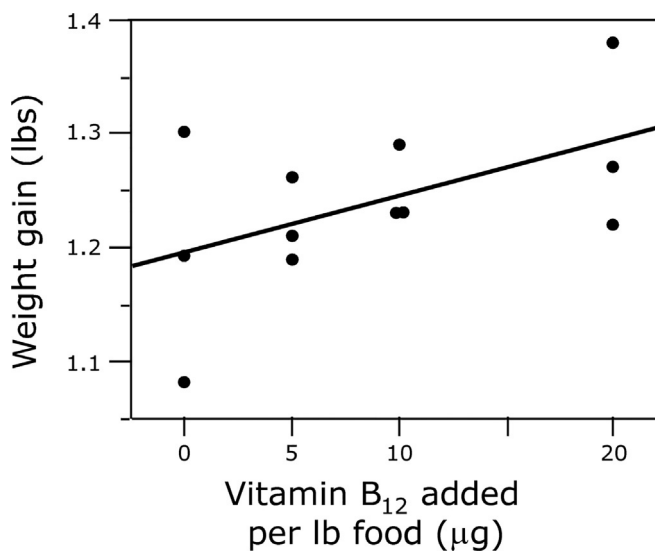


Fig. 27.21 Vitamin B₁₂ data shown as a regression.

Confidence Limits

The next order of business is to set confidence limits on the slope, just like confidence limits for the mean of a univariate distribution. It is more complicated for a bivariate distribution because there are two ways in which the slope can vary (Fig. 27.22).

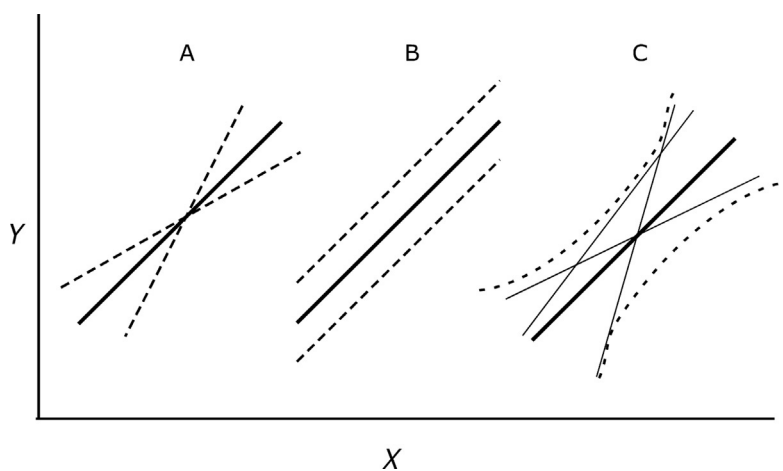


Fig. 27.22 Diagram of confidence limits for a line of best fit. *Thick solid line*—observed sample slope. *Thin solid lines*—other possible slopes. *Dashed lines*—possible variations.

Panel A shows that the population slope might be steeper or less steep, pivoting about the same central point. Panel B shows that the population line of best fit might be higher or lower than the sample value with the same slope. Both of these changes can occur independently; the population slope might be higher and steeper, higher and less steep, lower and steeper, or lower and less steep. When these choices are combined the confidence limits appear as curved biconcave lines around the sample line. Any straight line that can be fitted to lie completely within those boundaries is a possible population slope, as shown by the thin lines in panel C.

An example of 95% limits for the platelet count-bleeding time data shown in Fig. 27.10 is given in Fig. 27.23.

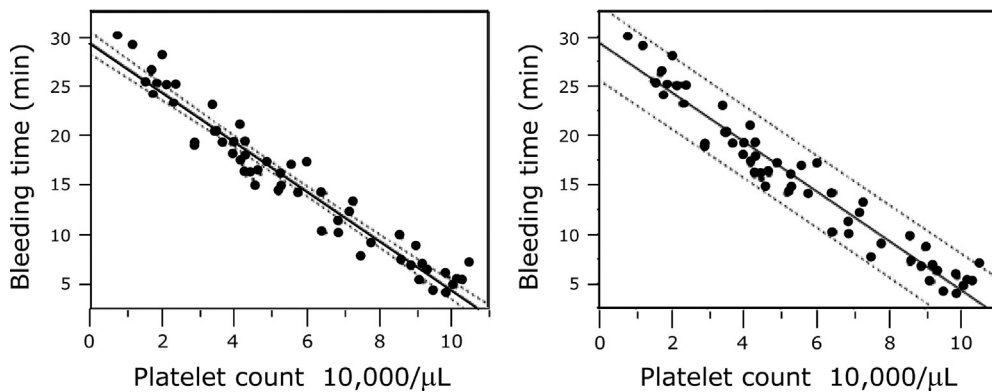


Fig. 27.23 Left panel: 95% confidence limits (*dashed lines*) for line of best fit. Right panel: 95% confidence limit for points.

These calculations can be done https://www.statstodo.com/CorReg_Pgm.php (for slope only) and https://www.xycoon.com/simple_linear_regression.htm.

A crucial distinction must be made between confidence limits for the line (which is a mean that changes as X changes) and confidence limits for individual points. For example, to use the data shown in Fig. 27.10 to determine if an individual value was abnormal and outside expected limits, add another component of variability, as shown in the right-hand panel. The points are the same in both panels.

The outer boundaries for individuals are set so that not $>5\%$ of individual points lie outside the boundaries. These boundaries are further from the line of best fit than are the inner boundaries because individual values vary more than means do. The computations and the regression line with confidence limits for individual points can be performed online at http://www.xycoon.com/simple_linear_regression.htm; the latter requires considerable experience with Excel as well as downloading the free module StatPlus:mac LE. The confidence

interval for the intercept can be calculated at <https://easycalculation.com/statistics/regression-intercept-interval.php>, <http://danielsoper.com/statcalc3/calc.aspx?id=27>.

Problem 27.3 Plot the 95% confidence limits for the data from problem 27.1.

To report how much steeper or less steep the slope could be, calculate s_b^2 , the variance of the slope, as.

$\frac{MS_W}{\sum(X_i - \bar{X})^2} = \frac{3.1}{404.29} = 0.0077$. Then the standard deviation of the slope is $\sqrt{0.0077} = 0.088$. (This is presented in Table 27.5, bottom line.)

The 95% confidence limits of the slope are $b \pm t_{0.95, n-2} s_b$, or $-2.52 \pm 0.0088 \times 2.004 = -2.54$ to -2.50 .

Sometimes we want to know how much the mean of Y related to X can vary, that is, we calculate the square root of $s_{Y.X}^2$. This value is obtained from $s_Y^2 - X/N = 3.1/50 = 0.2662$, the square root of which is 0.062. This is comparable to the standard deviation of the mean for a univariate distribution.

It is from these values that the confidence limits for the line of best fit and for individuals are computed. The curved confidence boundaries shown in Figs. 27.22 and 27.23, however, cannot be derived from measurements at a single point, but for the reasons given in Fig. 27.22 are narrowest at the mean of X and become wider on each side of it. They are a function of $\sum(X_i - \bar{X})^2$, and this means that the boundaries get wider as they move away from the mean of X . This is not unreasonable for values above the mean of X because these larger values should vary more. At the lower end, however, there is a mathematical artifact. Examine the data in Fig. 27.24, redrawn from Buckberg et al. (1971).

The dashed lines on either side of the solid line of identity indicate 20% above and below the line of identity. Most measurements fall within the range of +20% to -20%. The smaller values of X and Y show less variation than do the larger values, something expected when the coefficient of variation remains constant. Had confidence limits about the regression line been calculated, as described before, they would have widened out at the lower end (Fig. 27.25). The degree to which the lines flare out depends on the correlation between the two variables, being minimal if all the points are near the line and flaring widely if the points are widely scattered about the line.

In any linear regression, it is important not to extrapolate beyond the observed limits. The relationship might change direction or even curvature if measurements are made outside the observed limits.

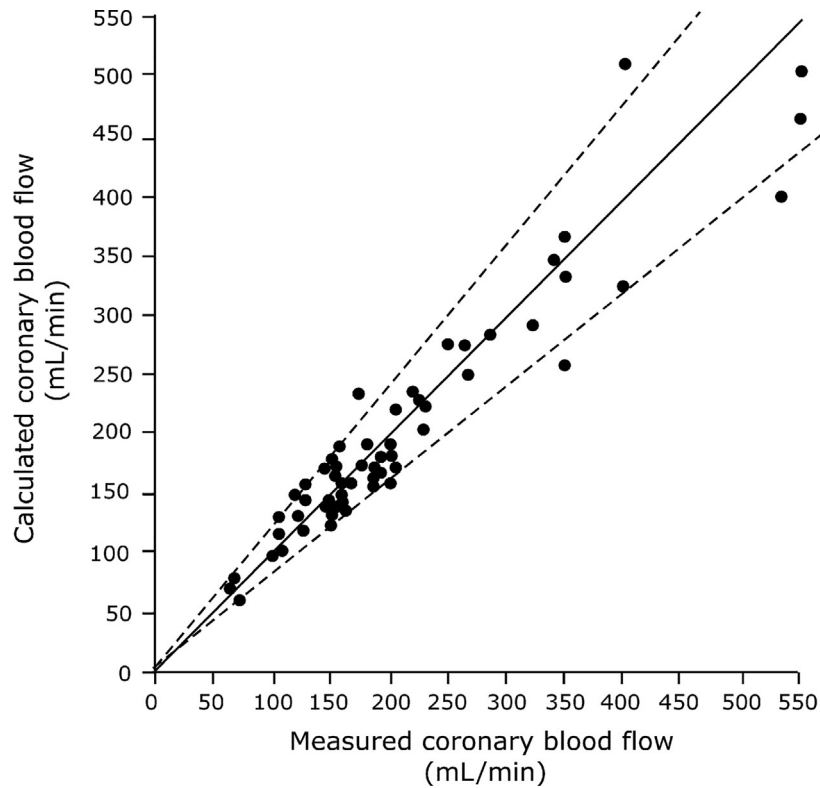


Fig. 27.24 Measured vs calculated coronary blood flow in sheep and dogs.

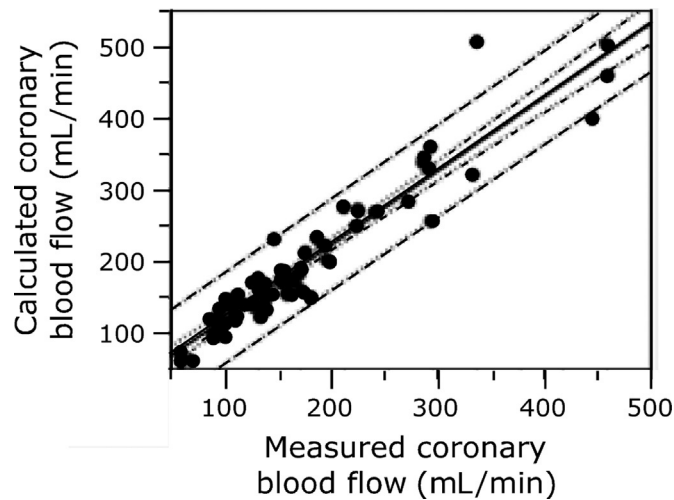


Fig. 27.25 Regression of measured vs calculated coronary blood flow, with 95% confidence limits. The limits at the lower end are artifactually wide.

Comparison Methods (Basic)

Some comparisons involve two ways of measuring the same thing, for example, comparing a new quicker or cheaper test against an older very reliable but complex method of measuring the concentration of a particular chemical. Because neither value is dependent on the other, but rather both rely on the same underlying phenomenon (the amount of material present), some authorities consider it wrong to designate one as X and the other as Y (Bland and Altman, 1986, 1999; Ludbrook, 1997). They pointed out that in these comparison measurements the slope of the regression line and the correlation between the two methods were not the main objects of the study. Rather, it was to show how much the two methods differed from one another. Bland and Altman (1999) and Ludbrook (1997) argued that the classical Model I regression provides the degree of association between the two variables, merely tests their linear relationship, and may not emphasize important differences between the methods. Furthermore, because both the X and Y variables have error but neither is dependent on the other, there are really two regression lines, one minimizing the vertical Y differences from regression and one minimizing the horizontal X differences from regression.

What such a comparison should show is whether there is a systematic difference between the two measurements, a difference that might be fixed (constant at all values of X) or proportional (increasing as X increases). Frequently these methods obtain some type of average of the slopes of Y on X and X on Y . Bland and Altman recommended plotting the difference between the two measurements on the Y -axis against the mean of the two measurements on the X -axis. This is based on the fact that the X - Y difference is uncorrelated with either X or Y when X and Y are both measurements of the same object (Bland and Altman, 2003). They also calculated what they termed the 95% limits of agreement as mean difference \pm standard deviation of the difference, a range expected to include about 95% of the observations. The plot can be made using Excel; see <http://medlabstats.com/bland/BLAND-AND-ALTMAN-PLOTS-IN-EXCEL.pdf>.

An example is provided in Fig. 27.26, using the microsphere data of Buckberg et al. (1971) that were plotted in Figs. 27.24 and 27.25.

Problem 27.4 Use the data set from Nagy et al. above to construct a Bland-Altman plot.

Bland and Altman pointed out that certain requirements were needed, the most important of which were that the differences were constant over the range of measurements and

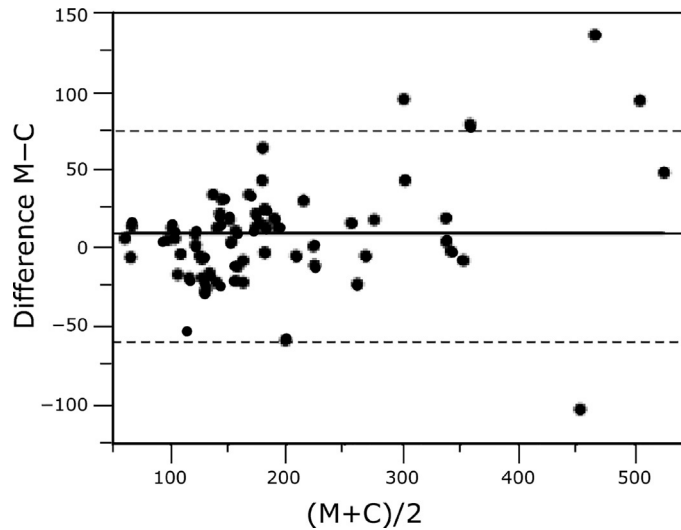


Fig. 27.26 Bland-Altman diagram of comparison between a standard method (C) and a microsphere method (M) of calculating coronary blood flow. The mean difference (*solid line*) is shown and is close to zero. The two *dashed lines* show 2 standard deviations above and below the mean and encompass approximately 95% of the observations. This depiction makes it easier to see differences.

were normally distributed. These can be checked visually on the scattergram, or more formally if needed. If the differences increase as the size of the measurement increases, then the mean might be correct but the limits of agreement would be too wide at small measurements and too narrow at large ones. To deal with this problem they advocated taking the logarithms of the measurements. If this is done, then the limits of agreement refer to proportions of the measurement rather than the original units. Other ways of dealing with this problem as well methods for determining the variability of the deviations are presented as follows.

Cautionary Tales

1. It is essential to examine the *XY* scatterplot before assessing regression or correlation. Examine [Fig. 27.27](#)

In the publication, the correlation coefficient (a measure of agreement between *X* and *Y*) was given as 0.21, with $P < 0.005$. The only way I can interpret their data is to state that the two variables are virtually completely unrelated, no matter what the test of the null hypothesis shows, and what the slope indicates is unclear. Not everything with a low P value has meaning.

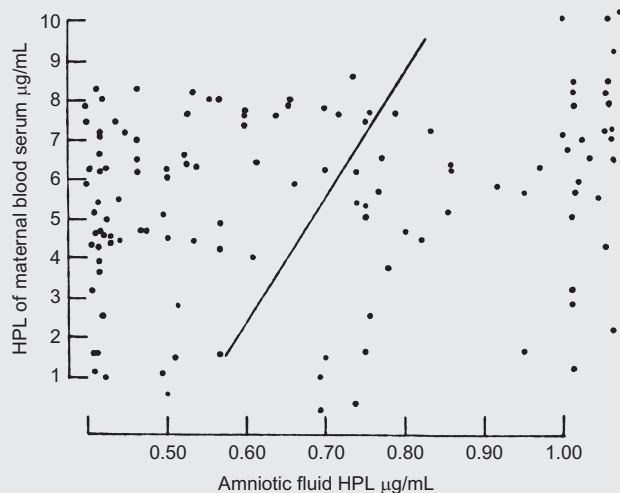


Fig. 27.27 Plot of amniotic fluid human placental lactogen (hPL) levels on X-axis vs maternal serum hPL level on the Y-axis (Lolis et al., 1977). (Reproduced with permission of the publishers.)

2. The importance of examining the graphs was stressed in an instructive example devised by Anscombe (1973). He created 4 data sets (Table 27.7).

Table 27.7 Anscombe's data sets

Data set	1-3	1	2	3	4	4
Variable	x	y	y	y	x	y
Obs. no. 1 :	10.0	8.04	9.14	7.46 :	8.0	6.58
2 :	8.0	6.95	8.14	6.77 :	8.0	5.76
3 :	13.0	7.58	8.74	12.74 :	8.0	7.71
4 :	9.0	8.81	8.77	7.11 :	8.0	8.84
5 :	11.0	8.33	9.26	7.81 :	8.0	8.47
6 :	14.0	9.96	8.10	8.84 :	6.0	7.04
7 :	6.0	7.24	6.13	6.08 :	8.0	5.25
8 :	4.0	4.26	3.10	5.39 :	19.0	12.50
9 :	12.0	10.84	9.13	8.15 :	8.0	5.56
10 :	7.0	4.82	7.26	6.42 :	8.0	7.91
11 :	5.0	5.68	4.74	5.73 :	8.0	6.89

TABLE. Four data sets, each comprising (11) (x, y) pairs.

Each of the four data sets yields the same standard output from a typical regression program, namely

Number of observations (n) = 11

Mean of the x 's (\bar{x}) = 9.0

Mean of the y 's (\bar{y}) = 7.5

Regression coefficient (b_1) of y on x = 0.5

Equation of regression line: $y = 3 + 0.5x$

Sum of squares of $x - \bar{x}$ = 110.0

Regression sum of squares y = 27.50 (1 d.f.)

Residual sum of squares of y = 13.75 (9 d.f.)

Estimated standard error of b_1 = 0.118

Multiple R^2 = 0.667

The first three sets have the same X variables, shown in the second column headed 1-3.
Reproduced with permission of the American Statistical Association.

Despite the fact that all of these data sets have identical equations and derived values, they are hugely different as shown by their plots (Fig. 27.28)

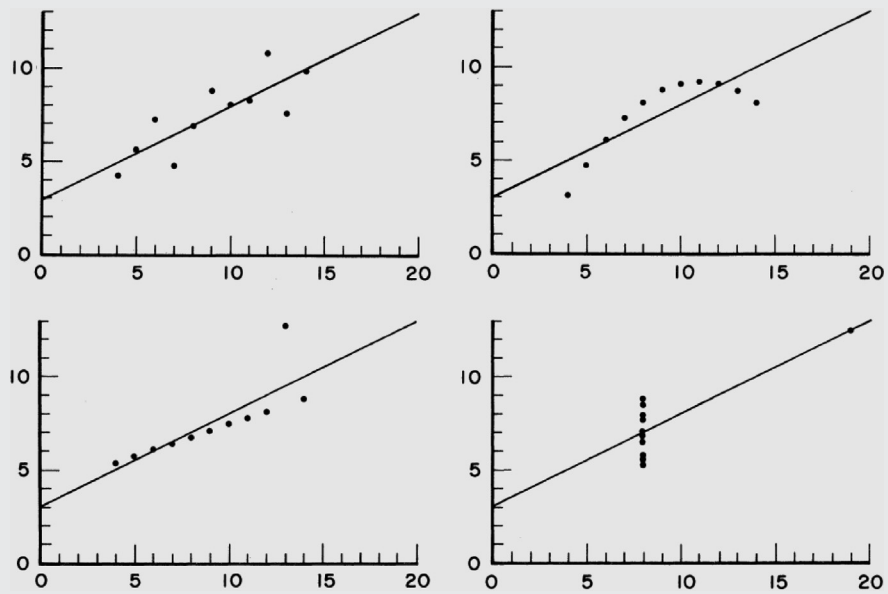


Fig. 27.28 Plots of Anscombe's data sets. Despite the identity of the regression equations (indicated by the *solid lines*), all four sets demonstrate different distributions of the XY values. (Reproduced with permission from the American Statistical Association.)

3. It is unwise to rely on a linear regression formula and a nonzero correlation coefficient to evaluate a relationship. Consider Fig. 27.29, a prototype of many in the literature:

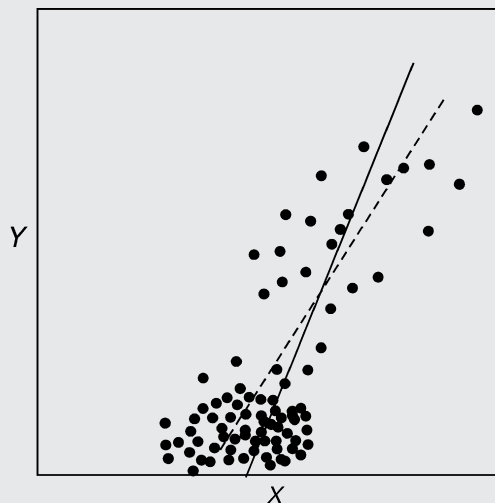


Fig. 27.29 "Inverted comet" figure.

There is a dense cluster of points at small values of Y , based on the ease of obtaining normal values. These dominate the graph and may give a slope (solid line) that is unrepresentative of the data; the dashed line might be more representative. Because of a large number of data points the correlation is high, but with the wide scatter the ability to predict Y from X is poor. The graph is not useless, but its use is limited. See Fig. 1 in [Nakauchi et al. \(1981\)](#), and Fig. 5 in [Olivíé et al. \(1995\)](#) as examples.

4. All data points must be represented in the figure and the calculations. [Mackler et al. \(1979\)](#) related iron deficiency and plasma phenylalanine concentration in rats. Instead of plotting all the points, they grouped averages for hemoglobin concentrations (g/100 mL blood): 4.5–5, 5–5.5, 5.5–6, 6–6.5, and 6.5–7. Each average had different numbers of measurements. Although the authors did not give a regression equation, they did state that the relationship was linear.

In a letter to the editor, [Bryant \(1980\)](#) pointed out that pooling different numbers for the means masked the variability. In the original publication the authors presented a table with the individual measurements, (a most unusual occurrence) and these are plotted in [Fig. 27.30](#):

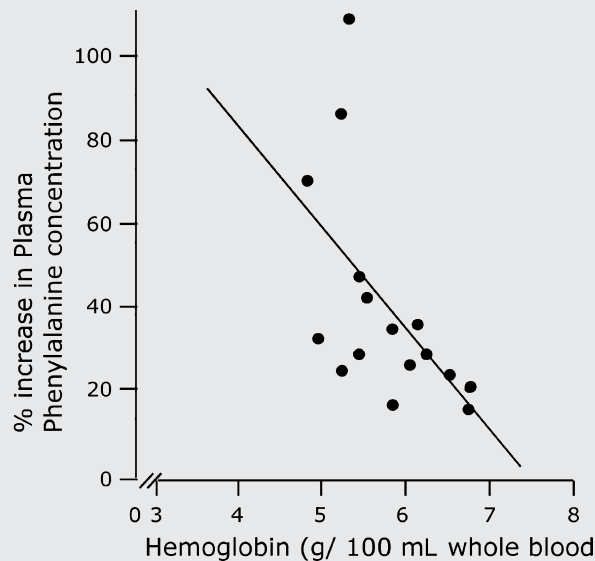


Fig. 27.30 XY plot with individual points.

There is an inverse correlation between X and Y , but this is not linear, and perhaps there is a threshold at about 5.5 g/100 mL hemoglobin, or a curved relationship. Taking averages with different sample sizes gave a false impression of linearity.

5. It is unwise to extrapolate a value of Y from a value of X that is much smaller than the lowest or much bigger than the highest X value used in constructing the regression line. Unless there is evidence for extended linearity based on other work, linear extrapolation may lead to incorrect predictions. In the platelet bleeding time discussed before, the linear relationship between the variables was absent at very low or very high platelet counts ([Fig. 27.31](#)).

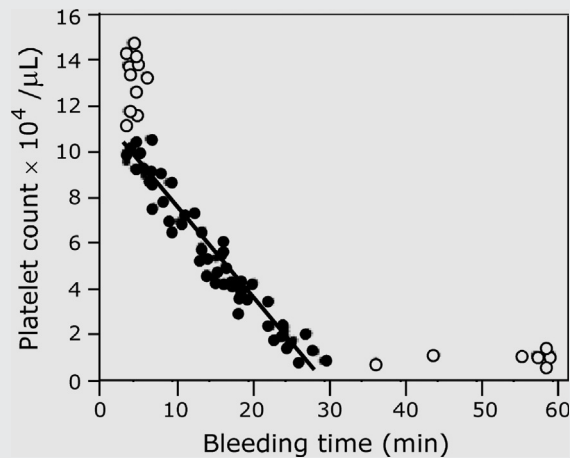


Fig. 27.31 Platelet count vs bleeding time. Note departure from linearity at the extremes, shown as open circles. Dark circles and line indicate linear range for the data. (Based on data from Harker, L.A., Slichter, S.J., 1972. The bleeding time as a screening test for evaluation of platelet function. *New Engl. J. Med.* 287, 155–159.)

ADVANCED OR ALTERNATIVE CONCEPTS

Heteroscedasticity

If the scattergram suggests heteroscedasticity, it may be worth testing this formally before deciding on the need for transformations. A plot of residuals will confirm the appearance of heteroscedasticity. Many tests of heteroscedasticity have been developed. You do not need them all, but they are frequently mentioned and used so that an understanding of what they are is useful. Two are basically ANOVA tests:

The Goldfeld-Quandt test divides the data set into two groups (omitting c measurements—about 20%—in the middle of the distribution), regresses Y on X for each group, calculates the residual sum of squares (SS_{res}) for each group, and does an F -test $\frac{SS_{\text{res(larger)}}}{SS_{\text{res(smaller)}}} = F_{(n_1-c, n_2-c)}$, where n_1 and n_2 are the numbers in the lower and upper data sets. A high value for F suggests heteroscedasticity (Example 27.1).

Example 27.1

For the data shown in Figs. 27.6 and 27.17, the residual sum of squares (SS_{res}) was 10,636 for the lower 34 X values, and 52,710 for the upper 34 X values. Their ratio of 4.96 is assessed by $F_{(17,17)}$, with $P \approx 0.001$. The two error terms are not similar, and the hypothesis of heteroscedasticity is supported.

The modified Levene test divides the data into halves, regresses Y on X for each half, estimates the median of each residual group, determines for each residual group separately the absolute deviations from the median, calculates the mean absolute deviation in each group, and then compares these two mean deviations by t -test or ANOVA. A large t or F value indicates that the null hypothesis can be rejected. An advantage of the Levene test is that if the regression appears to be curvilinear as well as heteroscedastic, the data set can be divided into three or more groups, and an ANOVA can be performed on the absolute deviations (Example 27.2).

Example 27.2

Dividing the data into two groups, the mean absolute deviations of the residuals are 18.9 for lower values of X and 34.6 for the higher X values. By t -test the difference was 15.7 with $P=0.0003$.

These tests may lose power if the Y distributions are not normal and numbers are small, and cannot readily be extended to more than one X variable, so other tests are often used. Those most often used are tests by Park, Glejser, and Breusch and Pagan (the last test described independently by Cook and Weisberg). They all have similar forms in that an auxiliary regression is performed between the residuals (ϵ_i) obtained by regressing Y on X .

1. Park test. $\log \epsilon_i^2 = c + \beta_1 \log X_1 + v_i$, where v_i is the new error term in the auxiliary regression. This is tested by the Lagrange Multiplier (LM) method, in which Nr^2 is approximately equal to chi-square with degrees of freedom equal to the number of independent variables, here 1. The regression equation was $Y = 3.26 + 0.077X$, and r^2 was 0.168658. The metric was $85 \times 0.168658 = 14.34$ with 1 Df, so that $P = 0.0002$, again attesting to heteroscedasticity.
2. Glejser test. $|\epsilon_i| = c + \beta_1 X_1 + v_i$, where $|\epsilon_i|$ is the absolute value of the deviation. The resultant linear regression equation was $Y = 1.0057X - 5.85$, with $r^2 = 0.36$. Using the LM method, the chi-square approximation was $85 \times 0.36 = 30.6$, also with 1 Df and $P \leq 0.00001$.
3. Breusch-Pagan test: This test is used more often than the others. The basis of the test is to determine if the squared residuals are constant at each value of X . If the sum of squares due to regression is small relative to the residual (or error) sum of squares, then the null hypothesis cannot be rejected.

To perform the test,

1. Regress Y on X , determine the residuals ϵ_i , square these residuals, and scale them by dividing each squared residual by the mean of the squared residuals

$$\text{Scaled } \epsilon_i^2 = \frac{\epsilon_i^2}{\frac{\sum \epsilon_i^2}{N}}.$$

This results in a mean squared residual of 1 that is necessary for evaluating the expression.

2. Then perform auxiliary regression (A) of these scaled squared residuals against X_i .
3. Take the resultant auxiliary regression sum of squares (model SS_A or SS_{regA}), divide it by 2, and then use that as an approximation to chi-square with 1 degree of freedom.
4. The same result can be obtained by regressing the unscaled squared residuals against X_i , dividing by 2, and then scaling at the end by dividing by square of the mean squared residual:

$$\frac{SS_{\text{regA}}}{2(\bar{e}^2)^2}. \text{ An alternative expression is } \frac{SS_{\text{regA}}}{2\left(\frac{SS_{\text{res}_o}}{N}\right)^2}, \text{ where } SS_{\text{res}_o} \text{ is the residual sum}$$

of squares of the original regression of Y on X . The two expressions in the parentheses are identities. Any of these three forms may appear in the literature.

As an example, consider the scattergram and residual plot shown in Figs. 27.6 and 27.17.

The regression equation (numerical data not shown) is $\hat{Y}_i = 0.31 + 0.40X_i$ with $r^2 = 0.4566$. The mean squared residual was 63,014.23, and the residual sum of squares was 74,987.74, with $N = 85$.

Regressing the scaled squared residuals against X_i yields 61.21 for the regression sum of squares. This divided by 2 gives 30.6 as an estimate of chi-square with 1 degree of freedom, for $P = 3.18e-8$, confirming our belief that variance was not independent of X .

With the unscaled squared residuals, the expression in parentheses would be $\left(\frac{SS_{\text{res}_o}}{N}\right)^2 = \left(\frac{74987.74}{85}\right)^2 = 778292.2$, and this divided into the regression sum of squares for the unscaled squared residuals of 47,635,724 gives 61.205 as before. This is then divided by 2 to give the metric.

An alternate way to approximate the chi-square value is to use the LM method and multiply the value of r^2 by N . In this data set r^2 for regressing the squared residuals (scaled or unscaled) was 0.356746. This multiplied by 85 gives 30.32, close to the estimate for chi-square obtained before.

For this test to be interpreted, the relationship between Y and X must be linear.

The Comparison Problem (Advanced)

Because the limits of agreement shown before for the Bland-Altman plot are point estimates, Bland and Altman (1999, 2003) also advised setting confidence intervals about

these limits. They used as an approximation for the standard error of the limits $\sqrt{\frac{3s^2}{N}}$, where s was the standard deviation of the differences, and N the total number of comparisons. Thus if, as in an example that they used, the mean difference of 231 comparisons was 0.2 mm with $s = 3$ mm and the 95% limits of agreement being -5.8 to $+6.1$ mm, then

the standard error would be $\sqrt{\frac{3(3)^2}{231}} = 0.34$, and the 95% confidence interval is $\pm 1.96 \times 0.34 = 0.67$, so that the 95% confidence intervals for the limits will be: lower limit $= -5.8 \pm 0.67 = -5.13$ to -6.47 and for the upper limit $= 6.1 \pm 0.67 = 5.43$ to 6.77 . Ludbrook (2010) has taken this further. Instead of the term 95% limits of agreement he prefers to use the concept of 95% tolerance limits with 95% confidence. Ludbrook modified the plot of difference versus mean to show the mean difference (as in the usual Bland-Altman plot) and two pairs of limit lines; the inner pair are the upper and lower 95% confidence limits for the population, and the outer pair are the 95% tolerance limits with 95% confidence. For the inner limits he proposed using a slightly more accurate formula: $\bar{X}_{\text{diff}} \pm (t_{0.05, N-1}) s_{\text{diff}} \sqrt{1 + \frac{1}{N}}$. For the example used before, the 95% population confidence limits are $0.21.96 \times 3 \sqrt{1 + \frac{1}{231}} = 0.25.89 = 6.09$ to -5.69 . These are slightly different from those calculated from the simpler formula. For the tolerance limits Ludbrook used mean difference $\pm k s_{\text{diff}}$, where k is taken from the tolerance tables referred to in Chapter 7. For $N = 231$, k is ~ 2.131 , and the tolerance limits are $0.2 \pm 2.131 \times 3 = 6.56$ to -6.16 . If the differences increase in proportion to the size of the measurement, then instead of taking logarithms, Ludbrook recommends carrying out Model I regression of differences versus mean values, and then calculating the hyperbolic 95% limits for points; these are essentially the tolerance limits.

The Bland-Altman comparison method assumes that the two methods are giving approximately the same results for each measured datum. However, it is possible that one method might measure proportionately more or less than the other; for example, the new method might measure twice as much as the standard method. If that occurred, the slope of an X-Y plot much greater or < 1 would cast doubt on the new method, although the regression could be used as a calibration curve.

Comparing Two or More Lines

If two (or more) sets of linearly related X and Y variables are drawn at random from a population, the sets will probably have different regression lines of best fit, both in terms of slope and elevation. Did these lines come from the same population? The method used is termed analysis of covariance or ANCOVA, and it is a two-step process. In the first step the procedure determines if the slopes are substantially different. If they are very different, it may not make sense to continue (Fig. 27.32, left panel).

If the slopes are very different, then pooling them may make little sense and predicting Y from X will be incorrect everywhere except where the two lines cross. Below the crossing point Y values of group 1 for any X value are lower than those of group 2, but above

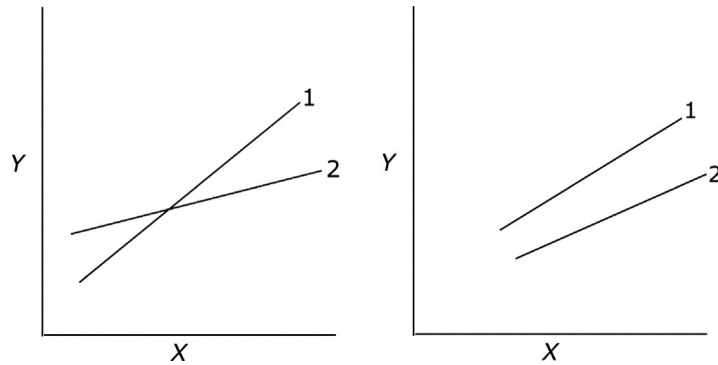


Fig. 27.32 Different slopes.

the crossing the Y values of group 1 are higher than those of group 2. On the other hand, examining the slopes over a restricted range of X may have value (Fig. 27.32, right panel), especially because any pair of nonparallel lines must cross somewhere.

Why do we need to know about common slopes and positions? Consider Fig. 27.33.

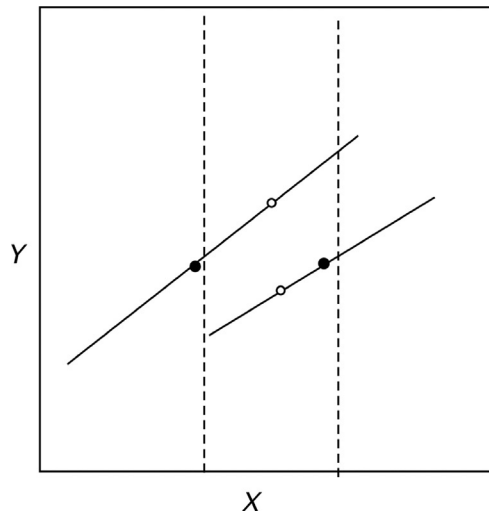


Fig. 27.33 Diagram showing two lines with the same mean values of Y (black dots) if the covariate X is ignored. If the two lines are compared at the same values of X (shown between the vertical dashed lines) then the mean Y values (open circles) are different.

Comparing the lines without correction for the covariate is a common error (Vickers, 2001; Vickers and Altman, 2001).

The computations should be done by computer programs that will calculate the mean values adjusted for the covariates. Online calculations may be done easily at <http://vassarstats.net/vsancova.html>, and http://www.statstodo.com/Compare2Regs_Pgm.

<http://vassarstats.net/textbook/index.html> and at <http://www.biostathandbook.com/ancova.html>.

As an example, consider Fig. 27.34 that shows two regression lines. The means of groups A and B are similar, 36.7 and 33.3, respectively. These values do not take account of the covariate X that is quite different for the two groups.

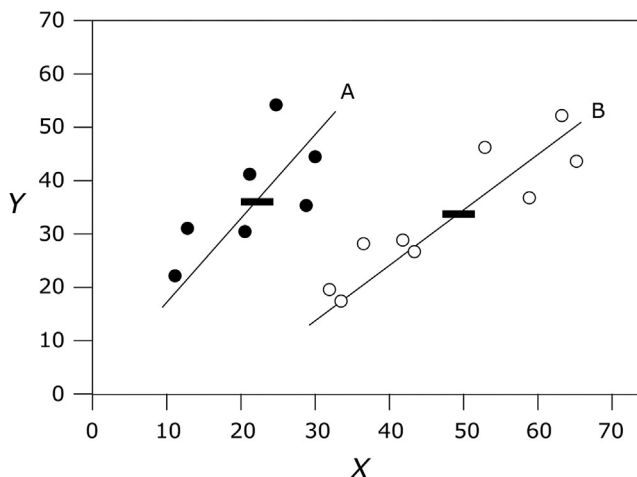


Fig. 27.34 Example for ANCOVA.

Stage 1. Compare the slopes for homogeneity. This is done by calculating a pooled (weighted) slope as

$$b_{\text{pooled}} = \frac{\sum (X_{i1} - \bar{X}_1)^2 b_1 + \sum (X_{i2} - \bar{X}_2)^2 b_2}{\sum (X_{i1} - \bar{X}_1)^2 + \sum (X_{i2} - \bar{X}_2)^2}.$$

If the slopes are similar, then there will be little difference between the SS calculated for each pooled slope (SS_{pooled}) and for each actual slope, so that SS_B will be small relative to SS_W . This is presented in Table 27.8 and Fig. 27.35 (Stage 1).

Table 27.8 Test of homogeneity of regression

Source	SS	Df	MS	F	P
SS_{pooled} (regressions)	11.65	1	11.65	0.26	0.6194
SS_W (residual)	529.25	12	44.10		
SS_T (each regression separately)	540.90	13			

There is no reason to reject the null hypothesis of equality of regression lines. As in one-way ANOVA, the three rows are, respectively, SS_B , SS_W , and SS_T .

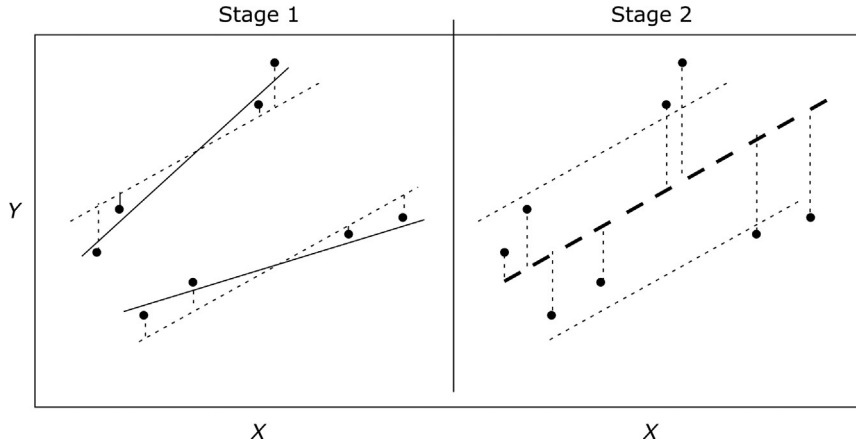


Fig. 27.35 Stages 1 and 2 of ANCOVA. As shown by the *dashed vertical lines*, the sum of squared deviations is greater from each pooled slope (*dashed lines*) than from the sum of each individual slope (*solid lines*) in the left panel, and greater from the common line (*thick dashes*) than from the sum of each separate line (*thin dashed lines*) in the right panel.

Stage 2. To determine if one line is markedly above or below the other, compare the total SS for all the points with the SS_W from both pooled slopes together. A large difference between them indicates that the null hypothesis of a common line can be rejected (Table 27.9 and Fig. 27.35, Stage 2).

Table 27.9 Analysis of covariance table

Source	SS	Df	MS	F	P
SS_B (adjusted means)	1021.03	1	1021.03	24.54	0.000264
SS_T (each regression separately)	540.90	13	41.61		
SS_T (all points)	1561.93	14			

The adjusted means are now very different, and we may safely reject the null hypothesis. As in one-way ANOVA, the three rows are, respectively, SS_B , SS_W , and SS_T , although SS_W and SS_T are not the same as in stage 1.

Stage 3. Calculate the means adjusted for equal covariates from

$$\text{adjusted } \bar{Y}_1 = b_{\text{pooled}}(\bar{X}_1 - \bar{X}_T)$$

(Table 27.10).

Table 27.10 Adjusted means

Means	Observed	Adjusted
A	36.71	49.58
B	33.33	23.33

Usually the lack of overlap is less marked than in this example, but unless the X values are identical for the two groups, adjustments have to be made.

The same method can be applied to comparing several different slopes, but unless there are specific hypotheses there are the same multiple comparison problems as occur with ANOVA without previous planning.

Outliers

In univariate distributions if a measurement is apparently far removed from the remaining measurements it will produce inefficient estimates of the mean and standard deviation, and thus make it more difficult to show differences between two groups. The same problem occurs in bivariate distributions and is important because unusual values can greatly alter the calculated regression line and the correlation coefficient (Fig. 27.36).

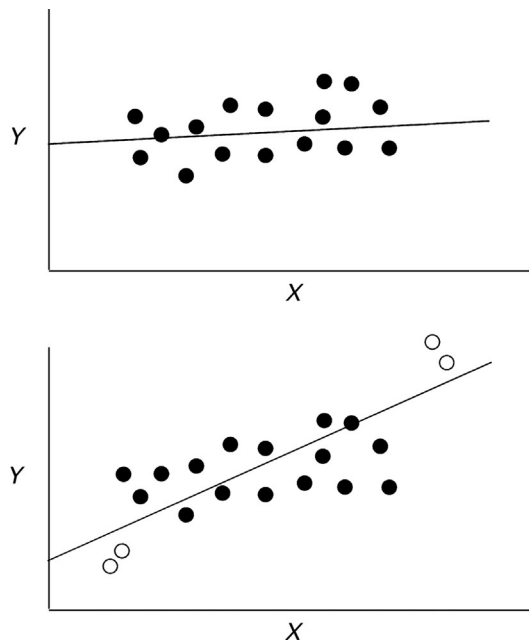


Fig. 27.36 Outliers. The upper panel shows a data set with little relationship between X and Y . The lower panel shows the same data points with four outliers (*open circles*) added. The slope of Y on X is now much different from 0, due entirely to the influence of a small number of measurements.

The problem is more complicated for bivariate than for univariate distributions because the point with unusual effect on the calculations is not always the point that appears to be most different from the rest.

Comparable to the univariate distribution, an outlier in regression can affect the mean and the standard deviation. The mean in a regression equation, however, has two components: the slope and the mean of the fitted value for Y (or the intercept). Any outlier may have a predominant effect on the slope, the mean of fitted Y , or the standard deviation from regression, or all three.

Leverage

Think back to the description of determining the confidence limits for a slope (Fig. 27.21). The reason why the slope is encased by concave lines was because as the X value moves farther and farther from the mean of X , the squared deviation gets larger and larger. Therefore points far from the mean have more effect than points near the mean, just as a weight placed on a beam on a fulcrum has more effect when far from the fulcrum than when close to it.

Consider Fig. 27.37.

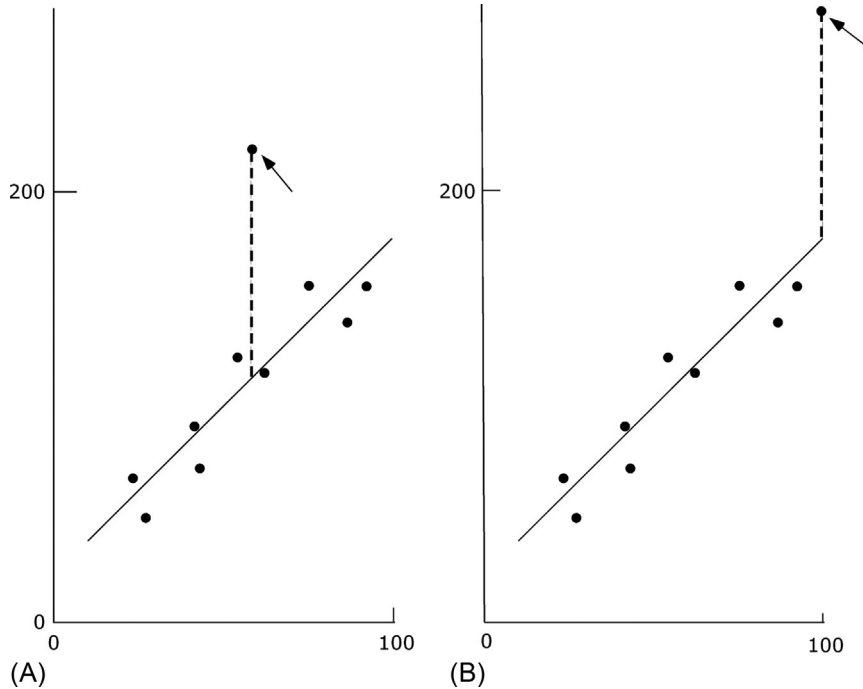


Fig. 27.37 Two XY plots identical except for the outliers (marked with an *arrow*) that have the same deviations from the regression line (*dashed line*).

The regression results are presented in Table 27.11.

Table 27.11 Regression constants for groups A and B compared with results when the aberrant point is omitted

Group	Intercept	Slope	MS residual
A	31.8	1.545	1569.5
B	−7.0	2.181	1178.4
No outlier	21.8	1.525	231.7

The aberrant point in B has a much greater effect on the slope and intercept. Both aberrant points increase the residual mean square that is not a method for distinguishing them.

The discussion to follow is based on the superb description given by [Glantz and Slinker \(2001\)](#). The effect of the position of an outlier on the regression line is known

as leverage. In simple linear regression the leverage is $h_{ij} = \frac{1}{N} + \frac{(X_i - \bar{X})^2}{\sum (X_j - \bar{X})^2}$, and that is

the way in which the standard deviation of \hat{Y} is calculated for each value of X (see earlier).

Ideally, all the points should have similar effect on the regression line, with an average value of $(k+1)/N$, where k represents the number of independent variables (only 1 for a simple linear regression) and N is the sample size. The average leverage in the previous example is $h_{ij} (1+1)/9 = 0.22$.

If any point has a leverage value more than twice the expected value, it should be subjected to further investigation.

There is a more accurate way of assessing leverage, and that is to make use of Studentized residuals. A Studentized residual is defined as

$$r_1 = \frac{e_i}{s_{Y.X} \sqrt{(1 - h_{ij})}}$$

These are standardized residuals that allow for the effects of the leverage factor h_{ij} . This expression is called the internally Studentized residual because it includes all the data points. There is also an externally Studentized residual in which each point in turn is omitted from the calculation of the residual, a technique known as the jackknife ([Chapter 34](#)):

$$r_{-1} = \frac{e_i}{s_{Y.X-1} \sqrt{(1 - h_{ij})}}.$$

These externally Studentized residuals should have an approximately normal distribution with most of the values within 2 standard deviations of the mean of zero, and these relationships can be shown in box plots or stem and leaf diagrams. They can also be tested in a normality plot constructed by plotting the ordered residuals against the cumulative frequencies derived from the equation $\frac{i - 3/8}{N + 1/4}$ where i is the rank of the residual and N is the number of residuals.

Leverage refers not specifically to unusual deviations from the regression line, but rather concentrates on the disproportionate effect of those deviations if they are far from the mean of X .

Influence

It is possible for a point to have little leverage, but because it is so far removed from the other data points it has a marked effect on the regression line. To deal with this problem there is a statistic called Cook's distance that incorporates all aspects of an excessive deviation. Each data point is removed in turn from the data set, the new regression line is computed, and the intercept is plotted on the X -axis against the slope on the Y -axis.

If all the points are reasonably distributed around the line, the relationship of intercept to slope will be similar for each revised data set, and all the points will be near each other. If removal of one particular point, however, produces a marked difference in location of the intercept-slope plot, then it is likely to warrant examination as an outlier. To judge the significance of such an outlying point, Cook's distance D is calculated as

$$D_i = \frac{r_i^2}{k+1} \times \frac{h_{ij}}{1-h_{ij}}.$$

D is a function of the internally Studentized residual (a function of the deviation of the point from the line) and the leverage effect. Cook's D has a distribution similar to the F distribution, specifically $F_{k+1, N-k-1}$. In general, a value of $D > 1$ suggests that the point(s) be considered carefully, and a value > 4 suggests a potentially serious outlier.

As a specific example, look at Fig. 27.38, redrawn from Butter et al. (2001).

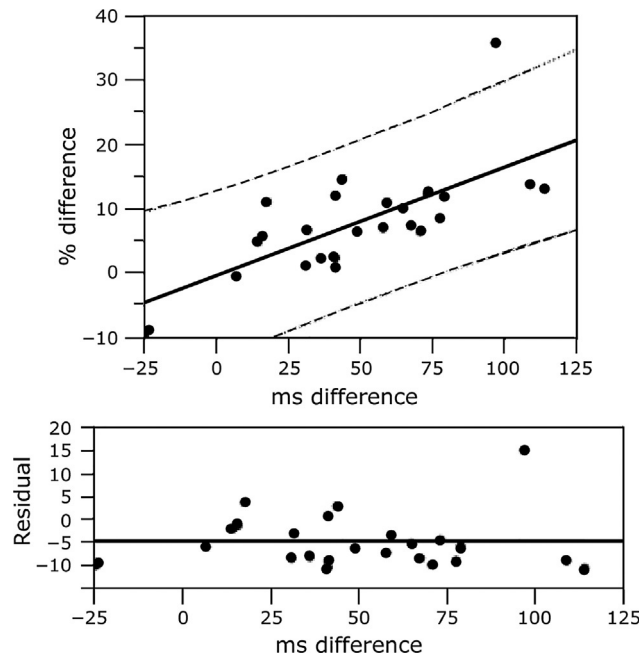


Fig. 27.38 Upper panel: Plot of intrinsic conduction delay difference between free wall and anterior wall (ms) vs % dP/dt difference (% difference) with biventricular pacing. The line of best fit and 95% confidence limits for points are shown. Lower panel: Residual plot.

There is one outlying value beyond the 95% limits for individual values. The expected value for leverage is $2/24 = 0.083$. There are two points that exceed $0.083 \times 2 = 0.166$. Neither of these is the apparent outlier; they come from the two with the largest values of X , even though they do not look very far from the regression line. However, they are pulling the line downwards. Cook's D shows that only the distance

for the apparently high point warrants examination, with a value of 0.90, the other two having values of 0.15 and 0.06.

If the regression analysis is repeated by removing the apparent high point with a big Cook's distance (97, 35.7) and then again with the full set but removing the two points with high leverage (109.1, 13.7 and 11.2, 13), the changes are presented in [Table 27.12](#).

Table 27.12 Three sets of regression results

Group	Intercept	Slope
All values	−0.4765	0.1688
Remove Y 35,7	0.7011	0.1271
Remove X 109, 114	−1.8074	0.2083

There are substantial differences depending on which points are omitted. Inasmuch as none of the tests for undue influence showed marked changes, there is no reason to do anything with these data other than to make sure that they are reliable. No matter what statistical tests may show, there is no substitute for common sense. Furthermore, removing data points can, as shown, make big differences to the regression constants, and should never be done without very good cause.

Ratio Measurements and Scaling Factors

Many anatomic or physiologic variables change with age or body size, for example, resting or maximal oxygen consumption or cardiac output, glomerular filtration rate, left ventricular mass, local blood flows, and so on. Two approaches are often used to determine if any subject's measurement of one of these variables is abnormal. One is to make a series of normal measurements that cover a range of ages or body sizes. The other is to normalize measurements by taking a ratio to body weight, surface area, or some other base, in order to obtain a single number to use in assessing normality.

If the regression relating the measurement Y to the value of the base X is linear and passes through zero, then we can use simplified arithmetic by using the ratio Y/X as a constant factor ([Fig. 27.39](#)).

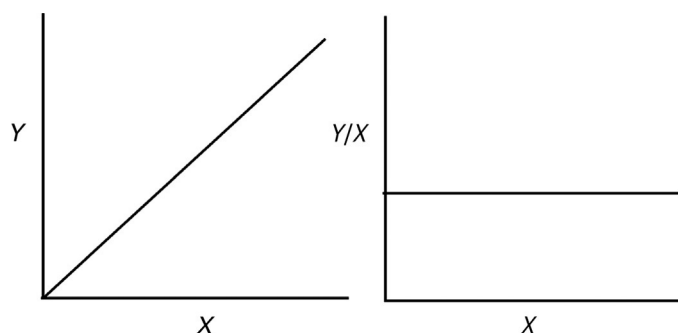


Fig. 27.39 If Y and X are linearly related and the regression line passes through zero, then we can use the common ratio of Y/X that is independent of variations in X to determine if a variable Y is normal.

Tanner, whose growth charts are well known, analyzed this issue with special reference to calculations of cardiac output (Tanner, 1949a, b). He pointed out that if the regression line between cardiac output and a measure of body size did not go through zero, then errors of judgment would result (Fig. 27.40).

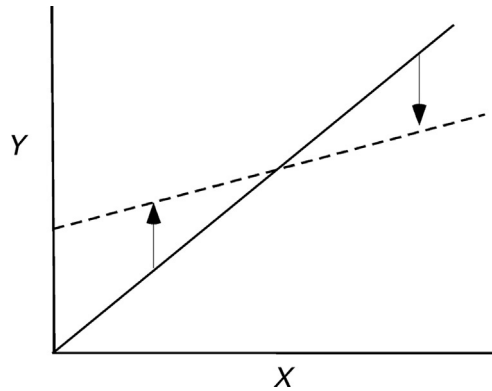


Fig. 27.40 Ratio vs nonratio measurements.

The solid line passing through zero implies a constant ratio between Y and X ; doubling X will double Y . If, on the other hand, the true (linear) relationship is shown by the dashed line, then at low values of X the constant ratio method underestimates the value of Y (upward arrow), and at high values of X the ratio method overestimates Y (downward arrow). If the two slopes are close the error may be unimportant, and the same conclusion can be made if the working range of X is small on either side of the point at which the two lines cross; the error increases when remote from the crossing point. The ratio method may underestimate cardiac output in infants and overestimate it in very large subjects (Hoffman, 2018), although these incorrect estimates probably have little clinical importance except in critical decisions based on pulmonary vascular resistance in infants. In general, incorrect use of a ratio leads to excessive standard deviations from regression and more difficulty comparing groups.

Many studies of oxygen consumption, cardiac output, and glomerular filtration rate, for example, have shown that a simple linear ratio is not accurate. For example, Armstrong and Welsman (1994) studied peak oxygen consumption versus age in boys and girls from 7 to 16 years old. Fig. 27.41 (left panel) shows the relationship between peak oxygen consumption on treadmill exercise versus age, and Fig. 27.41 (right panel) shows that correcting for body weight does not provide a constant range independent of age. Dallaire et al. (2015) showed recently that predicting z scores for pediatric echocardiography from BMI and age was unsatisfactory, and that polynomial models of the type $y = \sqrt{bx + c}$ or $y = ax^2 + bx + \sqrt{cx + d}$ gave the best predictive values.

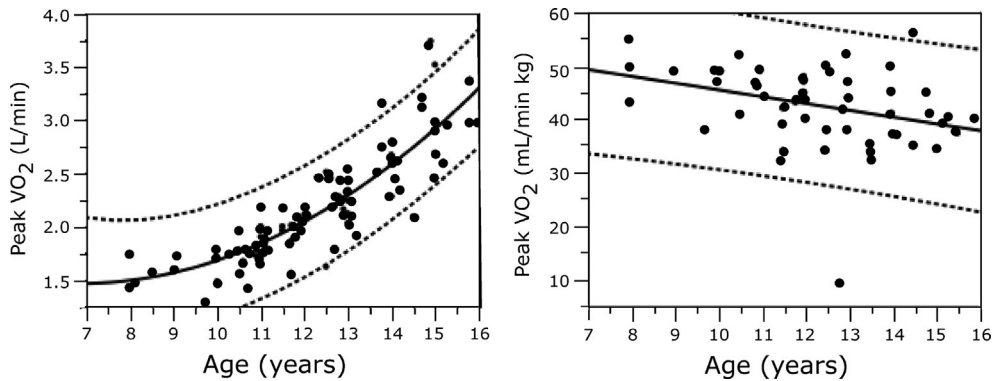


Fig. 27.41 Left panel: Relationship of peak oxygen consumption to age. Right panel: Relationship of peak oxygen consumption per unit mass to age. Data for girls were similar but less markedly curved. (Data taken from Figures in publication, but only for boys).

Many studies of scaling factors have been done, and most of them have found that linear scaling is almost never satisfactory. On the contrary, allometric scaling (using mass or height or body surface area raised to some fractional power) has provided greater constancy for predicting normal values (Chantler et al., 2005; Dewey et al., 2008, 2009; Neilan et al., 2009). The base used (weight, height, etc.) varies with the function being studied.

REFERENCES

- Anscombe, F.J., 1973. Graphs in statistical analysis. *Amer Stat* 27, 17–21.
- Armstrong, N., Welsman, J.R., 1994. Assessment and interpretation of aerobic fitness in children and adolescents. *Exerc. Sport Sci. Rev.* 22, 435–476.
- Bland, J.M., Altman, D.G., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 8, 307–310.
- Bland, J.M., Altman, D.G., 1999. Measuring agreement in method comparison studies. *Stat. Methods Med. Res.* 8, 135–160.
- Bland, J.M., Altman, D.G., 2003. Applying the right statistics: analyses of measurement studies. *Ultrasound Obstet. Gynecol.* 22, 85–93.
- Box, G.E.P., Hunter, W.G., Hunter, J.S., 1978. Statistics for experimenters. An introduction to design. In: *Data Analysis, and Model Building*. John Wiley & Sons, New York.
- Bryant, R.C., 1980. Masked variability in regression analysis. *Pediatr. Res.* 14, 352.
- Buckberg, G.D., Luck, J.C., Payne, D.B., Hoffman, J.I., Archie, J.P., Fixler, D.E., 1971. Some sources of error in measuring regional blood flow with radioactive microspheres. *J. Appl. Physiol.* 31, 598–604.
- Butter, C., Auricchio, A., Stellbrink, C., Fleck, E., Ding, J., Yu, Y., Huvelle, E., Spinelli, J., 2001. Effect of resynchronization therapy stimulation site on the systolic function of heart failure patients. *Circulation* 104, 3026–3029.
- Chantler, P.D., Clements, R.E., Sharp, L., George, K.P., Tan, L.B., Goldspink, D.F., 2005. The influence of body size on measurements of overall cardiac function. *Am. J. Physiol. Heart Circ. Physiol.* 289, H2059–H2065.
- Dallaire, F., Bigras, J.L., Prsa, M., Dahdah, N., 2015. Bias related to body mass index in pediatric echocardiographic Z scores. *Pediatr. Cardiol.* 35, 667–676.

- Delanaye, P., Mariat, C., Cavalier, E., Krzesinski, J.M., 2009. Errors induced by indexing glomerular filtration rate for body surface area: Reductio ad absurdum. *Nephrol. Dial. Transplant.* 24, 3593–3596.
- Dewey, F.E., Rosenthal, D., Murphy Jr., D.J., Froelicher, V.F., Ashley, E.A., 2008. Does size matter? Clinical applications of scaling cardiac size and function for body size. *Circulation* 117, 2279–2287.
- Glantz, S.A., Slinker, B.K., 2001. *Primer of Applied Regression and Analysis of Variance*. McGraw-Hill, Inc, New York.
- Harker, L.A., Slichter, S.J., 1972. The bleeding time as a screening test for evaluation of platelet function. *New Engl. J. Med.* 287, 155–159.
- Hoffman, J.I.E., 2018. The ratio fallacy, with special reference to the cardiac index. *Pediatr. Cardiol.* 39, 805–809.
- Jamal, F., Strotmann, J., Weidemann, F., Kukulski, T., D'hooge, J., Bijnens, B., Van De Werf, F., De Scheerder, I., Sutherland, G.R., 2001. Noninvasive quantification of the contractile reserve of stunned myocardium by ultrasonic strain rate and strain. *Circulation* 104, 1059–1065.
- Lolis, D., Konstantinidis, K., Papevangelou, G., Kaskarelis, D., 1977. Comparative study of amniotic fluid and maternal blood serum human placental lactogen in normal and prolonged pregnancies. *Am. J. Obstet. Gynecol.* 128, 724–726.
- Ludbrook, J., 1997. Comparing methods of measurement. *Clin. Exp. Pharmacol. Physiol.* 24, 193–203.
- Ludbrook, J., 2010. Confidence in Altman-bland plots: a critical review of the method of differences. *Clin. Exp. Pharmacol. Physiol.* 37, 143–149.
- Mackler, B., Person, R., Miller, L.R., Finch, C.A., 1979. Iron deficiency in the rat: effects on phenylalanine metabolism. *Pediatr. Res.* 13, 1010–1011.
- Montgomery, D.C., Peck, E.A., 1982. *Introduction to Linear Regression Analysis*. John Wiley and Sons, New York.
- Mosteller, F., Tukey, J.W., 1977. *Data Analysis and Regression. A Second Course in Statistics*. Addison-Wesley, Reading, CA.
- Nagy, A.I., Venkateshvaran, A., Dash, P.K., Barooah, B., Merkely, B., Winter, R., Manouras, A., 2014. The pulmonary capillary wedge pressure accurately reflects both normal and elevated left atrial pressure. *Am. Heart J.* 167, 876–883.
- Nakauchi, H., Okumura, K., Tango, T., 1981. Immunoglobulin levels in patients on long-term hemodialysis. *N. Engl. J. Med.* 305, 172–173.
- Neilan, T.G., Pradhan, A.D., King, M.E., Weyman, A.E., 2009. Derivation of a size-independent variable for scaling of cardiac dimensions in a normal paediatric population. *Eur. J. Echocardiogr.* 10, 50–55.
- Olivié, M.A.A., Garcia-Mayor, R.V., Leston, D.G., Sousa, T.R., Dominguez, A.S., Alvarez-Nono, R., Cortizas, J.A., 1995. Serum insulin-like growth factor (IGF) binding protein-3 and IGF-I levels during childhood and adolescence. A cross-sectional study. *Pediatr. Res.* 38, 149–155.
- Richardson, D., Catron, D.V., Underkofler, L.A., Maddock, H.M., Friedland, W.C., 1951. Vitamin B₁₂ requirement of male weanling pigs. *J. Nutr.* 44, 371–381.
- Senzaki, H., Isoda, T., Paolucci, N., Ekelund, U., Hare, J.M., Kass, D.A., 2000. Improved mechanoenergetics and cardiac rest and reserve function of in vivo failing heart by calcium sensitizer Emd-57033. *Circulation* 101, 1040–1048.
- Tanner, J.M., 1949a. The construction of normal standards for cardiac output in man. *J. Clin. Invest.* 28, 567–582.
- Tanner, J.M., 1949b. Fallacy of per-weight and per-surface area standards, and their relation to spurious correlation. *J. Appl. Physiol.* 2, 1–15.
- Tukey, J.W., 1977. *Exploratory Data Analysis*. Addison-Wesley Publishing Co, Menlo Park, CA.
- Vickers, A.J., 2001. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. *BMC Med. Res. Methodol.* 1, 6.
- Vickers, A.J., Altman, D.G., 2001. Statistics notes: analysing controlled trials with baseline and follow up measurements. *BMJ* 323, 1123–1124.

CHAPTER 28

Variations Based on Linear Regression

TRANSFORMING THE Y VARIATE

We may need transformations to stabilize the variance of the Y variate. These transformations have the added advantage that they often restore normality to the distribution of the Y variate. Common transformations are as follows:

1. If the Y data are counts and seem to be from a Poisson distribution, \sqrt{Y} or $\sqrt{Y} + \sqrt{Y+1}$ usually works.
2. For a continuous distribution, the one most often encountered, because of the likelihood that the coefficient of variation (standard deviation/mean) is constant, then the variance of the residuals is $\text{Var } e_i = k^2 X_i^2$ where k is the constant of proportionality. Divide both sides of the regression equation ($Y_i = c + bX_i + e_i$) by X_i to get

$$\frac{Y_i}{X_i} = \frac{c}{X_i} + b + \frac{e_i}{X_i}.$$

Define the new transformed variables as $Y'_i = \frac{Y_i}{X_i}$, $X'_i = \frac{1}{X_i}$, $e'_i = \frac{e_i}{X_i}$ and $c' = \frac{c}{X_i}$ so that the new regression equation in transformed variables is

$$Y'_i = c' + bX'_i + e'_i.$$

This new equation will have a constant variance, to be verified by examining residuals (see below).

INVERSE PREDICTION

Regression logic places the independent X value on the horizontal axis and the dependent Y value on the vertical axis, and the least squares principle considers only the vertical deviations of each Y from the regression line. The standard deviation from regression is used to predict the variation of Y at any X value. Sometimes, however, we need to do the inverse.

Hirschfeld et al. (1975) studied how to predict pulmonary vascular resistance from echocardiographically obtained time intervals. Their data are summarized in Fig. 28.1, based on a letter to the editor by Silverman and Hoffman (1976).

The upper panel shows the results obtained when plotting pulmonary vascular resistance (PVR) obtained at cardiac catheterization against a simultaneously obtained ratio of right ventricular preejection period (RPEP) to ejection time (RVET). The line of best fit and the 95% confidence limits for individual data are shown, but the individual

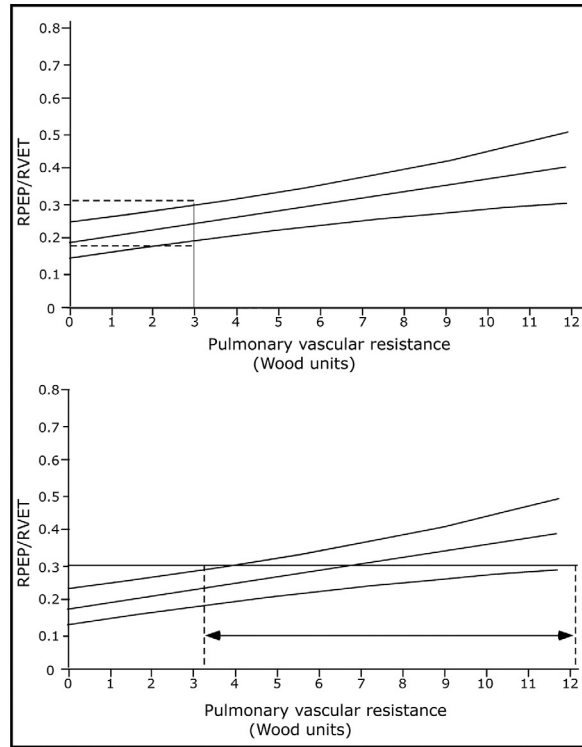


Fig. 28.1 Illustration of inverse prediction (see text).

points are not. To predict the ratio of right ventricular preejection period to ejection time from PVR, draw a vertical line from any value of PVR and predict the range in which 95% of these time ratios would fall by examining the projection of the dashed horizontal lines on the Y-axis. (We could calculate these limits from the basic equations.) That, however, is not what we want. Instead we want to predict the PVR from the noninvasive echocardiographic method. To do this from the same regression figure shown in the bottom panel, the logic requires drawing a horizontal line from a given ratio and determining the range within which 95% of the PVR measurements would fall from the projection of the dashed vertical lines on the X-axis. As seen from the figure, this range is so wide as to be of little clinical use. Furthermore, although we could determine the range by eye, the way in which regression calculations are carried out does not permit us to determine the horizontal deviations of the Y values from the regression line. Instead, a special inverse prediction calculation is needed (Zar, 2010).

The predicted value of $X(\hat{X})$ for any $Y(Y_i)$ is $\hat{X} = \frac{Y_i - c}{b}$. This is just altering the basic regression formula.

The confidence limits of X for a given value of Y may be determined by:

$$\bar{X} + \frac{b(Y_i - \bar{Y})}{K} \pm \frac{1}{K} \sqrt{\left(s_{Y.X}^2 \left[\frac{(Y_i - \bar{Y})^2}{\sum (X_i - \bar{X})^2} + K \left(1 + \frac{1}{N} \right) \right] \right)}$$

where

$$K = b^2 - t^2 s_b^2.$$

Other formulas have also been proposed (see <https://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/calibrat.htm>). Neter et al. (1996) recommended calculating the new $X(\hat{X})$ as $\hat{X} = \frac{Y_{\text{new}} - c}{b}$, and then calculating the variance of the predicted X as $s^2 = \frac{\text{MSE}}{b^2} \left[1 + \frac{1}{N} + \frac{(\hat{X} - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]$.

In the lower panel of Fig. 28.1, even with correct calculations, the limits of X for a given value of Y are very wide. This per se is not a difficulty due to inverse prediction, but to the lack of precision in the data. Gaddum, working in the field of bioassays, developed an index of precision termed $\lambda = s/b$. To see its effect, consider Fig. 28.2.

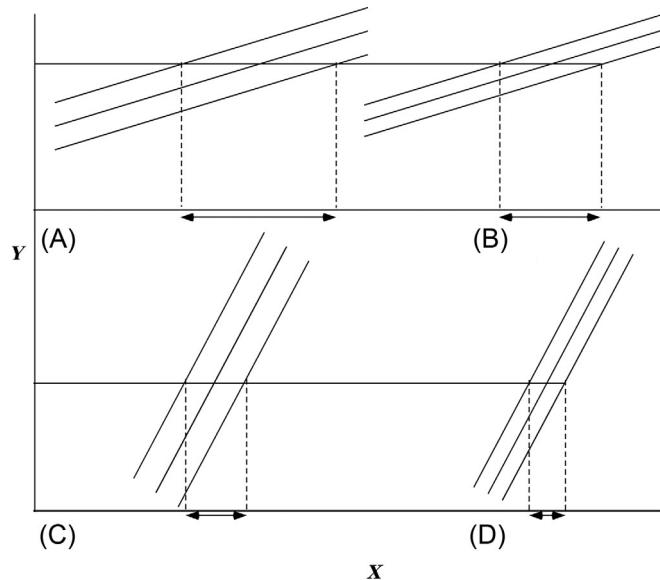


Fig. 28.2 Illustration of Gaddum's lambda index. In (A) the slope is not steep and the confidence limits (drawn for simplicity as *parallel lines*) are quite wide. The confidence limits of X for the given value of Y (shown by the paired vertical *dashed lines*) are wide, as shown by the *double-headed arrow*. These confidence limits for X are narrower if the confidence limits are narrower (B), or the slope is steeper (C), and are smallest for the steepest slope and narrowest confidence limits (D).

LINE OF BEST FIT PASSES THROUGH ZERO

It is common in biology and medicine to minimize variation by normalizing a measurement to body weight, body surface area, and so on. If we measure in children of different ages absolute cardiac outputs/l/min of 1, 3, and 10, and then base these outputs on a body surface area of 1 m^{-2} , the outputs might be 2.9, 3.2, and 4.7 L/min m^{-2} , and are less variable than the original data. Similarly, it may be possible to use a simple surrogate measurement instead of a complicated one provided one can be converted to the other. For example, it is relatively difficult to measure left ventricular volume, but it might be possible to find a power function of a simple linear measurement. For all these normalizing or calibration procedures we need to know if the relationship is linear and passes through zero, if it is linear but does not pass through zero, or if it curvilinear. These choices are represented in Fig. 28.3.

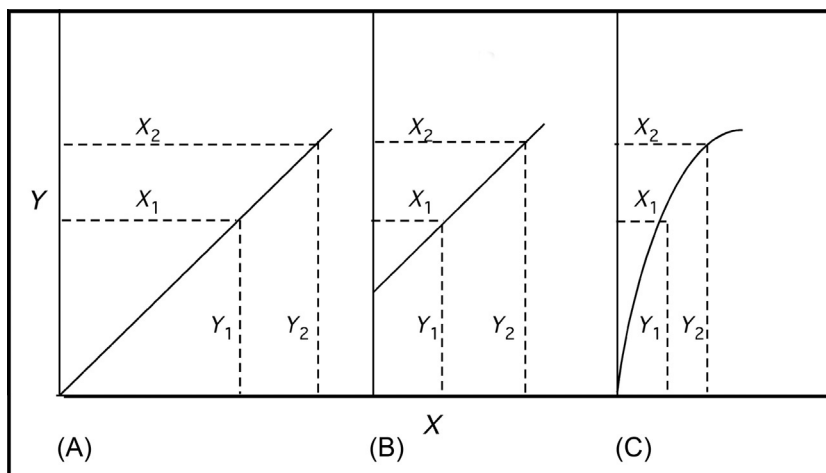


Fig. 28.3 Types of calibration or normalizing curves. (A) *Straight line through origin.* (B) *Straight line not through origin.* (C) *Curved line through origin.*

In panel A the ratio of Y_i to X_i is constant at any value of X and Y . The constant ratio can be used as a multiplier or divisor, for example, cardiac output is $\sim 3.5 \text{ L/min m}^{-2}$ body surface area. For the other panels there is no constant ratio, and the line or curve must be used to determine the X value for any Y value.

For a straight line through the origin, there are three possible scenarios; the variance from regression of Y , $s_{Y \cdot X}^2$, is constant, is proportional to X , or is proportional to X^2 . These lead to simplified calculations (Table 28.1).

Table 28.1 Summary of revised formulas

	b	c	SS_{reg}	$s_{y,x}^2$	s_y^2
Model 1	$\frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$	$\bar{Y} - b\bar{X}$	$\sum (Y_i - \bar{Y})^2 - \frac{[\sum (X_i - \bar{X})(Y_i - \bar{Y})]^2}{\sum (X_i - \bar{X})^2}$	$\frac{SS_{\text{reg}}}{N-2}$	$\frac{s_{y,x}^2}{\sum (X_i - \bar{X})^2}$
Model 2 $s_{y,x}^2 = k$	$\frac{\sum X_i Y_i}{\sum X_i^2}$	0	$\sum (Y_i)^2 - \frac{(\sum X_i Y_i)^2}{\sum X_i^2}$	$\frac{SS_{\text{reg}}}{N-1}$	$\frac{s_{y,x}^2}{\sum (X_i)^2}$
Model 2 $s_{y,x}^2 \propto X^2$	$\frac{\sum Y}{\sum X} = \frac{\bar{Y}}{\bar{X}}$	0	$\sum \left(\frac{Y_i^2}{X_i}\right) - \frac{(\sum Y_i)^2}{\sum X_i}$	$\frac{SS_{\text{reg}}}{N-1}$	$\frac{s_{y,x}^2}{\sum (X_i)^2}$
Model 2 $s_{y,x}^2 \propto X^2$	$\frac{\sum \left(\frac{Y}{X}\right)}{N} = \frac{\sum R}{N}$	0	$\sum (R_i - \bar{R})^2$	$\frac{SS_{\text{reg}}}{N-1}$	$\frac{s_{y,x}^2}{N}$

Model 1 Normal linear regression. Model 2 Regression with zero intercept.

Some care is needed in applying the ratio formula of the bottom row. A set of data might provide a regression plot with an intercept that is not zero, but with confidence limits that include zero. Unless the intercept is very small it is better not to use the ratio formula because then the ratio will not be truly constant. The method is best applied when theoretically the line must go through the origin, as in a chemical reaction in which absence of the input chemicals means absence of the output reaction products.

ERRORS IN THE X VARIATE

One of the requirements for regression is that the X variate be measured without error, but that is impossible. Both X and Y must have measurement errors (ϵ), although it is Y that is our main interest. In reality,

Measured $Y = Y^* = \text{true } Y + \epsilon_Y$ and measured $X = X^* = \text{true } X + \epsilon_X$.

The errors in Y and X are each assumed to be normally distributed around zero and to be independent of each other (Strike, 1981; Altman and Bland, 1983).

With measurement errors we actually calculate

$$b = \frac{\sum (X_i^* - \bar{X})(Y_i^* - \bar{Y})}{\sum (X_i^* - \bar{X})^2} = \frac{\text{Covariance } X^* Y^*}{\text{Variance } X^*}.$$

Now because ϵ_Y and ϵ_X are independent

$$\text{Covariance } X_i^* Y_i = \text{Covariance } X_i Y_i$$

and

$$\text{Variance } X_i^* = \text{Variance } X_i + \text{Variance } \varepsilon_{X_i}$$

$$b = \frac{\text{Covariance } X_i Y_i}{\text{Variance } X_i + \text{Variance } \varepsilon_X}.$$

Unless X is measured with very little error, we will not calculate the correct slope of the relationship between true X and true Y . This error may be of little importance when for example relating height to age, because the error in measuring age can be very small, if necessary not more than 1 day in 365, or 0.27%. If the likely error in X is large, it can materially affect the estimates of slope and correlation.

A procedure to deal with this problem was published in 1949 by Bartlett. (A similar approach had been recommended earlier by Nair and colleagues.) Bartlett divided the data set into three equal sized groups; if N was not divisible by 3, then the first and third groups were to be of equal size. Then the slope was calculated as

$$b = \frac{\bar{Y}_3 - \bar{Y}_1}{\bar{X}_3 - \bar{X}_1},$$

where 1 and 3 represent the lowest and highest thirds of X , respectively. The regression equation is as usual: $\hat{Y}_i = bX_i + (\bar{Y} - b\bar{X})$. If X has minimal error, the slopes calculated by this method and the classical method are the same, but Bartlett's method gives a better estimate of slope if X has large measuring errors.

More complex ways of handling errors of the X variate are discussed by [Ludbrook \(2012\)](#) but they require specialized programs.

BREAK POINTS

In many chemical or physiological processes there is a linear relationship between X and Y until a threshold is reached and the slope of the line suddenly changes. Example of this is the rise in blood lactate when oxygen supply cannot meet oxygen demand ([van der Hoeven et al., 1997](#)) or the anaerobic ventilatory threshold is reached ([Orr et al., 1982](#)). Methods for deciding where the breakpoint is, also known as piecewise linear regression, have been described [Hudson \(1966\)](#), [Mellits \(1968\)](#), [Jones and Molitoris \(1984\)](#) the easiest description to follow being that of [Jones and Molitoris \(1984\)](#) (Fig. 28.4).

They describe two straight lines:

$\hat{Y}_i = c_1 + \beta_1 X_i$ when $X \leq X_c$ and $\hat{Y}_i = c_2 + \beta_2 X_i$ when $X > X_c$. X_c is the critical point at which the two lines meet. Because the lines cross at X_c ,

$$c_2 = c_1 + \beta_1 X_c - \beta_2 X_c$$

This leads to two new equations:

$$\hat{Y}_i = c_1 + \beta_1 X_i \text{ when } X < X_c,$$

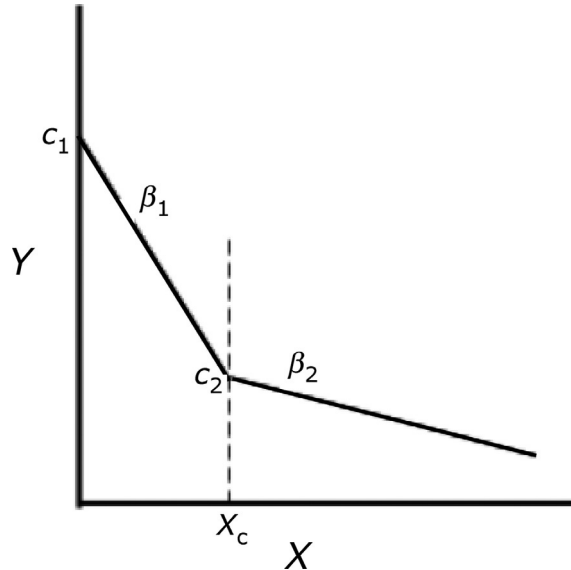


Fig. 28.4 Diagram of break point analysis.

$$\hat{Y}_i = c_1 + \beta_1 X_c + \beta_2 (X_i - X_c) \text{ when } X \geq X_c.$$

Programs search over different values of X_c so that the residual sum of squares is minimized. (If the break point is known, the two lines can be fitted by multiple regression methods.) There is a free program called SegReg (<https://www.waterlog.info/segreg.htm>) that will calculate the lines, but it works only with Windows operating systems.

Jones and Molitoris (1984) also gave a method for determining the approximate confidence intervals for the break point, and Graybill and Iyer (1994) described how to determine confidence limits for the slopes and intercepts.

A more refined but complicated method is to use joinpoints—a free program can be downloaded from <https://surveillance.cancer.gov/joinpoint/download>. (For Windows only.)

RESISTANT LINES

There are ways of determining if one or more points have undue leverage or influence, but how to deal with such aberrant points is not settled. One approach is to create a resistant line using median values (Velleman and Hoaglin, 1981; Emerson and Hoaglin, 1983; Selvin, 1995).

The basis of the method resembles Bartlett's test (see above) but rather than means uses medians that are less affected by outliers. The X values are divided into three portions

with equal numbers of observations ($N/3$) in each: if after dividing by 3 there is one extra number the middle group has one extra measurement, and if there are two extra numbers they go into the outer groups. Then for each group determine the median of X and the median of Y . As an example, the data of Butter et al. from Fig. 27.6 are presented in Fig. 28.5.

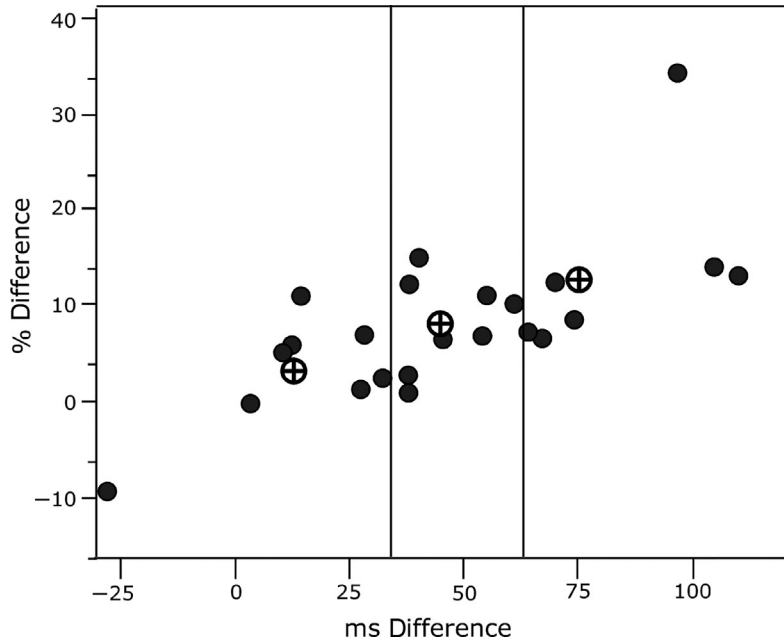


Fig. 28.5 Data from Butter et al. (2001) showing X values divided into thirds by the vertical lines, and the median values of X and Y in each third as open circles with an internal cross.

The slope b_1 can be calculated from

$$b = \frac{Y_R - Y_L}{X_R - X_L}$$

where X and Y are the medians for the right (R) and left (L) groups. For the data shown in Fig. 28.5, the slope is $b = \frac{12.10 - 3.55}{78.45 - 16.95} = 0.1390$. This is compared to the slope of 0.1688 in the standard regression; the outlier in the upper right part of the graph was probably responsible for inflating the slope estimate.

An average intercept c is calculated from each of the three groups:

$$c = \frac{(Y_L - bX_L) + (Y_M - bX_M) + (Y_R - bX_R)}{3}.$$

For this data set, c is calculated as

$$\begin{aligned} c &= \frac{(3.55 - 0.139 \times 16.95) + (8.50 - 0.139 \times 46.55) + (12.1 - 0.139 \times 78.45)}{3} \\ &= \frac{1.19395 + 2.02955 + 1.19545}{3} = 1.4730 \end{aligned}$$

c was -0.4765 by the standard regression, again possibly because of the one outlier.

The new equation, $\hat{Y} = c + X_i$ is resistant to outliers and is often a better estimate of the relationship between X and Y in the population, but it is necessary to examine residuals to determine if they are normally distributed about zero. In standard least squares regression, the sum of the deviations from the regression line is always zero, and a plot of residuals against X has zero slope. This is not necessarily true for the resistant line method, and if the plot of residuals against X has a slope, then the slope of the original regression line must be corrected.

Velleman and Hoaglin (1981) describe an iterative procedure for minimizing the residuals and obtaining the correct slope. The residuals, plotted against X , are divided into the same three groups and the medians taken as before. If the slope is now essentially zero, nothing further is needed. If the slope of the residuals (b') is not zero, it is added to the previous slope, new residuals are calculated, and the process continues until a zero residual slope is obtained. Velleman and Hoaglin (1981) give a way of shortening the iterations by a weighting method.

The final equation becomes

$$\hat{Y} = 0.1111X_i - 0.2887.$$

APPENDIX

Derivation of Formulas for a Straight Line Passing Through Zero

The line of best fit passes through the point $0,0$ rather than \bar{X}, \bar{Y} . Because the slope is the tangent of the line, the slope is now $\frac{Y_i - 0}{X_i - 0}$ instead of $\frac{Y_i - \bar{Y}}{X_i - \bar{X}}$. Therefore, whenever there is an expression such as $X_i - \bar{X}$ or $Y_i - \bar{Y}$ in the formula, it is replaced by X_i or Y_i , respectively. The divisor of $N - 1$ for estimating the variance from regression is due to the fact that $0,0$ is one fixed point, whereas the point \bar{X}, \bar{Y} uses two degrees of freedom.

For model 2 with variance proportional to X or X^2 , we weight each expression by $1/X$ or $1/X^2$.

REFERENCES

- Altman, D.G., Bland, J.M., 1983. Measurement in medicine: the analysis of method comparison studies. *Underst. Stat.* 32, 307–317.
- Butter, C., Auricchio, A., Stellbrink, C., Fleck, E., Ding, J., Yu, Y., Huvelle, E., Spinelli, J., 2001. Effect of resynchronization therapy stimulation site on the systolic function of heart failure patients. *Circulation* 104, 3026–3029.
- Emerson, J.D., Hoaglin, D.C., 1983. Resistant lines for y versus x. In: Hoaglin, D.C., Mosteller, F., Tukey, J.W. (Eds.), *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons, New York.
- Graybill, F.A., Iyer, H.K., 1994. *Regression Analysis: Concepts and Applications*. Duxbury Press, Belmont, CA.
- Hirschfeld, S., Meyer, R., Schwartz, D.C., Kofhagen, J., Kaplan, S., 1975. The echocardiographic assessment of pulmonary artery pressure and pulmonary vascular resistance. *Circulation* 52, 642–650.
- Hudson, D.J., 1966. Fitting segmented curves whose joint points have to be estimated. *J. Am. Stat. Assoc.* 61, 1097–1129.
- Jones, R.H., Molitoris, B.A., 1984. A statistical method for determining the breakpoint of two lines. *Anal. Biochem.* 141, 287–290.
- Ludbrook, J., 2012. A primer for biomedical scientists on how to execute model II linear regression analysis. *Clin. Exp. Pharmacol. Physiol.* 39, 329–335.
- Mellits, E.D., 1968. *Statistical Methods*. In: Growth, H. (Ed.), Cheek, D. Lea & Febiger, Philadelphia.
- Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W., 1996. *Applied Linear Statistical Models*. Irwin, Chicago.
- Orr, G.W., Green, H.J., Hughson, R.L., Bennett, G.W., 1982. A computer linear regression model to determine ventilatory anaerobic threshold. *J. Appl. Physiol. Respir. Environ. Exerc. Physiol.* 52, 1349–1352.
- Selvin, S., 1995. *Practical Biostatistical Methods*. Wadsworth Publishing Company, Belmont, CA.
- Silverman, N.H., Hoffman, J.I., 1976. Letter: echo assessment of PVR. *Circulation* 54, 525–526.
- Strike, P.W., 1981. *Medical Laboratory Statistics*. John Wright and Sons, Ltd., Bristol.
- Van Der Hoeven, M.A., Maertzdorf, W.J., Blanco, C.E., 1997. Mixed venous oxygen saturation and biochemical parameters of hypoxia during progressive hypoxemia in 10- to 14-day-old piglets. *Pediatr. Res.* 42, 878–884.
- Velleman, P.F., Hoaglin, D.C., 1981. *Applications, Basics and Computing of Exploratory Data Analysis*. Duxbury Press, Boston, MA.
- Zar, J.H., 2010. *Biostatistical Analysis*. Prentice Hall, Upper Saddle River, NJ.

CHAPTER 29

Correlation

BASIC CONCEPTS

Introduction

Correlation is closely related to regression and asks how closely X and Y are related rather than how much does Y change when X changes. Correlation is calculated so that if all the points lie on the line the value is 1, if X and Y are completely unrelated then the value is 0, and if they are partially related the value is somewhere between 0 and 1. If the slope of the line is negative, so that Y gets smaller as X gets bigger, then the correlation varies between 0 and -1 . The sample correlation coefficient (also known as Pearson's product-moment correlation coefficient) has two symbols: r if only two variables are related, and R if more than two variables are involved; the population coefficient is ρ .

The square of the correlation coefficient r^2 , termed the coefficient of determination, is defined as $\frac{SS_{\text{reg}}}{SS_{\text{total}}}$, where SS_{reg} is the sum of squares due to the regression relationship and SS_{total} is the sum of squared deviations from the mean of the Y array. It indicates the proportion of the total variability of Y that can be accounted for by the regression relationship. SS_{reg} is $\frac{\sum[(X_i - \bar{X})(Y_i - \bar{Y})]^2}{\sum(X_i - \bar{X})^2}$ and SS_{total} is $\sum(Y_i - \bar{Y})^2$

$$\text{so that } r^2 = \frac{SS_{\text{reg}}}{SS_{\text{total}}} = \frac{\sum[(X_i - \bar{X})(Y_i - \bar{Y})]^2}{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}$$
$$\text{and } r = \frac{\sum[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

If all the points lie on the line then the total variability of Y is explained by variability of X , so that $SS_{\text{reg}} = SS_{\text{total}}$, and $r = 1$. If the points are completely unassociated, then SS_{reg} is zero and r is zero.

The formulas show that both regression and correlation calculations make use of the same derived values (means, deviations from means) but in different ways.

These considerations can be developed further by a figure (Fig. 29.1).

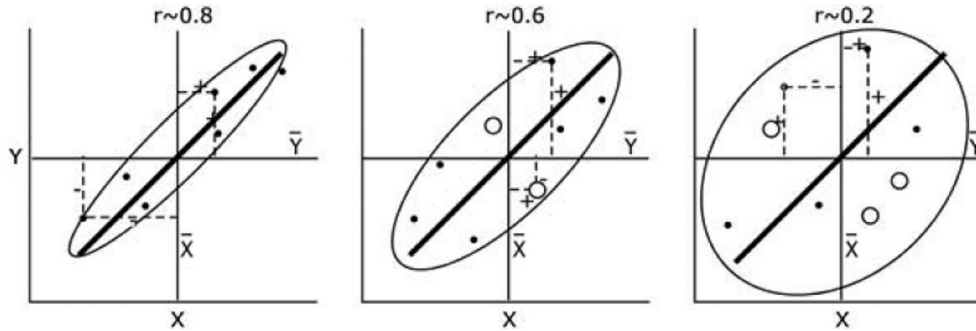


Fig. 29.1 Examples of 3 different correlation coefficients. The line of best fit is the *thick solid line* that passes through the point where the mean of X and the mean of Y cross. The individual XY points are shown in *solid black circles* for the upper right and lower left quadrants and in *open circles* for the other quadrants. The *dashed vertical lines* show the deviations from the mean of Y , and the *dashed horizontal lines* show the deviations from the mean of X . *Plus* and *minus* signs next to the *dashed lines* indicate positive or negative deviations, respectively. r is the correlation coefficient. The ellipses are drawn by eye to encompass all the data points.

The line of best fit has the same slope for all three panels, but the fit of points about the line is close for the left panel, moderate for the middle panel, and poor for the right panel. As shown by the dashed deviation lines, in the left panel all the points are in the upper right or lower left quadrants, so that the deviations from the means are both positive or both negative and their product is positive. In the middle panel a few points have one positive and one negative deviation, the product of which is negative, so that when all the product deviations from the mean are summed their total is less than the total in the left panel. Finally, in the right panel there are about as many negative as positive product deviations from the mean, and their sum is close to zero.

The ellipses each have a long and a short axis. With a high r the long axis: short axis ratio is high (about 5 here), with a moderate r the ratio is lower (about 2.2) and with a low r almost 1. The change in ratios in these data sets is due to a change in the short axis because the long axis is constant in this example.

Correlation is independent of the units used.

Online programs for the correlation coefficient are often included in those for linear regression (Chapter 27) but may be calculated separately at <https://www.easycalculation.com/statistics/correlation.php>, <https://www.easycalculation.com/statistics/r-squared.php>, https://www.statstodo.com/CorReg_Pgm.php, and <http://www.alcula.com/calculators/statistics/correlation-coefficient/>.

Problem 29.1 Calculate the correlation coefficient for the data in Problem 27.1.

P Values and Confidence Limits

The sample correlation coefficient r is an estimate of the population correlation coefficient ρ (rho). If the variables X and Y are normally distributed and ρ is zero, then for $N > 6$, the quantity $\frac{r}{\sqrt{\frac{1-r^2}{N-2}}}$ is distributed approximately normally like t with $N-2$

degrees of freedom. For example, if $r = 0.8288$ and $N = 51$,

$$t = \frac{0.8288}{\sqrt{\frac{1-0.8288^2}{51-2}}} = \frac{0.8288}{0.0799} = 10.37.$$
 Therefore $P < 0.0001$, allowing us to reject the null

hypothesis that $r = 0$ (the two variables are unrelated). The P value of the correlation coefficient is determined readily from tables or simple interactive online programs, for example, <http://vassarstats.net/>, (see Correlation and Regression) and <http://www.danielsoper.com/statcalc3/calc.aspx?id=44>, and programs such as <http://vassarstats.net/rho.html> and <https://www.easycalculation.com/statistics/regression-coefficient-interval.php> that also allow calculation of 95% or 99% confidence limits that for the previous example are 0.717–0.899 (95%) or 0.671–0.914 (99%).

To test r against any value for $\rho \neq 0$, the distribution is skewed but can be made approximately normal by a Z (zeta) transformation introduced by Fisher:

$$Z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right).$$

(This Z is not to be confused with the z transform for a normal distribution.) Z is normally distributed with a mean of $\frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right)$ and has a standard error of

$$\sigma = \sqrt{\frac{1}{N-3}}$$

for any value of $N > 20$.

To compare any r with $\rho \neq 0$, transform both r and ρ into their respective Z values. To test whether $r = 0.8288$ ($N = 51$) could have come from a population with $\rho = 0.6075$, the two conversions yield 1.1843 and 0.7049, respectively. Then

$$z = \frac{1.1843 - 0.7049}{0.1443} = 3.3971.$$

$P = 0.000341$, and the observed correlation coefficient of 0.8288 is unlikely to have come from a population with a correlation of 0.6075.

To compare two observed r values, adapt the previous equation by dividing the difference between the two zeta values by the sum of the two standard deviations, calculated

$$\text{by } \sqrt{\frac{1}{N_1-3} + \frac{1}{N_2-3}}.$$

To compare $r_1 = 0.8288$ ($N = 51$) with $r_2 = 0.6075$ ($N = 42$), calculate Z_1 and Z_2 as

$$Z_1 = 0.5 \ln \frac{1 + 0.8288}{1 - 0.8288} = 1.1843 \quad \text{and}$$

$$Z_2 = 0.5 \ln \frac{1 + 0.6075}{1 - 0.6075} = 0.7049.$$

The standard deviation of the difference is

$$s_{z1} - s_{z2} = \sqrt{\frac{1}{48} + \frac{1}{39}} = 0.2156.$$

Therefore $z = \frac{1.1843 - 0.7049}{0.2156} = 2.2236$, and so $P = 0.0131$. Although this difference suggests that we can reject the null hypothesis, it is less so than if $r = 0.8288$ is compared with a population value of the same amount (see previously). This comparison may be performed online at <http://vassarstats.net/index.html>, (see Correlation and Regression) <http://www.quantitativeskills.com/sisa/statistics/correl.php?r1=0.8288&r2=0.6075&r3=51&n=42&CI=95&proc=1>.

Fisher's Z transformation is the basis of many other tests of the correlation coefficient (Kleinbaum et al., 1988). For example, to combine several correlation coefficients from a number of small samples, first demonstrate their homogeneity. This can be done readily at <http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/MultiCorr.htm> or <http://vassarstats.net/> (see Correlation and Regression) or to determine if the correlation coefficient is different from some value for ρ that is not zero, or if two experimental correlation coefficients are substantially different from each other. Methods for determining confidence limits and other comparisons are presented by Zar (2010).

The value of r^2 indicates the percentage of change in Y that can be associated with a change in X . If r is 0.8, then 64% of the change in Y is associated with (and possibly caused by) a change in X . If r is 0.5, then only 25% of the change in Y has been explained, and if r is 0.2, then only 4% of the change in Y can be explained by a corresponding change in X . The importance of the correlation coefficient is easy to overestimate. Just as in the t -test the difference (the effect size) may be small, medium, or large, so can the correlation coefficient be small (<0.25), medium ($0.25-0.7$), and large (>0.7) (Cohen, 1988). This is a more important judgment than rejection of the null hypothesis, because any trivial difference may lead to rejection of the null hypothesis if sample size is large enough.

Finding a correlation coefficient that is incompatible with zero is equivalent to finding that the slope b is incompatible with zero. Either test can be used.

Sample Size and Power

From the formula for Z previously, $z = \frac{Z_i - Z_\rho}{\sigma_z}$, where z (lower case) is the normal z distribution, Z (upper case) is defined as before, and Z_ρ is a theoretical value from a population with correlation coefficient ρ :

$$Z_\rho = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right).$$

What is the power of detecting a correlation coefficient of 0.3 with a sample of 20 measurements? Begin with

$$Z_\rho = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right) = \frac{1}{2} \ln \left(\frac{1+0.3}{1-0.3} \right) = 0.3095.$$

$$\sigma_z = \sqrt{\frac{1}{20-3}} = 0.2425.$$

Then $z = \frac{0.3095}{0.2425} = 1.2762$ ($P=0.1009$). This is interpreted as stating that if the true value of the correlation coefficient is 0.0, then 0.1009 of the area under the curve will be as big as 0.3 or more. This is thus not good evidence that the correlation coefficient differs from zero, but what is the power of the test to show a difference of 0.3?

Consider the two normal curves shown in Fig. 29.2.

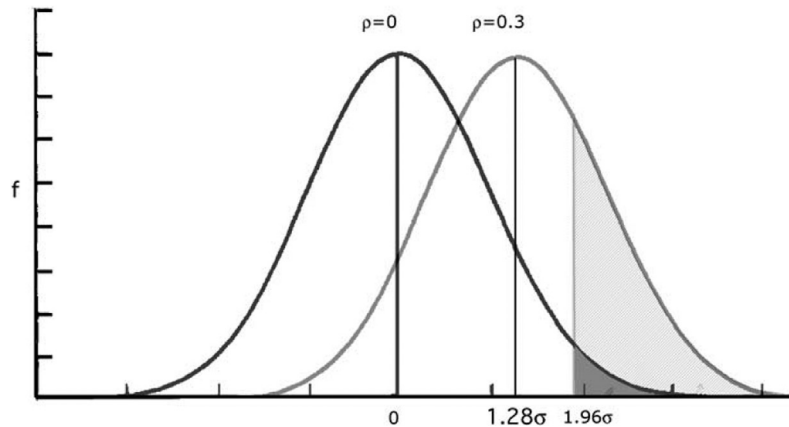


Fig. 29.2 Superimposition of two normal curves with the same standard deviation, but one with a mean of $\rho=0$ and the other with a mean of $r=0.3$.

This concept can be simplified to the equation: Power is the area of the standard normal distribution to the right of

$$z_{1-\beta_{\text{upper}}} = z_{\alpha/2} - \frac{Z_\rho}{\sqrt{\frac{1}{N-3}}}.$$

For the example given before, then $1.96 - \frac{0.3095}{0.2425} = 1.96 - 1.28 = 0.68$. The power is the area beyond this value of z .

To determine the needed sample size for any correlation coefficient, rearrange the previous equation to solve for N :

$$N = \left(\frac{z_{\alpha/2} - z_{1-\beta_{\text{upper}}}}{Z\rho} \right)^2 + 3$$

(Glantz, 2005).

To compare two observed correlation coefficients, the sample size equation becomes

$$N = 2 \left(\frac{Z_{\alpha} - Z_{\beta(1)}}{z_1 - z_2} \right)^2 + 3.$$

Here α and β are the Type I and Type II errors, respectively.

Sample size or power can be calculated online for a single value of r at http://www.statstodo.com/SSizCorr_Pgm.php and <http://www.sample-size.net/correlation-sample-size/>.

Cautionary Tales

There is nothing wrong with the concept of correlation, but often a great deal wrong with how it is used. Some statisticians disparage the use of correlation coefficients; Winsor and Tukey, for example, belonged to an informal society for the suppression of correlation coefficients!

1. The correlation coefficient is always higher than its square, and thus may confer an impression of greater importance than is justified.
2. In most studies, a low correlation coefficient does not provide much added information for the investigator to use. The one exception is in epidemiology where a disease that has many causes is being studied. Under these circumstances, no one X variable can be expected to have a very high correlation coefficient and finding a correlation coefficient of 0.1 might still indicate the need to consider that variable as one of the underlying causes of the disease.
3. Although the correlation coefficient can be calculated for linear or curvilinear regression, failure to consider linearity may result in incorrect calculations (Fig. 29.3).

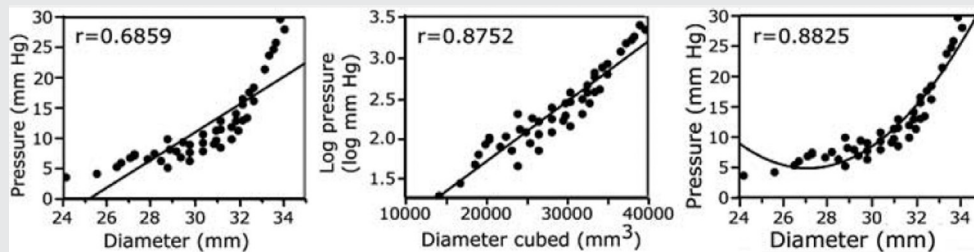


Fig. 29.3 Left panel: original data (Senzaki et al, 2000) fitted by *straight line*. Middle panel: transformed data fitted by a *straight line*. Right panel: original data fitted by *second-order polynomial*. Redrawn from Wolters Kluwer.

Fitting a straight line to a set of XY points with a curved relationship results in an incorrectly low correlation coefficient. Furthermore, although the correlation in the left panel of 0.6859 is quite high, it is no guarantee of linearity.

4. Although no specific criteria have been set for how the X values should be distributed, there are potential problems if the points are bunched at the low end and the high end, with none in the middle (Edwards, 1984). Fig. 29.4 shows some data simplified from an actual problem of growth of the right ventricle in fetal sheep.

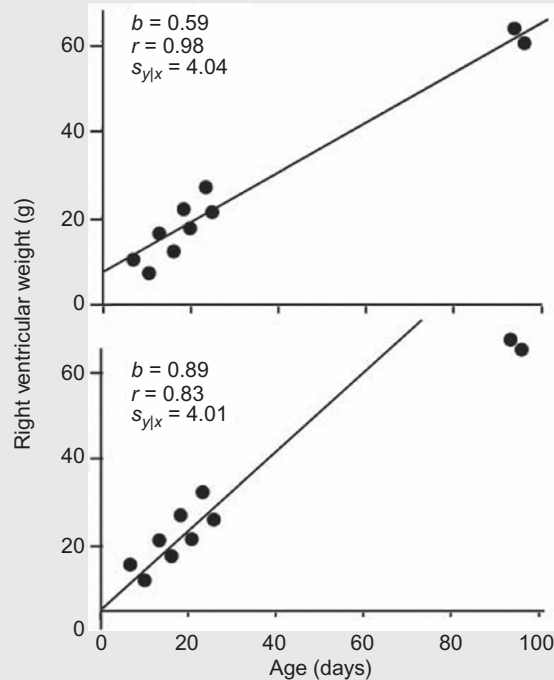


Fig. 29.4 Plot of index of age (X : days) vs right ventricular weight (Y :g).

The upper panel shows the effect of fitting a straight line to all the data. The bottom panel shows that if only the initial points are fitted, the slope is steeper and the correlation coefficient lower. If there are only two bunches of points, then it is always possible to join them by a straight line. It is possible that had there been measurements at intervening values of X the relation might have been shown to be curvilinear. An extreme example of this problem was shown in one study when the investigators plotted the relation between two inflammatory mediators and found a correlation of 0.77, even though there were 7 points near coordinates of 0, 0, and one point at coordinates 650, 850, with nothing between them.

Even with intermediate values of X , if there are a disproportionate number of measurements at one end of the curve, these will weight the line and the correlation coefficient (see Fig. 27.28).

Cautionary Tales—cont'd

5. One or a few outlying points can unduly influence not only the slope of the line but also the correlation coefficient that is not a resistant statistic. Consider the uncorrelated XY data of [Table 29.1](#)

Table 29.1 Artificial XY data set

X	1	2	3	4	5	6	7	8	9	10	11
Y	4	8	10	1	11	7	2	3	34	6	9

The correlation coefficient is 0.03. After changing the value of Y_1 from 4 to -40 or $+40$, the correlation coefficient becomes 0.46 or -0.52 , all because of the influence of a single point. This point was demonstrated nicely by [Curran-Everett \(2010\)](#).

6. There are other concerns with the correlation coefficient. Consider [Fig. 29.5](#).

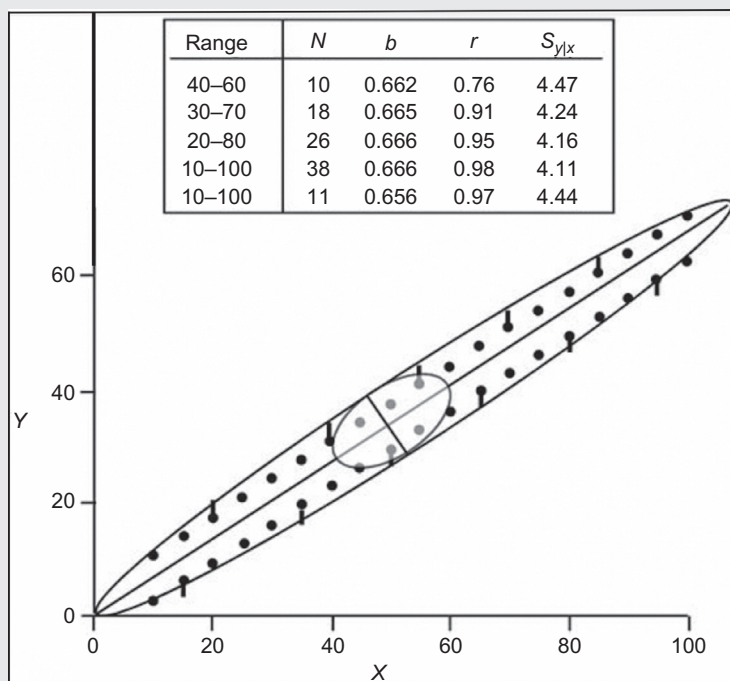


Fig. 29.5 Effect of changing the range of the X variate.

This is an artificially constructed figure in which each point is exactly 4 units above or below the line. If all the points from $X=10$ to $X=100$ are considered, the slope is 0.666 and the correlation coefficient is 0.98; the standard deviation from regression ($s_{y \cdot x}$) is 4.11, an estimate of the population value of 4. If the range over which X is examined is reduced to 20–80, 30–70, or 40–60, the slope and the standard deviation from regression change little, but the correlation coefficient decreases from 0.98 to 0.95, then 0.91, and then 0.76, respectively. Comparing the two ellipses shows the reason

for this. The ellipse that covers the full range of 10–100 has a high long: short axis ratio, but the ellipse that covers the range 40–60 has a much smaller ratio, even though its short axis (a function of the standard deviation from regression) has not changed. To show that the reduced number of points used in the calculation is not the cause for the change, the final calculation was done with only 11 points (shown by the short vertical lines attached to some of the points) and it produced almost the same results as were obtained for the calculation that used all 38 points.

This artifact is one of the reasons why the standard deviation from regression should always be given, even if the correlation coefficient is given as well.

The mathematical basis of this artifact was discussed by [van Belle \(2002, p. 57\)](#). It is possible to write the formula for the square of the correlation coefficient as

$$r^2 = \frac{1}{1 + \frac{(n-2)s_{y,x}^2}{b^2(X_i - \bar{X})^2}}.$$

The bigger the range of the X variate for a given sample size n the closer r is to 1. Conversely, if measurements are made over a restricted range, the correlation coefficient becomes smaller ([Curran-Everett, 2010](#)).

The effect of the range of X can be shown also by Chatillon's balloon trick ([Chatillon, 1984](#)) (Fig. 29.6).

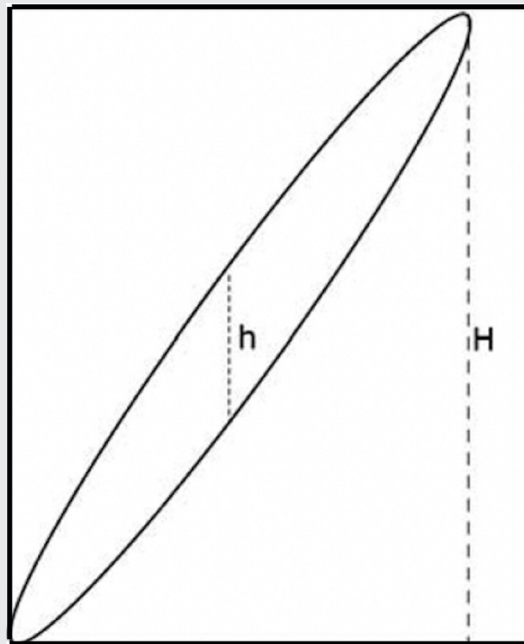


Fig. 29.6 Chatillon's balloon.

Cautionary Tales—cont'd

If an ellipse is drawn around the points (not shown), the length of a vertical through the center of the ellipse is h , and the total vertical height of the ellipse is H , then

$$r \approx \sqrt{1 - \left(\frac{h}{H}\right)^2}. \text{ If the range of } X \text{ is narrowed, then } H \text{ is reduced relative to } h,$$

and the correlation coefficient becomes smaller.

7. Because the correlation coefficient is the square root of the ratio of SS_{reg} to SS_{total} , if SS_{total} decreases with no change in SS_{reg} , the correlation coefficient decreases (Fig. 29.7).

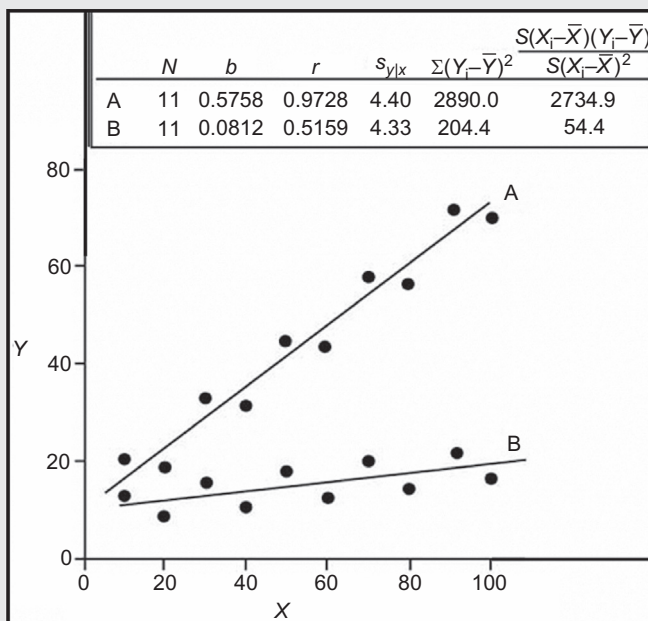


Fig. 29.7 Artificial example of effect of change in slope. The points are all exactly 4 units above or below the lines.

With the steeper slope of 0.5758 (line A) and standard deviation from regression of 4.40 the correlation coefficient is 0.9728. Without any major change in the standard deviation from regression (4.33) but with a slope of 0.0812, the correlation coefficient for line B is 0.5159. As shown in the second last column, this change in r has been due to a marked decrease in the sum of squared deviations from the mean of Y (here shown as $\sum(Y_i - \bar{Y})^2$).

This effect of rotating a set of correlated points has been shown formally. It can also be explained by Chatillon's balloon.

It is therefore imprudent to rely on a number for the correlation coefficient without considering the slope, its linearity, the range of X values, and the distribution of the X values. It is not wrong to mention the correlation coefficient in the publication, but its value is small compared to the vital information provided by the standard deviation from regression.

8. A little recognized requirement for the correlation coefficient is that the X values are normally distributed. We do not often pay attention to this requirement because regression and correlation are quite robust, but the requirement is violated if the X variables are determined in advance (Campbell and Machen, 1993) as, for example, when drug doses are allocated for a dose-response analysis. Even if the dose-response curve is constant, changing the actual doses used may change the correlation coefficient.

Even if the X variables are chosen at random, an abnormal distribution of the X variate can lead to a correlation coefficient that has a large standard error and wide confidence limits, and so makes it harder to compare two or more groups. One way around this problem is to use robust regression lines.

9. Measurement errors reduce the basic correlation between Y and X (Cohen and Cohen, 1975). If Y and X have random errors of ε_Y and ε_X , respectively, each error with a mean of zero, then the correlation coefficient becomes

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left(\sqrt{\sum (X_i - \bar{X})^2 + \sum \varepsilon_X^2}\right) \left(\sqrt{\sum (Y_i - \bar{Y})^2 + \sum \varepsilon_Y^2}\right)}}$$

r can only achieve its highest value when $\sum \varepsilon_X^2 = \sum \varepsilon_Y^2 = 0$. Because all measurements are made with some error, the “true” correlation coefficient is always less than that calculated; this is known as attenuation.

Ordinal Numbers

If the X or Y variables are ordinal numbers, classical regression cannot be done. For example, an investigator examines aortas at autopsy examination and grades the degree of atherosclerosis as 1+ to 6+ and wants to compare aortic atherosclerosis with a recent serum cholesterol concentration observed in those patients. Because the degree of atheroma (Y variable) is ordinal, it makes no sense to calculate a slope. Instead, calculate a modified correlation coefficient with Spearman’s or Kendall’s test. Both of these are ranking tests that give a correlation statistic that varies between +1 and −1. They can be used also if the numbers are ratio numbers but the bivariate population is far from normal; as in many distribution-free tests, they are relatively insensitive to outliers.

Spearman’s Test

For Spearman’s rank correlation procedure (Spearman’s ρ or r_s) the X and Y variables are each ranked from smallest to largest, the difference in ranks (d) is calculated as $d_i = \text{rank of } X_i$

minus rank of Y_i , and then squared to give d^2 . (The results will be the same if the data are ranked from largest to smallest.) Then Spearman's rank correlation coefficient (r_s or ρ) is

$$r_s = 1 - \frac{6 \sum d^2}{N(N^2 - 1)}$$

The null hypothesis for r_s can be assessed from online programs such as <http://www.wessa.net/rankcorr.wasp>, http://vassarstats.net/corr_rank.html, and http://www.statstodo.com/Spearman_Pgm.php. All allow entry of raw data or ranks.

If there are no tied ranks, the same correlation coefficient can be obtained from the classical formula for Pearson's correlation coefficient. If there are tied ranks, adjustments to the formula are needed, although the differences are minor unless there are very many ties (Zar, 2010).

If this test and classical linear regression are done on the same set of ratio numbers that fit the requirements for regression analysis, the rank correlation test is about 91% as powerful as the parametric test.

As an example, consider the ranking of ice cream flavors by two judges (Table 29.2).

Table 29.2 Ice cream ratings

Ice cream number	Judge A score	Judge A rank	Judge B score	Judge B rank	Rank difference (d)	d^2
1	5	3.5	7	6	-2.5	6.25
2	8	9	8	7	2	4
3	9	10	5	3	7	49
4	4	2	5	3	-1	1
5	6	5.5	9	8.5	-3	9
6	7	7.5	4	1	6.5	42.25
7	5	3.5	10	10	-6.5	42.25
8	7	7.5	5	3	4.5	20.25
9	3	1	6	5	-4	16
10	6	5.5	9	8.5	-3	9
						$\Sigma d^2 = 206$

For each judge separately, rank the scores from best (score = rank 10) to worst (rank 1). For judge A, ice cream 3 gets a score of 9 and a rank of 10, whereas ice creams 6 and 8, getting an equal score of 7, share ranks. Here $r_s = -0.25$, $P = 0.51$.

Kendall's Tau Test

Kendall's τ test is more tedious to calculate by hand, but can be done by interactive online programs http://www.wessa.net/rwasp_kendall.wasp#output, <http://calculator>.

vhex.net/calculator/statistics/kendall-tau-correlation, <http://www.real-statistics.com/correlation/kendalls-tau-correlation/>, or standard statistical programs. The X data are ranked in order from lowest to highest value. Then the corresponding ranked Y data (raw data, or ranked, usually the latter) are examined to determine if they are concordant c (arranged in the same order) or discordant d (arranged in the opposite order). For example, consider the set in Table 29.3.

Table 29.3 X and Y ranks

X	Y rank	Discordant	Concordant
1	3	2	7
2	2	1	7
3	5	2	5
4	1	0	6
5	10	5	0
6	8	3	1
7	4	0	3
8	7	1	1
9	6	0	1
10	9	0	0
		14	31

In the Y column, for each row count the number of ranks that are smaller below it (=discordant). For the first Y (Y_1) there are 0; for Y_2 there is 1, for Y_7 there is 1, and so on. These discordant ranks sum to 14. Now count the number of ranks below it that are higher. For Y_1 there are 6, for Y_2 there are 7, and so on and sum these. The sum of concordant ranks is 31. Then calculate $\tau = \frac{N_c - N_d}{N_c + N_d} = \frac{N_c - N_d}{\frac{1}{2}k(k-1)}$, where k = number of pairs.

In the previous example this gives $\tau = \frac{31 - 14}{31 + 14} = 0.3777$ $p = 0.1524$ (two-tailed).

There are different formulas with identical results. Corrections for multiple ties may be needed. The total sum of ranks can also be calculated from $\binom{k}{2} = \frac{k!}{2!(k-2)!}$. If $k = 10$, $\binom{k}{2} = \frac{10!}{2!8!} = \frac{3628800}{2 \times 40320} = 45$.

Both Spearman's ρ and Kendall's τ test if there is a correlation, but one cannot be converted into the other. For a given set of data Spearman's ρ tends to be bigger than Kendall's τ .

Problem 29.2. The table presents data on marks for English and Mathematics for each of 11 students.

Perform Spearman's r_s and Kendall's tau tests.

English	Mathematics
56	64
77	68
54	43
75	67
60	61
60	55
54	57
78	78
72	63
59	67
44	71

ADVANCED AND ALTERNATIVE CONCEPTS

Partial Correlation

O'Neill et al. (1983) studied patients with cystic fibrosis and wanted to predict maximal static expiratory pressure (P) from one or more physical and pulmonary function measurements. One of these was total lung capacity (T), and the correlation $r_{PT}=0.5996$. Their patients' ages ranged from 3 to 23 years, and there is a relation between age (A) and total lung capacity, with $r_{AT}=0.4687$. How much of the relationship between P and T was due to the change in age?

One way to test this would be to measure P and T in a narrow age range, but this would require a much bigger sample size to cover several age ranges. Instead calculate the partial correlation coefficient between P and T independent of age from:

$$r_{PT.A} = \frac{r_{PT} - r_{PA}r_{TA}}{\sqrt{(1 - r_{PT}^2)(1 - r_{TA}^2)}}.$$

The numerator is the difference between the bivariate correlation coefficient between the two variables under consideration and the product of the remaining two bivariate correlation coefficients. The denominator is the square root of the product of 1 minus the first of the remaining correlation coefficients squared and 1 minus the second of the remaining correlation coefficients squared.

In their study the bivariate coefficients were $r_{PT}=0.5996$, $r_{PA}=0.6135$, and $r_{TA}=0.4687$. What is the relationship of P and T independent of age? Calculate the partial correlation coefficient

$$r_{PT.A} = \frac{0.2495 - (0.3332 \times 0.5029)}{\sqrt{(1 - 0.3332^2)(1 - 0.5029^2)}} = 0.1005.$$

The correlation between P and T of 0.5996 was in part because both of these variables increased with age and removing the effect of age reduced the correlation. We have to be careful that the effect of age is directionally the same at different ages. If the bivariate correlation were negative in younger subjects and the opposite way in older subjects, partial correlation would be difficult to interpret. These calculations can be done online at <http://vassarstats.net/par.html> and <http://www.wessa.net/partcorr.wasp>.

With four variables A , B , C , and D , it is possible to examine the correlation coefficient between any two variables if the other two are held constant by extending the previous equation. Individual partial correlation coefficients between sets of three variables are calculated, and then the four-part partial correlation coefficient is

$$r_{AB.CD} = \frac{r_{AB.D} - r_{AC.D}r_{BC.D}}{\sqrt{(1 - r_{AC.D}^2)(1 - r_{BC.D}^2)}}.$$

These calculations can be done with the free online program <http://vassarstats.net/index.html> (see Correlation and Regression).

The confidence limits can be determined (Snedecor and Cochran, 1989). In addition, partial correlations can be determined using Kendall's τ , and Conover (1980) also asserts that partial correlations can be determined using Spearman's ρ coefficient.

Cautionary Tales

There are complexities involved in evaluating multiple and partial correlation coefficients. An excellent discussion of these issues of particular relevance to the social sciences is given by Cohen and Cohen in their Chapter 3 (Cohen and Cohen, 1975). As an example, consider the relationship between Y and two explanatory variables X_1 and X_2 . The multiple correlation coefficient $R_{Y.X_1X_2}$ is

$$R_{Y.X_1X_2} = \sqrt{\frac{r_{YX_1}^2 + r_{YX_2}^2 - 2r_{YX_1}r_{YX_2}r_{X_1X_2}}{1 - r_{X_1X_2}^2}}.$$

Assume that there is no correlation between Y and X_2 , so that $r_{YX_2} = 0$, but there is positive correlation between Y and X_1 as well as between X_1 and X_2 . Substitute $r_{YX_2} = 0$ in the above equation to get

$$R_{Y.X_1X_2}^2 = \frac{r_{YX_1}^2}{1 - r_{X_1X_2}^2}.$$

The denominator is < 1 , so that

$$R_{Y.X_1X_2}^2 > r_{YX_1}^2.$$

This produces a paradoxical result. Despite no correlation between Y and X_2 , the inclusion of X_2 in the multiple regression increases the correlation between Y and X_1 .

Spurious Correlation

This is a subtler trap than any of the issues described before and was elegantly discussed by Archie (1981) under the heading of Mathematical Coupling of Data. Coupling occurs when two variables are related when they have a common component, when one variable is contained in the other, or a third dependent variable is common to both variables.

Fig. 29.8 shows the results of plotting a random number (obtained from <http://www.random.org/>) against the difference between the second and the previous random number.

This was discussed in detail by Oldham (1962). By choosing at random numbers for X_1 and X_2 , and then plotting X_1 against $(X_2 - X_1)$, Oldham showed that there would necessarily be a regression line with a slope of -1 and a correlation coefficient of

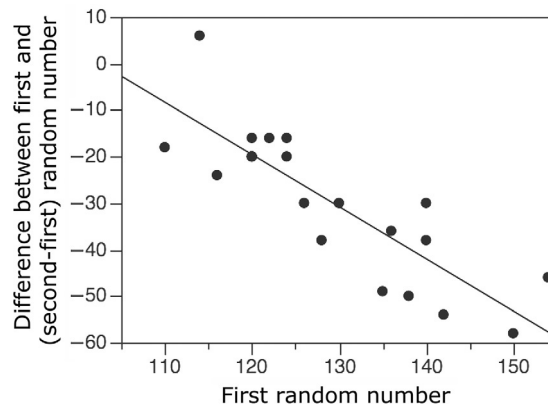


Fig. 29.8 A random number on the X-axis is plotted against the difference between a second random number and the first one on the Y-axis. Difference = $115.21 - 1.12$ first number, $r^2 = -0.70$.

-0.707 , despite the fact that the numbers were chosen at random. This leads to serious errors of interpretation. As one way to escape from this dilemma, Oldham recommended comparing Y_1 and Y_2 where $Y_1 = \frac{X_1 + X_2}{2}$ and $Y_2 = X_1 = X_2$, instead of comparing X_1 with $X_2 - X_1$, which will produce a spurious correlation. This method avoids spurious correlations between X_1 and X_2 , because the sums and differences are not correlated. Other possible combinations are described in his publication. It is permissible to plot X_1 against X_2 and avoid this particular problem.

The theoretical basis for Oldham's results can be seen by considering the relation between total X and its components. Thus if $X_{iT} = X_{iA} + X_{iB}$, where X_{iT} is the total and is correlated with one of its components, say X_{iA} , then the partial correlation r_{TA} between these two variates is

$$r_{TA} = \frac{s_A + r_{AB}s_b}{\sqrt{s_B^2 + 2r_{AB}s_Bs_A + s_A^2}}.$$

If r_{AB} is really zero, this expression simplifies to

$$r_{TA} = \frac{1}{\sqrt{1 + \frac{s_B^2}{s_A^2}}},$$

and if in addition the variances of X_A and X_B are the same, as they were in Oldham's random selection, this becomes $\frac{1}{\sqrt{2}} = 0.7070$.

Archie discussed in detail different types of coupling. In one of his examples he derived a plot of cardiac output against oxygen consumption, for example, and even though he used random numbers for the arterio-venous oxygen difference, he obtained a linear relationship between the two variables with a correlation coefficient of 0.75. Oxygen consumption is the product of the arterio-venous oxygen difference and the cardiac output, so that in effect cardiac output appeared on both axes. Similarly, cardiac output and heart rate are highly correlated because heart rate is a major determinant of cardiac output (cardiac output = heart rate x stroke volume).

Statistical analyses of these types of relationships in the literature produce faulty conclusions, not because of incorrect statistics but because they have been applied incorrectly to mathematically coupled variables. Everyone involved in scientific research should study Archie's publication and its examples. A recent survey of dental research has emphasized the erroneous deductions drawn by ignoring these problems (Tu et al., 2004a, b).

Included as spurious are the problems of comparing a change in some measurement after an intervention with the first measurement; these problems are ubiquitous and serious. An example is a study of the relationship between pre- and postdialysis concentrations of norepinephrine in uremic patients (Musso et al., 1989). The slope obtained by plotting the difference in concentrations against the initial concentration was close to -1 , the theoretical value if there was in reality no correlation (Fig. 29.9, left panel). In a subsequent letter Boer (1990) showed that if predialysis and postdialysis values were plotted against each other, the slope of the relationship was close to zero (Fig. 29.9, right panel). (The investigators actually plotted the difference on the X-axis against the initial value, but this had almost no effect on the results.)

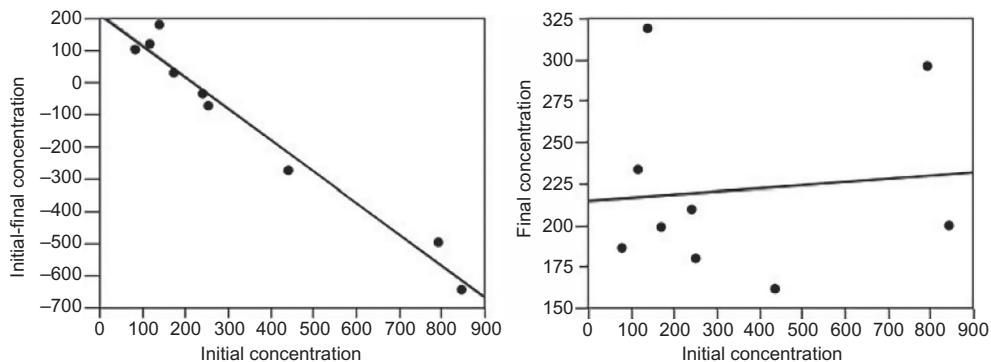


Fig. 29.9 Left panel: Initial values plotted against difference between initial and final concentrations. Right panel: Initial plotted against final concentrations.

The slope of almost -1 is similar to that obtained by Oldham who used random numbers (Fig. 29.9). In the left panel, the slope was -0.9811 and r^2 was 0.961 ($P < 0.0001$), and in the right panel the slope was 0.0189 , with r^2 of 0.1308 ($P = 0.79$). What did the experiment show? It seems simplest to regard the right panel as indicating no relationship between initial and final values. The left panel shows spurious correlation because it plots initial norepinephrine concentration against (initial minus final norepinephrine concentrations). Another example is a study of carbamazepine dosage purporting to show a high correlation between the metabolite-drug ratio and the drug dose-concentration ratio (van Belle and Friel, 1986). The previous finding of a high correlation (0.89 – 0.94) that would allow accurate estimates of patient compliance was shown to be spurious, and corrected analyses were given.

Another fallacy in these studies comes from ignoring the pervasive phenomenon of regression to the mean. Any subject's measurements vary from time to time. For example, systolic blood pressure is not constant. If subject A has a high pressure at time t_0 , it is likely that it will be lower at time t_1 . Conversely, a subject whose initial pressure is low is likely to have a later pressure that is higher. This phenomenon is shown by the data of Musso et al. in Table 29.4.

Table 29.4 Norepinephrine concentrations before and after dialysis

Initial value	Final value	Initial-final value
81	186	−105
119	233	−114
140	319	−179
174	199	−25
245	209	36
254	180	74
439	162	277
794	296	498
847	200	647

The high values decreased and the low values increased, and this is more in keeping with regression to the mean than an effect of dialysis. The right-hand panel in Fig. 29.8, showing no relation between the initial and final norepinephrine concentrations, suggests that this conclusion is correct.

Regression to the mean explains why a baseball pitcher who breaks records in one season is unlikely to repeat it, or why a hedge fund manager whose investments exceed the norm for 5 years may not necessarily do the same in the next 5 years. To reach the peak attainment a large number of factors must be favorable, and this is like the quincunx: a ball may end up in an extreme bin once, but its chances of doing this twice in a row are very low.

Ratios and Scaling Factors

When the independent, dependent, or both variables are made into ratios, the possibility of spurious association looms large. Think about the innumerable measurements taken as a ratio to body surface area (cardiac output, glomerular filtration rate, oxygen uptake), or the incidence per million population, to realize that regressions using ratios are common. One of the earliest warnings about this practice were made by [Tanner \(1949\)](#) and were reinforced by many other studies. Although some disagree with the criticisms, ([Firebaugh and Gibbs, 1985](#)) it would be wise to proceed very cautiously. Positive, negative, or zero correlations may all be incorrectly attained.

[Kronmal \(1993\)](#) recommended using multiple regression equations to avoid the ratio problem, for example, rather than relating $FEV1/height^2$ to age, as done in some studies (e.g., in men $FEV1/height^2 = 1.42 - 0.008 \text{ Age}$), it would be better to calculate the multiple regression equation ([Chapter 30](#))

$$FEV1 = 1.61 - 0.023 \text{ Age} + 0.899 \text{ height}^2 \text{ (for men).}$$

Others have recommended this approach, or else use of analysis of covariance to avoid these problems. [Williams et al. \(1984\)](#), [Vickers \(2001\)](#), [Vickers and Altman \(2001\)](#), [Nevill et al. \(1995\)](#) advised using logarithmic functions. For example, they showed that when relating running speed (Y) to oxygen uptake (X) and body weight (Z), spurious correlation was avoided by calculating

$$Z = 84.3 Y^{1.01} X^{1.03}$$

[Kronmal \(1993\)](#) summed up the field by quoting Neyman: “Spurious correlations have been ruining empirical statistical research from times immemorial.”

Intraclass Correlation

Reliability

One way of using the intraclass correlation (ICC) is to assess reliability. ([Fleiss, 1986](#); [Everitt, 1989](#)). Consider measuring some variable (e.g., blood pressure, attitude to some social issue) and repeating the measurement several times. Then the measured value X_i may be thought of as having two components: a “true” or “error-free” value T_i and a random error ε_i . Thus

$$X_i = T_i + \varepsilon_i.$$

Assume that a variable such as blood pressure or attitude may vary from moment to moment around some “true” value μ , with variance σ_T^2 , and that the random error is unrelated to the value of T and has a mean of zero and a variance of σ_ε^2 . Then the variance of X will be $\sigma_X^2 = \sigma_T^2 + \sigma_\varepsilon^2$. The relative magnitude of these two components of variability

is $R = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_\varepsilon^2}$. R is a form of intraclass correlation and is termed the coefficient of

reliability. R varies between 0 and 1 and can be interpreted directly as a proportion of variance due to random error, unlike the usual interpretation in which r^2 expresses the proportion of variance explained by the regression relationship.

One consequence of unreliability is that it lowers the calculated correlation between two variables. In estimating the correlation between two variables T and U what is actually measured are $X_i = T_i + \varepsilon_i$, and $Y_j = U_j + \varepsilon_j$, where ε_i and ε_j are unrelated, and the true correlation between T and U is ρ_{TU} . The correlation between the measured values X and Y is

$$\rho_{XY} = \rho_{TU} \sqrt{R_X R_Y},$$

where R_X and R_Y are the reliabilities of the two measured variables. If the reliability of each measurement is 0.7 and 0.8, respectively

$$\rho_{XY} = \rho_{TU} \sqrt{0.7 \times 0.8} = 0.75 \rho_{TU}$$

so that the true correlation has been reduced by 25%.

Rater Agreement

More often, ICC is used to assess the agreement among several raters. Consider Fig. 29.10.

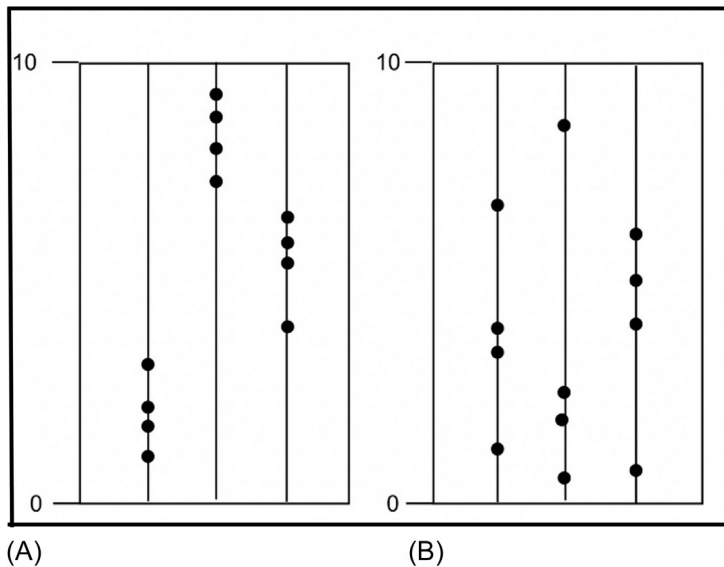


Fig. 29.10 Diagram representing three raters (*vertical lines, classes*) rating maximal expiratory effort on a scale of 0–10 (*dots*) of four different patients. (A) Marked variability among raters. (B) Variability within raters.

In panel A, the variability is mainly between the raters, and for each rater there is high similarity among degrees of effort. The ICC is high. In panel B, however, the degrees of effort are not clustered, so that ICC will be low. See [Uebersax \(2008\)](#).

In assessing the differences between multiple raters, [Shrout and Fleiss \(1979\)](#) showed that there were three different models that depended on the experimental design. In each there are n targets—often ratings of some quality—and r raters. In class 1 each target is rated by a different set of r raters who are randomly selected. In class 2 the r raters are chosen randomly (population sample) and each judge rates each target. Class 3 is similar to class 2 except that we draw conclusions about only those r judges (the population), not judges in general. Each class is analyzed by a different variation of ANOVA (one-way for class 1, two-way for classes 2 and 3); there are additional models that depend on how the data will be used (for simple discussion, see [Koo and Li \(2016\)](#)). The different values for ICC are similar. Statistical consultation is advised.

The model is based on the relationship:

$X = \text{True value} + \text{rater effect} + \text{error}$. Each of these components has variability. The variability of the true value is s_B^2 , the variability due to error is s_W^2 , and the variability due to rater differences is s_r^2 . These variances are assumed to be independent, and the rater and error terms are assumed to each have a mean of zero. Then the variance of X is estimated by $s_B^2 + s_W^2 + s_r^2$, and the ICC is
$$\text{ICC} = \frac{s_B^2}{s_B^2 + s_W^2 + s_r^2}.$$

If there were no variability within each rater, then all the variability would be between the raters (classes), and ICC would be 1. In practice ICC is never 1, but the higher it is, the more the variability is due to intraclass variability than to intersubject variability.

As a simple example I have modified data from [Everitt \(1989\)](#) on ratings of vital capacity measured by four observers on each of six patients ([Table 29.5](#)).

Table 29.5 Table of vital capacity

Subject	Rater 1	Rater 2	Rater 3	Rater 4
1	3350	3510	4000	3750
2	1320	1320	1590	1630
3	2100	2690	2660	2540
4	900	1150	1010	1940
5	3100	3170	3270	3030
6	1700	1800	1400	1250

Looking at the data, it is obvious that each rater gives approximately the same measurement for each patient, so that there will be little difference between raters and most of the variability is related to differences between patients. If we perform a two-way ANOVA without replications, we get the following results ([Table 29.6](#)).

Table 29.6 Two-way NOVA

Source	Df	SS	MS	F
Total	23	21,238,250		
Between subjects	5	19,793,450	3,958,690	50.88
Between raters	3	277,817	92,606	1.19
Error	15	1,166,983	77,799	

This confirms that most of the variability is associated with the subject differences.

We need one further step before calculating ICC. The between-subjects mean square (s_B^2) is estimated by $\frac{s_B^2 - S_W^2}{r}$ and the between-raters mean square by $\frac{s_r^2 - S_W^2}{n}$. Therefore the estimated value of s_B^2 is $\frac{3,958,690 - 77,799}{4} = 970,223$, and the estimated value of the between raters mean square is $\frac{92,606 - 77,799}{6} = 2468$. Now we can calculate the ICC as $\frac{970,223}{970,223 + 2468 + 77,799} = 0.9236$.

This calculation can be performed online at https://www.statstodo.com/IntraclassCorrelation_Pgm.php.

We can also determine how much of the variability is due to differences among raters as $ICC_k = \frac{s_k^2}{s_k^2 + s_B^2 + s_W^2}$. For the previous data, this will be $ICC = \frac{970.223}{970.223 + 2468 + 77,799} = 0.012$ confirming the high agreement among raters.

APPENDIX

1. Spearman's rank correlation coefficient is merely the Pearson correlation coefficient with ranks replacing the numerical values

$$r_s = \frac{\sum \left[X_{\text{rank}} - \left(\frac{N+1}{2} \right) \right] \left[Y_{\text{rank}} - \left(\frac{N+1}{2} \right) \right]}{N(N-1)(N-2)}.$$

12

Related Expressions

The equations for slope (b) and correlation coefficient (r) can be displayed in several equivalent forms. It may be useful to know some of these if you want to recalculate results from a report that does not give the original data

$$r = \frac{s_x}{s_y} b \quad \text{and} \quad b = \frac{s_y}{s_x} r.$$

REFERENCES

- Archie Jr., J.P., 1981. Mathematic coupling of data: a common source of error. *Ann. Surg.* 193, 296–303.
- Boer, P., 1990. Misleading statistics: predialysis norepinephrine and its change after hemodialysis. *Nephron* 55, 78–80.
- Campbell, M.J., Machen, D., 1993. *Medical Statistics. A Commonsense Approach*. John Wiley & Sons, Chichester.
- Chatillon, G., 1984. The balloon rules for a rough estimate of the correlation coefficient. *Amer Stat* 1.
- Cohen, J., 1988. *Statistical Power Analysis for Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Cohen, J., Cohen, P., 1975. *Applied Multiple Regression/Correlation Analysis in the Social Sciences*. John Wiley & Sons, New York.
- Conover, W.J., 1980. *Practical Nonparametric Statistics*. John Wiley & Sons, New York.
- Curran-Everett, D., 2010. Explorations in statistics: correlation. *Adv. Physiol. Educ.* 34, 186–191.
- Edwards, A.L., 1984. *An Introduction to Linear Regression and Correlation*. W.H. Freeman and Co, New York.
- Everitt, B.S., 1989. *Statistical Methods for Medical Investigations*. Oxford University Press, New York.
- Firebaugh, G., Gibbs, J.P., 1985. User's guide to ratio variables. *Am Sociolog Rev* 50, 713–722.
- Fleiss, J.L., 1986. *The Design and Analysis of Clinical Experiments*. John Wiley & Sons, New York.
- Glantz, S.A., 2005. *Primer of Biostatistics*. McGraw-Hill, New York.
- Kleinbaum, D.G., Kupper, L.L., Muller, K.E., 1988. *Applied Regression Analysis and Other Multivariable Methods*. PWS-KENT Publishing Company, Boston.
- Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* 15, 155–163.
- Kronmal, R.A., 1993. Spurious correlation and the fallacy of the ratio standard revisited. *J. Royal Statist. Soc. Ser. A (Stat. Soc.)* 3, 379–392.
- Musso, N.R., Deferrari, G., Pende, A., Vergassola, C., Saffioti, S., Gurreri, G., Lotti, G., 1989. Free and sulfoconjugated catecholamines in normotensive uremic patients: effects of hemodialysis. *Nephron* 51, 344–349.
- Nevill, A.M., Holder, R.L., McShane, P., Kronmal, R.A., 1995. Letters to the editor. Spurious correlations and the fallacy of the rationstandard revisited. *J Roy Stat Soc. Ser A* 158, 619–625.
- Oldham, P.D., 1962. A note on the analysis of repeated measurements on the same subjects. *J Chron Dis* 15, 969–977.
- O'Neill, S., Leahy, F., Pasterkamp, H., Tal, A., 1983. The effects of chronic hyperinflation, nutritional status, and posture on respiratory muscle strength in cystic fibrosis. *Am. Rev. Respir. Dis.* 128, 1051–1054.
- Senzaki, H., Isoda, T., Paolucci, N., Ekelund, U., Hare, J.M., Kass, D.A., 2000. Improved mechanoenergetics and cardiac rest and reserve function of in vivo failing heart by calcium sensitizer EMD-57033. *Circulation* 101, 1040–1048.
- Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* 86, 420–428.
- Snedecor, G. W. & Cochran, W. G. 1989. *Statistical Methods*, Ames, Iowa, Iowa State University Press.
- Tanner, J.M., 1949. Fallacy of per-weight and per-surface area standards, and their relation to spurious correlation. *J. Appl. Physiol.* 2, 1–15.
- Tu, Y.K., Clerehugh, V., Gilthorpe, M.S., 2004a. Ratio variables in regression analysis can give rise to spurious results: illustration from two studies in periodontology. *J. Dent.* 32, 143–151.
- Tu, Y.K., Maddick, I.H., Griffiths, G.S., Gilthorpe, M.S., 2004b. Mathematical coupling can undermine the statistical assessment of clinical research: illustration from the treatment of guided tissue regeneration. *J. Dent.* 32, 133–142.
- Uebersax, J., 2008. *Statistical Methods for Rater Agreement*. Available at: <http://www.john-uebersax.com/stat/agree.htm>.
- Van Belle, G., 2002. *Statistical Rules of Thumb*. Wiley Interscience, New York.
- Van Belle, G., Friel, P.N., 1986. Problem of spurious correlation in the evaluation of steady-state carbamazepine levels using metabolite data. *Therap Drug Monitoring* 8, 177–183.

- Vickers, A.J., 2001. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. *BMC Med. Res. Methodol.* 1, 6.
- Vickers, A.J., Altman, D.G., 2001. Statistics notes: analysing controlled trials with baseline and follow up measurements. *BMJ (Clin. Res. Ed.)* 323, 1123–1124.
- Williams, G.W., Forsythe, S.B., Textor, S.C., Tarazi, R.C., 1984. Analysis of relative change and initial value in biological studies. *Am. J. Phys.* 246, R122–R126.
- Zar, J.H., 2010. *Biostatistical Analysis*. Prentice Hall, Upper Saddle River, NJ.

CHAPTER 30

Multiple Regression

BASIC CONCEPTS

Introduction

This subject is complicated, and inexperienced investigators should not perform these analyses on their own. Nevertheless, all investigators and even most readers should be aware of the issues. Few dependent variables are related to only one independent variable, and nonlinear relationships are common. The intent of this chapter is to provide basic understanding of when and how to proceed with these analyses and what difficulties to take into account.

Frequently, a dependent variable Y is a function of more than one explanatory variable, so that the general equation is

$\hat{Y}_i = c + b_1X_1 + b_2X_2 + \dots b_kX_k$ or $Y_i = c + b_1X_1 + b_2X_2 + \dots b_kX_k + \varepsilon_i$ for the dependence of Y on k different X variables. In principle the same procedure is used as for simple bivariate regression. An equation is sought that minimizes the sum of the squared deviations from regression $\sum \varepsilon_i^2$ and maximizes the (multiple) correlation coefficient, here termed R .

To minimize $\sum \varepsilon_i^2$, the coefficients are calculated by formulas similar to those used for bivariate regression (Edwards, 1984; Glantz and Slinker, 2001). The computations should be done with computer programs. Despite the complexity, there are several freeware online programs. The site at <http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/MultRgression.htm> allows 4 X variables but only 16 samples. The sites at <http://www.xuru.org/rt/MLR.asp#Manually>, http://www.wessa.net/rwasp_multipleregression.wasp and <http://vassarstats.net/> (see Correlation and Regression) allow many more samples and dependent variables. Many of these programs allow you to copy and paste from your data sheet. These free programs will usually not perform the subsidiary analyses discussed later.

Multiple Linear Regression With Two X Variables

This is the simplest form of multiple regression and illustrates the main features. The basic regression equation is $\hat{Y}_i = c + b_1X_1 + b_2X_2$ or $Y_i = c + b_1X_1 + b_2X_2 + \varepsilon_i$. For two independent X variables the equation defines a plane or surface. Imagine this as a set of bivariate relations relating Y to X_1 with X_2 held constant, and Y to X_2 holding X_1 constant. The requirements for multiple regression are similar to those for bivariate

regression—ratio numbers, independence, normal distributions, equal variances—but these are more difficult to evaluate with several independent variables.

Once the regression equation has been obtained, can we reject the null hypothesis that each of the two coefficients comes from a population in which each coefficient is zero? Even though two independent variables have been measured, one of them might not offer any predictive information. To test the individual coefficients, use the formula

$$s_{b_1} = \sqrt{\frac{s_{Y \cdot X}^2 (\text{or } MS_{\text{res}})}{\sum (X_{1i} - \bar{X}_1)^2 (1 - r_{X_1 X_2}^2)}}$$

for b_1 , and a corresponding formula for b_2

$$s_{b_2} = \sqrt{\frac{s_{Y \cdot X}^2 (\text{or } MS_{\text{res}})}{\sum (X_{2i} - \bar{X}_2)^2 (1 - r_{X_1 X_2}^2)}}$$

These formulas use the squared correlation between X_1 and X_2 , $r_{X_1 X_2}^2$. If X_1 and X_2 are completely uncorrelated so that $r^2 = 0$, the expression for the standard deviation of b becomes the same as that for a bivariate linear relationship. If X_1 and X_2 are highly correlated, for example, $r = 0.9$, then the variance of b_1 or b_2 becomes 10 times larger, making it much more difficult to reject the null hypothesis that $b_i = 0$. The variance of b_i is also large if $s_{Y \cdot X}^2$ is big or $\sum (X_i - \bar{X}_i)^2$ is small, so that the correlation between explanatory variables is not the only factor to be considered.

Perform a t -test on each coefficient. Then

$$t_1 = \frac{b_1}{s_{b_1}} \quad \text{or} \quad t_2 = \frac{b_2}{s_{b_2}}$$

t is evaluated with $N - 3$ degrees of freedom—one for each variable.

In the multiple regression equation, the coefficient c is the intercept, and just as in simple linear regression it represents the value of \bar{Y} when all the X variables are zero. The coefficients (b_1 or b_2) show how much \bar{Y} changes for a one-unit change in one variable when the other is kept constant and are termed partial regression coefficients.

The previous example can be extended to more independent variables. In a study relating systolic blood pressure to age, weight, height, subcutaneous fat thickness, and arm size Whyte (1959) observed.

Systolic pressure (mm Hg) = $165 + 0.35 \text{ weight (lbs)} - 0.01 \text{ age (years)} - 1.55 \text{ height (inches)} - 0.09 \text{ fat (mm)} + 0.81 \text{ arm size (cm)}$.

It is also possible to calculate standardized coefficients, represented by b^* , as

$$b^* = b_1 \frac{s_{X_1}}{s_Y},$$

where b_i is the regression coefficient in question, s_X is the standard deviation of variate X_i , and s_Y is the standard deviation of variate Y . This is interpreted as the change in standard

deviation of predicted Y for a 1 standard deviation change in X_i if the remaining X variates remain constant. The value for b^* ranges from $+1$ to -1 and indicates the strength of the association between Y and X_i without the need to consider units. The regression equation obtained by Whyte (above) can be changed into standardized units to give

$$\begin{aligned}\text{Systolic pressure (mm Hg)} = & 0.513 \text{ weight (lbs)} - 0.004 \text{ age (years)} \\ & - 0.272 \text{ height (inches)} \\ & - 0.086 \text{ fat (mm)} + 0.133 \text{ arm size (cm)}.\end{aligned}$$

(The coefficients for fat and arm thickness were too low to consider further.)

In this form the coefficients indicate the relative importance of the X variables. Weight is therefore the single most important variable because it has the highest standardized coefficient of 0.513. Neither equation tells us how good the prediction equation is. This has to be determined by the multiple correlation coefficient that, at 0.24, was small.

Multiple Collinearity

This is an important complicating factor. The expression for the standard deviation from regression shows that s_b is smallest when r_{X_1, X_2}^2 is smallest. The more closely X_1 and X_2 are correlated with each other the smaller the component $1 - r_{X_1, X_2}^2$ will be, so that the denominator of the expression for s_b becomes smaller, and s_b becomes bigger. The regression analysis will not be able to distinguish the separate contributions of each of the two correlated variables.

High degrees of correlation are common. When they occur, the resulting analyses become erratic, with unpredictable consequences. The R^2 in predicting Y is still accurate, with perhaps a low P value, but the contributions of individual predictors may be incompatible with a high predictive value. In addition, the confidence limits will be very wide, and adding or subtracting a single value can materially alter the regression equation.

A different random sample from the populations of Y , X_1 and X_2 should give similar even if not identical coefficients, and we can envisage a stack of planes that are close to each other and approximately parallel. If X_1 and X_2 are highly correlated, however, other samples could have wildly different coefficients and planes. This is the serious consequence of multicollinearity. [Slinker and Glantz \(1985\)](#) and [Glantz and Slinker \(2001\)](#) give examples of this problem and explain it well. Other helpful descriptions of multicollinearity appear in the books by [Kleinbaum et al. \(1988\)](#) and [Neter et al. \(1996\)](#).

The multiple regression technique described before can be extended to many X variates.

Prerequisites for Multiple Regression

These are ratio numbers, normality of distribution, independence of measurements, linearity, and homoscedasticity, as well as absence of multicollinearity. The requirements may be more difficult to determine by casual inspection than in simple bivariate

regression because a single scattergram may not suffice to show problems. Normality can be tested but is not important if sample size is large. Independence of observations is often obvious, but if, for example, a time-dependent relationship is present, a test such as the Durbin-Watson test (Chapter 31) can be done. Of more importance are linearity and homoscedasticity, because in their absence the tests of hypotheses may be erroneous. Linearity and homoscedasticity can be examined by plotting residuals of predicted against observed dependent variables for the whole data set.

This global test, however, does not mean that all individual relationships are linear. One way to test this is to do residual analyses for the bivariate relationship of the dependent variable to each of the independent variables. Examples are shown in the Advanced section later.

Determining the “Best” Regression

After determining the regression equation relating Y to several X variates, decide which of the X variates adds substantially to the prediction of Y . Ideally, the model should include all the relevant X variates that determine Y . If these are all present there is a *correctly specified* model, one in which the various components are estimated without bias. If some important determinants are left out, the model is *underspecified*, leading to incorrect and biased coefficients and increased variability, as shown by comparing linear with quadratic regression described in Chapter 27. If there are *extraneous* factors unrelated to the Y or any of the other X variates, they do not affect the results except that it takes more work to collect and analyze the data. Finally, if redundant factors are present, the model is *overspecified* and risks multicollinearity problems.

With many X variables, there are several ways of proceeding. One is to examine all possible combinations of the X variables, known as all subsets regression. The program determines R^2 and residual sums of squares for all possible combinations of the X variates. If there are many variables, and especially if powers of the variables and cross-products of the variables are included, it provides so many results that selecting the best one is difficult. Some criterion is needed to indicate which of the many equations gives a useful relationship. One way of telling if additional X variates contribute to the regression is to calculate R^2 that increases as more variates are introduced. In applying this technique to a regression with many independent X variates, R^2 tends to increase with each added factor, even if in the population no change in correlation has occurred. R^2 has a slight positive bias; the expected value of R^2 by chance is $k/(N-1)$. To allow for this source of error most programs calculate an adjusted R^2 :

$$\text{Adjusted } R^2 = 1 - (1 - R^2) \frac{N-1}{N-k-1},$$

where N is the sample size and k the number of regressors. If k is small relative to N then the adjusted R^2 and R^2 will be similar, but if k is large relative to N the two forms of R^2 will differ. There is a convenient online calculator for this adjustment at <http://danielsoper>.

[com/statcalc3/calc.aspx?id=25](http://www.statcalc3/calc.aspx?id=25), <http://www.cedu.niu.edu/~walker/calculators/adjr2.asp>, or <https://www.easycalculation.com/statistics/adjusted-r-squared.php>.

The better the equation fits the data the higher the value for R^2 and the lower the residual sums of squares will be.

Another method for assessing the number of variables needed to avoid underspecification is the C_p statistic developed by Mallows in 1964. C_p is the ratio between the residual mean square calculated from a subset of the variates to the total residual mean square with a correction. Formally it is

$$C_p = \frac{MS_{res,p}}{MS_{res,k}} - (N - 2p).$$

N is the total sample size, k is the total number of variates, and p is the reduced number of X variates being examined. Because the model includes the intercept, p is one more than the number of variates included. The closer C_p is to p , the better specified is the equation. The statistic should not be used to choose one “best” model, but rather to choose models to exclude because their values of C_p are much greater than p . Small differences for C_p should not be used to decide which models to include.

Some programs include the Akaike “an information criterion (AIC)” based on maximal likelihood principles. $AIC = 2k + N \ln \left(\frac{2\pi SS_{res}}{N} \right) + 1$, where k is the number of parameters (here X variates plus the equation constant). Because the AIC is a ranking rather than an absolute index, the constant 2π and $+1$ values can be removed, leaving

$$AIC = 2k + N \ln \left(\frac{SS_{res}}{N} \right).$$

The AIC attempts to find the minimum number of free parameters that explains the data, so that the lower the value of AIC, the more likely is it that superfluous parameters are excluded. As for C_p , the AIC should be used to select the few regressions with the lowest values to concentrate on, rather than being used to select the best single combination of parameters.

Other criteria have been proposed, one of which is Schwarz’s Bayesian Information Criterion (BIC). This may be written as $BIC = N \ln \left(\frac{SS_{res}}{N} \right) + k \ln N$, where N is the number of data points, and k is the number of X variates. It is similar to the AIC except for the logarithm of N in the second term. This expression penalizes excess parameters more heavily than does the AIC, and it also penalizes larger sample sizes that have the disadvantage of making even small differences incompatible with the null hypothesis. Like the AIC, the smaller the value for BIC, the more useful the model is likely to be.

There is no perfect way of selecting the best regression. Probably the combination of good fit and biological sense allows for effective data fitting.

If there are too many variables for all subsets regression, some alternatives are forward stepwise regression, backwards stepwise regression, and mixed regression. Forward

stepwise regression starts with the highest bivariate correlation, say between Y and X_3 . Then it adds another variable, perhaps one with the next highest bivariate correlation, and tests by ANOVA to determine if there has been a reduction in the residual sum of squares and thus an increase in adjusted R^2 . The process continues until no further substantial change in adjusted R^2 occurs as indicated by a large P value. Backward stepwise regression begins with all the variables included, and then recalculates the adjusted R^2 after omitting the X variate with the lowest bivariate correlation with Y . If the change is unimportant (large P value), the inference is that the deleted X variate was not a factor. Then the next X variate is deleted, and the process continues until deleting an X variate produces a substantial change in adjusted R^2 , (small P value) so then that variate is retained and the final equation is determined. If there are many X variates these two stepwise regression techniques may reach different conclusions, and sometimes the order in which variates are included or excluded makes a difference. As an alternative, the mixed (sometimes called stepwise) regression technique is used. The procedure begins like the forward regression method, but after each addition of a variable a backward elimination is done to see if any of the variables previously included have become redundant. It is possible for variables to be added, removed, added in again, and so on, as the regression equation changes. In fact, different results can occur if variables are entered in a different order, or if variables are entered in sets. It is also possible for the forward, backward, and stepwise regression methods to end up with different sets of variables. The more variables there are in the full regression equation the more the potentiality for error. If the different regression methods selected a common set of variables, then these could very well be the important ones to consider. If they selected different sets of variables, then it is for the investigator to evaluate the usefulness of each variable as a predictor, based on prior knowledge, or even to select a likely subset of variables and repeat the experiment or observational study.

Which Is the Best Regression?

Numerous methods have been proposed to determine the “best” regression model, but the notion of “best” is not well defined. The investigator has to use common sense and not rely too much on complex and often poorly understood statistical techniques. An excellent discussion of the advantages and disadvantages of each selection method is presented by [Katz \(2006\)](#). He recommends the forward selection method for relatively small sample sizes or if there is concern about multicollinearity, the backward selection method if there are likely to be suppressor variables, and the best subset regression if the number of explanatory variables is small. Katz comments too that elimination of regressor variables is usually better justified by biological insights than statistical techniques.

One difficulty with all of these complex procedures is that they are designed to minimize the errors for that particular set of data. Despite determining coefficients, there is no

guarantee that another data set will find the same coefficients or even the same set of variables. To guard against this contingency, use a criterion sample. Split the existing data set at random into two portions, sometimes called the training and holdout (or validation) portions and compare the resulting equations. There are several ways of doing this. One simple approach is to compute the predicted values in the training and holdout groups separately, and then for each group calculate r^2 between observed and predicted Y values. The difference $r^2(\text{training}) - r^2(\text{holdout})$ is called shrinkage and is usually positive. If it is small (e.g., about 0.10), then the fitting process was probably reliable and it is reasonable to pool the two data sets and determine a final regression equation. Greater shrinkage suggests unreliability and suggests caution.

In one such study (Thursz et al., 1995) the effect of various MHC class I and class II antigens in the clearance of hepatitis B virus (HBV) was examined in 1344 children, about one-third of whom were positive for HBV core antibody. Many apparent associations were detected. When these associations were assessed in 235 adults exposed to HBV, only a few of these associations were confirmed. Other examples of this validation procedure are readily found (Richardson et al., 2001, Haricharan et al., 2009).

Residual analysis is important, especially if sample size is <100 . One approach is to plot the residuals (predicted-observed values) for the full model against the dependent variable and each independent variable. As with any other regression residuals, linearity, normality, and homoscedasticity can be tested (Glantz and Slinker, 2001). If any residual plots seem to depart substantially from the ideal, it may be worth transforming one of the variables.

Nonlinear Regression: Polynomial Regression

If in a bivariate regression the relationship is clearly alinear, then linear regression is inefficient. If there is a known mechanism, for example, a logarithmic or exponential relationship, then either the Y or the X variate can be transformed, and a linear regression can be done on the transformed variables. If no suitable transformation is known, then fitting a power function, a specific form of multiple regression, may be valuable. These functions are of the form

$$\hat{Y}_i = c + b_1 X_1 + b_2 X_i^2 + b_3 X_i^2 + \dots b_k X_i^k.$$

A function including X^2 is a quadratic, X^3 is cubic, X^4 is quartic, X^5 is quintic, and so on. A quadratic expression produces a parabola, and because parabolas have different curvature in different portions and because curvature varies with the parabola's focal length, this is often the most general form of curved regression used. These computations are computer intensive and can be done online at <http://www.xuru.org/rt/PR.asp>, <http://polynomialregression.drque.net/online.php>, and <https://arachnoid.com/polysolve/>.

Although in theory any number of powers can be used, it seldom makes sense to go beyond a cubic form.

Conclusions About Regression Methods in General

The key to effective multiple regression methods is to design the study correctly. Apart from asking material questions and making the required measurements accurately, care must be taken to have an adequate sample size. Unlike the usual power calculations, what is at stake here is the number of variables examined relative to the total number of subjects. [Altman \(1992\)](#) recommends no more than $N/10$ variables, where N is the total sample size. [Katz \(2006\)](#) recommends determining the power of the various bivariate relationships; if any of these have inadequate power then the multiple regression is also likely to be underpowered. Commercial programs such as Power and Precision or PASS are available.

Fitting the data by a regression equation should be regarded as the beginning of the analysis and not an end in itself. After checking for outliers and errors, the resulting equation has to be useful. The purpose of the modeling is not merely to produce an equation that fits the data, but to be able to predict future values of the dependent Y variable and to assess the relative importance of the explanatory factors.

As summarized by [Chatfield \(1988\)](#), the objectives in model building include:

1. To provide a parsimonious summary or description of one or more sets of data.
2. To provide a basis for comparing several sets of data.
3. To confirm or refute a theoretical relationship suggested a priori.
4. To describe the properties of the random or residual variation.....to assess the uncertainty in any conclusion.
5. To provide predictions which act as a “yardstick” or norm, even when the model is known not to hold for some reason.
6. To provide physical insight into the underlying physical process.

Chatfield does not consider “trying lots of different models until a good-looking fit is obtained” to be one of the purposes of model building and points out that the “choice between models which fit data approximately equally well should be made on grounds external to the data.” The more complex the statistical analysis, for example, multiple regression, the more his advice should be taken to heart.

One useful test is to evaluate the coefficients to see if they make sense from a physical or biological point of view. Test the model by inserting reasonable values for the independent variable(s) to make sure that the predicted value of Y appears to be reasonable, and that for example an impossible negative value is not calculated.

A second is to add new data and then make sure that the revised prediction equation is little altered from the original equation. This is essential, because any form of regression

modeling is designed to optimize the prediction for that data set, and it is important to know if it remains stable and predictive when more data are added.

ADVANCED CONCEPTS AND EXAMPLES

For those who want to see the principles mentioned before in practice, detailed examples are presented later. It is not essential to read this section which many will find difficult.

Multiple Regression With Many Independent X Variates

Table 30.1 features a number of variables in patients with cystic fibrosis (O'Neill et al., 1983). These variables were considered those most likely to predict PE_{\max} , a measure of malnutrition, and can be tested to determine the most useful explanatory or predictive model.

Table 30.1 Cystic fibrosis data

Subject	Age	Gender	Height	Weight	BMP	FEV1	RV	FRC	TLC	PE_{\max}
1	7	0	109	13.1	68	32	258	183	137	95
2	7	1	112	12.9	65	19	449	285	134	85
3	8	0	124	14.1	64	22	441	268	147	100
4	8	1	125	16.2	67	41	234	146	124	85
5	8	0	127	21.5	93	52	202	131	104	95
6	9	0	130	17.5	68	44	308	155	118	80
7	11	1	139	30.7	89	28	305	179	119	65
8	12	1	150	28.4	69	18	369	198	103	110
9	12	0	146	25.1	67	24	312	194	128	70
10	13	1	155	31.5	68	23	413	225	136	95
11	13	0	156	39.9	89	39	206	142	95	110
12	14	1	153	42.1	90	26	253	191	121	90
13	14	0	160	45.6	93	45	174	139	108	100
14	15	1	158	51.2	93	45	158	124	90	80
15	16	1	160	35.9	66	31	302	133	101	134
16	17	1	153	34.8	70	29	204	118	120	134
17	17	0	174	44.7	70	49	187	104	103	165
18	17	1	176	60.1	92	29	188	129	130	120
19	17	0	171	42.6	69	38	172	130	103	130
20	19	1	156	37.2	72	21	216	119	81	85
21	19	0	174	54.6	86	37	184	118	101	85
22	20	0	178	64	86	34	225	148	135	160
23	23	0	180	73.8	97	57	171	108	98	165
24	23	0	175	51.1	71	33	224	131	113	95
25	23	0	179	71.5	95	52	225	127	101	195

Gender: 0, male; 1, female. *BMP*, body mass index as a percentage of age-specific normal median value; *FEV1*, forced expiratory volume in 1 second; *FRC*, functional residual capacity; PE_{\max} , maximal static expiratory pressure (cm H₂O); *RV*, residual volume; *TLC*, total lung capacity. PE_{\max} is the dependent variable.

Linearity and homoscedasticity can be examined by plotting residuals of predicted against observed PE_{\max} for the whole data set (Fig. 30.1).

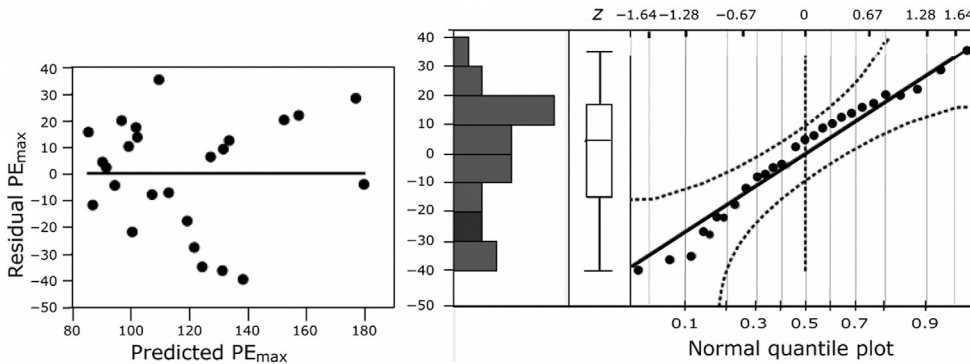


Fig. 30.1 Left panel: Residual plot with all variables. Right panel: Quantile plot for Residual PE_{\max} . The plots do not show any marked alinearity or heteroscedasticity.

This global test, however, does not mean that all individual relationships are linear. One way to test this is to do residual analyses for the bivariate relationship of the dependent variable to each of the independent variables (Fig. 30.2).

Although none of these individual bivariate regressions appear to be grossly alinear, there is an advantage to testing the linearity of all of the independent variates simultaneously. Some programs offer tests such as Ramsey's RESET test, also referred to as the Powers of Y (POY) test or a mis-specification test. If the original equation can be summarized as $Y_{ii} = c + \sum b_i X_i + \varepsilon_i$, where $b_i X_i$ represents all the possible explanatory variables, then Ramsey's auxiliary equation is $Y_i = c + \sum b_i X_i + b_{i+1} \hat{Y}^2 + b_{i+2} \hat{Y}^3 \dots + \varepsilon_i$; power > 4 is seldom justified. The power functions are applied to the predicted values of Y . The null hypothesis is that the coefficients of the extra terms are all zero. If the hypothesis is rejected, then nonlinear terms must be included in the final analysis. To test the null hypothesis, the multiple correlation coefficients for the original data set R_O^2 and the auxiliary data set R_A^2 are compared by

$$\frac{\frac{R_A^2 - R_O^2}{k-1}}{\frac{1 - R_A^2}{N-k}} \sim F_{(k-1, N-1)},$$

where N is the sample size and k is the number of auxiliary parameters.

When this equation was calculated for the cystic fibrosis data,

$$\hat{Y}_i = -437.31 + 4.93\text{Age} - 7.19\text{Gender} + 3.33\text{Height} - 12.17\text{Weight} + 6.03\text{BMP} \\ - 4.79\text{FEV1} - 0.086\text{RV} + 0.96\text{FRC} - 0.040\text{TLC} + 0.017\text{PE}_{\max}^2 + 0.000023\text{PE}_{\max}^3$$

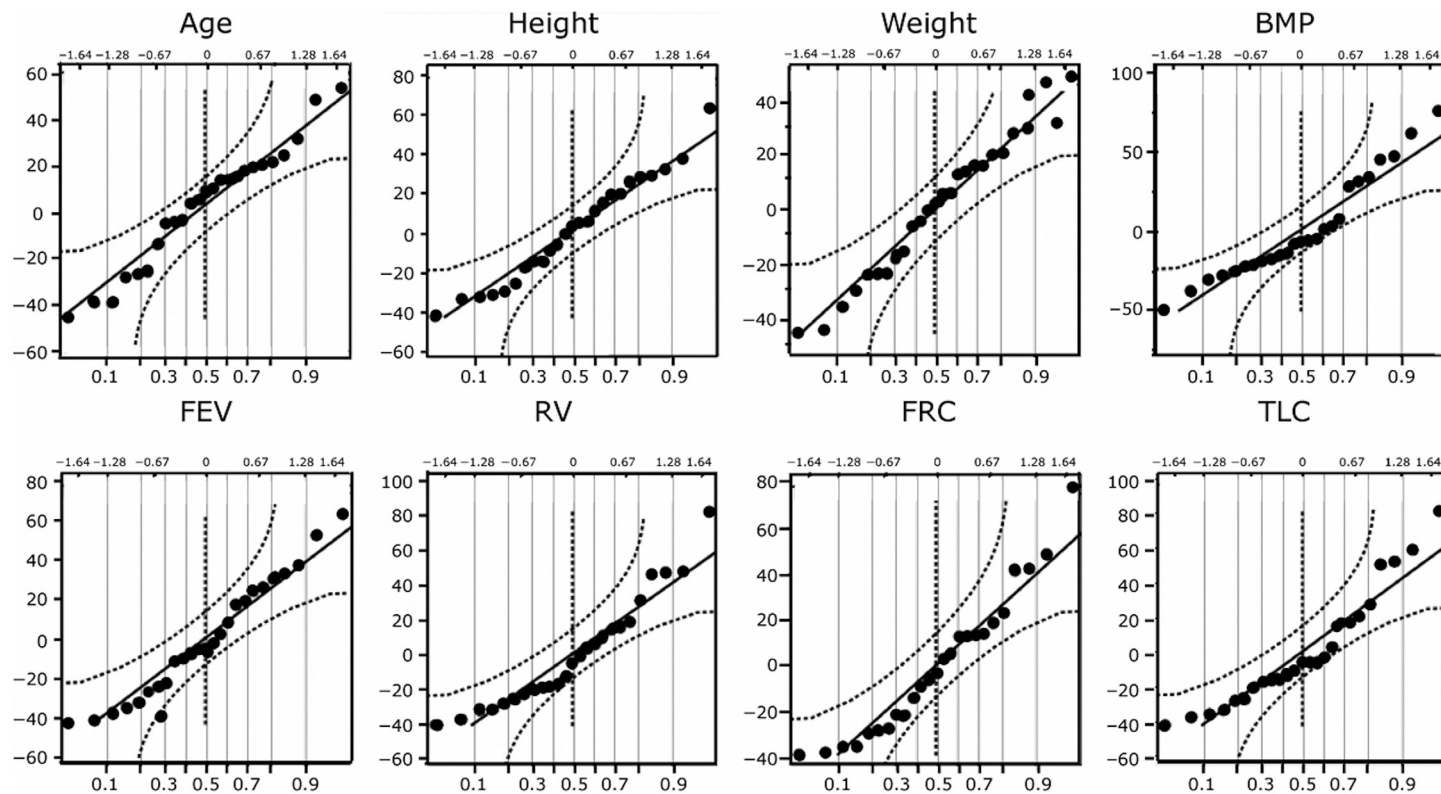


Fig. 30.2 Residual distributions of PE_{\max} for different variables, and normal quantile plots for bivariate regressions.

and $R_A^2 = 0.7591$. The adjusted R_O^2 without the power functions was 0.6349. Then the previous formula gives

$$\frac{\frac{0.7591 - 0.6349}{10}}{\frac{1 - 0.7591}{14}} = 0.72.$$

This value for F is too low to reject the null hypothesis, so there is no evidence that a linearity is a problem; this confirms what is in Fig. 30.2. Thursby and Schmidt suggested that a similar test based on powers of the explanatory (POX) variables might be better. The Ramsey reset test can be performed online at <http://www.wessa.net/esteq.wasp>.

Other programs provide tests for heteroscedasticity, such as the tests by White, Glesjer, or Breusch-Pagan-Godfrey (Chapter 27). These tests involve regressing the residuals or their logarithms against the independent variables, and some of the tests include power functions of the variables and their cross-products.

White's test is important for assessing heteroscedasticity. It uses an auxiliary regression of the form

$$\begin{aligned} \widehat{\varepsilon}_i^2 = & c + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_k X_k \\ & + \beta_{k+1} X_1^2 + \beta_{k+2} X_2^2 + \dots \beta_{2k-1} X_k^2 \\ & + \beta_{2k} X_1 X_2 + \beta_{2k+1} X_1 X_3 + \dots \beta_{k(k-1)} \frac{X_{k-1} X_k}{2} + v_i \end{aligned}$$

where v_i is a normally distributed error term independent of ε . In other words, regress the square of the residuals (ε) against a constant, the original explanatory variables, the squares of the original explanatory variables, and the cross-products of the explanatory variables.

As an example, consider the regression of PE_{\max} against Weight, FEV1, and BMP. This results in the equation

$$\hat{Y}_i = 126.01 + 1.53\text{Weight} + 1.11\text{FEV1} - 1.46\text{BMP}$$

From this, obtain the residuals and derive the auxiliary equation

$$\begin{aligned} \hat{\varepsilon}_i^2 = & -27295 - 6.42\text{Weight} + 644.01\text{BMO} + 166.02\text{FEV1} + 0.21\text{Weight}^2 \\ & - 2.18\text{FEV1}^2 - 4.13\text{BMP}^2 + 0.11\text{Weight} \times \text{BMP} + 0.20\text{Weight} \times \text{FEV1} + 2.69\text{BMP} \times \text{FEV1}. \end{aligned}$$

This equation has $R_A^2 = 0.355858$. To evaluate this, multiply it by N to get $24 \times 0.355858 = 8.54$. This is evaluated like chi-square with degrees of freedom equal to the number of coefficients excluding the constant. There are 9 such coefficients, and a chi-square of 8.54 with 9 degrees of freedom, so that $P = 0.4807$. There is thus no reason to consider heteroscedasticity, again confirming the results in Fig. 30.2.

Analysis of Cystic Fibrosis Example

Begin by examining the regressions between PE_{\max} and each of the other explanatory variables (Table 30.2).

Table 30.2 Cystic fibrosis data

Variable	<i>b</i>	R^2	Adj R^2	RMS res	se	<i>t</i>	<i>P</i>
Age	4.0547	0.3764	0.3492	26.9736	1.0884	3.73	0.0011
Gender	−19.045				13.176	−1.45	0.16
Height	0.9319	0.3591	0.3312	27.3448	0.2596	3.59	0.0015
Weight	1.1867	0.4035	0.3776	26.3794	0.3009	3.94	0.0006
BMP	0.6392	0.0527	0.01149	33.2443	0.5652	1.13	0.2698
FEV1	1.3539	0.2056	0.1710	30.4440	0.5550	2.44	0.0228
RV	−0.1227	0.0996	0.0604	32.110	0.769	−1.59	0.1244
FRC	−0.2859	0.1662	0.1300	31.1883	0.1335	−2.14	0.0431
TLC	−0.3579	0.0330	−0.0091	33.5880	0.4041	−0.89	0.3849

As R^2 gets smaller, the residual root mean square (RMS res) becomes bigger, because it is derived from the residual sum of squares that is closer to the total sum of squares of the dependent variable that was 26832.640. For the dichotomous variable gender, the coefficient −19.045 is the difference between the mean values for males and females, with females having a lower value than males.

The bivariate correlations in Table 30.2 are useful only to determine the best single explanatory variable but give no information about combining explanatory variables. The next step is to examine the bivariate correlations between each of the explanatory *X* variables to get more information about the model, and in particular determine the possibility of multicollinearity. (Table 30.3).

Table 30.3 Bivariate correlation coefficients

	PEmax	Age	Gender	Height	Weight	BMP	FEV1	RV	FRC
Age	0.6135								
Gender	−0.289	−0.167							
Height	0.5992	0.9261	−0.1680						
Weight	0.6352	0.9059	−0.1990	0.9207					
BMP	0.2295	0.3778	−0.138	0.4413	0.6725				
FEV1	0.4534	0.2945	−0.528	0.3167	0.4483	0.5455			
RV	0.3156	0.5519	0.271	0.5695	0.6215	0.5824	0.6659		
FRC	0.4077	0.6378	0.184	0.6387	0.6157	0.4369	0.6588	0.9136	
TLC	0.1816	0.4694	0.024	0.4571	0.4185	0.3649	0.4430	0.5891	0.6870

r values > 0.9 shown in bold type.

An extension of this matrix, suggested by Weisberg (1985) (see his Fig. 6.2) is the matrix plot developed by Chambers et al. (1983). The correlation matrix presented in Table 30.3 is

supplemented by a grid of bivariate plots, one for each pair of variates. This shows more clearly the form of the relationship and whether there are any obvious outliers.

Detecting Multicollinearity

Multicollinearity should be suspected if R^2 is high but the individual coefficients have unexpected values or unusually high standard errors. If any two variables, X_j and X_k , have a correlation coefficient above 0.8 you should be suspicious, and multiple collinearity is almost certain to be a problem if the correlation coefficient is over 0.95. This check is less useful if there are more than two X variables, because it is possible for redundancy to be present by virtue of the combination of several variables, not just two.

As an example, examine the regression between PE_{\max} and both age and height. This yields the equation $\hat{PE}_{\max} = 17.8600 + 2.7278\text{Age} + 0.3397\text{Height}$, with adjusted R^2 of 0.3271. The standard errors of the coefficients of age and height are, respectively, 2.9326 and 0.6900, both large.

Now regress PE_{\max} on age, height, and weight to get

$$\hat{PE}_{\max} = 64.3793 + 1.5553\text{Age} - 0.0728\text{Height} + 0.8689\text{Weight}.$$

The adjusted R^2 has increased slightly to 0.3277, and the standard errors of the coefficients are, respectively, 3.1489, 0.8016, and 0.8601. Although we have gained little predictive ability by adding in weight (and did not expect to because weight and height are highly correlated), there have been big changes in the standard errors and big changes in the coefficients; for example, the coefficient for height has changed from 0.3397 to -0.0728 , and for age from 2.7278 to 1.5553. Similarly, if another highly correlated variate, BMP, is added, the coefficients and standard errors change unpredictably (Table 30.4). Such changes suggest that multicollinearity is involved.

Table 30.4 Effect of multicollinearity

	Intercept	Age	Height	Weight	BMP	RV	FEV1	FRC
Coefficient	17.8600	2.7178	0.3397					
se		2.9326	0.6900					
Coefficient	64.6550	1.5765	-0.0761	0.8695				
se		3.1436	0.8028	0.8592				
Coefficient	274.63	-3.0832	-0.6985	3.6338	-1.9621			
se		3.6566	0.8008	1.5354	0.9318			
Coefficient	45.3677	4.1756				0.01289		
se		0.3337				0.0784		
Coefficient	-52.4425	4.5416				0.1621	1.5742	
se		1.1944				0.0900	0.6032	
Coefficient	-51.9147	4.5314				0.1633	1.5716	-0.0053
se		1.3052				0.1501	0.6356	0.2958

If this is true then regression involving less highly correlated variables should be more stable. Regressing PE_{\max} on age and RV, and then with age, RV, and FEV1, gives

$$\hat{PE}_{\max} = 45.3677 + 4.1756\text{Age} + 0.01289\text{RV},$$

with adjusted R^2 being 0.3205, and the coefficient standard errors being respectively 1.3337 and 0.0784, and

$$\hat{PE}_{\max} = -52.4425 + 4.5416\text{Age} + 0.1612\text{RV} + 1.5742\text{FEV1},$$

with adjusted R^2 of 0.4625, and standard errors of 1.1944, 0.0900, and 0.6032, respectively. The coefficients remain stable when FRC is included, because it is not as highly correlated with the other variates.

The coefficients and standard errors in the two examples are compared in [Table 30.4](#)

In the upper part of the table the coefficients for weight and height all alter unpredictably as new variates are added to age, whereas in the lower part of the table the coefficients are more stable.

To detect this type of multicollinearity without having to calculate all these combinations, consider the expression for the standard error of a coefficient of X_j when there are more than two X variables:

$$s_{bj} = \sqrt{\frac{MS_{\text{res}}}{\sum (X_j - \bar{X}_j)^2 (1 - R_j^2)}}.$$

This resembles the expression for the standard error if there are only two X variables except that R_j^2 is the multiple correlation coefficient between all the X variables. The expression $(1 - R_j^2)$ is referred to as tolerance. If R_j^2 is high then the tolerance is low, and values below 0.1 demand close inspection of the data to avoid problems due to multicollinearity. If R_j^2 is zero, then the standard error $s_{bj_{\min}}$ is

$$s_{bj_{\min}} = \sqrt{\frac{MS_{\text{res}}}{\sum (X_j - \bar{X}_j)^2}}.$$

Squaring the previous two equations to obtain variances, then when there is no redundant information

$$s_{bj}^2 = \frac{1}{1 - R_j^2} s_{bj_{\min}}^2.$$

The term $\frac{1}{1 - R_j^2}$ is known as the variance inflation factor (VIF) and is the reciprocal of tolerance as defined before. With no redundant information the variance inflation factor is 1. The more redundant information there is the higher the value of VIF. If it is >10 there are almost certain to be problems due to multicollinearity, and any VIF >4 requires serious consideration.

A high VIF indicates redundant information. It would be reasonable to remove variables with $VIF > 10$ and then repeat the regression on the reduced set of variables. This approach, however, does not allow for possible interaction between the explanatory variables, and another way to identify these redundant variables is by *auxiliary regression*. One at a time each X variate is regressed against the remaining $k - 1$. X variates to produce an expression

$$\hat{X}_j = c + b_1X_1 + b_2X_2 + \dots b_{j-1}X_{j-1} + b_{j+1}X_{j+1} + \dots b_kX_k.$$

If none of the variables are correlated with X_j , then none of the coefficients will be substantially different from zero. Any that are correlated will have coefficients that are much different from zero. Furthermore, if the multiple correlation coefficient for any of the auxiliary regressions exceeds that for the primary regression (involving Y), multicollinearity should be suspected (Kleinbaum and Klein, 2010).

If in the example used before we regress any X variate against the remaining X variates, the R^2 and the coefficients (unadjusted) are presented in Table 30.5.

Most of the coefficients are not substantially different from zero, those with $P < 0.05$ are in bold type, and those with $P < 0.00625$ (using a Bonferroni correction) are shown in bold italic type. For the prediction equation of PE_{\max} using all the X variables, the multiple correlation coefficient was 0.6349, much less than some of the results from auxiliary regression. Multicollinearity is a therefore problem with this data set.

All the regressions with coefficients substantially different from zero are associated with VIF values ≥ 8 . There are two clusters of redundant variables. Age, height, weight, and BMP are highly interrelated, and so are RV and FRC. These interrelationships make good physiological sense and suggest that it would be advantageous to eliminate some of the redundant variables before carrying out the final multiple regression analysis.

Determining the Correct Model

Testing whether the individual coefficients are substantially different from zero gives some guidance, as shown in the last two columns of Table 30.2. Unfortunately, these tests do not tell us whether retaining those variates with substantial coefficients is the best choice. The question is to how to tell when an increase in adjusted R^2 is large enough to denote that the factor should be retained in the final model. Several methods have been suggested (see Glantz and Slinker, 2001).

All Subsets Regression

In the cystic fibrosis example, the regression program starts with the 9 X variables. Then it examines the regression for the 9 combinations with one of the X variates removed, then the regression for all the 36 combinations of 7 variables, 84 combinations of 6 variables, and so on. The procedure is computer intensive. More importantly, the investigator faces the decision about which of the many hundreds of regression equations to select. The program used here, JMP, highlights the best fit for each set of variables—9, 8, 7, and so on (Table 30.6).

Table 30.5 Auxiliary regression

R^2	c	Age	Gender	Height	Weight	BMP	FEV1	RV	FRC	TLC	VIF
0.9557	31.90		0.5565	−0.0496	<i>0.3578</i>	<i>−0.1533</i>	<i>−0.1113</i>	0.0107	−0.0392	−0.0384	22.6
0.5710	−101.43	2.1501		0.2495	−0.7814	0.0740	0.7887	−0.0866	0.2463	0.1146	2.3
0.9378	202.45	−1.2572	2.4411		<i>1.5868</i>	−0.5526	−0.5097	0.0496	−0.1836	−0.1364	16.1
0.9816	−99.97	<i>1.8638</i>	−1.3051	<i>0.3258</i>		<i>0.4265</i>	<i>0.2736</i>	−0.0249	0.0895	0.0981	61.0
0.8753	155.35	<i>−2.7796</i>	0.6548	−0.3949	<i>1.4846</i>		−0.1638	−0.0183	−0.0170	−0.1800	8.0
0.8118	141.30	<i>−2.3150</i>	<i>4.7810</i>	−0.4181	<i>1.0930</i>	−0.1880		0.0586	<i>−0.2748</i>	−0.1154	5.3
0.9034	−220.23	6.7744	−13.5236	1.2325	−3.0183	−0.6373	1.7758		<i>1.9248</i>	−0.0650	10.4
0.9393	229.65	−4.7719	7.5830	−0.8812	2.0917	−0.1142	<i>−1.4465</i>	<i>0.3718</i>		0.1241	16.5
0.6024	264.05	−3.8785	4.0312	−0.5430	1.9002	−1.0020	−0.5597	−0.0104	0.1028		2.5

The empty squares show the variates used as the dependent variable. Boxes show variables with high VIF. Bold type— $P < 0.05$; Bold italic type— $P < 0.00625$.

Table 30.6 Small section of results of all subsets analysis

Model	Number	RSquare	RMSE	AICc	BIC
Age,Gender(1-0),Height,Weight,BMP,FEV1,RV,FRC,TLC	9	0.6349	25.5545	262.525	255.625
Age,Height,Weight,BMP,FEV1,RV,FRC,TLC	8	0.6339	24.7794	256.005	252.479
Age,Gender(1-0),Height,Weight,BMP,FEV1,RV,FRC	8	0.6330	24.8093	256.065	252.539
Gender(1-0),Height,Weight,BMP,FEV1,RV,FRC,TLC	8	0.6282	24.9702	256.388	252.863
Age,Gender(1-0),Weight,BMP,FEV1,RV,FRC,TLC	8	0.6270	25.0121	256.472	252.947
Age,Gender(1-0),Height,Weight,BMP,FEV1,RV,TLC	8	0.6259	25.0491	256.546	253.020
Age,Gender(1-0),Height,Weight,BMP,FEV1,FRC,TLC	8	0.6109	25.5451	257.526	254.001
Age,Gender(1-0),Height,Weight,BMP,RV,FRC,TLC	8	0.6098	25.5798	257.594	254.069
Age,Gender(1-0),Height,BMP,FEV1,RV,FRC,TLC	8	0.5825	26.4615	259.289	255.763
Age,Gender(1-0),Height,Weight,FEV1,RV,FRC,TLC	8	0.5727	26.7685	259.865	256.340
Age,Height,Weight,BMP,FEV1,RV,FRC	7	0.6308	24.1394	250.498	249.467
Height,Weight,BMP,FEV1,RV,FRC,TLC	7	0.6282	24.2248	250.674	249.644
Age,Weight,BMP,FEV1,RV,FRC,TLC	7	0.6269	24.2658	250.759	249.729

The highlighted sets give the best predictions for 9, 8, and 7 variables.

In addition, the program plots out the residual mean square for each combination, as shown in Fig. 30.3.

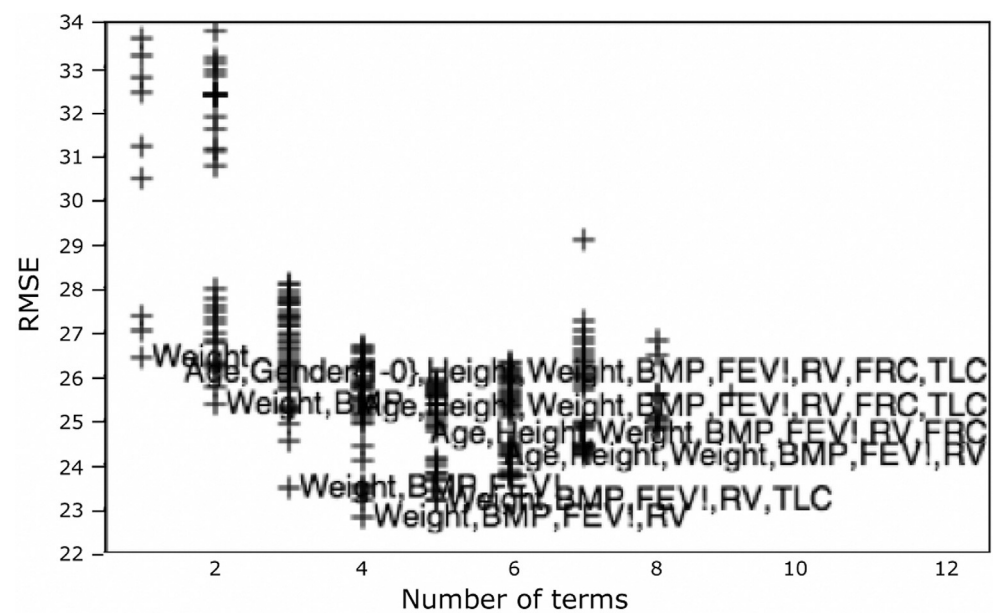


Fig. 30.3 All subsets regression. For each number of variates, the components providing the lowest RMSE are listed.

The investigator is not required to pick the lowest of these values, and a decision about which is the best should not be made from this test. What it does do is to provide a short list of acceptable candidates for the investigator to decide which combinations make the best sense.

To shorten the calculations, certain sequential techniques have been developed. These techniques have in common successive addition or removal of variates until some specified criterion is met. In statistical programs the final results are usually given immediately, but it is possible to follow step by step, thus gaining insight into the process.

Forward Selection

The forward selection method begins with the variate with the largest bivariate regression sum of squares between Y and X . Then the X variate with the next largest regression sum of squares is added. This addition almost always increases R^2 , increases MS_{reg} , and decreases MS_{res} . Then test the incremental effect. Some programs will do calculations for each of the possible second variates, add the variate that achieves the highest F ratio, then evaluate the effect of adding in each of the remaining variates as the third variate, put in the variate with the highest F ratio, and so on, until the addition of the next variate produces an F ratio below some critical value, often termed F_{enter} or F_{in} . Frequently F_{in} is set at 4, but this is not a universal default. Other programs may examine the effect of adding the first of the remaining variates, and if this achieves a F_{in} above the critical value leave that variate in and proceed to the next.

As an example, look at the successive steps applied to the cystic fibrosis data.

After the first step (Table 30.7)

Table 30.7 First forward step

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC	
16004.693	23	26.379087	0.4035	0.3776	3.5083272	2	239.6338	242.1476	
Current Estimates									
Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"		
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	63.568172	1	0	0.000	1		
<input type="checkbox"/>	<input type="checkbox"/>	Age	0	1	213.7819	0.298	0.59073		
<input type="checkbox"/>	<input type="checkbox"/>	Gender{1-0}	0	1	788.7532	1.140	0.29714		
<input type="checkbox"/>	<input type="checkbox"/>	Height	0	1	36.62306	0.050	0.82434		
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Weight	1.18686368	1	10827.95	15.561	0.00065		
<input type="checkbox"/>	<input type="checkbox"/>	BMP	0	1	1895.191	2.955	0.09965		
<input type="checkbox"/>	<input type="checkbox"/>	FEV1	0	1	954.2458	1.395	0.2502		
<input type="checkbox"/>	<input type="checkbox"/>	RV	0	1	270.9908	0.379	0.5445		
<input type="checkbox"/>	<input type="checkbox"/>	FRC	0	1	12.22795	0.017	0.89798		
<input type="checkbox"/>	<input type="checkbox"/>	TLC	0	1	225.0707	0.314	0.58102		
Step History									
Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC
1	Weight	Entered	0.0006	10827.95	0.4035	3.5083	2	239.634	242.148

Because the bivariate F ratio is high, Weight has been selected and is associated with the highest regression sum of squares (SS)—10827.95. Then the program adds in BMP that has the next highest regression sum of squares to give Table 30.8.

Table 30.8 Second step

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC	
14109.501	22	25.324711	0.4742	0.4264	2.6061802	3	239.3401	242.2156	
Current Estimates									
Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"		
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	124.486719	1	0	0.000	1		
<input type="checkbox"/>	<input type="checkbox"/>	Age	0	1	666.7947	1.042	0.31905		
<input type="checkbox"/>	<input type="checkbox"/>	Gender{1-0}	0	1	824.6327	1.304	0.26642		
<input type="checkbox"/>	<input type="checkbox"/>	Height	0	1	679.7208	1.063	0.31429		
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Weight	1.63666146	1	11309.67	17.634	0.00037		
<input type="checkbox"/>	<input checked="" type="checkbox"/>	BMP	-0.9987454	1	1895.191	2.955	0.09965		
<input type="checkbox"/>	<input type="checkbox"/>	FEV1	0	1	2552.977	4.639	0.04302		
<input type="checkbox"/>	<input type="checkbox"/>	RV	0	1	17.88261	0.027	0.87188		
<input type="checkbox"/>	<input type="checkbox"/>	FRC	0	1	27.82682	0.041	0.84054		
<input type="checkbox"/>	<input type="checkbox"/>	TLC	0	1	91.90814	0.138	0.71431		
Step History									
Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC
1	Weight	Entered	0.0006	10827.95	0.4035	3.5083	2	239.634	242.148
2	BMP	Entered	0.0997	1895.191	0.4742	2.6062	3	239.34	242.216

The adjusted R^2 has increased, and C_p , AIC, and BIC are lower. The F ratio is borderline, but BMP has increased adjusted R^2 from 0.3776 to 0.4264 and is retained.

To see what effect BMP has made, perform an ANOVA (Table 30.9).

Table 30.9 SE—Standard error, SS—Sum of squares, MS—mean square

Weight					Weight and BMP				
R^2		0.4035			0.4742				
R^2 adj		0.3776			0.4263				
RMSE		26.48			25.32				
Mean		109.12			109.12				
N		25			25				
ANOVA									
Source	Df	SS	MS	F	Source	Df	SS	MS	F
Model	1	10827.95	10827.95	15.58	Model	2	12723.14	6361.57	9.92
Error	23	16004.69	695.9	$P=0.0006$	Error	22	14109.50	641.34	$P=0.0008$
Total	24	26832.64			Total	24	26382.64		
Parameter estimates									
Term	Estimate	SE	t	P	Term	Estimate	SE	t	P
Intercept	63.57	12.70	5.01	<0.0001	Intercept	124.49	37.48	3.32	0.0031
Weight	1.19	0.30	3.94	0.0006	Weight	1.64	0.40	4.20	0.0004
BMP	—				BMP	−0.0087	0.58	−1.72	0.0997

Comparing the models with 1 and 2 independent variables (left vs right panels), including two independent variables increases the model SS from 10827.947 to 12723.139, and reduces the residual SS (labeled Error) from 16004.693 to 14109.501. To determine if the reduction in error SS is due to chance, perform a one-way ANOVA, and test the reduction in the residual SS by relating it to the smaller of the two residual mean squares (Table 30.10).

Table 30.10 ANOVA after second step

Source of variation	SS	Df	Mean square	F
Weight (W) and BMP	12723.14	2		
Weight alone	10827.95	1		
Difference	1895.19	1	1895.19	2.96
SS _{res} due to W, BMP	14109.50	22	641.34	

The reduction in SS_{res} is small because $F < 4$, but nevertheless the program retains BMP because of substantial decreases of C_p , AIC, and BIC. It is also too soon to stop the process. The difference of 1895.19 is the same as the SS due to BMP in Table 30.8.

Next FEV1 is added, with a further increase in adjusted R^2 and decrease of C_p (Table 30.11).

Table 30.11 Third step

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC	
11556.524	21	23.458702	0.5693	0.5078	0.6967521	4	237.5081	240.4445	
Current Estimates									
Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"		
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	126.007181	1	0	0.000	1		
<input type="checkbox"/>	<input type="checkbox"/>	Age	0	1	570.6859	1.039	0.32024		
<input type="checkbox"/>	<input type="checkbox"/>	Gender(1-0)	0	1	3.329585	0.006	0.94024		
<input type="checkbox"/>	<input type="checkbox"/>	Height	0	1	517.2305	0.937	0.34459		
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Weight	1.53296348	1	9748.56	17.715	0.00039		
<input type="checkbox"/>	<input checked="" type="checkbox"/>	BMP	-1.4591073	1	3493.923	6.349	0.01991		
<input type="checkbox"/>	<input checked="" type="checkbox"/>	FEV!	1.10877314	1	2552.977	4.639	0.04302		
<input type="checkbox"/>	<input type="checkbox"/>	RV	0	1	1174.58	2.263	0.14815		
<input type="checkbox"/>	<input type="checkbox"/>	FRC	0	1	815.3789	1.518	0.23218		
<input type="checkbox"/>	<input type="checkbox"/>	TLC	0	1	645.6409	1.183	0.28959		
Step History									
Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC
1	Weight	Entered	0.0006	10827.95	0.4035	3.5083	2	239.634	242.148
2	BMP	Entered	0.0997	1895.191	0.4742	2.6062	3	239.34	242.216
3	FEV!	Entered	0.0430	2552.977	0.5693	0.6968	4	237.508	240.445

The F ratio is adequate, being >4 . An ANOVA gives (Tables 30.12 and 30.13).

Table 30.12 Three variables

Weight, BMP and FEV1

R^2		0.5693		
R^2 adj		0.5078		
RMSE		23.45		
Mean		109.12		
N		25		
Source	Df	SS	MS	F
Model	3	15976.12	5002.04	9.25
Error	21	11556.52	550.31	$P=0.0004$
Total	24	26832.64		
Term	Estimate	SE	t	P
Intercept	126.01	34.72	3.63	0.0016
Weight	1.53	0.36	4.21	0.0004
BMP	-1.46	0.58	-2.52	0.0199
FEV1	1.11	0.51	2.15	0.0430

Table 30.13 ANOVA after third step

Source	SS	Df	MS	F
With weight, BMP, and FEV1	15276.12	3		
Weight and BMP alone	12723.14	2		
Difference	2552.98	1	2552.98	4.64
SS_{res} due to W, BMP	11556.52	21	550.31	

The decrease in SS_{res} is retained because the $F > F_{\text{in}}$.

The variate with the next highest regression sum of squares is RV, and including this gives Table 30.14.

No added advantage is gained by adding other variates, so the process terminates (Tables 30.15 and 30.16).

The reduction in SS_{res} after adding TLC is negligible

The full model is

$$\hat{P}_{\text{max}} = 63.8243 + 1.7438 \text{Weight} + 1.5461 \text{FEV1} + 0.1252 \text{RV} - 1.3702 \text{BMP},$$

with an adjusted R^2 of 0.5357. This is the same set as the all subsets analysis selected as having the lowest RMSE (Fig. 30.1). These are not the same variates that would have been selected by inspecting the individual coefficients in isolation in Table 30.2. Testing the residuals showed satisfactory distribution.

Table 30.14 Fourth step

Current Estimates								
Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"	
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	63.8242644	1	0	0.000	1	
<input type="checkbox"/>	<input type="checkbox"/>	Age	0	1	169.2978	0.315	0.58121	
<input type="checkbox"/>	<input type="checkbox"/>	Gender{1-0}	0	1	1.665115	0.003	0.95655	
<input type="checkbox"/>	<input type="checkbox"/>	Height	0	1	173.641	0.323	0.57636	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Weight	1.74379041	1	10902.83	21.003	0.00018	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	BMP	-1.3702183	1	3047.545	5.871	0.02501	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	FEV!	1.54608863	1	3709.674	7.146	0.01461	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	RV	0.12519417	1	1174.58	2.263	0.14815	
<input type="checkbox"/>	<input type="checkbox"/>	FRC	0	1	0.483083	0.001	0.97659	
<input type="checkbox"/>	<input type="checkbox"/>	TLC	0	1	190.5608	0.355	0.55818	

Step History									
Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC
1	Weight	Entered	0.0006	10827.95	0.4035	3.5083	2	239.634	242.148
2	BMP	Entered	0.0997	1895.191	0.4742	2.6062	3	239.34	242.216
3	FEV!	Entered	0.0430	2552.977	0.5693	0.6968	4	237.508	240.445
4	RV	Entered	0.1481	1174.58	0.6131	0.8981	5	238.337	240.984

Table 30.15 After entering TLC, there is an increase in C_p , AIC, and BIC, so that no advantage had been obtained

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
10191.384	19	23.160065	0.6202	0.5202	2.6062834	6	241.7957	243.7396

Current Estimates								
Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"	
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	44.3404922	1	0	0.000	1	
<input type="checkbox"/>	<input type="checkbox"/>	Age	0	1	56.05601	0.100	0.75599	
<input type="checkbox"/>	<input type="checkbox"/>	Gender{1-0}	0	1	25.36652	0.045	0.83454	
<input type="checkbox"/>	<input type="checkbox"/>	Height	0	1	84.01574	0.150	0.70343	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Weight	1.7624741	1	11065.04	20.629	0.00022	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	BMP	-1.3802475	1	3089.674	5.760	0.0268	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	FEV!	1.57582816	1	3826.201	7.133	0.01511	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	RV	0.10532008	1	719.4996	1.341	0.26114	
<input type="checkbox"/>	<input type="checkbox"/>	FRC	0	1	72.63393	0.129	0.72344	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	TLC	0.20693936	1	190.5608	0.355	0.55818	

Step History									
Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC
1	Weight	Entered	0.0006	10827.95	0.4035	3.5083	2	239.634	242.148
2	BMP	Entered	0.0997	1895.191	0.4742	2.6062	3	239.34	242.216
3	FEV!	Entered	0.0430	2552.977	0.5693	0.6968	4	237.508	240.445
4	RV	Entered	0.1481	1174.58	0.6131	0.8981	5	238.337	240.984
5	TLC	Entered	0.5582	190.5608	0.6202	2.6063	6	241.796	243.74

Table 30.16 ANOVA

Source	SS	Df	MS	F
With weight, BMP, and FEV1, RV and TLC	16641.26	5		
Weight BMP, FEV1 and RV	16450.70	4		
Difference	190.56	1	190.56	0.37
SS _{res} due to W, BMP, FEV1, RV	10381.95	20	519.10	

Backward Elimination

The backward selection program starts with all the variates included in the regression (Table 30.17). Then the effect of removing a variable on MS_{res} is assessed for each variable, and the variable with the least effect on increasing MS_{res} is removed if it does not increase the F ratio for removal, F_{out} . As for the forward selection process, F_{out} is often but not always set at 4. The process continues until removal causes a substantial change in MS_{res} , when that variate is left in and no further removals are done. Eventually the reduced model that provides the regression equation is

$$\hat{PE}_{max} = 126.3336 + 1.5365 \text{Weight} + 1.10863 \text{FEV1} - 1.4653 \text{BMP}.$$

The adjusted R^2 was 0.5086.

Comparing the forward and backward selection methods, RV was selected only in the forward method, and as a result there are some differences in the coefficients of the variates. Which of these procedures should we accept? That answer depends on the purpose of the study. If the objective was to obtain the best possible prediction of PE_{max} , then choose the model with the higher adjusted R^2 , namely, the model that includes RV (Adjusted R^2 0.5369 vs 0.5086). Remember, though, that another random sample from the same population might give a slightly different prediction. If the objective is to understand what factors are important in predicting PE_{max} , then statistics will not answer the question. For that, you need physiological and experimental insights. If the inclusion or exclusion of RV is potentially important for understanding the process, further experiments will be needed.

Both of these techniques work reasonably well but have a tendency to stop too soon. Neither of them guards against revealing redundancy when a subsequent variable is added or subtracted.

Nonlinear Regression: Polynomial Regression

In Chapter 27, Figs. 27.12–27.15, some data from a study by Jamal et al. (2001) were used to show the difference between a linear and a quadratic fit. The curved regression line seems to be a better fit with more evenly scattered residuals, but the question is if this could have occurred by chance or whether the curvilinear relationship is better. We also need to ask if higher powers would produce better fits. Fig. 30.4 shows fitting up to 5

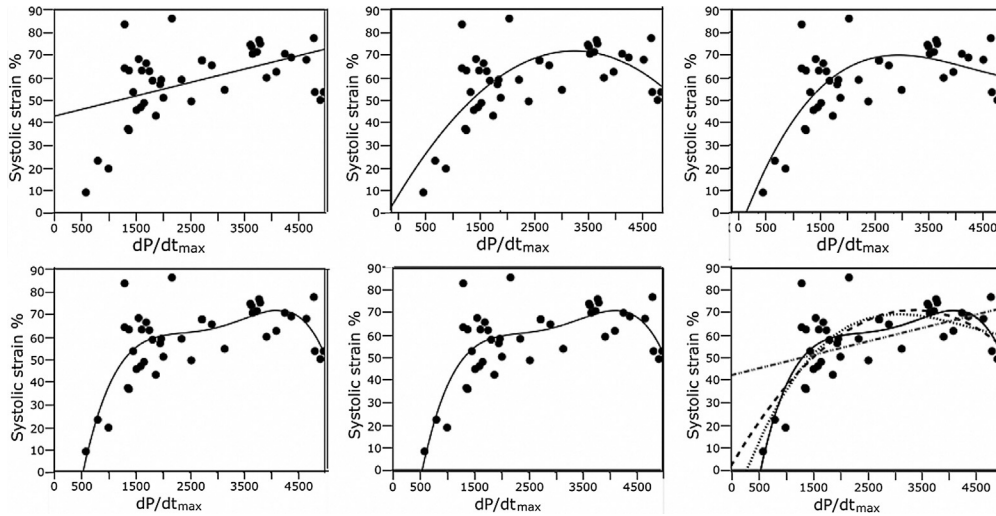


Fig. 30.4 Data taken from [Jamal et al. \(2001\)](#). Upper 3 panels show linear, second-order (quadratic), and third-order (cubic) fits. Lower 3 panels show fourth-order (quartic) and fifth-order (quintic) fits, and finally all power functions superimposed. In this last panel the second-order fit is the *dashed curve*, third-order fit is the *dotted curve*, and the fourth- and fifth-order fits are superimposed.

Table 30.17 Power fits to data from linear (column 1) to quintic (column 5)

Power	1	2	3	4	5
c	42.37	2.01	-18.79	-85.40	-91.16
b_1X	0.005945	0.04097	0.06970	0.21089	0.22706
b_2X^2		$-6.0479e^{-6}$	-0.0000174	-0.0001134	-0.000129
b_3X^3			$+1.3236e^{-9}$	$+2.6682e^{-9}$	$+3.3393e^{-8}$
b_4X^4				$-2.268e^{-12}$	$-3.3578e^{-12}$
b_5X^5					$+9.841e^{-17}$
R^2	0.2264	0.4783	0.4994	0.5657	0.5658
$S_{Y.X}$	14.4559	12.0182	11.9231	11.2511	11.4005
SS_{reg}	2506.92	5297.30	5530.49	6264.47	6265.90
Df	1	2	3	4	5
MS_{reg}	2506.92	2648.65	1843.50	1566.12	1253.18
SS_{res}	8567.88	5777.50	5544.31	4810	4808.90
Df	41	40	39	38	37
MS_{res}	208.97	144.44	142.16	126.59	129.97
SS_{total}	11074.80	11074.80	11074.80	11074.80	11074.80

power functions, and [Table 30.22](#) presents some of the regression results for the different power functions ([Table 30.17](#)).

The results of analyzing the different functions are presented in [Table 30.17](#).

Before analyzing these data in detail, examine the general effects of increasing power functions on the results.

1. There are changes in the intercept c and the linear coefficient b_1 .
2. The second-order coefficients are all 1,000-fold smaller than the first-order coefficients, and the third-, fourth-, and fifth-order coefficients are smaller still. This does not mean that they can be neglected; these coefficients are multiplied by large numbers.
3. R^2 , a measure of fit, increased substantially from linear to quartic, and then did not improve with the quintic fit. This change went hand in hand with an increase in SS_{reg} (sum of squares due to regression) because $R^2 = SS_{\text{reg}}/SS_{\text{total}}$.
4. Increasing the power function reduced the standard deviation from regression S_{YX} and is to be expected because the standard deviation from regression is what is left over when the variability due to regression is accounted for. The standard deviation from regression did not improve with the quintic fit.

Determine if the changes are consistent with the null hypothesis by ANOVA. First determine if the improved results obtained by a quadratic fit over a linear fit allow us to reject the null hypothesis of no improvement with the quadratic function (Table 30.18).

Table 30.18 ANOVA for quadratic vs linear fit

Source	SS	Df	MS	F
Quadratic	5297.30	2		
Linear regression	2506.92	1		
Quadratic after linear	$5297.3 - 2506.92 = 2790.38$	1	2790.38	19.31
$SS_{\text{res}} (X^2)$	5777.50	40	144.44	$P < 0.0001$

The change in SS_{reg} due to the higher power gives one mean square, and that is compared with the minimal residual sum of squares that comes from the higher power fit. The F ratio is very high, so reject the null hypothesis that these data fit a linear model. Then perform the next ANOVA to compare the quadratic with the cubic fits (Table 30.19).

Table 30.19 ANOVA to compare cubic with quartic

Source	SS	Df	MS	F
Cubic	5530.49	3		
Quadratic	5297.30	2		
Cubic after quadratic	233.19	1	233.19	1.64
$SS_{\text{res}} (X^3)$	5544.31	39	142.16	$P = 0.2079$

The improvement is minor, so that a cubic fit does not improve on a quadratic fit. Nevertheless, proceed with the quartic comparison (Table 30.20).

Table 30.20 Comparison of quartic and cubic functions

Source	SS	Df	MS	F
Quartic	6264.47	4		
Cubic	5530.49	3		
Quartic after cubic	733.57	1	733.57	5.80
$SS_{\text{res}}(X^4)$	4810.0	38	126.58	$P=0.021$

The quartic is substantially better than the cubic, suggesting that the quartic is indeed better fitting.

If the purpose of the regression analysis was to determine the best fit, then a quartic regression is the best choice. As a model for understanding the subject, however, quartic regression may make less physiological sense than quadratic regression.

More information about curvilinear regression can be found in the book by [Hamilton \(1992\)](#).

If the data appear to fit an exponential model

$$\hat{Y} = \alpha\beta^x e,$$

deal with this by realizing that this is an intrinsically linear model because taking logarithms of both sides gives

$$\log \hat{Y} = \log \alpha + X \log \beta + \log e.$$

Logarithms to base 10 or base e are equally effective.

Splines

Some data cannot be fitted by any simple power function. Interpolating splines, in particular, cubic splines, can fit a smoothed curve connecting each data point. A cubic spline is a piecewise cubic polynomial whose coefficients change in different portions of the data set according to certain rules. As a result, a complex smooth curve (a Bezier curve) can be fitted to the points. The procedure is best implemented by a computer program, but understandable online tutorials are available. (Lancaster, [McKinley and Levine, n.d.](#)) Online programs for creating these splines are found at <http://www.akiti.ca/CubicSpline.html>. A simple applet for demonstrating interpolating splines can be found at <http://www.math.ucla.edu/~baker/java/hoefer/Spline.htm>.

Sometimes a smoothing nonparametric function called a LOWESS (or LOESS) function is used to produce a moving average in a cloud of points. This is best done with a commercial program but can also be done in Excel by using the directions given in <http://peltiertech.com/loess-smoothing-in-excel/>.

Correcting for Multicollinearity

If it is apparent that two variables are virtually measuring the same thing, one of them should be eliminated. If, for example, cardiac output is measured after infusing different volumes of blood, there is little to be gained by regressing cardiac output against total blood volume and infused volume because these volumes will be highly correlated. If the variables cannot be eliminated, a procedure known as centering is used. To do this, the mean of each X variate is subtracted from its components to give $X_i - \bar{X}_i$, and the regression is calculated with these transformed variates. Sometimes this deviation is divided by the standard deviation to give a z score. Both of these transformations give a mean of zero that helps to avoid multicollinearity problems. A particularly clear explanation of why this is useful is presented by [Glantz and Slinker \(2001\)](#). In addition, the transformation may be done in such a way as to simplify the numbers, and this helps the computation because even high-powered computers may have difficulties with matrix inversion (used in the computation) of large numbers.

Multicollinearity is very likely to be present in polynomial regression of the form

$$\hat{Y}_i = c + b_1 X_1 + b_2 X_1^2 + b_3 X_1^3 + b_k X_1^k,$$

and also when interactions are present as in

$$\hat{Y}_i = c + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2.$$

With centered values for the calculations presented in relation to [Fig. 30.1](#) there is no change to the fitted curve, the residuals, and the correlation coefficient, but there are differences of the X coefficients. This can be seen by comparing the cubic polynomial with and without centering ([Fig. 30.5](#)).

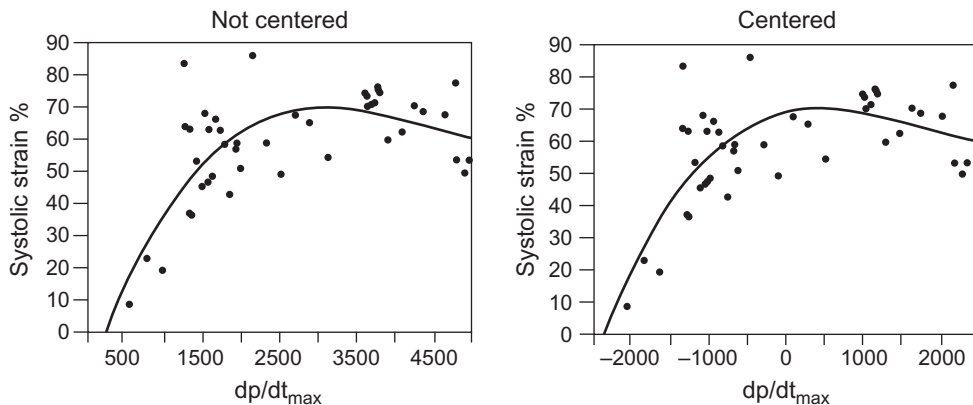


Fig. 30.5 Cubic regression without (left panel) and with (right panel) centering.

The accompanying Table 30.21 gives the basic data for these curves.

Table 30.21 ANOVA for centered vs noncentered cubic curves

Uncentered					Centered				
R^2	0.4994				0.4994				
R^2 adj	0.4609				0.4609				
RMSE	11.92				11.92				
Mean	58.1				58.1				
N	43				43				

ANOVA									
Source	Df	SS	MS	F	Source	Df	SS	MS	F
Model	3	5530.49	1843.5	12.97	Model	3	5530.49	1843.5	12.97
Error	39	5544.31	142.16	$P = < 0.0001$	Error	39	5544.31	142	$P = < 0.0001$
Total	42	11074.80			Total	42	11074.80		

Parameter estimates									
Term	Estimate	SE	t	P	Term	Estimate	SE	t	P
Intercept	-18.79	19.08	-0.98	0.3309	Intercept	68.47	2.90	23.61	<0.00001
dP/dt_{\max}	0.070	0.024	2.93	0.0057	dP/dt_{\max}	0.0055	0.0032	4.20	0.00015

Only the parameter estimates have changed. The regression equations are shown in Table 30.22. Systolic strain % =

Table 30.22 Centered vs uncentered regression equations

	Intercept	dP/dt_{\max}	dP/dt_{\max}^2	dP/dt_{\max}^3
Uncentered	-18.79	+0.070	-0.0000174	+1.32e ⁻⁹
Centered	68.37	+0.0054	-6.89e ⁻⁶	+1.32e ⁻⁹

The fitted curves are identical as are the values for R^2 , MS_{res} , mean response, and the ANOVA. Differences are noted in the intercepts and coefficients, and particularly in the much wider standard errors of the coefficients without centering. (The intercept occurs when X is zero, and so is not at the left vertical axis line for the centered graph.) Centering also emphasizes the central part of the distribution where the greatest accuracy occurs (Glantz and Slinker, 2001).

Online programs for polynomial regression can be found at <http://www.xuru.org/rt/NLR.asp>, http://web.cecs.pdx.edu/~eas199/A/notes/14/polynomial_curve_fit_notes.pdf (an Excel program), and <http://polynomialregression.drque.net/online.php>.

Other ways of avoiding multicollinearity are described by Katz (2006) whose book on multivariate analysis is easy to read and instructive. It is wise to omit a variable if it seems to have no plausible biological function (unless it is being tested for this specifically), or if it is an intermediate explanatory variable. For example, in a study of explanatory variables

for essential hypertension, it is unnecessary to include pro-renin, renin, and angiotensin. Angiotensin alone is a useful predictor of response. A second technique is to use a categorical scale with an “and/or” response. Rather than include renin and angiotensin as ratio numbers, specify that a renin concentration in excess of X_1 units or an angiotensin concentration in excess of Y units counts as a positive response. A third technique is to construct a multivariable scale in which closely related questions are coded to fit 1– N groups (the same for each variable). For example, if in predicting a coronary artery calcium score by using BMI, insulin resistance, and fasting blood glucose, all of which are related, assign each factor a score of 1–5 (e.g., BMI <20, 20–25, 26–30, 31–35, >35), and then use the cumulative score as a single explanatory factor. Finally, if the variables are unrelated, it is still possible to define a combined score; as examples see Apgar score in Table 3.3, or create a score for degree of exercise, weight, use of beta blockers, and a measure of psychological stress as a combined predictor of myocardial infarction.

Principles of Nonlinear Regression

Nonlinear fits are frequently used in biochemical and pharmacological systems that examine rate constants, feedback controls, cooperativity, and other complex interactions. They may be simple exponential curves, or else equations such as those defining Michaelis-Menten kinetics:

$$v = \frac{V_{\max} S}{K_m + S}.$$

Sometimes complex equations with partial derivatives are involved.

Some simple curved XY relationships can be transformed to make them linear. A polynomial smoothing spline was used by Sherrill et al. (1989) to model pulmonary and somatic growth functions.

The principle of least squares is used for these nonlinear fits, and the procedure requires several iterations. At first, guesses are made about the values of the different parameters. The program then adjusts these estimates, and if the residual sum of squares is diminished then a second iteration with new values is done. The procedure continues until negligible improvement occurs in the residual sums of squares. There are a variety of considerations and techniques used in the fitting process, and many of these are described by Hamilton (1992) and Motulsky and Ransnas (1987) (Motulsky was the originator of the statistics program Prism that is particularly well adapted to nonlinear solutions) and by Garfinkel and Fegley (1984). Online programs can be found at <http://statpages.org/nonlin.html>, <http://www.colby.edu/chemistry/PChem/scripts/lsfitpl.html>, and or indicator <http://www.xuru.org/rt/NLR.asp>. Although these programs perform complex calculations, they should be used with care. An excellent online tutorial is provided at <http://graphpad.com/curvefit/>.

Dummy or Indicator Variables

It is possible to introduce a qualitative (nominal or ordinal) factor into regression. This dummy or indicator variable allows us to identify different categories of a qualitative variable.

Coding with dummy variables can be performed by reference cell coding (more common) or by effect cell coding. In reference cell coding, commonly used for comparisons with a designated control group, assign zeros to the control group and combinations of 0 and 1 to the other groups. With effect cell coding, the various groups are compared with their average. If there are two categories, designate one of them (which one is immaterial) as 1, and the other as -1 for effect coding, or as 0 (basal state) and 1 for reference coding, and then perform multiple regression:

$$\hat{Y}_i = c + b_1 X_{1i} + b_2 X_{2i},$$

where X_1 is the independent variable and X_2 is the indicator variable. (Different programs have different defaults, JMP being one of the few that uses effect coding as the default. JMP treats 1 and 0 as if they were 1 and -1 .)

The models imply that the slope of the relationship is the same in both groups that differ only in their mean value or intercept, but the two data sets might also differ by mean and slope, that is, that there is interaction. This is handled by using as a model

$$\hat{Y}_i = c + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2.$$

Considering variable X_1 , for which $X_2 = 0$,

$$\hat{Y}_i = c + b_1 X_1.$$

Considering variable X_2 ,

$$\hat{Y}_i = c + b_1 X_1 + b_2(1) + b_3 X_1(1) = (c + b_2) + (b_1 + b_3) X_1.$$

This shows that the change from group 1 to group 2 in intercept is reflected by the parameter b_2 , and in slope by the parameter b_3 . There is a nice description with examples at https://www.sagepub.com/sites/default/files/upm-binaries/21120_Chapter_7.pdf.

There is no reason to restrict the number of groups to 2. For example, compare the relationship between age and cholesterol in four states by setting up 3 more variables (Table 30.23).

Table 30.23 Indicator variables for four states

State	X_2	X_3	X_4
Iowa	0	0	0
Nebraska	1	0	0
California	0	1	0
Texas	0	0	1

Then the prediction equation is

$$\hat{Y}_i = c + b_1 \text{Age} + b_2 X_2 + b_3 X_3 + b_4 X_4 + \varepsilon_i.$$

For Iowa, the equation becomes $\hat{Y}_i = c + b_1 \text{Age} + \varepsilon_i$, and for California it becomes $\hat{Y}_i = c + b_1 \text{Age} + b_3 X_3 + \varepsilon_i$.

As another example, relate blood pressure to the dose of norepinephrine in control patients, diabetic patients, and patients who have had a myocardial infarct. Then code the control group as (0,0), the diabetic group as (1,0), and the infarct group as (0,1). The equation is then

$$\hat{Y}_i = c + b_1 \text{Dose} + b_2 \text{Diabetic} + b_3 \text{Infarct}.$$

The intercept alone, c , represents the blood pressure in the absence of agonist, diabetes, and infarction.

If the agonist is infused, then if the subject is a control, the relationship is $\hat{Y}_i = c + b_1 \text{Dose}$, because both diabetics and infarct patients are coded zero.

If the patient is a diabetic, the equation is $\hat{Y}_i = c + b_1 \text{Dose} + b_2 \text{Diabetic}$ and the difference between control and diabetic subjects at any dose is b_2 .

If the patient has had an infarct, the equation is $\hat{Y}_i = c + b_1 \text{Dose} + b_3 \text{Infarct}$, and the difference between control and infarct subjects at any dose is b_3 .

Effect cell coding, on the other hand, allows comparison of the mean of a group from the overall mean of all the groups. To achieve this, code as presented in Table 30.24.

Table 30.24 Effect cell coding model

Group code	Group 2	Group 3	Group 4
Iowa (I)	-1	-1	-1
Nebraska (N)	1	0	0
California (C)	0	1	0
Texas (T)	0	0	1

When this is done, the regression equation is $\hat{Y}_i = c + b_1 I + b_2 N + b_3 C + b_4 T$. Because of effect coding, the intercept c is the overall mean value of the measured variable. The coefficient b_1 shows the difference between the mean for Iowa and the overall mean, coefficient b_2 shows the difference between the mean for Nebraska and the overall mean, and so on.

An excellent discussion of dummy (categorical) variables is presented by Katz (2006) and by te Grotenhuis and Thijs at <https://arxiv.org/pdf/1511.05728.pdf>.

The examples discussed before all had a single observation of the dependent variable for each subject and are referred to as *between-subjects* designs. It is, however, common to use a number of subjects and then make a series of measurements on each of them; this is

referred to as a *within-subjects* design. The repeated measurements may be categorical, for example, measurement of left ventricular volume 1. Before anesthesia, 2. After anesthesia with closed chest, 3. Open chest, closed pericardium, and 4. Open chest, open pericardium. Alternatively, the repeated measurements may be serial times, doses, pressures, and so on. If repeated measurements are done, it would be a mistake to pool them all without considering how homogeneous they are. For example, examine Fig. 30.6.

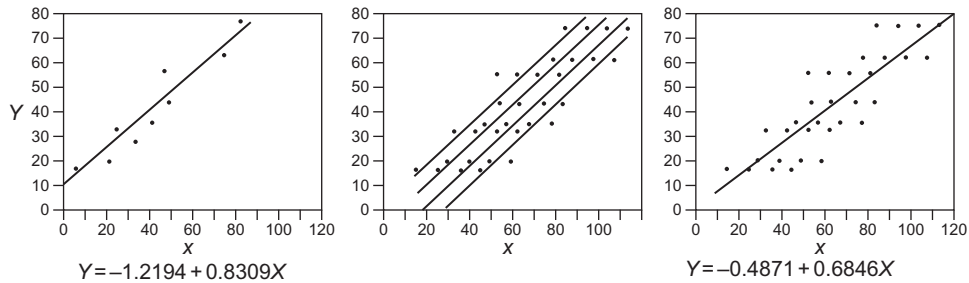


Fig. 30.6 Diagram to show problem of repeated measures.

The left panel shows a linear XY relationship in a single subject with 8 measurements. (It resembles the regression obtained when relating arterial oxygen tension measured directly in blood to arterial oxygen tension measured with a skin electrode.) The center panel shows three other subjects with identical slopes but different intercepts. If differences among subjects were ignored and all 32 measurements were pooled (right panel), the pooled regression line could have a substantially different slope, as shown by the regression equations.

In repeated measures, the analysis must allow for differences among subjects. [Glantz and Slinker \(2001\)](#) recommend dummy variables to account for differences among the subjects. This can be done by defining two dummy variables:

$$D_1 \begin{cases} 1 & \text{if subject 1} \\ 0 & \text{if subject 2} \\ -1 & \text{if subject 3} \end{cases} \quad D_2 \begin{cases} 0 & \text{if subject 1} \\ 1 & \text{if subject 2} \\ -1 & \text{if subject 3} \end{cases}$$

The regression model then becomes

$$\hat{Y}_i = c + b_1 X_1 + b_2 D_1 + b_3 D_2.$$

Then the regression equations are, respectively

$$\begin{aligned} \hat{Y}_i &= c + b_1 X_1 + b_2 \cdot 1 + b_3 \cdot 0, \text{ (subject 1)} \\ \hat{Y}_i &= c + b_1 X_1 + b_2 \cdot 0 + b_3 \cdot 1, \text{ (subject 2) and} \\ \hat{Y}_i &= c + b_1 X_1 + b_2 \cdot -1 + b_3 \cdot -1. \text{ (subject 3).} \end{aligned}$$

For each subject designated by the dummy variable, the appropriate number is inserted into the regression formula. With this procedure, the slope for each subject is shown to be zero, and there are differences only in the intercepts.

Multivariate Analysis

This is not the same as multiple regression, in which a single dependent variable is related to a set of independent explanatory variables. In contrast, multivariate analysis, also termed MANOVA, has two or more dependent variables that may not themselves be independent of each other.

As an example, consider a study, described by [Everitt and Hay \(1992\)](#) to determine whether postpartum depression is associated with delayed intellectual development of the infants. Two groups are examined: mothers without postpartum depression and mothers with depression, and at several months of age a variety of psychological tests are performed. As part of the study, comparisons are made between three components of IQ scores: perceptual, verbal, and quantitative. It certainly would be possible to perform separate *t*-tests on each of these components, but this not only raises problems due to multiplicity but it also loses some information because these components are interrelated. In addition, doing single variable comparisons results in a loss of power. A simple ANOVA is not appropriate, because the components of a factor may be correlated.

Another example might be testing the effect of several drugs on systolic and diastolic blood pressures.

The correct analysis is to use Hotelling's T^2 test. This test requires calculation of a variance-covariance matrix and is implemented in most standard statistics programs. It can also be performed by the free program BrightStats. <http://www.brightstat.com/>. and in Excel at <http://www.real-statistics.com/multivariate-statistics/hotellings-t-square-statistic/one-sample-hotellings-t-square/> (you will need the Real Statistics Resource Pack).

Multivariate methods are used in discriminant analysis and in logistic regression.

Longitudinal Regression

Regression studies in which several members of a group are each followed for many time periods are considered as longitudinal studies. Unlike the repeated measures described before in which the variability among subjects is allowed for in examining the relationship of the dependent to the independent variables, in longitudinal studies the whole set of outcome variables is of interest; an example is the assessment of growth curves or glucose tolerance curves in two or more experimental groups. Frequently the mean values at each time period these data are joined to form average curves, and multiple *t*-tests are calculated at each time point. Not only must these tests be corrected for

multiplicity, but in fact the curve may be misleading because the average curve may not reflect any individual curve. In addition, successive measurements in each subject are likely to be highly correlated, some being hypo- and others hyper-responders. As a consequence, the successive t -tests are not independent, and these time-varying covariances involving the individuals' responses make the usual multivariate methods unsuitable.

If formal analyses are indicated, there are many proposed solutions, one of the most general being that developed by Zeger and Liang. Their method is based on estimating a common correlation between all repeated measurements on each subject. Dummy variables to indicate a specific group or treatment can be incorporated. The method is robust, can accommodate linear or curvilinear relations, but requires specialized general estimating equations.

REFERENCES

- Altman, D.G., 1992. *Practical Statistics for Medical Research*. p. 611.
- Chambers, J.M., Cleveland, W.S., Kleiner, B., et al., 1983. *Graphical Methods for Data Analysis*. Wadsworth, Belmont, CA.
- Chatfield, C., 1988. *Problem Solving. A Statistician's Guide*. Chapman & Hall, London.
- Edwards, A.L., 1984. *An Introduction to Linear regression and Correlation*. W.H. Freeman and Co., New York.
- Everitt, B.S., Hay, D., 1992. *Talking about Statistics. A Psychologist's Guide to Design & Analysis*. Edward Arnold, London.
- Garfinkel, D., Fegley, K.A., 1984. Fitting physiological models to data. *Am. J. Physiol.* 246, R641–R650.
- Glantz, S.A., Slinker, B.K., 2001. *Primer of Applied Regression and Analysis of Variance*, second ed. McGraw-Hill, Inc., New York
- Hamilton, L.C., 1992. *Regression with Graphics. A Second Course in Applied Statistics*. Duxbury Press, Belmont, CA.
- Haricharan, R.N., Barnhart, D.C., Cheng, H., Delzell, E., 2009. Identifying neonates at a very high risk for mortality among children with congenital diaphragmatic hernia managed with extracorporeal membrane oxygenation. *J. Pediatr. Surg.* 44, 87–93.
- Jamal, F., Strotmann, J., Weidemann, F., et al., 2001. Noninvasive quantification of the contractile reserve of stunned myocardium by ultrasonic strain rate and strain. *Circulation* 104, 1059–1065.
- Katz, M.H., 2006. *Multivariate Analysis. A Practical Guide for Clinicians*. Cambridge University Press, Cambridge.
- Kleinbaum, D.G., Klein, M., 2010. *Logistic Regression. A Self-Learning Text*. Springer, New York.
- Kleinbaum, D.G., Kupper, L.L., Muller, K.E., 1988. *Applied Regression Analysis and Other Multivariable Methods*. PWS-KENT Publishing Company, Boston.
- McKinley, S., Levine, M., Cubic Spline Interpolation. Available: <https://mse.redwoods.edu/darnold/math45/laproy/Fall98/SkyMeg/splinepres/sld006.htm>.
- Motulsky, H.J., Ransnas, L.A., 1987. Fitting curves to data using nonlinear regression: a practical and non-mathematical review. *FASEB* 1, 365–374.
- Neter, J., Kutner, M.H., Nachtsheim, C.J., et al., 1996. *Applied Linear Statistical Models*. Irwin, Chicago.
- O'Neill, S., Leahy, F., Pasterkamp, H., et al., 1983. The effects of chronic hyperinflation, nutritional status, and posture on respiratory muscle strength in cystic fibrosis. *Am. Rev. Respir. Dis.* 128, 1051–1054.
- Richardson, D.K., Corcoran, J.D., Escobar, G.J., et al., 2001. SNAP-II and SNAPPE-II: Simplified newborn illness severity and mortality risk scores. *J. Pediatr.* 138, 92–100.

- Sherrill, D.L., Morgan, W.J., Taussig, L.M., et al., 1989. A mathematical procedure for estimating the spatial relationships between lung function, somatic growth, and maturation. *Pediatr. Res.* 25, 316–321.
- Slinker, B.K., Glantz, S.A., 1985. Multiple regression for physiological data analysis: the problem of multicollinearity. *Am. J. Physiol.* 249, R1–12.
- Thursz, M.R., Kwiatkowski, D., Allsopp, C.E., et al., 1995. Association between an MHC class II allele and clearance of hepatitis B virus in the Gambia. *N. Engl. J. Med.* 332, 1065–1069.
- Weisberg, S., 1985. *Applied Linear Regression*. John Wiley & Sons, New York.
- Whyte, H.M., 1959. Blood pressure and obesity. *Circulation* 19, 511–516.

CHAPTER 31

Serial Measurements: Time Series, Control Charts, Cusums

INTRODUCTION

Frequently observations are obtained serially over time: hourly or daily body temperatures, daily blood creatinine concentrations, weekly hemoglobin concentrations, weekly check of a standard sodium salt concentration in a laboratory, annual birth rates in a country, weekly rainfall, or the Dow-Jones index. Fig. 31.1 shows the weekly incidence over several years of influenza-associated illness in children.

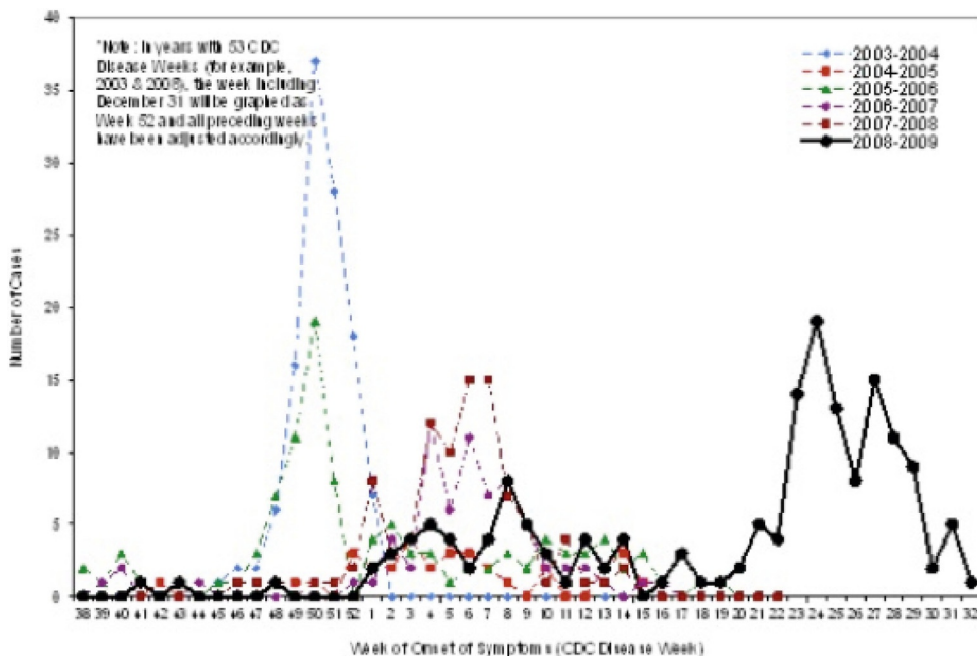


Fig. 31.1 Weekly incidence of influenza, from graph supplied to California physicians by the California State Department of Public Health on 8/20/09. Note increased incidence of influenza due to the H1N1 strain of swine flu at right of figure. (Reproduced with permission of the County of Marin Department of Health and Human Services.)

The mean value and variability of such serial measurements are of interest, and so is the trend toward increasing and decreasing. Analysis of time series is important in engineering and many textbooks deal with the subject. Various methods of smoothing the graphs and exposing underlying regularities in the apparently erratic data are used (Chatfield, 1980; Diggle, 1990; Hamilton, 1992). Oscillatory behavior is common in biological studies, ranging from electroencephalograms to calcium cycling in cells to pulsatile secretion of gonadotrophic hormones. An easy introduction to the subject is provided by Pollard (1977), Neter et al. (1978), Wonnacott and Wonnacott (1981), Mendenhall and Sincich (1986), Everitt (1989), Montgomery (1990). Biological oscillations may be less regular than those occurring in engineering studies and are correspondingly difficult to analyze (Merriam and Wachter, 1982; Filicori et al., 1984; Veldhuis et al., 1986). It is even possible to find artifactual oscillations that are induced by the mathematical process used, as discussed in an article entitled “Biological clock in the unicorn (Cole, 1957).”

SERIAL CORRELATION

A question about data collected serially over time is whether consecutive measurements are correlated with each other. Serial correlation may seriously underestimate the standard deviation from regression and all the assessments of variability derived from it, leading to a false estimate of accuracy.

Wald-Wolfowitz Runs Test

There are several ways of determining serial correlation. A simple method for nominal scale data is the Wald-Wolfowitz runs test. Consider a study in which 20 men (M) and 20 women (F) are enrolled. To evaluate randomness, define a run as a series of one kind without interruption. If the order of entry into the study was MMFFFMFFFMMMMMFMMMFFMFFMMMFFMMMFFMMMFFMMFF there are 17 runs: 2M, 3F, 1M, 3F, and so on (even a single member constitutes a run). A table of probabilities for this test gives $P > 0.20$, so that there is no evidence against randomness. The test can be done online at <http://www.quantitativeskills.com/sisa/statistics/ordinal.htm>.

If selection is randomized, it is unlikely that the first 20 subjects will be of one gender, and the next 20 the other gender; this sequence has only 2 runs, and $P < 0.001$. An exact alternation of men and women is also unlikely; a set with 40 runs also has $P < 0.001$.

With ≥ 20 members of each group, use the normal approximation.

The expected mean μ if the null hypothesis is true is

$$\mu = \frac{2n_1n_2}{N} + 1,$$

where n_1 and n_2 are the numbers in each group, and $N = n_1 + n_2$.

The variance σ^2 is calculated from

$$\sigma^2 = \frac{2n_1n_2(2n_1n_2 - N)}{N^2(N - 1)}$$

The value for W , the standardized estimate of the difference between the number of runs observed R and the number expected from the null hypothesis, is estimated as

$$W = \frac{R - \frac{2n_1n_2}{N} - 1}{\sqrt{\frac{2n_1n_2(2n_1n_2 - N)}{N^2(N - 1)}}}$$

The value for W is obtained from the z tables.

Some authorities recommend a correction for continuity in the numerator. If $R < \mu$, use $+0.5$; if $R > \mu$, use -0.5

$$W = \frac{\left(R - \frac{2n_1n_2}{N} - 1\right) \pm 0.5}{\sqrt{\frac{2n_1n_2(2n_1n_2 - N)}{N^2(N - 1)}}}$$

where the demarcating lines $| |$ indicate the absolute value of the argument. An extension to more than two groups is described by [Zar \(2010\)](#).

Up-and-Down Runs Test

A slightly different procedure, the up-and-down procedure, is used to determine if the alternation is between measurements on a ratio, interval, or ordinal scale. If a series of temperatures are taken successively, it would be suspicious if the temperatures alternated regularly or if a steady rise occurred for half the temperatures followed by a consistent fall. This is similar to the runs problems. To examine the up-and-down problem, the method of first differences has been used ([Edgington, 1961](#)). Determine if the difference between X_n and X_{n+1} is positive or negative, and then determine the number of runs. For total number < 50 , there are tables of critical values ([Durbin and Watson, 1950](#); [Zar, 2010](#)), and for larger numbers there is a normal approximation. The expected mean value of the number of runs R if the numbers are random is

$$\mu_R = \frac{2N - 1}{3},$$

and the standard deviation σ_R is

$$\sigma_R = \sqrt{\frac{16N - 29}{90}}.$$

Therefore

$$z = \frac{R - \frac{2N - 1}{3}}{\sqrt{\frac{16N - 29}{90}}}.$$

Table 31.1 presents a set of daily temperatures.

Table 31.1 Daily temperatures

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
T	37.6	37.8	37.9	36.5	37.3	37.7	36.6	36.9	37.6	36.9	36.2	36.8	37.4	37.1	36.8
Sign Δ		+	+	−	+	+	−	+	+	−	−	+	+	−	−

The number of runs is 8 (2+, 1−, 2+, 1−, 2+, 2−, 2+, 2−). From the runs table the probability of ≤ 8 runs in 15 consecutive observations is 0.2216. There is no reason to reject the null hypothesis that the temperatures were random. The test can be done with <http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Randomness.htm> and with Excel—see <http://www.youtube.com/watch?v=YWlod6Jdu-k>, but is easier to do by hand.

Rank von Neumann Ratio

A more powerful test than the up-and-down test just described is to associate the measured values of X_i with ranks r_i (from lowest to highest) and then calculate the rank von Neumann ratio ν as

$$\nu = \frac{\sum_{i=2}^N (r_i - r_{i-1})^2}{\frac{N(N^2 - 1)}{12}}.$$

The critical values for ν are given by Bartels (1982). Just as the Durbin-Watson test (below) the ratio varies from 0 to 4.

Applying this method to the data of Table 31.1 gives Table 31.2.

Table 31.2 Daily temperatures

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
T	37.6	37.8	37.9	36.5	37.3	37.7	36.6	36.9	37.6	36.9	36.2	36.8	37.4	37.1	36.8
ΔT		0.2	0.1	−1.4	0.8	0.4	−0.9	0.3	0.7	−0.6	−0.7	0.5	0.6	−0.3	−0.3
Rank		8	7	1	14	10	2	9	13	4	3	11	12	5.5	5.5

Then the rank ratio is

$$\nu = \frac{(7-8)^2 + (1-7)^2 + (14-1^2) \dots + (5.5-5.5)^2}{\frac{15(15^2-1)}{12}} = \frac{540.25}{280} = 1.93.$$

This value yields $P > 0.10$, so that the null hypothesis of randomness cannot be rejected. Tied ranks are averaged, and the formula is unreliable with many ties.

These tests can be used for more than time series. For example, they are recommended for assessing the independence of residuals in a regression equation or an analysis of variance.

Ratio Measurements

If the data consist of ratio measurements, one of the tests commonly used is the Durbin-Watson test for first-order autocorrelation. For example, if blood pressures are measured serially throughout the day, there would be positive autocorrelation because of lower pressures at night than in the daytime. The coefficient of autocorrelation is termed ρ , and the null hypothesis is that $\rho = 0$.

Autocorrelation affects the ability to estimate accurately the standard deviation from regression. The requirements for linear regression assume that the residuals (error terms) have zero mean, constant variance, and are uncorrelated, as well as being normally distributed. If these requirements are not met and there is positive autocorrelation, then the standard errors of the regression coefficient tend to be too small.

To assess autocorrelation, the Durbin-Watson test examines successive residuals about the regression line. The test statistic d is given by

$$d = \frac{\sum_{i=2}^N (e_i - e_{i-1})^2}{\sum_{i=1}^N e_i^2}.$$

where e_i represents the residuals. The statistic can be calculated online at <http://www.wessa.net/slr.wasp>, <https://www.easycalculation.com/statistics/durbin-watson-test-calculator.php> and <http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Trend.htm>, and tables of critical values at <http://berument.bilkent.edu.tr/DW.pdf>, <http://www.statisticshowto.com/durbin-watson-test-coefficient/>, <http://www.math.nsysu.edu.tw/~lomn/homepage/class/92/DurbinWatsonTest.pdf>, and <http://eclectic.ss.uci.edu/~drwhite/courses/Durbin-Watson.htm>.

Autocorrelation, if present, is usually positive, (Fig. 27.18) and then the differences between any two consecutive residuals will be small, and the sum of the squared differences will be small relative to the sum of the squared residuals. d has a maximal value of 4.

The tables have both upper and lower critical values that need to be judged. The usual null hypothesis H_0 is that the coefficient of autocorrelation $\rho = 0$, and the usual alternative (for positive autocorrelation) is that $\rho > 0$. Then if $d < d_L$ (the lower of the two tabulated values) conclude that d is unlikely to have been due to chance and reject the null hypothesis at level α ; if $d > d_U$ d is negligible and do not reject H_0 ; and if $d_L < d < d_U$, the test is inconclusive.

In general, $d = 2$ means no autocorrelation, $d < 1$ suggests marked positive autocorrelation, and $d > 3$ suggest marked negative autocorrelation, although the latter is rare (Fig. 27.19). To test the alternative hypothesis $\rho < 0$, repeat the above tests but use $4 - d$ instead of d .

CONTROL CHARTS

Quality control refers to determining the limits within which a process functions. A *process* is a series of operations that produces an outcome, such as an accurate measurement of a blood constituent. In measuring a blood constituent, for example, the process includes pipetting out the correct initial amount of blood, adding an exact amount of a diluent with precisely known property, perhaps adding an accurately measured amount of a chemical reagent with precisely known composition and concentration, and estimating the absorption of the resulting solution in an accurately calibrated spectrophotometer. A process is regarded as being in statistical control if repeated samples act like random samples from a stable probability distribution. Ideally, repeating a measurement of a standard sodium concentration of 140 mMol should give the same value. However, repeated measurements of the same standard sodium solution will vary. The problem then is to determine stability of the signal in the face of noise, that is, natural variability. When is variability excessive and indicates bias? If the samples are shown to be out of statistical control, then the source must be determined so that corrective action can be taken.

Early in the 20th century Shewhart, working at the Bell Telephone Laboratories, introduced the notion of control charts. These in general take the form of a chart in which the value of the measurement is plotted against time, usually in regular increments.

A typical control chart is given in Fig. 31.2.

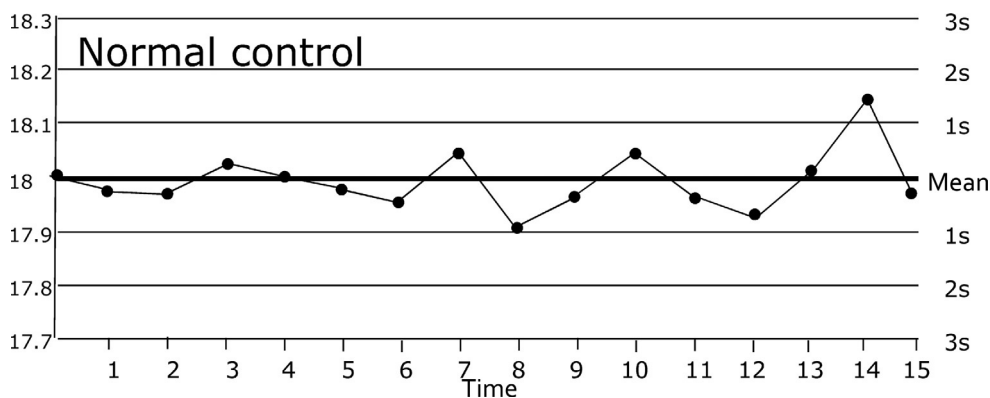


Fig. 31.2 Control chart. s , Standard deviation.

Each point is a measurement made at successive times. The thick centerline is the mean value, determined by theory or prior observation. The lines above and below the centerline are drawn at values corresponding to 1, 2, and 3 standard deviations above the mean, based on prior observations. The values of 17.7 to 18.3 are artificial values used here for illustration. In an actual control chart, they are determined by the subject being measured: for example, the mean might be 140 mMol for a normal serum sodium concentration and the standard deviation determined by observation of repeated measurements.

Determining the correct standard deviation depends in part on the number of “control” measurements used in its calculation. Although a large sample size yields a sample standard deviation closer to the population value, it is the short-term variation that we want. One recommendation is that the control period be divided into successive subgroups (natural or artificially determined), the range of each subgroup is calculated, then the mean range is determined, and a working standard deviation is derived by dividing the mean range by Hartley’s constant that depends on the number of subgroups and their sample size (Caulcutt, 2004). The constants are found at <http://dfx.nl/userfiles/file/pdf/Control%20Chart%20Constants%20and%20Formulas.pdf> (or look up under Hartley’s constant) and are used for dividing the mean range to get the upper and lower control limits directly. Frequently a range of 2 is chosen. Then consecutive ranges $|Y_1 - Y_2|$, $|Y_2 - Y_3|$, ... are averaged and divided by the constant 1.28 to provide an estimated standard deviation (Caulcutt, 2004).

The control chart in Fig. 31.2 shows the accuracy and precision of a process. For it to be more useful, a series of rules have been created. These specify sets of one or more unusual events that help in deciding when a process is out of control enough to warrant further investigation. The Western Electric Company (1956) defined some of these events, and a fuller set were defined by Nelson’s rules (Nelson, 1984). Nelson’s rules are as follows:

1. 1 point is >3 standard deviations from the mean.
2. 9 or more consecutive points are on the same side of the mean.
3. 6 or more consecutive points are increasing or decreasing.
4. 14 or more points in a row alternately increase and decrease.
5. 2 or 3 out of 3 consecutive points are >2 standard deviations of the mean on the same side of the mean.
6. 4 or 5 out of 5 consecutive points are >1 standard deviation of the mean on the same side of the mean.
7. 15 points in a row are all within 1 standard deviation on either side of the mean.
8. 8 consecutive points are all >1 standard deviation on either side of the mean.
9. Each of these rules is illustrated in Fig. 31.3.

In some publications the lines 3 standard deviations above and below the mean are termed, respectively, the upper control limit (UCL) and the lower control limit (LCR). Companies and laboratories are free to determine their own rules, and sometimes the control limits are determined not by standard deviation units but by practical considerations of what tolerance should be allowed.

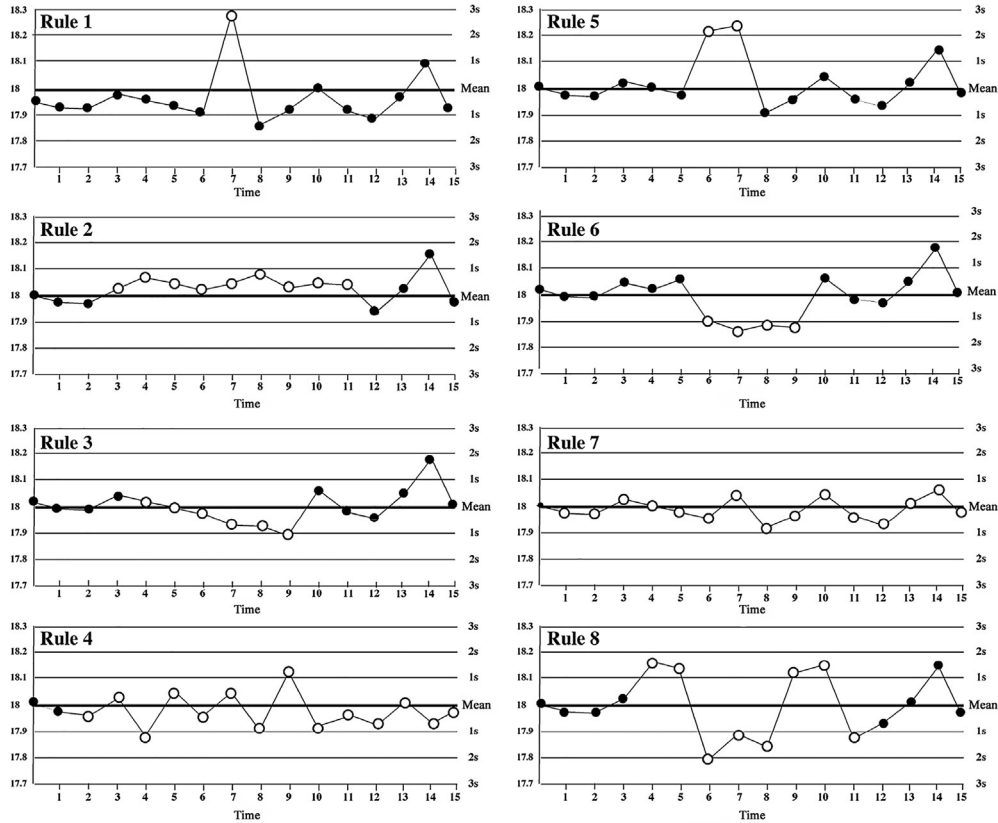


Fig. 31.3 Nelson's rules. The *open circles* indicate the abnormal points that fit each rule.

It is also possible to have control charts for the means, ranges, or standard deviations of small groups. Each of these has use in a particular field. Many other types of control charts are described by Roberts (1966), Hogg and Ledolter (1987), and Oakland (2003). Control charts with step-by-step instructions can be created in Excel at (<http://www.vertex42.com/ExcelTemplates/control-chart.html>).

CUMULATIVE SUM TECHNIQUES

In medical practice, as well as laboratory control, there may be a quicker way to determine when a process is changing. This is a particularly challenging problem because often the signal is almost undetectable in the noise level. Consider Fig. 31.4, taken from an article entitled "Why don't doctors use Cusums?" by Chaput de Saintonge and Vere (1974).

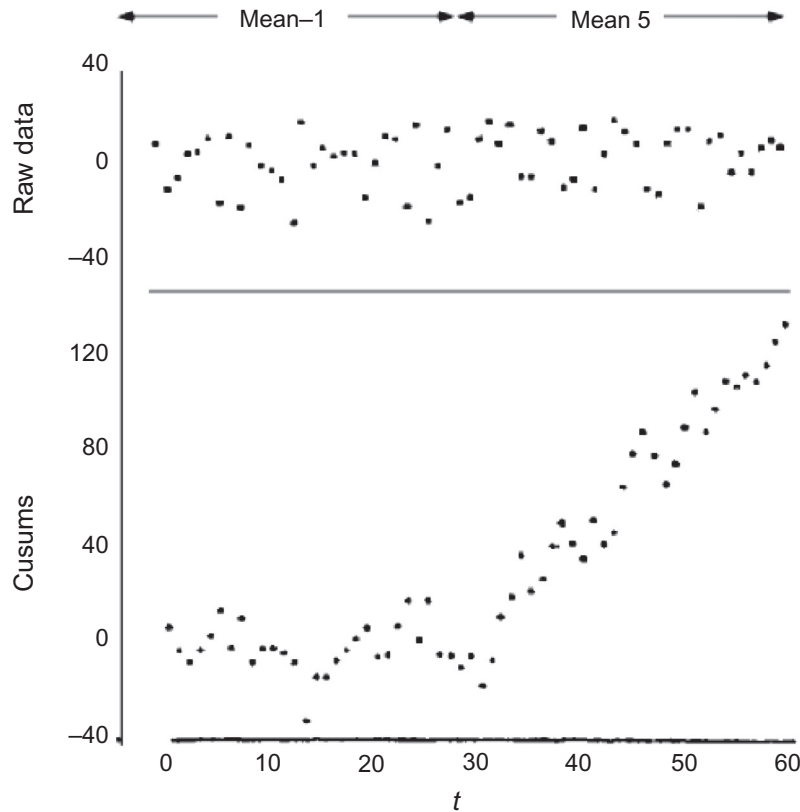


Fig. 31.4 Upper panel. Artificial example of a low signal:noise ratio. The first 30 points show randomly scattered values about a mean of -1 . The second portion shows similar random variation after the mean has been altered to 5 . The change is undetectable by eye. Lower panel. Cusum of the raw data.

The cumulative sum technique could not be simpler. By eye, a mean value is selected for the first few points. The difference between the next point (X_1) and the mean is d_1 . Then the difference $d_2 = X_2 - X_1$ is added to d_1 to obtain $\sum d_i$. Then d_3, d_4, \dots, d_n are added sequentially, and $\sum d_i$ is plotted against sample number. If the values of X continue to scatter above and below the mean, some of the deviations will be positive, some will be negative, and the cumulative sum of the deviations tends to remain near zero. If, on the other hand, the values of X tend to increase or decrease consistently, even though the trend is hidden by the scatter, the cumulative sum (Cusum) of the deviations becomes increasingly positive or negative, respectively. The results of the procedure on the data shown in Fig. 31.4 are given in the lower panel (Chaput de Saintonge and Vere, 1974).

Another example is from a study by Davey et al. (1986) of the discharge rate of a single gamma motor neuron in the cat hind limb before and after facilitation by episodic stretch (Fig. 31.5).

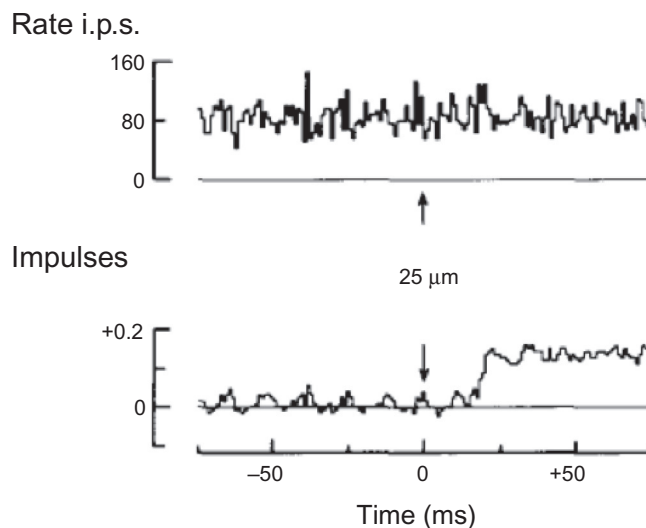


Fig. 31.5 Discharge rate of motor neuron after facilitation (*arrow*). Basic record of rate in impulses per second above, Cusums below. It is doubtful that any change could have been detected by eye.

Instructions and sample charts can be found at https://www.statstodo.com/CUSUM_Exp.php.

Problem 31.1 The following gives daily white cell counts (1000/c.mm) in a patient. Can you tell when the white count has begun to increase?

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
WBC	9.4	8	9.3	11.7	12.2	10.2	8.1	11.5	9.2	10.4	9	11.5	10.5	9.4	10.1
Day	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
WBC	9.4	10.6	10.3	8.5	10.8	10.9	9.3	12.3	11.5	10.6	11.1	10.4	11.6	11.3	10.5

How can we tell when a deviation from zero slope is significant enough to warrant further investigation? The most commonly used method is to use a V mask, illustrated in Fig. 31.6.

The V mask is based on the desired average run length (ARL), that is the average length of the Cusum while the failure or deviation rate is acceptable. The average run length equals the number of patients or measurements seen before the Cusum exceeds the control limit. A wide mask angle reduces the number of false alarms, whereas a narrow angle allows early detection of a change that could lead to corrective action. The angle of the V depends on how many units ahead of the last Cusum point the apex is placed, how many standard deviations from the mean are acceptable, and how many units of divergence from the mean can be tolerated before some action is taken (Woodward and Goldsmith, 1964). Some of these issues are more important in industry than

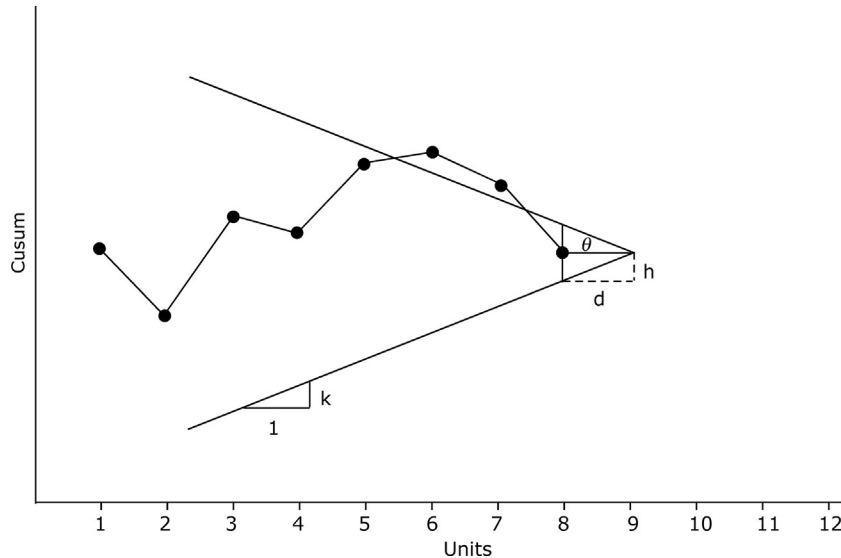


Fig. 31.6 Application of V mask to the Cusum. The apex of the mask is placed one or more units ahead of the last measurement. The appearance of Cusum values that cross the upper line shows an upward slope that warrants investigation.

medicine. For example, in Medicine, we often need to take action early, for example, to stabilize blood pressure or control ventilation, so that a one-unit lead and a short run time in which the Cusum departs from the mean are preferred.

V masks are implemented in most statistical programs.

To calculate the V mask, determine a deviation δ in standard deviation units that should be detected. The V mask has a half angle θ that the slope of the upper or the lower arm makes, as defined by the rise or fall of the slope (k) for a 1-unit change. This is the same as the vertical height h divided by the distance d that is the number of units the apex is ahead of the last point so that $d = h/k$.

As a rule of thumb, if k is half of δ , h should be about 4 or 5. If a high sensitivity to small changes or slow trends is wanted, $h = 8$ and $k = 0.25$ are appropriate, and if large changes or rapid shifts need to be detected, then $h = 2.5$ and $k = 1$ are appropriate. As an example, if $d = 8$ and $\tan \theta = 0.25$, a change of 2σ in performance will be detected on the average after 4 measurements, a one σ change after 8 measurements, and a 0.5σ change after 18 measurements (Marshall, 1979). The selection of a suitable V mask depends on what the investigator wants to do with the data. Alternative ways of assessing changes from the process mean have been described (Everitt, 1989).

Although some investigators have not found Cusums to be useful in biomedical studies (Mitchell et al., 1980) this is not the predominant view (Chaput de Saintonge and Vere, 1974; Westgard et al., 1977; Chatfield, 1980; Rowlands et al., 1980; Diggle, 1990; O'Brien and Christie, 1997; Grunkemeier et al., 2003; Chang and McLean, 2006, and these methods may be particularly useful in infectious diseases with erratic

swinging fevers. [Walters and Griffin, 1986](#); [Kinsey et al., 1989](#)) Cusums are used extensively in industry. Clear descriptions of their use and evaluation are given by [Woodward and Goldsmith \(1964\)](#) and by [Chatfield \(1995\)](#).

Various extensions of the CUSUM method have been used to follow surgical complications (death, excessive blood transfusion, etc.) so as to detect a change in their incidence. One variation is the Exponential Weighted Moving Average that resembles the CUSUM method except that for each sum the previous observations are given exponentially decreasing weights ([Smith et al., 2013](#)). The second variation allows for the possibility that successive patients may have very different associated problems by creating a risk adjusted CUSUM ([Steiner et al., 2000](#)). Consultation with a statistician is recommended in deciding which variation to use.

SERIAL MEASUREMENTS

It is common for a group of animals or patients to be given a stimulus, and then to have serial observations made at successive time intervals of some response, for example, glucose or catecholamine concentrations, blood pressures, tumor size. [Chapter 26](#) discussed repeated measures analysis, and this certainly enters into analyses of the results of such experiments. There are, however, other problems to deal with. The most important is that the successive measurements are not independent of each other but are serially correlated.

As pointed out by [Matthews et al. \(1990\)](#) two types of responses are typically seen. One is the peaked response, where the response rises above control values to a maximum, (or falls below them to a minimum) and then falls back to baseline. In the other there is a monotonic rise or fall over the duration of the study. Frequently several subjects are in each group. An example of the peaked response can be seen in [Fig. 31.1](#).

The first comment about such a figure is that joining the means of the results at different time periods gives a curve that may conceal the individual responses ([Fig. 31.7](#)).

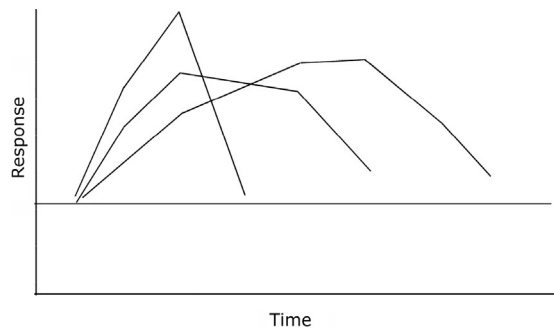


Fig. 31.7 Serial responses. Hypothetical results of three patients.

It is possible to construct a mean curve, but instead of conveying useful information it conceals information. It is even more difficult to analyze if there are missing values, or else values are obtained opportunistically at different times for each subject. Many statisticians have recommended measuring the area under the curve for each subject. Then each subject's responses are summarized by a single number, and the sets of areas under the curves can be compared by a standard two-sample test.

On the other hand, the question of interest may be the time to reach the maximum in the two groups, and once again the mean curves may be misleading. Matthews et al. (1990) recommend plotting the maximal value against the mean time for each group, and then either compare the mean times to maximum or perhaps the slopes of the resulting plots. Alternatively, the question might be what the maxima or the minima are, and these are again easy to compare as sets of summary figures.

For monotonic curves, the analysis also depends upon the question being asked. This might be the rate of change, estimated by (linear) regression lines, the maximal values at a given time after the stimulus, or the time taken to reach a given change from the baseline. Whatever the question, it is best approached by summarizing each subject's response in a single measurement, rather than blending all the subjects in a group into one curve that might not be representative of the variability of the subjects.

REFERENCES

- Bartels, R., 1982. The rank version of von Neumann's ratio test for randomness. *J. Am. Stat. Assoc.* 77, 40–46. or https://www.researchgate.net/publication/230639951_The_Rank_Version_of_von_Neumann%27s_Ratio_Test_for_Randomness.
- Caulcutt, R., 2004. Control charts in practice. *Significance* 1, 81–84.
- Chang, W.R., Mclean, I.P., 2006. CUSUM: a tool for early feedback about performance? *BMC Med. Res. Methodol.* 6, 8.
- Chaput De Saintonge, D.M., Vere, D.W., 1974. Why don't doctors use cusums? *Lancet* 1, 120–121.
- Chatfield, C., 1980. *The Analysis of Time Series. An Introduction*. Chapman & Hall, London.
- Chatfield, C., 1995. *Statistics for Technology. A Course in Applied Statistics*. Chapman & Hall, London.
- Cole, L.C., 1957. Biological clock in the unicorn. *Science* 125, 874–876.
- Davey, N.J., Ellaway, P.H., Stein, R.B., 1986. Statistical limits for detecting change in the cumulative sum derivative of the peristimulus time histogram. *J. Neurosci. Methods* 17, 153–166.
- Diggle, P.J., 1990. *Time Series. A Biostatistical Introduction*. Clarendon Press, Oxford.
- Durbin, J., Watson, G.S., 1950. Testing for serial correlation in least squares regression, I. *Biometrika* 37, 409–428.
- Edgington, E.S., 1961. Probability table for number of runs of signs of first differences in ordered series. *J. Am. Stat. Assoc.* 56, 156–159.
- Everitt, B.S., 1989. *Statistical Methods for Medical Investigations*. Oxford University Press, New York.
- Filicori, M., Butler, J.P., Crowley, W.F., J., 1984. Neuroendocrine regulation of the corpus luteum in the human. Evidence for pulsatile progesterone secretion. *J. Clin. Invest.* 73, 1638–1647.
- Grunkemeier, G.L., Wu, Y.X., Furnary, A.P., 2003. Cumulative sum techniques for assessing surgical results. *Ann. Thorac. Surg.* 76, 663–667.
- Hamilton, L.C., 1992. *Regression with Graphics. A Second Course in Applied Statistics*. Duxbury Press, Belmont, CA.

- Hogg, R.V., Ledolter, J., 1987. *Applied Statistics for Engineers and Physical Scientists*. Macmillan Publishing Company, New York.
- Kinsey, S.E., Giles, F.J., Holton, J., 1989. Cusum plotting of temperature charts for assessing antimicrobial treatment in neutropenic patients. *BMJ (Clin. Res. Ed.)* 299, 775–776.
- Marshall, R.A.G., 1979. The analysis of counter performance by cusum techniques. *J Radioanal. Chem.* 54, 87–94.
- Matthews, J.N., Altman, D.G., Campbell, M.J., et al., 1990. Analysis of serial measurements in medical research. *Br. Med. J. (Clin. Res. Ed)* 300, 230–235.
- Mendenhall, W., Sincich, T., 1986. *A Second Course in Business Statistics: Regression Analysis*. Dellen Publishing Co., San Francisco.
- Merriam, G.R., Wachter, K.W., 1982. Algorithms for the study of episodic hormone secretion. *Am. J. Physiol.* 243, E310–E318.
- Mitchell, D.M., Collins, J.V., Morley, J., 1980. An evaluation of cusum analysis in asthma. *Br. J. Dis. Chest* 74, 169–174.
- Montgomery, D.C., 1990. *Introduction to Statistical Quality Control*. Wiley, New York.
- Nelson, L.S., 1984. Technical aids. *J Qual. Technol* 16, 238–239.
- Neter, J., Wasserman, W., Whitmore, G.A., 1978. *Applied Statistics*. Allyn and Bacon, Inc., Boston.
- Oakland, J.S., 2003. *Statistical Process Control*. Butterworth-Heinemann, Amsterdam.
- O'Brien, S.J., Christie, P., 1997. Do CuSums have a role in routine communicable disease surveillance? *Public Health* 111, 255–258.
- Pollard, J.H., 1977. *A Handbook of Statistical and Numerical Techniques*. Cambridge University Press, Cambridge.
- Roberts, S.W., 1966. A comparison of some control chart procedures. *Technometrics* 8, 411–430.
- Rowlands, R.J., Wilson, D.W., Nix, A.B., et al., 1980. Advantages of CUSUM techniques for quality control in clinical chemistry. *Clin. Chim. Acta* 108, 393–397.
- Smith, I.R., Garlick, B., Gardner, M.A., et al., 2013. Use of graphical statistical process control tools to monitor and improve outcomes in cardiac surgery. *Heart Lung Circ.* 22, 92–99.
- Steiner, S.H., Cook, R.J., Farewell, V.T., et al., 2000. Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics* 1, 441–452.
- Velduis, J.D., Weiss, J., Mauras, N., et al., 1986. Appraising endocrine pulse signals at low circulating hormone concentrations: use of regional coefficients of variation in the experimental series to analyze pulsatile luteinizing hormone release. *Pediatr. Res.* 20, 632–637.
- Walters, S., Griffin, G.E., 1986. Resolution of fever in *Staphylococcus aureus* septicaemia—retrospective analysis by means of Cusum plot. *J. Inf. Secur.* 12, 57–63.
- Western Electric Company, 1956. *Statistical Quality Control Handbook*. Western Electric Co., Indianapolis, IN.
- Westgard, J.O., Groth, T., Aronsson, T., et al., 1977. Combined Shewhart-cusum control chart for improved quality control in clinical chemistry. *Clin. Chem.* 23, 1881–1887.
- Wonnacott, T.H., Wonnacott, R.J., 1981. *Regression: A Second Course in Statistics*. John Wiley & Sons, New York.
- Woodward, R.H., Goldsmith, P.L., 1964. *Cumulative Sum Techniques*. Oliver and Boyd, Edinburgh.
- Zar, J.H., 2010. *Biostatistical Analysis*. Prentice Hall, Upper Saddle River, NJ.

CHAPTER 32

Dose-Response Analysis

GENERAL PRINCIPLES

An increasing stimulus (dose or concentration of an agent, rate of nerve stimulation, degree of stretch, etc.) is given to a responding system that may be a bacterial culture, insect larvae, lymphocytes in culture, a muscle strip, a blood vessel, or even a whole organism in which the response or effect may be blood pressure, temperature, white cell count, percentage of organisms killed, amount of cytokine released, and so on. At extremely low stimuli there may be no response. Then, as the stimulus increases, a response occurs and usually increases to reach some maximum value, after which increasing stimuli might have no further effect. (It is possible for the response to decrease after some maximal stimulus has been given.) The basic response curve is usually tripartite—a horizontal portion of no response at low stimuli, a rising response at higher stimuli, and then another horizontal maximal response at the highest stimuli. This gives an S-shaped or sigmoid curve (Fig. 32.1).

In the top panel, the dose increases from left to right, and the response increases from bottom to top. In the bottom panel, the response decreases as dose increases because of inhibition. Frequently, using a logarithmic scale for doses produces an approximately linear intermediate (nonhorizontal) phase that may be easier for analysis. In addition, the logarithmic scale expands the regions of low dosage where rapid changes are occurring and compresses the scale at higher doses where the response is changing more slowly. The response may be continuous, such as a gradual increase in muscle tension as the agonist concentration increases, or may be quantal, as when a particular effect is or is not reached. The quantal response is an ungraded response; it may be a fixed effect such as reduction of coughing below x times per hour or may be an all-or-none response such as living or dying.

The typical dose-response curve is a plot of the logarithm of the dose on the X-axis and the response on the Y-axis. Frequently the responses and/or the doses are transformed into response or dose metameters. Judicious selection of transformation for the dose can be selected so that the dose metameters are simple numbers such as -1 and $+1$ for two dose levels; -1 , 0 , and $+1$ for 3 dose levels, and so on (Finney, 1952; Finney, 1964; Colquhoun, 1971). The response is obtained by direct measurement in whatever units are appropriate to the subject, but then is converted into percentages based on no effect as 0% and full effect as 100% . Almost all responses are the result of binding of an agonist to a receptor, and the degree of binding is known as the affinity, that is, how chemical interactions result in occupancy of the receptors. This is not the

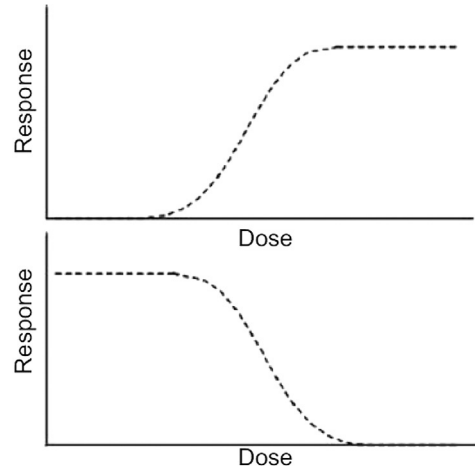


Fig. 32.1 Basic dose-response curve.

same as efficacy, which refers to the response induced by that binding. This in turn is distinguished from potency that refers to the amount of agonist required to produce a given effect and is a function of affinity and efficacy; potency is often defined as the effect of 1 unit of a standard preparation. The concentration of agonist required to produce 50% of the maximal effect is known as the ED50, and in some experiments the concentration of agonist required to kill 50% of the animals or cells is termed the LD50. If an agonist shifts the curve to the left (curve B in Fig. 32.2) it has increased potency because the ED50 is seen at a lower concentration; the maximal response and thus the maximal efficacy may not change. Some of these features are shown in Fig. 32.2.

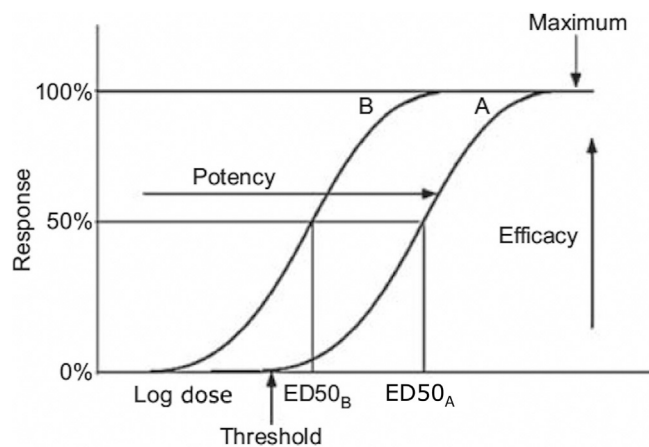


Fig. 32.2 Some typical features in a dose-response curve. ED50 is the median effective dose.

The slope of the middle portion of the curve often supplies information. The concentrations at which the response first begins (threshold) and first reaches maximum are also of interest but may be difficult to define accurately. Sigmoid dose-response curves are of major importance in analyzing receptor-binding information. Analysis of sigmoid curves is mathematically complex and requires good programs to achieve it. The analysis has to handle a complex equation

$$Y_i = \frac{a - d}{1 + \left(\frac{X_i}{c}\right)^b},$$

where Y is the response, X is the dose, a is the response when $X=0$, d is the maximal response, c is the ED50, and b is a steepness factor. Another more general formulation for the sigmoid dose-response curve is

$$\hat{Y} = \text{Bottom} + \frac{\text{Top} - \text{Bottom}}{1 + 10^{\text{LogEC50} - X}}.$$

See excellent online discussion by Motulsky at http://www.graphpad.com/guides/prism/6/curvefitting/index.htm?reg_classic_dr_variable.htm.

Multiple nonlinear regression techniques are needed to solve these problems.

One general statistical program that features these analyses prominently is Prism, but all large programs allow dose-response analysis, and there are a host of specialized programs for this purpose, one of the best-known being Ligand (DeLean et al., 1978; Munson and Rodbard, 1980; Curran-Everett, 2005).

QUANTAL DOSE-RESPONSE CURVES

Some responses are all-or-none and are termed quantal. Examples are whether an insect does or does not survive a given dose of insecticide, or whether a patient does or does not respond to a painful stimulus when given a particular concentration or amount of an anesthetic. Analysis of such responses differs from other forms of dose-response analysis, because the investigator does not give increasing doses of the agonist and measure a corresponding response in any subject. Instead, a specific dose of agonist is given to a number of subjects (insects, patients) and the number that do or do not react (survive for insects, move with patients) are recorded. Then a higher dose is given to a new set of subjects, and so on. How any given individual responds to a particular dose is unknown, but the percentage responding at that dose can be determined. In general, the percentage responding plotted against the dose, or the log of the dose, produces an S-shaped curve (Fig. 32.2).

It is possible to straighten out the cumulative normal sigmoid curve by plotting the X variate against the NED (normal equivalent distribution) (Chapter 6). By converting a normal Gaussian curve into standard form

$$z = \frac{X_i - \mu}{\sigma}$$

any given area under the curve can be specified in z units. The lowest 2.5% under the curve occurs at -2 units (2σ below the mean), the lowest 15.87% at -1 units, 50% at 0 units, 84.2% at 1 unit, and so on. This scale is symmetrical about zero, and although in theory it extends to infinity above and below the mean, in practice only 0.0000002867 of the area is more than 5 units below or above the mean, so that virtually all the area lies between 5 units below and 5 units above the mean of 0. To avoid negative numbers 5 is added to the scale, and the new transformed NED numbers are termed probits. Therefore if the quantal dose–(log) response curve is approximately symmetrical and normal, a plot of the dose against the percent response transformed into probits (derived from tables) gives a straight line (Fig. 32.3).

$$\text{Probit} = 5 + \frac{X_i - \mu}{\sigma} = \left(5 - \frac{\mu}{\sigma}\right) + \left(\frac{1}{\sigma}\right) X_i$$

and this is the equation to a straight line with intercept $\left(5 - \frac{\mu}{\sigma}\right)$ and slope $\left(\frac{1}{\sigma}\right)$.

A program for converting percentages to probits is given at <http://userwww.sfsu.edu/efc/classes/biol710/probit/ProbitAnalysis.pdf>.

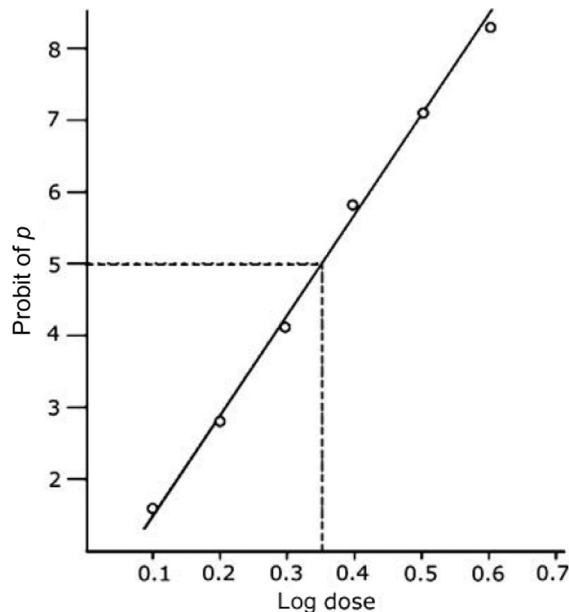


Fig. 32.3 Quantal dose-response probit plot.

If the points can be fitted reasonably well by a straight line, then the value of X corresponding to a probit of 5 (shown by the dotted lines) is the median log dose required for a 50% response. The reciprocal of the slope is the standard deviation, and this can be used to set confidence limits (in log units).

Precise fitting of the line is complex, because the binomial response results in larger variability for high proportions than for low proportions (heteroscedasticity). Fitting is done by weighted least squares methods and usually has several iterations.

Precision of estimate is greatest for the ED50 (median effective dose), but often a lower value is required. There is not much interest in knowing how much radiation is needed to kill or cause cancer in 50% of people, but much importance in knowing at what lower limit of radiation exposure the risk is very low. This means working at the low extreme of the curve, but here data are more difficult to assess and the number required to reach a firm conclusion may be very large. This is why we are still arguing about whether there is or is not a threshold for radiation damage.

REFERENCES

- Colquhoun, D., 1971. *Lectures on Biostatistics*. Clarendon Press, Oxford.
- Curran-Everett, D., 2005. Estimation of dose-response curves and identification of peaks in hormone pulsations: classic marriages of statistics to science. *Am. J. Physiol. Endocrinol. Metab.* 289, E363–E365.
- Delean, A., Munson, P.J., Rodbard, D., 1978. Simultaneous analysis of families of sigmoidal curves: application to bioassay, radioligand assay, and physiological dose-response curves. *Am. J. Phys.* 235, E97–102.
- Finney, D.J., 1952. *Probit Analysis*. Cambridge University Press, Cambridge.
- Finney, D.J., 1964. *Statistical Method in Bioassay*. Hafner Publishing Co, London.
- Munson, P.J., Rodbard, D., 1980. Ligand: a versatile computerized approach for characterization of ligand-binding systems. *Anal. Biochem.* 107, 220–239.

CHAPTER 33

Logistic Regression

INTRODUCTION

Until now the regression discussion has involved a continuous Y variate. Sometimes, however, the Y variate is discontinuous, and in particular has a dichotomous value: an outcome either happens or does not happen. Examples are the response of an organism to a toxin—it either lives or dies, survival of a premature infant related to birth weight, the presence or absence of a disease related to certain clinical or laboratory findings. Therefore Y can take on only one of two values—no (0) or yes (1). If many subjects are studied at different doses of a drug or different birth weights, then at each dose or weight a probability (P) of an event such as death or a disease can be assigned.

If in a typical regression equation $\hat{Y}_1 = c + b_1 X_1$ Y is replaced by P , then $P = c + bX_1$.

This is unsuitable for dichotomous results because P can vary only from 0 to 1, whereas Y can be any value, including values above 1 or being negative. To avoid a negative result, the equation might be written as $P = e^{c+bX}$, but this, although always positive, can be >1 . The desired criterion can be met by an expression of the form $P = \frac{e^{c+bX}}{1 + e^{c+bX}}$

that can never exceed 1. This can also be written as $P = \frac{1}{1 + e^{-(c+bX)}}$.

This is termed the logistic function that has an S shape (Fig. 33.1).

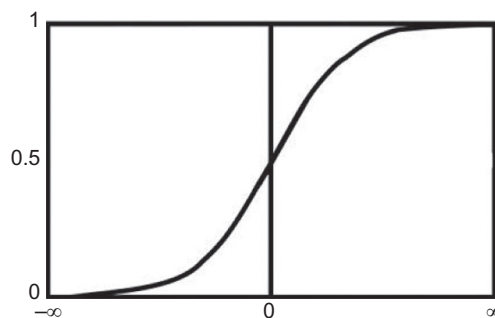


Fig. 33.1 Logistic function curve.

When the coefficient b in the equation is $-\infty$, the value of the function is zero, and when it is $+\infty$, the value is 1.

In normal regression c represents the intercept on the Y -axis when $X=0$, but in logistic regression it has a slightly different interpretation (see later), and β_0 is used in its place. Then the logistic function becomes

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}.$$

β_1 is the coefficient of the X variate.

P is the probability of success, and $1 - P$ is the probability of failure.

$$1 - P = 1 - \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} = \frac{1 + e^{-(\beta_0 + \beta_1 X)} - 1}{1 + e^{-(\beta_0 + \beta_1 X)}} = \frac{e^{-(\beta_0 + \beta_1 X)}}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

The odds of success are defined as $\frac{P}{1 - P}$, and

$$\frac{P}{1 - P} = \frac{\frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}}{\frac{e^{-(\beta_0 + \beta_1 X)}}{1 + e^{-(\beta_0 + \beta_1 X)}}} = \frac{1}{e^{-(\beta_0 + \beta_1 X)}} = e^{(\beta_0 + \beta_1 X)}.$$

Taking natural logarithms gives

$$\log_n \frac{P}{1 - P} = \log_n e^{(\beta_0 + \beta_1 X)} = \beta_0 + \beta_1 X.$$

This is the equation of a linear regression between X and $\log_n \frac{P}{1 - P}$.

The expression $\log_n \frac{P}{1 - P}$ is the logit transformation.

After deriving the linear equation, convert the log of the odds of success to the probability P of success, even though this latter relationship is a linear. These calculations can be performed online at <http://statpages.org/logistic.html>, <http://vassarstats.net/logreg1.html#down>, http://www.wessa.net/rwasp_logisticregression.wasp, but only for simple problems.

What meaning is attached to the different coefficients? If the X variate is zero then $P = \frac{1}{e^{-\beta_0}}$ or $\text{logit } P = \beta_0$. This is true, whether or not X can physically be zero. In the examples used later for artificial ventilation in neonates, it is impossible to have a neonate with zero gestational weight. In this instance, β_0 is the background risk of artificial ventilation in the absence of any explanatory factors; that is, it is the average risk for the whole studied population.

The β_1 coefficient also has a meaning. It is the change in $\text{logit } P$ for a one-unit change in the X variable. If there are multiple explanatory factors (X_1, X_2, X_3 , etc), β_i is the change in $\text{logit } P$ for a one-unit change in the X_i variable when the other variables are constant.

SINGLE EXPLANATORY VARIABLE

In a newborn nursery, the risk of needing artificial ventilation is a function of birth weight or gestational age. From a data set of 315 newborn infants, kindly supplied by Dr. Terri Slagle of the California Pacific Medical Center in San Francisco, examine these relationships by logistic regression. Fig. 33.2 show the results for the probability of needing ventilation, based on birth weight.

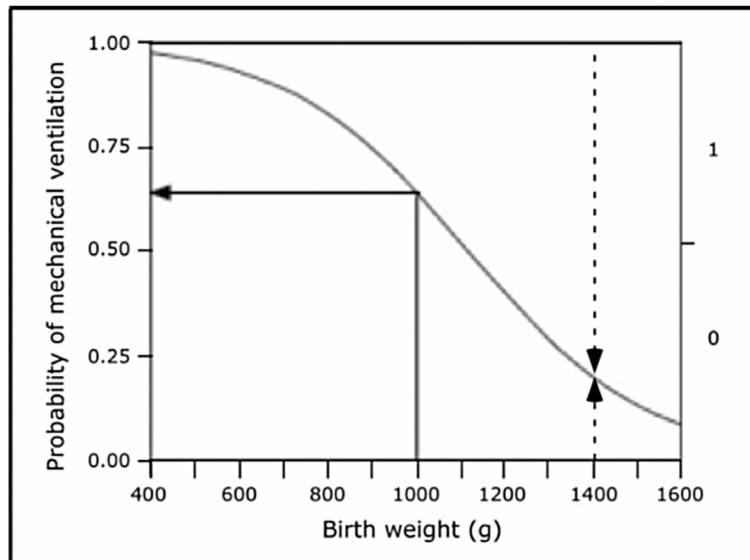


Fig. 33.2 Birth weight vs probability of artificial ventilation.

Interpret the curve as splitting the area into a portion with risk of ventilation and a portion with no risk of ventilation; the two vertical dashed lines show that at a birth weight of 1400g, the odds of artificial ventilation are about one-quarter the odds of no artificial ventilation.

At very low birth weights almost 100% of the neonates needed artificial ventilation, a percentage that decreased to very low values at high birth weights.

Table 33.1 gives the analytic results.

Table 33.1 The logistic function

Model	–Log likelihood	Df	Chi-square	Prob > chi-square
Difference	54.15	1	108.30	<0.0001
Full	163.93			
Reduced	218.08			
R ²	0.2483			
Parameter estimates				
Term	Estimate	Standard error	Chi-square	Prob > chi-square
Intercept	5.5080	0.6522	71.33	<0.0001
Birth weight (g) for log odds 0/1	0.004923	0.0005675	75.27	<0.0001

The programs use the method of maximum likelihood estimation to fit the curve. In each panel, the negative log likelihood is calculated for all the data points without regard to any explanatory factors (full) and this is compared with the negative log likelihood calculated when the factor is included (reduced). The difference resembles the within-groups mean square in ANOVA and is tested for ability to reject the null hypothesis. There is a substantial reduction in variability due to the explanatory factor.

As in any chi-square, the higher the value for given degrees of freedom the more likely are we able to reject the null hypothesis that birth weight did not affect the need for artificial ventilation. The R^2 of 0.2483 shows that the explanatory variable can account for no more than 24.83% of the variability.

The Parameter Estimates provide the prediction equation as

$$P = \frac{1}{1 + e^{-(5.5080 - 0.004923BW)}}.$$

and this can be used to calculate probabilities at any desired birth weight. For example, a neonate weighing 1000g at birth has a probability of needing artificial ventilation of

$$P = \frac{1}{1 + e^{-(5.5080 - 0.004923 \times 1000)}} = 0.6309, \text{ as in Fig. 33.1 (arrow).}$$

If the birth weight is 2000 g, then the probability is $P = \frac{1}{1 + e^{-(5.5080 - 0.004923 \times 1600)}}$, and this is 0.0856.

Repeat the previous calculations using gestational age as the explanatory variable (Fig. 33.3).

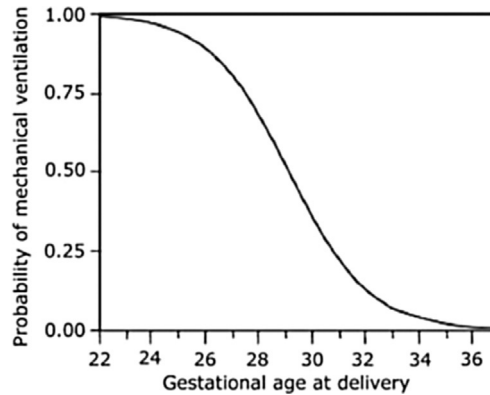


Fig. 33.3 Gestational age vs probability of artificial ventilation.

The analytic data are in Table 33.2.

Table 33.2 Gestational age (GA) in weeks

Model	–Log likelihood	Df	Chi-square	Prob >chi-square
Difference	76.11	1	152.21	<0.0001
Full	141.97			
Reduced	218.08			
R^2	0.3490			
Parameter estimates				
Term	Estimate	Standard error	Chi-square	Prob >chi-square
Intercept	19.40	2.1350	82.58	<0.0001
GA for log odds 0/1	–0.6662	0.07414	83.00	<0.0001

The fit is better as shown by an R^2 of 0.3490. The prediction equation is $P = \frac{1}{1 + e^{-(19.4011 - 0.6662GA)}}$. The probability of ventilation at a gestational age of 30 weeks is

$$\text{logit } P = 19.4011 - 0.6662 \times 30 = -0.5849.$$

Then taking antilogarithms of both sides gives

$$\frac{P}{1-P} = e^{-0.5849} = 0.5572.$$

Solving for P gives $P = 0.3572$ which fits nicely with the previous graph.

MULTIPLE EXPLANATORY VARIABLES

An advantage of logistic regression is that it allows the evaluation of multiple explanatory variables by extension of the basic principles. The general equation is

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}} = \frac{1}{1 + e^{-\left(\beta_0 + \sum \beta_i X_i\right)}}.$$

For example, consider predicting the probability of artificial ventilation from birth weight, gestational age, and maternal age. The statistics are presented in [Table 33.3](#).

Table 33.3 Probability of artificial ventilation

Model	–Log likelihood	Df	Chi-square	Prob >chi-square
<i>Birth weight</i>				
Difference	54.15	1	105.30	<0.0001
Full	163.92			
Reduced	218.07			
R^2	0.2483			
N	315			
<i>Gestational age</i>				
Difference	76.11	1	152.21	<0.0001
Full	141.96			
Reduced	218.07			
R^2	0.2490			
N	315			
<i>Maternal age</i>				
Difference	4.90	1	9.7936	0.0018
Full	213.17			
Reduced	218.07			
R^2	0.0225			
N	315			

All three explanatory factors on their own suggest rejecting the null hypothesis, with the best single predictor being gestational age. Although maternal age as a predictor is not due to chance, with $P=0.0018$, the R^2 value of 0.0225 shows that maternal age by itself can explain only 2.25% of the variability.

What happens if all the three explanatory factors are included in a single equation? Will it provide more information than any of the single regressions? Will it show if any of the variables are redundant? The results are presented in [Table 33.4](#).

Table 33.4 Regression with three variables

Model	=Log likelihood	Df	Chi-square	Prob >chi-square
Difference	79.77	3	159.7358	<0.0001
Full	138.30			
Reduced	218.07			
R^2	0.3662			
N	315			

Parameter estimates

Term	Estimate	SE	Chi-square	Prob >chi-square	Lower CL 95%	Upper CL 95%
Intercept	20.44	2.82	60.75	<0.0001	15.58	25.90
Birth weight (g)	−0.00087	0.00085	1.05	0.3062	−0.0025	0.00081
Gestational age (w)	−0.5994	0.1002	35.81	<0.0001	−0.81	−0.41
Maternal age (y)	−0.05919	0.02445	5.86	0.0155	−0.11	−0.12

g, grams; w, weeks; y, years.

The R^2 has increased slightly to 0.3662 from the highest single value of 0.3490 for gestational age alone. Birth weight is no longer a useful predictor; it has a small chi-square ($P = .3062$) and the confidence limits for its coefficient range from positive to negative. Therefore omit birth weight and produce a final regression with two variables (Table 33.5).

Table 33.5 Two variable regression

	−Log likelihood	Df	Chi-square	Prob >chi-square
Difference	79.37	2	158.69	<0.0001
Full	138.70			
Reduced	218.07			
R^2	0.3639			
Number	315			

Parameter estimates

Term	Estimate	Standard error	Chi-square	Prob >chi-square
Intercept	21.56	2.4171	60.75	<0.0001
	−0.67	0.0745	35.81	<0.0001
MA (y)	−0.06	0.02441	5.86	0.00134

BW, birth weight; GA, gestational age.

The final equation is

$$P = \frac{1}{1 + e^{-(21.5626 - 0.6704GA - 0.0604MA)}},$$

where GA is gestational age and MA is maternal age.

For gestational ages of 25, 30, and 35 weeks, and maternal ages of 20 and 40 years, the probabilities of needing artificial ventilation calculated from the formula are given in Table 33.6.

Table 33.6 Selected probabilities

Gestational age (weeks)	Maternal age (years)	Probability of artificial ventilation
25	20	0.9732
25	40	0.9158
30	20	0.5603
30	40	0.2758
35	20	0.1300
35	40	0.0139

In keeping with the coefficients that were determined, maternal age plays a small role in determining the need for artificial ventilation.

Alternatively, write

$$\log_e \frac{P}{1-P} = 21.5626 - 0.6704GA - 0.0604MA.$$

The value for β_0 of 21.5626 is the average risk of artificial ventilation independent of any explanatory variables. β_1 , the coefficient for gestational age, is 0.6704. The logit decreases by 0.6704 units for each week increase in gestational age if maternal age is constant.

APPROPRIATENESS OF MODEL

Is the statistical model, appropriate for the data? Other possible models, such as discriminant analysis and Hotelling's T^2 test have been used (Cupples et al., 1984). The logistic regression model has advantages over the other two in not needing normally distributed variables. Many different approaches have been used, (Glantz and Slinker, 2001; Lemeshow and Hosmer, 1982) and different programs implement these in different ways. In Table 33.3 the lack of fit is tested and the fit was found to be acceptable.

In previous chapters on regression several tests were discussed: plotting residuals or studentized residuals, leverage, Cook's distance, and others. Some of these tests can be applied to examining the results from logistic regression models (Glantz and Slinker, 2001). Furthermore, the model may contain interaction terms that also need to be evaluated.

Sample size and power calculations are as important here as in other statistical tests. The recommendation of Altman (1992) that the sample size should be at least 10 times the number of variables holds here as it does for multiple regression, with the exception that

the sample size refers to the number of positive events. If examining the outcome (death) of premature infants is related to 6 possible explanatory variables, then at least 60 deaths are needed for an adequate sample size with enough power.

Then, just like multiple regression methods, multicollinearity is a concern. For example, birth weight and gestational age are correlated. Will this interfere with the estimates? Check this to determine what the correlation is (Fig. 33.4).

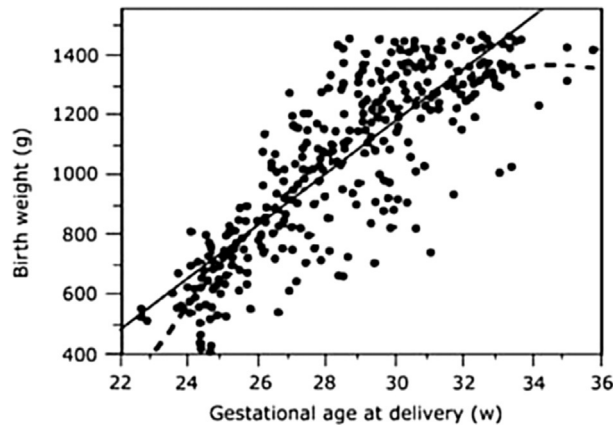


Fig. 33.4 Relationship between birth weight and gestational age.

For linear regression R^2 was 0.6632, and for quadratic regression it was 0.7034, and neither of these is high enough to suggest that multicollinearity will cause problems. If there is concern, then better estimates can be obtained by centering the data.

Because this is a form of multiple regression there is the choice of forward, backward, or step-wise regressions, with the same potential problems as discussed in [Chapter 30](#).

A self-learning book by [Kleinbaum and Klein \(2010\)](#) is excellent.

REFERENCES

- Altman, D.G., 1992. *Practical Statistics for Medical Research*. Chapman & Hall, London.
- Cupples, L.A., Heeren, T., Schatzkin, A., Colton, T., 1984. Multiple testing of hypotheses in comparing two groups. *Ann Int Med* 100, 122–129.
- Glantz, S.A., Slinker, B.K., 2001. *Primer of Applied Regression and Analysis of Variance*. McGraw-Hill, Inc, New York.
- Kleinbaum, D.G., Klein, M., 2010. *Logistic Regression. A Self-Learning Text*. Springer, New York.
- Lemeshow, S., Hosmer Jr., D.W., 1982. A review of goodness of fit statistics for use in the development of logistic regression models. *Am. J. Epidemiol.* 115, 92–106.

CHAPTER 34

Poisson Regression

INTRODUCTION

Poisson regression requires special statistical programs and is complex so that it is not featured in most basic textbooks. Nevertheless, it is often used in epidemiology, sociology, and psychology. In 2004–08 1568 articles using Poisson regression had appeared in the medical literature (Kleinman and Norton, 2009), so that investigators need to know when it is appropriate, how to interpret the results, and what precautions should be taken before using it.

Standard regression, linear or multiple, is used to predict a dependent variable that may be of any size and could be positive or negative from one or more explanatory variables that may be continuous ratio numbers or dummy variables. Logistic regression is used to predict a dichotomous variable, for example, survival or dying, from similar explanatory variables. What can be used for counted data that probably fit a Poisson distribution? Neither of the previous methods will be useful. The results are not dichotomous, and counts violate the requirements for standard regression in at least three ways (Coxe et al., 2009):

1. Poisson data are skewed, whereas standard regression requires a symmetric distribution of errors.
2. Standard regression can sometimes produce negative values that cannot occur in a Poisson distribution.
3. Whereas in standard regression the variance should be (reasonably) constant, in a Poisson distribution the variance increases in proportion to the mean.

Poisson regression overcomes these three problems by a logarithmic transformation that compensates for skewness, prevents a negative predicted value, and also includes the proportionality between variance and the mean.

If Y has a Poisson distribution, then a log-linear model can be constructed as $\ln \hat{Y} = a + \beta X$, and can be extended to several explanatory variables $\ln \hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_k X_k$.

In practice actual counts are needed. This can be handled by exponentiating both sides

$$e^{\ln \hat{Y}} = e^{(\alpha + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_k X_k)}.$$

Because $e^{\ln \hat{Y}} = \hat{Y}$, rewrite the equation as

$$\hat{Y} = e^{(\alpha + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_k X_k)}.$$

Now the predicted value of Y is in counts. This equation can be manipulated further to give

$$\hat{Y} = e^{\alpha} e^{\beta_1 X_1} e^{\beta_2 X_2} \dots e^{\beta_k X_k}.$$

If all the variables except X_i are held constant, a 1-unit change in X_i causes a change in predicted Y to $\hat{Y} e^{\beta_i}$.

Fitting the Poisson regression is done by maximum likelihood methods that are beyond the scope of this book. Examples of the use of these principles in epidemiological investigations, for example, of the risk of cancer and radiation exposure, are given by [Selvin \(1995\)](#).

Poisson regression is also used in comparing exposure rate ratios, defined as

$$\frac{\text{rate in exposed group}}{\text{rate in unexposed group}}.$$

Then the formula $\ln \hat{Y} = \alpha + \beta X$ is equivalent to $\log \text{rate} = \log \text{baseline} + \log \text{exposure rate ratio}$ ([Kirkwood and Sterne, 2003](#)).

In standard regression the departure from perfect prediction is assessed by the value of $1 - R^2$, but this cannot be calculated for Poisson regression. In its place, the maximum likelihood method calculates a *deviance* that represents variability. Deviance is a relative and not an absolute metric. A large deviance indicates a poor fit, whereas a small deviance indicates a better fit. The deviance is used to calculate a pseudo R^2 from ([Coxe et al., 2009](#))

$$R^2_{\text{deviance}} = 1 - \frac{\text{deviance}(\text{fitted mode})}{\text{deviance}(\text{intercept alone})}.$$

This pseudo R^2 varies from 0 to 1 and gets bigger as more explanatory variables are included. Variations on estimates of pseudo R^2 are also used ([Coxe et al., 2009](#)). The difference between two deviances obtained by adding in more explanatory variables is tested by chi-square, with degrees of freedom equal to the difference in the number of parameters tested ([Coxe et al., 2009](#)). More complex programs for assessing model adequacy, leverage, and outliers are available ([Coxe et al., 2009](#)).

SUITABILITY OF POISSON REGRESSION

Sometimes Poisson regression is inappropriate, for example, when the observed variance is much greater than the mean. This is termed overdispersion and is usually due to failure to include all the explanatory variables or to nonindependence between events, a condition referred to as a contagious distribution ([Chapter 19](#)). Poisson regression may be inefficient if there are too many zeros, something that is quite frequent. Both of these factors may be

present. For example, [Table 34.1](#) presents the frequency of 12-year-old children with decayed, missing, or filled teeth in an Iranian study ([Moghinbeigi et al., 2008](#)).

Table 34.1 Distribution of decayed, missing, and filled teeth (DMFT) in 12-year-old Iranian children

DMFT	Frequency
0	652
1	55
2	69
3	56
4	75
5	36
6	29
7	22
8	22
9	11
10	6
11	2
12	2
13	1
14	4
15	1
16	0
17	1
18	0
19	1

There were 1045 children, the mean number of DMF teeth was 1.598, and the standard deviation was 2.706. If this were a Poisson distribution, the number of children with different numbers of DMF teeth is presented in [Table 34.2](#) and compared with the observed numbers.

Table 34.2 Observed and expected numbers, based on Poisson distribution

Number of DMFT	Proportion from Poisson	Expected number from Poisson (E)	Observed number (O)	$O - E$	$(O - E)^2/E = \chi^2$
0	0.202	211.09	652	440.91	920.94
1	0.323	337.54	55	-282.54	236.50
2	0.258	269.61	69	-200.61	149.27
3	0.138	144.21	56	-88.21	53.96
4	0.055	57.47	75	17.53	5.35
5	0.018	18.81	36	17.19	15.71
6	0.005	5.22	29	23.78	108.33
>6	0.001	1.05	73	71.95	4930.29
Total	1.000	1045	1045	0.00	$\chi^2_T = 6420.35$

With the discrepancy shown, there is hardly any need to do a chi-square test to show that the two sets of numbers do not fit. This distribution of DMF teeth differs from a Poisson distribution in two respects: there is an excess of zeros, and the standard deviation exceeds the mean.

DETECTING OVERDISPERSION

Because overdispersion causes the standard deviation to exceed the mean, it needs to be tested for. Some statistics programs report in addition to the deviance the Pearson statistic (total chi-square) (Dallal, 2008). If the total chi-square is divided by the corresponding degrees of freedom, it provides an index of dispersion; if there is no overdispersion the ratio should be 1. (Remember that the 0.50 value is about the same as the degrees of freedom for any chi-square with >4 degrees of freedom.) Furthermore, the ratio deviance to degrees of freedom should also be 1. These two indexes do not necessarily agree (Dallal, 2008), but the chi-square method is favored.

Because the programs performing Poisson regression are complex and have to be selected with care, it is important to consider the advice given by Dallal (2008) "...virtually any sin that can be committed with least squares regression can be committed with Poisson and negative binomial regression. These include stepwise procedures and arriving at a final model by looking at the data."

CORRECTING FOR OVERDISPERSION

This is done most simply by using a scaling factor

$$\phi = \frac{\text{chi-square}}{\text{df}}.$$

The model then becomes a Poisson regression with mean μ and variance $\phi(\mu)$, and the standard deviation becomes $\sqrt{\phi(\mu)}$. The deviance of the new model becomes deviance/ ϕ , and being smaller, indicates a better fit.

In the example presented in Table 34.2,

$$\phi = \frac{6420.35}{6} = 1070.06, \text{ so that the standard deviation becomes } 32.72.$$

As an alternative, a negative binomial regression can be done (Coxe et al., 2009).

It may also be necessary to correct for excess numbers of zeros.

There are many possible variations for testing the assumptions and performing the analyses, and several different types of analysis that can be done. Apart from variants of the zero-inflated Poisson there are models for zero-inflated negative binomial regression that, in addition to the excess zeroes, can also take heterogeneity into account. In some data sets the events are very sparse, for example, the admission of children with

Kawasaki syndrome to hospital may occur on the average once in 10–12 days, and then a variation known as gamma regression has advantages. Consultation with a statistician is essential.

REFERENCES

- Coxe, S., West, S.G., Aiken, L.S., 2009. The analysis of count data: a gentle introduction to poisson regression and its alternatives. *J. Pers. Assess.* 91, 121–136.
- Dallal, G.E., 2008. Poisson Regression. Available: <http://www.jerrydallal.com/LHSP/Poisson.htm>.
- Kirkwood, B.R., Sterne, J.A.C., 2003. *Essential Medical Statistics*. Blackwell Science, Malden, MA.
- Kleinman, L.C., Norton, E.C., 2009. What's the risk? A simple approach for estimating adjusted risk measures from nonlinear models including logistic regression. *Health Serv. Res.* 44, 288–302.
- Moghinbeigi, A., Eshragian, M.R., Mohammad, K., McArdle, B., 2008. Multilevel zero-inflated negative binomial regression modelling for over-dispersed count data with extra zeros. *J. App. Stat.* 25, 1193–1202.
- Selvin, S., 1995. *Practical Biostatistical Methods*. Wadsworth Publishing Company, Belmont, CA.

SECTION VIII

Miscellaneous Topics

CHAPTER 35

Survival Analysis

BASIC CONCEPTS

Introduction

Statistical methods are used to determine time to failure in industry and have been adapted to medical purposes; the techniques are known as survival analysis. Survival may be defined as “the absence of a specific event after prolonged surveillance (Muenz, 1983).” Death is the prime example of an event, but other end points can be used: recurrence of a supraventricular arrhythmia after ablation, pacemaker failure, relapse after leukemia treatment, or readmission for congestive heart failure are all events. Survival analysis in these shows the rate at which failure or the event occurs. It might take many years before the arrhythmia returns, the pacemaker fails, or the leukemia returns. The question then becomes how to determine survival rates in a timely fashion? In two or more different techniques of ablation or cancer treatment, we would not want to wait for 50 years before deciding on the best technique.

The techniques used in Medicine are based on the Life table (survivorship) method. In general, the table or graph starts with the time of entry, t_0 , and 100% survival, and survival decreases with time. Occasionally the plot shows failures (deaths, events), and then the table starts at time t_0 and events 0%, and events increase with time.

The life table method probably started with the *Observations on the Bills of Mortality* published in 1662 by John Graunt (1620–74), and an article about life tables and annuities published in 1693 by the astronomer Edmund Halley (1656–1742). Details of the method can be found in textbooks of Epidemiology or Public Health, and many books (Selvin, 1991; Allison, 1995; Armitage et al., 2002). Survival analysis is a major tool used in clinical trials, and all the precautions needed for a successful trial need to be followed or else the statistical analysis will be fruitless. Perhaps the most easily understood reference sources for understanding such trials are the two excellent articles by Peto et al. entitled “Design and analysis of randomized clinical trials requiring prolonged observation of each patient (Peto et al., 1976, 1977).”

An early medical publication was the study of survival in systemic lupus erythematosus by Merrell and Shulman (1955). They emphasized the need to define the starting point from which the disease is followed and also observed that the accrual of patients over time meant that at any time point patients still alive would have been followed for different time periods, and that during the follow-up some patients will have become lost from the study without any information about their survival. These issues are illustrated in Fig. 35.1.

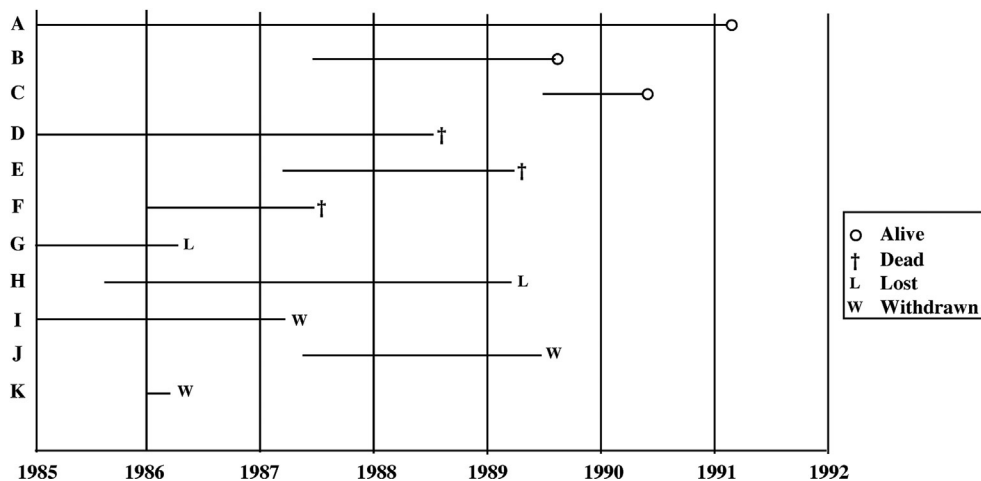


Fig. 35.1 Data set to illustrate differences in accrual data and fate.

Patients A, B, and C were all alive in 1989, 1990, and 1991, respectively, but A has been followed for 6 years, B for 2 years, and C for less than 1 year. Patients D, E, and F have all died, but at different times after entry into the study. Patients G and H were lost to the study, and could not be found, either because they had moved and left no new address or because they did not want to cooperate. Patients I, J, and K were all withdrawn at different times after entry because of death due to an unrelated disease, transfer to another treatment, or failure to maintain the treatment.

Merrell and Shulman developed a method of dealing with these problems, and Cutler and Ederer in 1958 extended the procedure to handle patients with different periods of follow-up (Fig. 35.2).

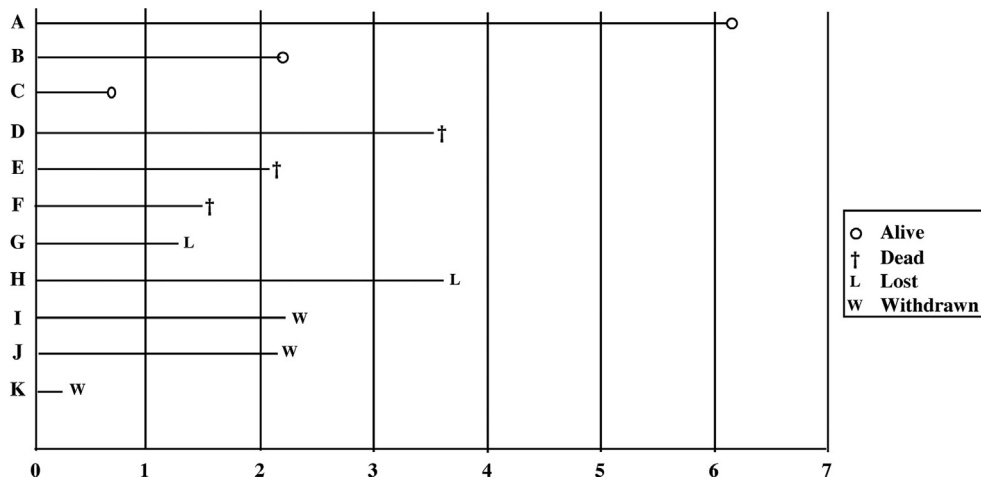


Fig. 35.2 Same data set as in Fig. 35.1 arranged so that all patients start at a common time, t_0 .

In 1958 Kaplan and Meier described a rigorous version of this method and termed it the product-limit method. The publications by [Anderson et al. \(1974\)](#) and [Grunkemeier and Starr \(1977\)](#) introduced these methods into surgical practice and revolutionized the way in which surgeons analyzed their results.

Basic Method

The basis of Berkson's actuarial method is to follow patients for a period, for example, 1 year, determine how many die in that year, and calculate the proportion p_1 who survive as

$$p_1 = 1 - \frac{\text{\#of deaths in period}}{\text{\#alive at beginning of period}}.$$

Then take those alive at the start of the second time period, determine how many die in this period, and calculate the probability p_2 of surviving the second period. Then the cumulative probability of surviving both periods is $p_1 \times p_2$, and so on. This method allows for subjects who have not yet completed the second period so that the base number for p_1 can be greater than the base number for p_2 .

[Table 35.1](#) shows how the data are presented.

Table 35.1 Basic life table

Start of period t_k	# Alive at start of period l_k	# Deaths in period d_k	# Alive < full period w_k	# Lost or withdrawn u_k	Adjusted number alive at start of period l'_k	Probability of dying in period q_k	Probability of surviving period p_k	Cumulative survival Πp_k
0-1	309	27	0	0	309	$27/309 = 0.0874$	0.9126	0.9126
1-2	282	14	6	2	278	$14/278 = 0.0504$	0.9496	0.8667
2-3	260	9	15	11	247	$9/247 = 0.0364$	0.9636	0.8352
3-4	225	5	13	6	215.5	$5/215.5 = 0.0232$	0.9768	0.8158

The symbols for the various columns vary in different texts.

The first column shows the period, (t_k). The second column shows how many patients entered the trial at the beginning of each period (l_k). The third column shows how many patients died in that period (d_k). When calculating the probability of dying, some patients did not complete a given period, either because they had been followed for less than a full period (w_k) or were lost to the study or withdrawn from it (u_k). The conventional approach is to assume that patients in columns w_k and u_k are evenly distributed across the period and so on the average contribute one-half of their number to be deducted from the total entering that period alive; the results would be biased if patients in columns D and E were treated as if all were alive or if all were dead. Therefore $l'_k = l_k - 0.5(w_k + u_k)$.

Then the probability of dying (q_k) is $q_k = \frac{d_k}{l'_k}$.

The next column shows the probability of surviving that period as $p_k = 1 - q_k$. This is then multiplied by the cumulative probability up to that period to give the probability of surviving period t_k given that the patient has already survived periods 1,2,3 ... t_{k-1} , and this is shown in the final column. Clear descriptions of this procedure are provided by [Anderson et al. \(1974\)](#) and [Kleinbaum \(1966\)](#).

The reasons for loss or withdrawal must be independent from the disease being studied. For example, in a study of breast cancer, a patient dying from a myocardial infarction is classified as withdrawn, not as death due to cancer. On the other hand, if the patient died from a hemorrhage probably caused by one of the treatments for cancer, this is not an independent event.

In the actuarial method the periods are usually years, but could be months, weeks, 5-year groups, or any other appropriate period. The Kaplan-Meier method is identical to the actuarial method except that the cumulative probability is recalculated whenever a failure occurs. Therefore graphs drawn by the actuarial method show evenly spaced probabilities as against irregularly spaced values with the Kaplan-Meier method.

[Fig. 35.3](#) shows a survival plot for the persistence of sinus rhythm in children with episodes of supraventricular tachycardia after ablation of pathways by two different sized cooling tips; they were followed for the duration of the study. (Courtesy of Dr. F. Collins and Dr. N. Chanani).

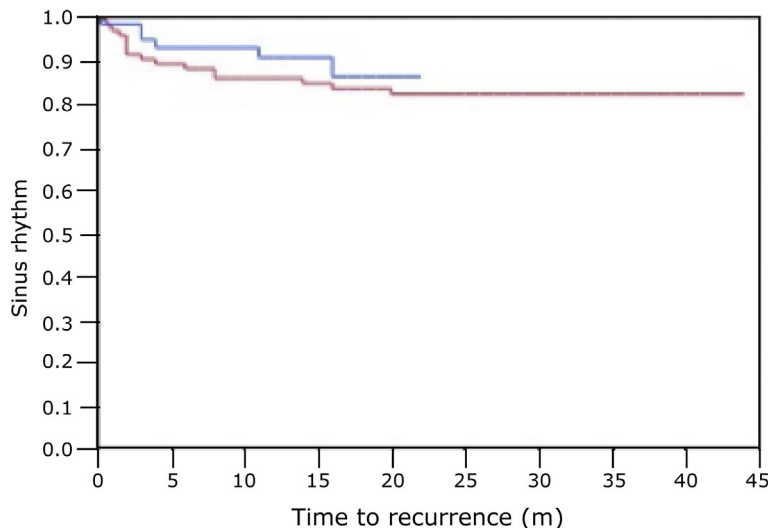


Fig. 35.3 Kaplan-Meier plot comparing small 4 mm tip (lower line) with larger 6 mm tip (upper line). It cannot be an actuarial plot because the intervals at which the lines change are irregular.

Patients entered the study at different times, and thus were followed for different periods. A partial data set is presented in Table 35.2.

Table 35.2 Partial data set for ablation study

	ID	Date of study	Group (tip size)	Censored	Time to recurrence (m)
1	1	10/27/05	0	1	23
2	2	6/16/05	0	0	3
3	3	5/16/05	0	1	28
4	4	6/15/04	1	1	39
5	5	8/1/05	1	0	16
6	6	6/6/05	0	1	28
7	7	4/20/06	0	1	17
8	8	2/13/06	1	1	19
9	9	10/17/05	1	1	23
10	10	5/23/06	1	0	0.75
11	11	8/16/05	0	1	25
12	12	12/9/04	1	1	34

Column 1 is the identifier (nominal variable), which could be an admission hospital number or an arbitrary number or letter. Column 2 is the date of study. Column 3 is tip size (nominal variable) with 0 being 4 mm and 1 being 6 mm diameter. Column 4 indicates censoring, with 0 being recurrence and 1 being no recurrence. (The choice of 0 or 1 for censored data is arbitrary. Programs allow for either.) Column 5 is the time to recurrence (continuous variable). (Other columns of data in the study are not reproduced here.) If the patient had a recurrence, indicated by a 0 in the censor column, then the time shows when recurrence occurred. If the patient had not yet had a recurrence when the data were analyzed, as indicated by a 1 in the censor column, the time shows the length of follow-up to that date. Some of these patients may develop recurrences in the future, and because they have not been followed until failure has occurred, they are described as right censored. They contribute information about how long patients can be followed without failure, but there is no way to know how much longer this state will continue. The censor variable 0 or 1 tells the program which patients have and have not had an event.

Some investigators provide details of the numbers of successes and failures at each time point. Fig. 35.4 gives a hypothetical example.

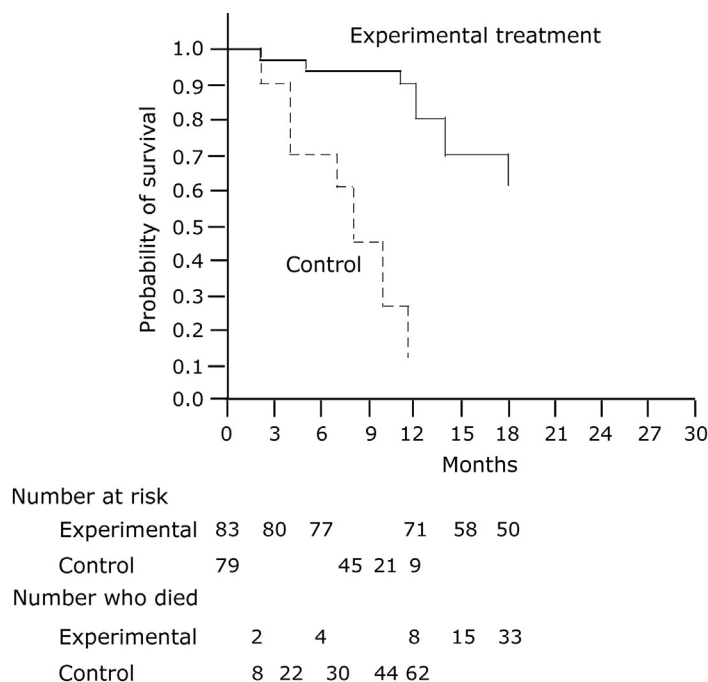


Fig. 35.4 Comparison of experimental and control treatments, with supplementary numbers provided.

Occasionally these figures display smoothed curves, but these are approximations to the actual data.

Online programs for these plots have two formats. One requires entry for each subject line by line, as in Table 35.2, and these can be done via an Excel file (see http://www.ehow.com/how_8369388_make-survivorship-curve-chart-excel.html, <http://in-silico.net/statistics/survivor>, with instructions for performing the test, https://www.statstodo.com/Survival_Pgm.php (but this needs membership for large data sets)). The other allows entry of the life table, as in Table 35.1, and can be found at <http://vassarstats.net/survival.html> and <http://eurekastatistics.com/kaplan-meier-survival-curve-grapher/>. Files must be set up as text files. None of these are easy to use.

Problem 35.1 Two groups of patients are randomized to receive either conventional therapy (C) or a new experimental therapy (E) for breast cancer. Use the Kaplan-Meier method to decide if the new treatment is better.

Experimental treatment			Conventional treatment		
Time of event t (months)	# at start of the month	# who died at time t	Time of event t (months)	# at start of the month	# who died at time t
1	25	0	1	49	0
5	24	0	3	48	0
6	23	2	5	47	0
9	21	0	6	46	8
10	20	2	8	38	2
12	17	4	9	36	0
13	12	1	10	36	2
15	11	0	12	34	6
16	10	0	13	28	0
20	9	0	15	25	0
24	8	1	16	22	0
27	7	0	18	21	1
32	5	1	20	18	1
44	5	0	22	17	0
			24	16	2
			27	13	0
			28	12	0
			30	9	1
			32	7	1
			33	6	0
			34	5	0
			36	4	0
			42	2	1
			44	2	0

Confidence Limits

The survival curve is a point estimate at each time period, and similar samples would have slightly different data. The confidence limits at each point can be calculated from

$$se = S(t) \sqrt{\sum_{j=1}^k \frac{d_j}{(I_j - d_j)}}$$

The results of these calculations are shown as se_{exact} in [Table 35.3](#).

Table 35.3 Calculations of standard error

Period	$se_{\text{(exact)}}$	Peto
0–1	0.0161	0.0156
1–2	0.0194	0.0196
2–3	0.0214	0.0226
3–4	0.0225	0.0247

In commercial programs, the standard errors are printed out for each group. A partial printout is given for the 4 mm tip in the ablation example (Table 35.4). Standard errors or confidence limits may be calculated online at https://statcom.dk/K-M_plot.php, http://iscc-serv2.imm.dtu.dk/%7Emerser/K-M_plot.php, <http://www.hutchon.net/Kaplan-Meier.htm>, and <http://vassarstats.net/survival01.html#next>.

The final column in Table 35.3 headed “Peto” gives a simple approximation described by Peto et al. (1977). For any value of the computed probability of survival Π_{p_k} (final column of Table 35.1), the standard error is approximately

$$SE\hat{\Pi}_{p_k} = \pi p_k \sqrt{\frac{1 - \pi p_k}{N}},$$

where N is the number of patients alive at the end of the year in question. The approximation is close to the exact value.

Table 35.4 Partial printout of data for the 4 mm tip used for ablation

4						
Time to recurrence (months)	Survival	Failure	SurvStdErr	Number failed	Number censored	At Risk
0.0000	1.0000	0.0000	0.0000	0	0	91
0.0000	1.0000	0.0000	0.0000	0	1	91
0.5000	0.9889	0.0111	0.0110	1	0	90
0.7500	0.9778	0.0222	0.0155	1	0	89
1.0000	0.9667	0.0333	0.0189	1	0	88
1.5000	0.9556	0.0444	0.0217	1	0	87
2.0000	0.9111	0.0889	0.0300	4	0	86
3.0000	0.9000	0.1000	0.0316	1	0	82
4.0000	0.8889	0.1111	0.0331	1	0	81
6.0000	0.8778	0.1222	0.0345	1	0	80
7.0000	0.8778	0.1222	0.0345	0	1	79
8.0000	0.8553	0.1447	0.0371	2	0	78
9.0000	0.8553	0.1447	0.0371	0	1	76
10.0000	0.8553	0.1447	0.0371	0	1	75

The standard error increases progressively.

Some programs allow confidence limits to be plotted on the graph.

Comparison of Different Survival Curves

This is most often done by the log-rank test of Mantel, although similar methods based on a modified Wilcoxon test can also be performed. The log-rank test examines the 2×2 table consisting of observed vs expected failures in each group whenever a failure occurs and tests the null hypothesis across all the tables with a Mantel-Haenszel test. The test can be extended to more than two groups.

The test statistic can be formalized as

$$\chi^2_{\log \text{rank}} = \sum_g \frac{(O_g - E_g)^2}{E_g}, \text{ where } O \text{ and } E \text{ are the observed and expected events}$$

in group g , calculated each time an event occurs.

The log-rank test places more weight on longer survival times, and the Wilcoxon tests place more weight on early survival times because they obtain a weighted average of each $O - E$ deviation by using the number of survivors in the group at each time. Different investigators have recommended different weighting systems. If the publication does not mention which form of Wilcoxon's test was used, it matters very little because they all give similar results. Furthermore, it would be folly to reject the null hypothesis at $p=0.045$ with one test but not reject it if another test had $P=0.067$.

In the ablation study described before, (Table 35.5).

Table 35.5 Comparison between groups
Tests between groups

Test	Chi-square	DF	P
Log-rank	0.6330	1	0.4262
Wilcoxon	0.8432	1	0.3588

In this example, neither test suggests rejecting the null hypothesis.

Caveats: If patients are accrued over several years, there must be stationarity, that is, all features of the patients and treatments must remain constant. Then it is unwise to extrapolate beyond the observed curves. It is possible for curve A to show better survival than curve B for 10 years, but then for survival to deteriorate in later years in group A and by 20 years be worse than the final survival for group B. Finally, because there will be smaller numbers of patients followed for the longest times, a single failure might have a large effect; for example, if at 5 years there are 100 patients, one failure changes the probability of survival little, but one failure out of 3 survivors at 10 years produces a large decrease in survival. Although this is obvious and would be made more obvious by examining the confidence limits, many reports do not include these limits.

Sample Size

Before beginning the study, an attempt should be made to estimate the likely sample size needed. This can be done online at <http://www.quesgen.com/SSSurvival.php>, https://www.statstodo.com/SSizSurvival_Pgm.php.

As for other sample size estimates, some guesses or provisional data must be obtained for the likely event rate and the proportion in each group.

ADVANCED CONCEPTS

Calculating the Log-Rank Test

The test involves pooling the two groups and arranging the survival times in rank order. Then a series of 2×2 tables is constructed each time there is a failure (Table 35.6).

Table 35.6 Components of 2×2 table

Group	Failure	No failure	Total exposed
A	O_A	$I'_A - O_A$	I'_A
B	O_B	$I'_B - O_B$	I'_B
Total	O_T	$I'_T - O_T$	I'_T

As in any chi-square table, the expected value for failure in group A (E_A) is

$$E_A = \frac{O_T I'_A}{I'_T}.$$

As a rule, because a new table is set up each time an event occurs, O_T is usually 1, and O_A and O_B are either 0 or 1. Therefore $E_A = \frac{I'_A}{I'_T} = p_A$.

The calculations are shown in the hypothetical example (Table 35.7).

Table 35.7 Data set for Mantel-Haenszel log-rank test

Time	I'_A	I'_B	I'_T	O_A	O_B	O_T	E_A	$O_A - E_A$	V	E_B	$O_B - E_B$
3	35	45	80	0	1	1	$\frac{35 \times 1}{80} = 0.4375$	-0.4375	0.2461	0.5625	0.4375
7	35	44	79	0	1	1	$\frac{35 \times 1}{79} = 0.4430$	-0.4430	0.2468	0.5670	0.433
8	35	43	78	0	2	2	$\frac{35 \times 2}{78} = 0.8974$	-0.8974	0.2442	1.1026	0.8974
11	35	41	76	1	0	1	$\frac{35 \times 1}{76} = 0.4605$	0.5395	0.2484	0.5395	-0.5395
29	34	41	75	0	1	1	$\frac{34 \times 1}{75} = 0.4533$	-0.4533	0.2478	0.5467	0.4533
54	34	40	74	1	2	3	$\frac{34 \times 3}{74} = 1.3784$	-0.3784	0.2416	1.6216	0.3784
Σ				2	7	9	4.0701	-2.0701	1.4749	4.9399	2.0601

I'_A , I'_B , and I'_T are the number of subjects in group A, group B, and the total, respectively; O_A , O_B , and O_T are the number of failures (deaths or other events) in group A, group B, and the total, respectively; $E_A = \frac{O_T I'_A}{I'_T}$.

The variance V of the difference $O_A - E_A$ is calculated as.

$$V = \frac{I'_A I'_B (I'_T - O_T)}{I'^2_T (I'_T - 1)} \quad (\text{Altman, 1992}). \text{ When a single event occurs, this reduces to}$$

$$V_{O_A} = \frac{(I'_T - 1) I'_A I'_B}{(I'_T)^2 (I'_T - 1)} = \frac{I'_A I'_B}{(I'_T)^2} = \frac{I'_A}{I'_T} \times \frac{I'_B}{I'_T} = p_A p_B = p_A (1 - p_A).$$

Then calculate chi-square as

$x^2 = \frac{[\sum (O_A - E_A)]^2}{V} = \frac{-2.0701}{1.4749} = -1.4036$. In this example, the 0.05 value of chi-square of 3.84 has not been reached, although the trend seems to be for more events to

occur in group B. The log-rank test examines cumulative $O-E$ differences for one group. If the two groups are not very different, then sometimes the $O-E$ differences will be positive (an event occurs in group A) or sometimes negative (an event occurs in group B). If these are random occurrences, then the sum will be small and indicate no substantial differences.

An alternative calculation that does not involve V is.

$$\chi^2 = \frac{(\sum O_A - \sum E_A)^2}{\sum E_A} + \frac{(\sum O_B - \sum E_B)^2}{\sum E_B}, \text{ where } \sum E_B = \sum O_T - \sum E_A.$$

This formula gives a slightly different answer

$$\chi^2 = \frac{-2.0701^2}{4.0701} + \frac{2.0601^2}{4.9399} = 1.9120.$$

This variant is regarded as more conservative than the previous one (Armitage et al., 2002).

Examples of the calculations are given by Armitage et al. (2002) and by Altman (1992). The log-rank test can be done with more than two data sets, and on data sets stratified into subgroups. The test can be performed online at http://iscc-serv2.imm.dtu.dk/%7Emerser/K-M_plot.php or <https://merser.shinyapps.io/Survival/>.

The Hazard Function

The survival function $S(t)$ is $S(t) = \frac{\text{number of patients surviving more than time } t}{\text{total number of patients in study}}$.

This is also termed the cumulative survival function because it gives the probability of surviving all time intervals from the start to the selected time point.

When following a group of patients over time, some patients die between times t_x and t_{x+1} . The instantaneous hazard function $h(t)$ is the risk of dying (or failure in general) in a very small time period Δt if the patient has survived to time t_x . It represents the risk of dying in subjects who have survived up to that time. It can be symbolized formally by

$$\widehat{h(t)} = \lim_{\Delta t \rightarrow 0} \frac{P(t_x \leq T \leq t_x + \Delta t \mid T \geq t_x)}{\Delta t}.$$

This is the conditional probability that a person who has survived for t_x periods ($T \geq t_x$) will die between the short period t_x and $t_x + \Delta t$. It is the slope of the survival curve at any given time t and can also be written as

$$\widehat{h(t)} = - \left[\frac{dS(t)/dt}{S(t)} \right].$$

The hazard function is a probability per unit time and is thus a rate.

There is also a cumulative hazard function that adds up all the instantaneous hazard functions to a given point. It can be calculated as $\Delta t = -\log_e S(t)$ and produces a curve that is almost a mirror image of the cumulative survival curve. See Fig. 35.5.

As an example, if in the month after a coronary bypass operation the hazard rate is calculated as 0.015, then on average 0.015 patients are expected to die in the first month. In a database of 1000 patients, 15 will die in the first month. This figure does not apply to subsequent months that might have different hazard functions.

Typically, in examining the results after cardiac surgery, curves as in Fig. 35.5 are shown.

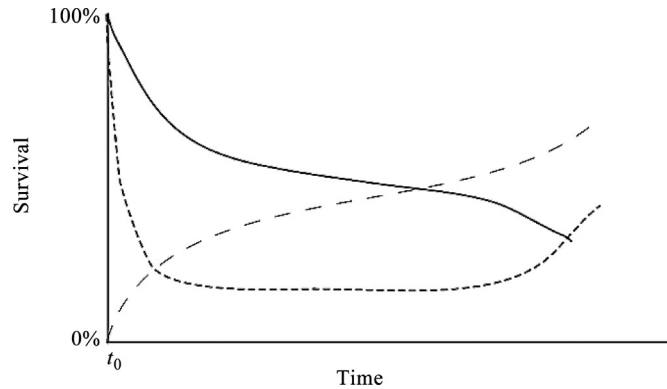


Fig. 35.5 Survival $S(t)$ (solid line), instantaneous hazard $h(t)$ (short dashed line), and cumulative hazard $\Delta(t)$ (long dashed line) curves. Immediately after surgery there is a rapid decrease in survival, confirmed by the steep initial hazard curve. After that there is a period of very slow decrease in survival where the hazard function is low and steady, and then the hazard function increases as late complications become manifest. The instantaneous hazard function in the middle of the figure is low, but the cumulative hazard function continues to rise because patients die from time to time.

These hazard functions can be fitted to different distributions that might throw some light on the underlying processes (Altman, 1992).

If the survival curve is approximately exponential, as it often is, then the hazard rate H is estimated from

$$H = \frac{d}{\sum f + \sum c}$$

where d is the number of failures, $\sum f$ is the sum of the failure times, and $\sum c$ is the sum of the times of the censored individuals. The reciprocal of H is the estimate of mean survival time

$$\bar{t} = \frac{1}{H}.$$

These two formulas can be used to set approximate confidence limits. The 95% confidence limits for mean survival time are $\bar{t} \pm 1.96 \times \text{se}_X$, and the standard error can be

approximated by $\sqrt{\frac{X^2}{d}}$.

Hazard Ratio

The log-rank test, like any other test of the null hypothesis, indicates whether survival between two groups is different enough to reject the null hypothesis, but does not

indicate how different they are. One way to show the difference is to state the survival in each group at comparable times, for example, 56% at 10 years in group A as against 32% at 10 years in group B. Another way is to compare observed and expected numbers in each group. If the survival is the same in the two groups, then the number of observed events will be proportional to the expected numbers and will be the same in each group. To examine this the hazard ratio is computed. This ratio R is

$$R = \frac{\sum O_A / \sum E_A}{\sum O_B / \sum E_B},$$

where O and E are the observed and expected numbers of events, and A and B indicate the groups being compared. For the example in Table 35.7, the hazard ratio is

$$R = \frac{2/4.0701}{7/4.9399} = 0.3468. \text{ If the hazard ratio is } >1, \text{ it indicates that the treatment group}$$

has a shorter survival than the control referenced group, and if it is <1 , it indicates that the group of interest is less likely to have a shorter time to the event than the reference group. The ratio does not quantify the magnitude of the difference.

It is possible to determine approximate confidence limits for the logarithm of the hazard ratio (Simon, 1986). Calculate

$$K = \frac{\sum O_A - \sum E_A}{\sum V}$$

an estimate of the logarithm of the hazard ratio with an approximate standard error of $\frac{1}{\sqrt{\sum V}}$.

For the data in Table 35.7

$$K = \frac{2 - 4.0701}{1.4749} = -1.4036. \text{ and } \frac{1}{\sqrt{\sum V}} = \frac{1}{\sqrt{1.4749}} = 0.8234.$$

The 95% confidence limits of K are $K \pm 1.96 \times 0.8234 = -1.4036 \pm 1.6139 = -3.0175$ to 0.2013 . Therefore the 95% confidence limits for the hazard ratio are $e^{-3.0175}$ to $e^{0.2013}$, or 0.0489 to 1.2230. These limits are wide, include 1, and confirm the findings that the two groups are not substantially different.

Spruance et al. (2004) emphasized that the hazard ratio quantifies the degree of difference between the groups but does not indicate what the absolute difference in duration of the illness is. (In a race, for example, the hazard ratio gives the probability that A will win but does not tell us by how much.) They wrote: “The hazard ratio is equivalent to the odds that an individual in the group with the higher hazard reaches the endpoint first.” In a trial of treatment to shorten the duration of symptoms in herpes zoster, for

example, the hazard ratio represents the odds that the time to remission of symptoms is less in a patient from the treatment than from the control group. The probability of getting better first is the odds of healing first divided by the probability of not healing first. Therefore in this setting,

$$\text{Odds of healing first (HR)} = \frac{\text{Probability of healing first}}{\text{Probability of not healing first}} = \frac{P}{1 - P}. \text{ This can be}$$

rearranged to give $\text{Probability of healing first} = \frac{\text{HR}}{1 + \text{HR}}.$

Thus a hazard ratio of 2 matches a 67% chance that the treated patient will heal first, and a ratio of 3 corresponds to a 75% chance of healing first. The actual difference in time to healing requires absolute numbers, such as the median ratio that may give values that are not the same as the hazard ratio. It is possible for a hazard ratio to be greater or less than the median ratio.

Cautionary Tales

[Hernán \(2010\)](#) pointed out that in epidemiologic studies mean hazard ratios are usually cited, but this practice ignores the possibility that the hazard ratios change with time. In the Women's Health Initiative, for example, in which hormone therapy was compared with a placebo, the successive annual hazard ratios for coronary heart disease were 1.81, 1.34, 1.27, 1.25, 1.45, and 0.70. Therefore there will be different averages, depending on how long the study continues and thus how many hazard ratios are averaged. Examining year-specific hazard ratios overcomes this problem but leads to another one related to selection bias. Some people are more prone to develop heart disease than are others. These may be detected early, leaving a pool of slightly less susceptible people to enter the next period. After several years the control and treated groups are no longer comparable and the hazard ratio may decrease to 1 or below 1.

[Singer \(1994\)](#) pointed out that if after an acute phase of decreased survival, the two survival curves were virtually parallel, this did not mean that each set had the same annual mortality. Because the lower of the two curves was based on a smaller number of survivors, the mortality rate was higher.

Cox Proportional Hazards Regression

What factors influence the survival curves? In assessing survival after a myocardial infarction, how important are covariates such as the age at time of infarction, body mass index, diabetes, serum LDL concentration, and blood pressure; in assessing survival with cirrhosis of the liver, how important are serum albumin, alkaline phosphatase, alcohol intake; in assessing survival with kidney disease, how important are age, gender, creatinine concentration, hemoglobin concentration, serum albumin? One way of assessing these variables is to realize that the survival function $S(t)$ has values between 0 and 1, and thus might be suitable for logistic multiple regression analysis.

However, logistic regression does not take survival times or censoring into account and is replaced by the Cox regression model. (Sir David Cox, b.1924, an eminent British statistician.) This is a robust nonparametric model. Cox's publication is the most highly cited reference in the entire literature of statistics and ranks among the top 100 publications in all of science (Altman, 1992). Cox regression helps to adjust for imbalance between the groups.

One way of analyzing the data is to assume that the hazard functions related to the two survival curves are proportional to each other, that is, if the hazard function for curve A is 20% less than that for curve B at a given time, it is approximately 20% less at other times too. More generally, $h_A(t) = kh_B(t)$ where k is a constant. The survival curves for these two groups can be related to the equation $S_B(t) = [S_A(t)]^k$ (Selvin, 1991; Glantz and Slinker, 2001).

If the proportional hazards requirement is met, then Cox regression is applicable. Cox regression is semiparametric, applies to many different survival functions, and its requirements are quite robust, but would certainly not apply if the two curves crossed.

Cox regression assumes that the ratio of the two hazard functions R is logistic, so that. $R = h_0(t)e^{\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k}$. The proportionality constant k can therefore be written as

$$k = (e^{\beta_1 X_1}) (e^{\beta_2 X_2}) \dots (e^{\beta_j X_j}).$$

When all the β coefficients are zero, $e^0 = 1$, and $h_0(t)$ is the baseline, time-dependent hazard function¹. Furthermore, if all the β coefficients except one ($e^{\beta_i X_i}$) are set to zero, the value of $e^{\beta_i X_i}$ represents the hazard function for that variable alone. The exponential components that are the potential explanatory variables are subject- but not time-dependent. (Although weight, age, serum lipoprotein concentrations do change with time, in this type of study each contributes only one fixed value.)

Taking logarithms of both sides gives

$$\log R = \log \frac{h_A(t)}{h_B(t)} = h_0(t) + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k.$$

Once the data are fitted to the model, the calculations estimate the coefficients β_1, β_2 , and so on, together with their standard errors so that the hypothesis that each coefficient is zero can be tested. (Precautions for dealing with multiple regression and multicollinearity are involved in these calculations.) Then the contributions of each variable can be assessed for the change in hazard. A positive coefficient for any variable, for example, 2, means that the variable makes survival worse relative to the reference group, and a negative coefficient, for example -0.35 , means that the factor improves survival.

As an example, consider the arrhythmia ablation data presented before, and add in age and number of cryoablation lesions as possible explanatory variables (Data not shown). The analysis is given in Table 35.8.

¹ This is the same for both groups.

Table 35.8 Cox regression for ablation data

#events				18					
# censored	124			102					
Total	148			120					
Model	−log likelihood	Df	Chi-square	Model	−log likelihood	Df	Chi-square		
Difference	0.3260	1	0.6521 <i>P</i> =0.4194	Difference	3.3948	3	6.7897 <i>P</i> =0.0789		
Full	106.0179			Full	80.3806				
Reduced	106.3439			Reduced	83.7784				
Parameter estimates									
Term	Estimate	SE	CL	CU	Term	Estimate	SE	CL	CU
Group	−0.1905	0.2420	−0.7068	0.2600	Group	−0.2716	0.3974	−1.2208	0.4192
					# cryo	0.1227	0.04806	0.0143	0.2101
					Age	0.0253	0.0651	−0.0828	0.1340
Effect likelihood ratio tests									
Source	Df	Chi-square	<i>P</i>	Source	Df	Chi-square	<i>P</i>		
Group	1	0.6521	0.4194	Group	1	0.5185	0.4715		
				#cryo	1	4.7764	0.0289		
				Age	1	0.2105	0.6464		
Unit risk ratios									
Term	Risk ratio	CL	CU	Reciprocal					
Group	0.7622	0.2950	1.5208	1.3120					
# cryo	1.1306	1.0144	1.2334	0.8845					
Age	1.0260	0.9205	1.1434	0.9750					

SE—standard error, CL—lower 95%confidence limit, CU—upper 95% confidence limit. #cryo—number of attempts to ablate focus.

The left-hand panel shows the results for survival time in the two tip sizes (group) with no other explanatory variables added. As shown by the chi-square and probability for the Model, the differences do not favor rejecting the null hypothesis. The parameter estimate (−0.1905) indicates that arrhythmias in the new group (6 mm tip) are 19% less likely to recur than in the old group (4 mm tip), but with a wide standard error. The risk ratio was 0.8265 with wide confidence limits.

In the right-hand panel, age and number of attempts at ablation (# cryo lesions) have been added. As shown under Model, chi-square has increased. Tip size has a larger effect (parameter estimate −0.2716 or −27% difference), age has no effect (small estimate, large standard error), but lesion number increases failure rate by 13% per unit increase and has a small *P* value (see lowest panel). As shown in the lowest section, a 1-unit increase in lesion number increases the hazard function by 13% (but with 95% confidence limits of 1%–23%). Because age is not an explanatory variable, delete it and leave in only tip size and lesion number (Table 35.9).

Table 35.9 Only 2 explanatory variables from ablation data

Total Model	120 −log likelihood		Df	Chi-square
Difference	3.28958		2	6.5792 <i>P</i> =0.0373
Full	80.48888			
Reduced	83.77846			
Parameter estimates				
Term	Estimate	SE	CL	CU
Group	−0.2454	0.3937	−1.1901	0.4357
% cryo	0.1198	0.0482	0.01223	0.2051
Effect likelihood ratio tests				
Source	Df	Chi-square	<i>P</i>	
Group	1	0.4289	0.5125	
#cryo	1	4.6431	0.0312	
Unit risk ratios				
Term	Risk ratio	CL	CU	Reciprocal
Group	0.7824	0.3042	1.5460	1.2781
#cryo	1.1272	1.0124	1.2277	0.8871

As Model shows, adding in lesion number has made an improvement as shown by the higher chi-square and low P value. The 6-mm tip size now has a 22% lower failure rate as compared with the reference group with 4-mm tip size (Risk ratio 0.78), and the lesion number increases failure rate by 12.7% (Risk ratio 1.127). $e^{-0.2453694}=0.782415$, and $e^{0.11975937}=1.127226$, that is, e to the coefficient gives the hazard ratio for that variable.

Knowing that the number of lesions affects the survival curves, keep that data in the final model and have a more accurate hazard ratio for tip size that is now not influenced by any effect that the number of lesions has on the outcome. If there were data with the same numbers of lesions in the 2 groups, then lesion number would no longer be an influencing or confounding factor and no allowance for it would be needed. More complex analyses involving interactions among the explanatory factors can also be performed.

Online calculations can be done at <http://statpages.org/prophaz.html>, but this is not very flexible. There are good descriptions of Cox regression in nephrology and hepatology (Christensen, 1987; Elashoff, 1983; Schlichting et al., 1983; van Dijk et al., 2008).

Competing Risks Analysis

If a group of elderly patients having replacement of the aortic valve with a mechanical prosthetic valve are followed, the survival curve drops off quite rapidly, but not all

the deaths are due to complications or aortic valve surgery. Older people have a higher incidence of cancer or renal disease that may cause death. Therefore the survival curve is due to all causes of death, not just postsurgical causes. Epidemiologists define two probabilities of dying: crude probability refers to death of an individual in the presence of multiple causes of death, whereas net probability refers to death from a particular cause when other causes of death are not present. Survival after aortic valve surgery would differ if some people did not die from cancer or accidents.

Two approaches are used to evaluate competing risks (Selvin, 1991). One approach assumes that the net probabilities can be described by exponential functions with hazard rates of λ_1 and λ_2 , net probability rates of $Q_1 = 1 - e^{-\lambda_1}$ and $Q_2 = 1 - e^{-\lambda_2}$, and that the probability of surviving an interval $= P_1 P_2 = (1 - Q_1)(1 - Q_2) = e^{-(\lambda_1 + \lambda_2)}$.

This implies that causes 1 and 2 are independent, an assumption probably true for aortic valve surgery and cancer, but perhaps not true for aortic valve surgery and renal failure.

Crude probability of death from either cause is $q = 1 - P_1 P_2 = 1 - e^{-(\lambda_1 + \lambda_2)}$.

Because $P_i = 1 - Q_i$,

$$Q_i = 1 - P_i = 1 - (1 - q) \frac{\lambda_i}{\lambda_1 + \lambda_2}.$$

The ratio of the two hazard rates $\frac{\lambda_i}{\lambda_1 + \lambda_2}$ is estimated by $\frac{d_i}{d_1 + d_2}$, where d_i is the number of deaths from cause 1 and $d_1 + d_2$ is the total number of deaths from all causes. This estimate is used to determine the net probability of death from cause 1 as

$$Q_i = 1 - \left(1 - \frac{d}{l}\right) \frac{d_i}{d_1 + d_2}$$

where l is the number of individuals at risk at the beginning of the interval. This is less than the crude rate based on d by virtue of the quantity in parentheses. For example, if $l = 150$, $d_1 = 5$, and $d_2 = 7$, the net death rate will be

$$\hat{Q}_i = 1 - \left(1 - \frac{12}{150}\right) \frac{5}{12} = 0.0341.$$

This is less than the crude death rate that is $12/150 = 0.08$.

An alternative that does not involve the exponential assumptions is to regard individuals at risk as dying from cause 1, dying from cause 2, or living through the interval. Considering those who die from cause 2 as being lost to follow-up, then deaths from cause 1 are underestimated because if subjects had not died from cause 2 they would be exposed to the risk of dying from cause 1. On the usual assumption that the lost individuals are exposed to risk for half the interval, the number exposed to risk of dying from cause 2 had been decreased by $0.5d_2$. Therefore corrected net probability of death from cause 1 is

$$\hat{Q}_i = \frac{d_i}{l - 0.5d_2}.$$

Using the numbers from the previous example, $\hat{Q}_i = \frac{5}{150 - 0.5 \times 12} = 0.0347$, similar to the value obtained by the previous formula.

The problem of competing risks is prominent in evaluating the results of heart valve replacement surgery. Younger patients who do not have many competing causes of death may have a postoperative survival curve that shows the value of the operation. Investigators have then used the same life table methods in older patients. Unfortunately, if patients have died it is not possible to tell if they would have had valve failure at a later stage, so that estimates of valve failure will be incorrect (Grunkemeier et al., 1977, 2007; Blackstone, 1999; Blackstone and Lytle, 2000; Grunkemeier and Wu, 2001; Austin et al., 2016).

There is much discussion in the medical literature about how to deal with competing risks (Grunkemeier et al., 1997; Blackstone, 1999; Blackstone and Lytle, 2000; Grunkemeier et al., 2001; Kaempchen et al., 2003; Bodnar and Blackstone, 2005; Grunkemeier et al., 2007). Consider a simple example of 200 patients followed after aortic valve replacement. If 50 of them die early from noncardiac causes before any valve failure, then there are only 150 patients who remain at risk for valve failure. If after 20 years 15 patients have needed valve replacement, then the risk of valve failure is 10% after 20 years. On the other hand, if none had died of noncardiac causes, we would have followed 200 patients and might (who knows?) have observed 25 valve (12.5%) failures. On the other hand, there might still have been 15 failures for a final percentage of 7.5% after 20 years. Both values are correct but based on different considerations.

Grunkemeier et al. (1977, 1997, 2001, 2007), Grunkemeier and Starr (1977), and Grunkemeier and Wu (2001) have argued in favor of determining cumulative incidences of each of the competing events, something that can be done simply from the primary data. This results in a set of cumulative curves (Fig. 35.6).

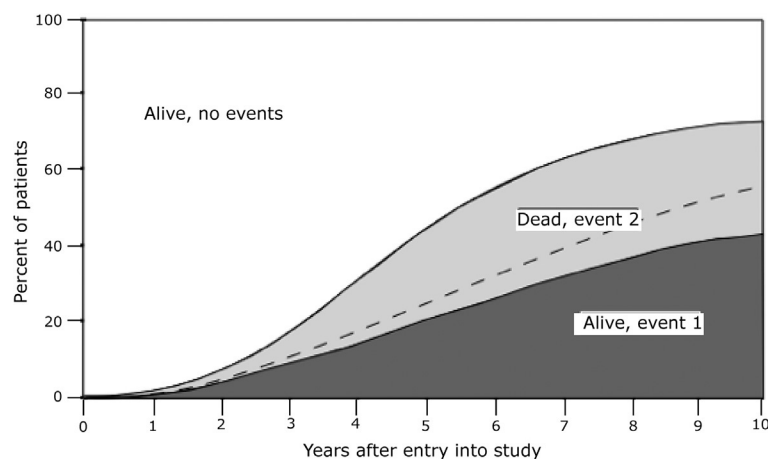


Fig. 35.6 Cumulative curves to show competing risks effect. These are the observed data. If the Kaplan-Meier method had been used with event 1 as the event of interest, then both those with event 2 (died) and those alive with no events are regarded as censored, and the dashed line shows that the incidence of event 1 is exaggerated.

REFERENCES

- Allison, P.D., 1995. *Survival Analysis Using SAS*. SAS Institute, Cary, NC (A Practical Guide).
- Altman, D.G., 1992. *Practical Statistics for Medical Research*. Chapman and Hall, London, p. 611.
- Anderson, R.P., Bonchek, L.I., Grunkemeier, G.L., Lambert, L.E., Starr, A., 1974. The analysis and presentation of surgical results by actuarial methods. *J. Surg. Res.* 16, 224–230.
- Armitage, P., Berry, G., Matthews, J.N.S., 2002. *Statistical Methods in Medical Research*. Blackwell, Oxford.
- Austin, P.C., Lee, D.S., Fine, J.P., 2016. Introduction to the analysis of survival data in the presence of competing risks. *Circulation* 133, 601–609.
- Blackstone, E.H., 1999. Actuarial and Kaplan-meier survival analysis: there is a difference. *J. Thorac. Cardiovasc. Surg.* 118, 973–975.
- Blackstone, E.H., Lytle, B.W., 2000. Competing risks after coronary bypass surgery: the influence of death on reintervention. *J. Thorac. Cardiovasc. Surg.* 119, 1221–1230.
- Bodnar, E., Blackstone, E.H., 2005. Editorial: an 'actual' problem: another issue of apples and oranges. *J. Heart Valve Dis.* 14, 706–708.
- Christensen, E., 1987. Multivariate survival analysis using Cox's regression model. *Hepatology* 7, 1346–1358.
- Elashoff, J.D., 1983. Surviving proportional hazards. *Hepatology* 3, 1031–1035.
- Glantz, S.A., Slinker, B.K., 2001. *Primer of Applied Regression and Analysis of Variance*. McGraw-Hill, Inc., New York.
- Grunkemeier, G.L., Starr, A., 1977. Actuarial analysis of surgical results: rationale and method. *Ann. Thorac. Surg.* 24, 404–408.
- Grunkemeier, G.L., Wu, Y., 2001. Actual versus actuarial event-free percentages. *Ann. Thorac. Surg.* 72, 677–8.s.
- Grunkemeier, G.L., Thomas, D.R., Starr, A., 1977. Statistical considerations in the analysis and reporting of time-related events. Application to analysis of prosthetic valve-related thromboembolism and pacemaker failure. *Am. J. Cardiol.* 39, 257–258.
- Grunkemeier, G.L., Anderson, R.P., Miller, D.C., Starr, A., 1997. Time-related analysis of nonfatal heart valve complications: cumulative incidence (actual) versus Kaplan-Meier (actuarial). *Circulation* 96, II-70-4; discussion II-74-5.
- Grunkemeier, G.L., Anderson, R.P., Starr, A., 2001. Actuarial and actual analysis of surgical results: empirical validation. *Ann. Thorac. Surg.* 71, 1885–1887.
- Grunkemeier, G.L., Jin, R., Eijkemans, M.J., Takkenberg, J.J., 2007. Actual and actuarial probabilities of competing risks: apples and lemons. *Ann. Thorac. Surg.* 83, 1586–1592.
- Hernan, M.A., 2010. The hazards of hazard ratios. *Epidemiology* 21, 13–15.
- Kaempchen, S., Guenther, T., Toschke, M., Grunkemeier, G.L., Wottke, M., Lange, R., 2003. Assessing the benefit of biological valve prostheses: cumulative incidence (actual) vs. Kaplan-Meier (actuarial) analysis. *Eur. J. Cardiothorac. Surg.* 23, 710–713 (discussion 713–4).
- Kleinbaum, D.G., 1966. *Survival Analysis. A Self-Learning Text*. Springer Verlag, New York.
- Merrell, M., Shulman, L.E., 1955. Determination of prognosis in chronic disease, illustrated by systemic lupus erythematosus. *J. Chronic Dis.* 1, 12–32.
- Muenz, L.R., 1983. Comparing survival distributions: a review for nonstatisticians. *Cancer Invest.* 1, 455–466.
- Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J., Smith, P.G., 1976. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br. J. Cancer* 34, 585–612.
- Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., Mcpherson, K., Peto, J., Smith, P.G., 1977. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *Br. J. Cancer* 35, 1–39.
- Schlichting, P., Christensen, E., Andersen, P.K., Fauerholdt, L., Juhl, E., Poulsen, H., Tygstrup, N., 1983. Prognostic factors in cirrhosis identified by Cox's regression model. *Hepatology* 3, 889–895.
- Selvin, S., 1991. *Statistical Analysis of Epidemiologic Data*. Oxford University Press, Oxford.

- Simon, R., 1986. Confidence intervals for reporting results of clinical trials. *Ann. Intern. Med.* 105, 429–435.
- Singer, R.B., 1994. Pitfalls of inferring annual mortality from inspection of published survival curves. *J. Insur. Med.* 26, 333–338.
- Spruance, S.L., Reid, J.E., Grace, M., Samore, M., 2004. Hazard ratio in clinical trials. *Antimicrob. Agents Chemother.* 48, 2787–2792.
- van Dijk, P.C., Jager, K.J., Zwinderman, A.H., Zoccali, C., Dekker, F.W., 2008. The analysis of survival data in nephrology: basic concepts and methods of cox regression. *Kidney Int.* 74, 705–709.

CHAPTER 36

Meta-analysis

INTRODUCTION

Meta-analysis is a set of techniques used to combine the results of a number of different reports to create a single, more precise estimate of an effect (Ferrer, 1998). The aims of meta-analysis are to increase statistical power and improve estimates of size of effect. There must be a valid reason to combine the studies. Although the frequency with which meta-analysis is used is increasing, meta-analysis has its detractors. If carefully performed it yields useful information, but a meta-analysis of badly designed studies produces erroneous statistics and may be misleading. Ignoring heterogeneity and combining apples and oranges is a pervasive error in meta-analysis (Eysenck, 1995) and techniques exist to assess them (Ferrer, 1998; Tang and Liu, 2000). Other forms of bias also interfere with effective meta-analysis (Egger and Smith, 1998).

Meta-analysis allows investigators to pool data from many trials that are too small by themselves to allow for secure conclusions. Although ideally any clinical trial should plan an adequate sample size, historically most trials have been underpowered. In 2002 a study of 5503 clinical trials (McDonald et al., 2002) identified 69% as having fewer than 100 subjects. Small trials make it more difficult to reject the null hypothesis because they lead to larger standard deviations and standard errors. There is also a risk of bias. A small trial with an effect probably due to chance might not be submitted for publication, whereas the same sized trial with an effect that was probably not due to chance (whether warranted or not) will probably be published (Stern and Simes, 1997). Egger et al. (2002) concluded that on average unpublished trials underestimate treatment effects by 10%. As another cause of bias, Stanbrook et al. (2006) found that clinical trials named with an acronym were more likely to be published in a major journal or to be cited than trials not named, independent of whether the results were positive or negative.

FOREST GRAPHS

Presentation of the results of the meta-analysis may be by tables and graphs (Bax et al., 2009). One common variety of graph, sometimes called a forest graph, is shown in Fig. 36.1.

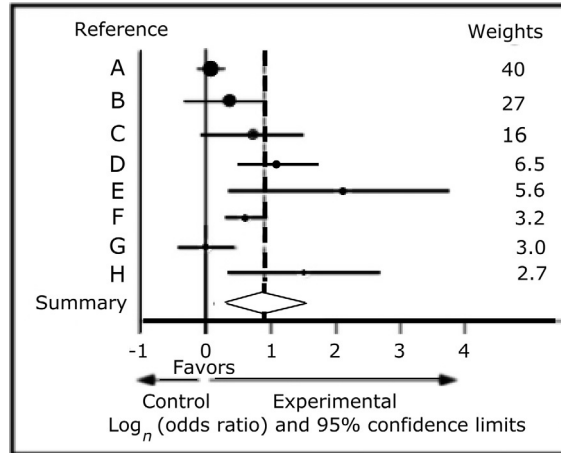


Fig. 36.1 Typical figure of response to a new treatment. The *dots* are the mean log odds ratio; the *horizontal lines* indicate 95% confidence limits. Figures in the right-hand column are the weights (see [Appendix](#)) assigned to each study; the studies are often arranged in descending weights. The *solid vertical line* at 0 indicates no difference between the groups, and the *dashed vertical line* indicates the weighted average ([Cuzick, 2005](#)). Some figures include the numbers of subjects in each group. The open diamond gives the weighted average and confidence limits of the individual samples.

The sample size or weight is sometimes indicated by the size of the symbol for the mean ([Lewis and Clarke, 2001](#)) to overcome the potential for smaller samples with wider confidence limits to feature more prominently than those with smaller limits.

As often happens, the largest studies show the least effects.

The calculations depend upon whether a fixed or a random effects model is used. The former implies that all the studies are estimating the same population mean, whereas the latter implies that there may be variability between the means as well as within each separate study ([Borenstein et al., 2010](#)). For example, if the effect of a beta-adrenergic blocker after myocardial infarction is tested on several small groups of similar composition, it is reasonable to consider that they all come from the same population and estimate the same population mean. If, on the other hand, the samples have different compositions, (age groups, associated diseases, etc.), then although the medication may be useful in all groups, it may affect each subpopulation differently. Therefore not only the variability within a sample but also the difference between the various population means need consideration. This additional source of variability changes the weighting factors and makes the confidence intervals wider. Meta-analysis programs take account of heterogeneity as well, and if necessary correct for it. Excellent articles featuring these issues are by [Hulley et al. \(2007\)](#) and [Higgins et al. \(2003\)](#) and the principles and practice of meta-analysis can be found in the book by [Borenstein et al. \(2009\)](#).

Higgins et al. (2003) recommended an index of heterogeneity called I^2 (see Appendix).

Forest plots and the relevant calculations may be produced online at <http://www.healthstrategy.com/meta/metainput.htm>, <http://www.singlecaseresearch.org/idea-center/forest-plot-in-excel>, and <https://www.evidencepartners.com/resources/forest-plot-generator/>.

There is also a well-documented commercial program called Comprehensive Meta-Analysis. It comes with an excellent tutorial. Both fixed and random effects are described by Cumming's ESCI at <http://thenewstatistics.com/itns/esci/>.

FUNNEL PLOTS

In the funnel plot the X -axis represents the mean result (that may be an odds or risk ratio, or a percent difference) and the Y -axis shows the sample size or an index of precision (Egger et al., 1997). Sterne and Egger (2001) recommended the inverse standard error for the Y -axis and the log of the odds ratio for the X -axis. (The symmetry of the plot may vary depending on whether sample size or inverse standard error is used as an index of precision Tang and Liu, 2000.)

Because there are usually more small than large samples, the points that represent each mean value are widely spread at the base and narrow as they move to the top, thus resembling an inverted funnel or a fir tree (Fig. 36.2).

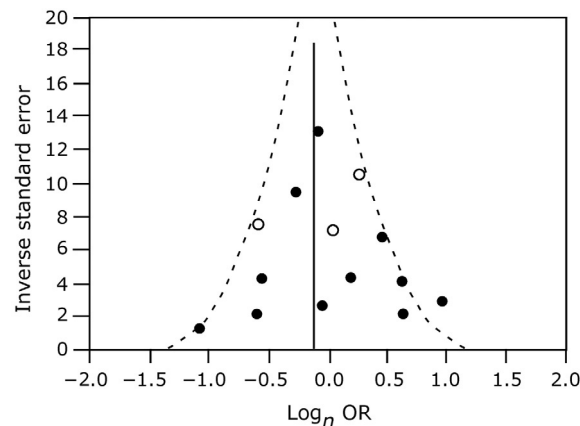


Fig. 36.2 Funnel plot. The *dashed lines* indicate the funnel. *Solid and open circles* can be used to differentiate subgroups.

The funnel plot is often used to assess bias (Ferrer, 1998; Tang and Liu, 2000; Song et al., 2002; Souza et al., 2007). Prominent causes of bias are publication bias, with studies giving positive results more frequently submitted for publication and more likely to be published (Roehr, 2012); English language bias—negative studies are less likely to be

published in English language journals, although this has not always been observed; and citation bias—studies with positive conclusions are cited more frequently and thus are more easily identified and incorporated in the database. Bias may be deliberate, as occurs when a Pharmaceutical Company deliberately withholds studies that do not favor their product (Eyding et al., 2010). The basis of assessing bias is that if all the studies give random assessments of the same unbiased mean value, the plot should be symmetrical. If the studies are biased, for example, by having too few small studies with positive results and large effect sizes, then the funnel plot becomes asymmetrical with a deficit near the bottom (Fig. 36.3).

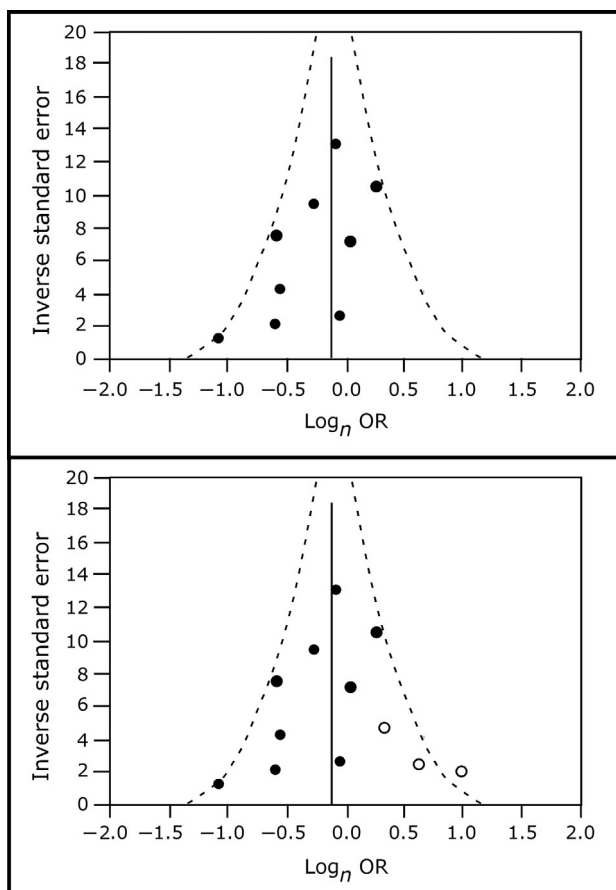


Fig. 36.3 Top panel: deficit of points with large effect and small sample size. Bottom panel: restoration of symmetrical funnel (added open circles).

The asymmetry may be judged by eye but can also be tested mathematically. One approach (Egger et al., 1997) is to fit the points by linear regression. In a symmetrical

funnel plot the intercept on the X-axis should be close to zero, whereas with asymmetry it may deviate considerably from zero. Egger et al. (1997) analyzed 37 meta-analyses published between 1993 and 1996 in the *Annals of Internal Medicine*, *BMJ*, *JAMA*, and *Lancet*. Twenty-six intercepts were well below zero, showing a deficiency in the lower right-hand corner of the funnel.

A technique called “trim and fill” was developed by Duval and Tweedie to deal with serious asymmetry by modeling the data as if they were symmetrically distributed, as they should be if all the samples are unbiased estimators of the same mean value. An iterative method is used in which some of the extreme values remaining are removed, and a new mean is calculated. If the distribution remains asymmetrical, another iteration is performed. After 2–4 cycles when the distribution is symmetrical the trimmed data are replaced and their theoretical counterparts on the other side of the axis of symmetry are inserted.

This technique may improve estimates when there are deficits due to publication bias, but deficits may be due to other factors, and different corrections may be required.

Funnel plots can be drawn by online programs at <http://www.kurtosis.co.uk/technique/funnel/index.htm>. They can also be drawn with standard XY plot software.

RADIAL PLOTS

To evaluate heterogeneity among the samples, Galbraith recommended a radial plot in which one axis plotted the z score (log odds ratio/standard error of log odds ratio) and the other axis plotted the log odds ratio on a curved line resembling a speedometer.

The Galbraith scale is more often represented as a classical XY plot, with the z score on the Y-axis, the inverse standard error on the X-axis, and two parallel lines indicating ± 2 standard errors (Fig. 36.4).

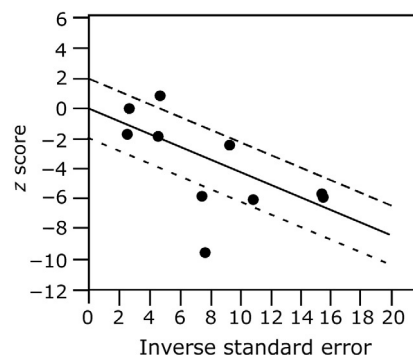


Fig. 36.4 Modified Galbraith plot. If the studies all estimate one fixed parameter, then the dots should fall between the two outer lines.

L'ABBÉ PLOTS

In the L'Abbé plot control vs treated outcome for each study are plotted on a graph that shows the line of identity (Fig. 36.5).

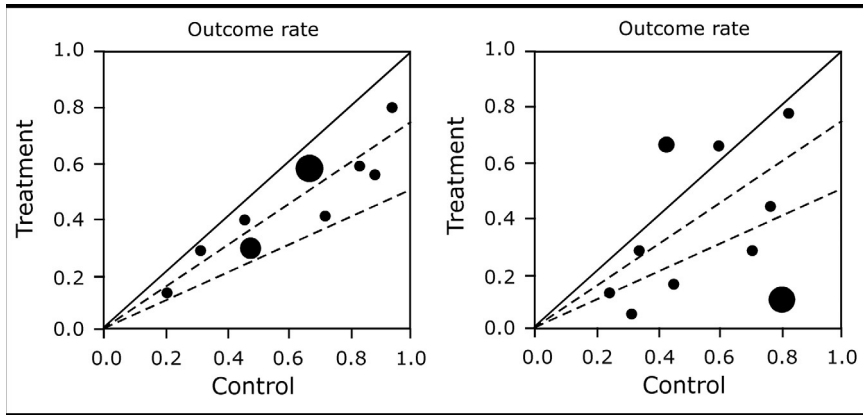


Fig. 36.5 Two L'Abbé plots. The two *dashed lines* in each plot indicate 25% and 50% reduction of effect. The left-hand plot shows homogeneity, but the right-hand plot shows excessive dispersion, casts doubt upon the homogeneity of the samples, and suggests that the studies may be measuring different things.

Descriptions of all these plots, with recommendations for their best use, are provided online by <https://pdfs.semanticscholar.org/8fa9/41074749a9f168f8c6e21e64219bf735346a.pdf>.

In some studies, the dots in the L'Abbé plot are circles of different sizes that represent the weights (size) of the samples. The circles may be colored to indicate subgroups.

Bax et al. (2009) also describe box plots, quantile plots, and standard residual histograms, among others. They compared the different plots and concluded that heterogeneity was shown best by the forest plot and the standard residual histogram, that the funnel plot might be best for assessing publication bias, but finally that many problems remained to be solved. They urged caution in interpreting any of these graphic displays.

When used carefully, meta-analysis is a useful technique for many purposes. A useful summary is provided by Anzures-Cabrera and Higgins (2010) and a useful online tutorial together with some online calculators is provided by Basu (2017).

Cautionary Tales

Few meta-analyses meet all the criteria for a good study, so that their conclusions are often suspect (Sacks et al., 1987, Chan and Altman, 2005).

One of the severest critics is Ioannidis, whose nontechnical Presidential lecture to the Society for Research Synthesis Methodology should be read by everyone who performs

or reads about meta-analysis (Ioannidis, 2010). He analyzed critically several meta-analyses performed to determine if steroids were helpful in bacterial meningitis. The earliest studies concluded that they were very beneficial, but as later studies were analyzed the conclusions became less certain, and the most recent analysis in 2010 showed no benefit at all. As the author pointed out, historically there has been a tendency for treatment effects to become smaller with repeated studies. This by itself is not a cause for alarm, because in science new data always have the ability to correct previous errors or misinterpretations. What was alarming was that the four most cited articles in the meningitis field were two early trials that found an implausibly large treatment effect, one that found an effect only in adults, and one a nonsystematic “expert” review. As Ioannidis wrote: “I suspect that the scientific literature is much littered with such wrecks that have not been removed from view.”

More criticisms were offered in publications by Shibata (2013) and LeLorier et al. (1997) who pointed out that about 30%–50% of large randomized clinical trials failed to confirm the results of previous meta-analyses. The probable reasons for the failure included small trial sizes used in the meta-analyses and, more importantly, clinical heterogeneity. The various difficulties of interpretation are described in detail for cardiac stem cell trials by Gyöngösi et al. (2016).

The testing of funnel plots for bias was criticized by Lau et al. (2006) because many of them had too small sample sizes and did not take heterogeneity into account. Furthermore, the pattern could vary with the choice of the X-axis (risk ratio, odds ratio, logarithms) and choice of the Y-axis (inverse variance, inverse standard error, sample size, etc.) These choices have not been standardized, and what might be a symmetrical funnel plot with one set of metrics may not be so with another.

APPENDIX

1. Each study in the fixed effects model is weighted by the inverse of its standard error, so that $W_i = \frac{1}{se_i^2} = \frac{N_i}{s_i^2}$, and the weighted mean $M = \frac{\sum W_i M_i}{\sum W_i}$. Its variance V_M is estimated as $V_M = \frac{1}{\sum W_i}$ and is used to set pooled confidence limits. The weighting is a little different in the random effects model.
2. The measure of heterogeneity Q is the total weighted sum of squares between the k studies and is calculated as $Q = \sum W_i M_i^2 - \frac{(\sum W_i M_i)^2}{\sum W_i}$. In the fixed effects model Q is the degrees of freedom $(k-1)$. One way of deciding if Q is unusually large is to calculate $I^2 = \frac{Q - df}{Q} \times 100\%$. Any value for I above zero indicates the added variability due to differences among the individual means, and if I is large, then a random effects model must be used. If there is any doubt, it is safer to use the random effects

model. Borenstein et al. (2017) who created the index have recently criticized it on the grounds that it does not indicate the heterogeneity of effect sizes but rather how much of the observed variance would remain if we eliminated sampling error.

REFERENCES

- Anzures-Cabrera, J., Higgins, J.P.T., 2010. Graphical displays for meta-analysis: an overview with suggestions for practice. *Res. Synth. Methods* 1, 66–68. or see, <https://pdfs.semanticscholar.org/8fa9/41074749a9f168f8c6e21e64219bf735346a.pdf>.
- Basu, A., 2017. How to conduct meta-analysis. Available: <http://www.pitt.edu/~super1/lecture/lec1171/001.htm>.
- Bax, L., Ikeda, N., Fukui, N., Yaju, Y., Tsuruta, H., Moons, K.G., 2009. More than numbers: the power of graphs in meta-analysis. *Am. J. Epidemiol.* 169, 249–255.
- Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein, H.R., 2009. *Introduction to Meta-Analysis*. John Wiley & Sons, Chichester.
- Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein, H.R., 2010. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res. Synth. Methods* 1, 97–111.
- Borenstein, M., Higgins, J.P., Hedges, L.V., Rothstein, H.R., 2017. Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Res. Synth. Methods* 8, 5–18.
- Chan, A.W., Altman, D.G., 2005. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ* 330, 753.
- Cuzick, J., 2005. Forest plots and the interpretation of subgroups. *Lancet* 365, 1308.
- Egger, M., Smith, G.D., 1998. Bias in location and selection of studies. *BMJ (Clin. Res. Ed.)* 316, 61–66.
- Egger, M., Davey Smith, G., Schneider, M., Minder, C., 1997. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 315, 629–634.
- Egger, M., Ebrahim, S., Smith, G.D., 2002. Where now for meta-analysis? *Int. J. Epidemiol.* 31, 1–5.
- Eyding, D., Lelgemann, M., Grouven, U., Harter, M., Kromp, M., Kaiser, T., Kerekes, M.F., Gerken, M., Wieseler, B., 2010. Rboxetine for acute treatment of major depression: systematic review and meta-analysis of published and unpublished placebo and selective serotonin reuptake inhibitor controlled trials. *BMJ* 341, c4737.
- Eysenck, H.J., 1995. Meta-analysis of best-evidence synthesis? *J. Eval. Clin. Pract.* 1, 29–36.
- Ferrer, R.L., 1998. Graphical methods for detecting bias in meta-analysis. *Fam. Med.* 30, 579–583.
- Gyongyosi, M., Wojakowski, W., Navarese, E.P., Moye, L.A., Investigators*, A., 2016. Meta-analyses of human cell-based cardiac regeneration therapies: controversies in meta-analyses results on cardiac cell-based regenerative studies. *Circ. Res.* 118, 1254–1263.
- Higgins, J.P., Thompson, S.G., Deeks, J.J., Altman, D.G., 2003. Measuring inconsistency in meta-analyses. *BMJ* 327, 557–560.
- Hulley, S.B., Cummings, S.R., Browner, W.S., Grady, D.G., Newman, T.B., 2007. *Designing Clinical Research*. Lippincott, Williams & Wilkins, Philadelphia.
- Ioannidis, J.P.A., 2010. Meta-research: The art of getting it wrong. *Res. Synth. Methods* 1, 169–184.
- Lau, J., Ioannidis, J.P., Terrin, N., Schmid, C.H., Olkin, I., 2006. The case of the misleading funnel plot. *BMJ* 333, 597–600.
- Leloirier, J., Gregoire, G., Benhaddad, A., Lapierre, J., Derderian, F., 1997. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *N. Engl. J. Med.* 337, 536–542.
- Lewis, S., Clarke, M., 2001. Forest plots: trying to see the wood and the trees. *BMJ* 322, 1479–1480.
- McDonald, S., Westby, M., Clarke, M., Lefebvre, C., 2002. Number and size of randomized trials reported in general health care journals from 1948 to 1997. *Int. J. Epidemiol.* 31, 125–127.
- Roehr, B., 2012. Routine screening for ovarian cancer harms more than it helps, says US authority. *BMJ* 345, e6203.
- Sacks, H.S., Berrier, J., Reitman, D., Ancona-Berk, V.A., Chalmers, T.C., 1987. Meta-analyses of randomized controlled trials. *New Engl. J. Med.* 316, 450–455.
- Shibata, M.C., 2013. What is wrong with meta-analysis? The importance of clinical heterogeneity in myocardial regeneration research. *Int. J. Clin. Pract.* 67, 1081–1085.

- Song, F., Khan, K.S., Dinnes, J., Sutton, A.J., 2002. Asymmetric funnel plots and publication bias in meta-analyses of diagnostic accuracy. *Int. J. Epidemiol.* 31, 88–95.
- Souza, J.P., Pileggi, C., Cecatti, J.G., 2007. Assessment of funnel plot asymmetry and publication bias in reproductive health meta-analyses: an analytic survey. *Reprod. Health* 4, 3.
- Stanbrook, M.B., Austin, P.C., Redelmeier, D.A., 2006. Acronym-named randomized trials in medicine—the ART in medicine study. *N. Engl. J. Med.* 355, 101–102.
- Stern, J.M., Simes, R.J., 1997. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ* 315, 640–645.
- Sterne, J.A., Egger, M., 2001. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J. Clin. Epidemiol.* 54, 1046–1055.
- Tang, J.L., Liu, J.L., 2000. Misleading funnel plot for detection of bias in meta-analysis. *J. Clin. Epidemiol.* 53, 477–484.

CHAPTER 37

Resampling Statistics

INTRODUCTION

Resampling occurs when observed data or a data generating mechanism such as a die or a computer are used to produce new samples unrelated to a hypothetical underlying population. These techniques are used when the form of the distribution is unknown or the samples are too small for standard methods to be used; for example, determining the interquartile interval for a very small data set.

The techniques are computer intensive. Because the calculations require special programs, resampling is usually out of reach for the average nonstatistician, but investigators will see the methods referred to in publications and need to understand the principles of resampling, and the strengths and weaknesses of the methods. For relatively understandable explanations of resampling methods, see the web description by [Berger \(2015\)](#) with excellent descriptions and examples, articles by [Diaconis and Efron \(1983\)](#), [Hesterberg et al. \(2003\)](#), [Boos and Stefanski \(2010\)](#), [Curran-Everett \(2009, 2012\)](#), and the books by [Mosteller and Tukey \(1977\)](#) and [Manly \(1997\)](#).

The major techniques include the bootstrap and the jackknife, Monte Carlo methods of random sampling, and permutation ([Efron, 1979, 1982](#); [Diaconis and Efron, 1983](#); [Efron and Tibshirani, 1986](#); [Ludbrook, 1994, 1995a, b, 2002](#); [Ludbrook and Dudley, 1994](#); [Manly, 1997](#); [Curran-Everett, 2009, 2012](#)). With these techniques computers can produce rapidly thousands of samples from a data set.

BOOTSTRAP

Samples are taken repeatedly, analyzed, and replaced many times, with no limit to the number of resampling runs done ([Boos and Stefanski, 2010](#)). The analyses produce robust estimates of point variables (mean, proportion, odds ratio, regression coefficient, etc.) with their standard errors and confidence limits.

[Simon \(1997\)](#) gives a simple example. Consider 20 people working in an office and assume that on the average one of them becomes ill on any day and stays at home. How often will three people be away on any given day? This estimate can be made by assuming that the data fit a Poisson process. An alternative estimate is to resample. Generate a batch of 20 random numbers between 1 and 20, and arbitrarily assign one number, say 9, to represent an ill worker. Count how many 9s are present in the batch of 20. Repeat the sampling 1000 times and determine how often three 9s appear in the sample. This number as a percentage of the total number of samples gives the required answer.

What is the variability of an estimate? Consider a set of experimental weight gains of 10 infant animals fed a particular formula: 50, 54, 46, 42, 51, 56, 61, 54, 48, 47 g/day. The mean value is 50.9 g/day. The usual method for calculating standard deviation gives $s = 5.53$, $se = 1.74$, 95% confidence limits 46.94–54.85 g/day. Similar estimates can be obtained by carrying out a bootstrap. Program the computer to select at random ten numbers from a data set including thousands of each of these 10 numbers. Any one selection may include any of these 10 numbers in any pattern; for example, one of each, or any 10 even numbers, or occasionally 10 of the same number. Determine the mean for each set of 10 numbers and find the 95% confidence limits. In this example they would be very similar to those calculated before; when I ran 5000 resamplings, the mean was also 50.9 and the standard error was 1.66. This was close to the previous value. The difference is that resampling does not involve any assumptions about the population from which those numbers came.

Sometimes it is difficult or impossible to know what the population distribution is, and therefore resampling the one data set provides data that are independent of any statistical theory. For example, determining the interquartile distance from a set of 10 numbers would be inexact with such a small set. Resampling allows the determination of thousands of interquartile distances and averages them. An excellent online tutorial is provided by Teknomo (2006). There is one caveat and that is that if the sample is unrepresentative of the population, a biased result may be obtained.

Other examples of calculations that would be difficult to achieve by standard statistical methods are determining the median difference between two independent samples or the mean and confidence limits for the product or ratio of the means of two data sets (Ludbrook, 1995a). Bootstrapping can be used also to determine regression coefficients when the data do not conform to the usual requirements for parametric regression analysis. The null hypothesis is that there is no correlation between the X and Y variates, so that in theory any X_i could be associated with any Y_i . Therefore the program draws at random all possible combinations of X and Y , calculates the resulting regression coefficient, and determines the probability that the observed regression coefficient could come from a set in which the regression coefficient is zero. These methods can be used to solve problems for which no parametric solutions apply (Manly, 1997).

There are free online calculators available at http://www.wessa.net/rwasp_bootstrapplot1.wasp, <http://www.real-statistics.com/non-parametric-tests/resampling-procedures/> (Excel) and there is an add-on for Excel at <http://www3.wabash.edu/econometrics/EconometricsBook/Basic%20Tools/ExcelAddIns/bootstrap.htm>.

As an example, consider the set of differences in weight growth in the peanut problem (Table 20.1) Repeated sampling (bootstrapping) for 1000 simulations gives the results in Fig. 37.1.

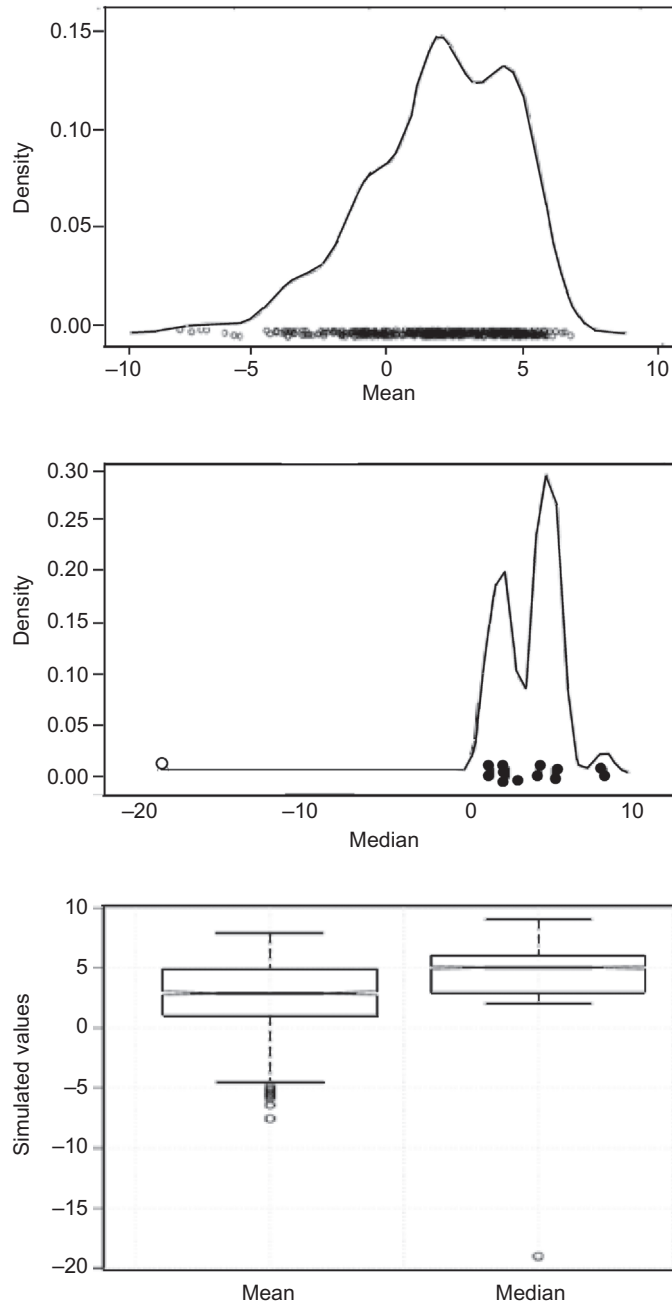


Fig. 37.1 Results of 1000 samplings from the 10 sets of differences. The upper panel and the left box plot (lower panel) show that the means of these 1000 samples are above zero more than 5% of the time, and confirm the conclusion from the t -test, but without regard to any of the requirements of the t -test. The median, being less sensitive to outliers, suggests that the roasted peanuts cause a consistent although small increase in growth.

PERMUTATIONS (OR RANDOMIZATION) TEST

Permutations share features with the bootstrap, but each data set is sampled until all the permutations are acquired; duplicate permutations are not allowed. If the data set is very large, the number of permutations becomes so big that even large computers take excessive time to cover them all, and then random sampling by the Monte Carlo method can be done to derive a reasonable set of permutations. This is the basis of the Wilcoxon tests, but these had to be done on ranks because of limited computer power. Today the same tests can be done using the individual values in the data set. For example, to assess the difference between the means of two samples, examine all possible permutations, calculate the difference between the means, and do this for every possible permutation. Then the probability of rejecting the null hypothesis is

$$p = \frac{\text{Number of permutations with mean difference} \geq \text{observed difference}}{\text{Total number of possible permutations}}.$$

As an example, consider group A with values 20, 11, 8, 7, 6 and group B with values 14, 8, 6, 5, 3. On the null hypothesis these all come from the same population, but the distributions are unsuitable for a *t*-test. Therefore take successive permutations of the 10 numbers, split them into two groups of 5 and examine the difference between the means. The sampling is repeated for each of the 252 possible permutations (Table 37.1).

Table 37.1 Three of the possible sets of permutations

Group A	Group B	Mean A-mean B
20, 11, 8, 7, 6	14, 8, 6, 5, 3	3.2
20, 14, 8, 7, 3	11, 8, 8, 7, 5	2.6
11, 6, 6, 5, 3	20, 11, 8, 8, 7	−4.6

If the sample numbers are large, the number of permutations is huge. For example, the number of permutations of two samples, each with $n = 10$, is $\frac{20!}{10! \times 10!} = 184,756$. The data can be analyzed by random sampling from the two samples and can be done online at <http://vassarstats.net/resamp1.html>.

Simple instructive examples of permutations are given by Mosteller and Rourke (1973), Ludbrook (1995a, b), and Curran-Everett (2012). Permutation tests can be done also at <http://www.real-statistics.com/non-parametric-tests/resampling-procedures/> (Fig. 37.2).

Other than the number of values sampled, what are the differences between bootstrapping and permutation tests? In general, bootstrapping is used to estimate the mean and standard deviation or confidence limits of the chosen distribution, whereas permutation tests are used to test the hypothesis of differences between the groups.

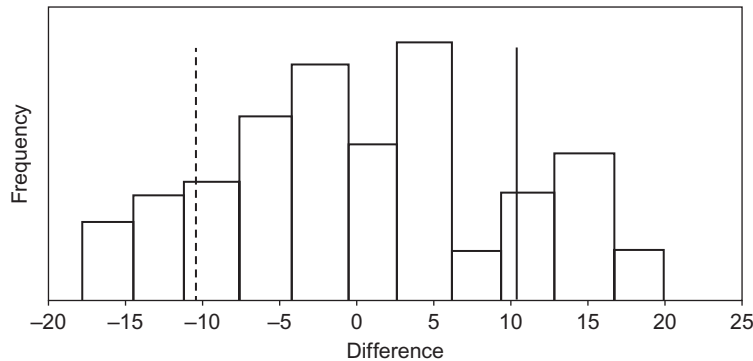


Fig. 37.2 Example of permutation test. About 2/3 of the differences are between 10 and -10 . Any greater differences are unlikely.

JACKKNIFE

This was one of the earliest resampling methods proposed. A sample is drawn and the statistics are recalculated after removing one data value at a time, starting with X_1 and ending with X_n ; the number of resamplings is finite. These repeated samples give independent estimates from which parameters, variances, and confidence limits may be derived. The procedure has the ability to remove bias.

The jackknife is a limited version of the bootstrap. Both methods can be used to determine confidence limits, but these are often wider than their parametric equivalents, and the best method of calculating them is still disputed. Outliers may affect the results. Nevertheless, for certain problems no other ways of determining confidence limits exist.

MONTE CARLO METHODS

Monte Carlo methods make random selections from the samples, based on an assumed model. Bootstrapping and permutation methods are specific types of more general Monte Carlo methods that can be applied to many types of data sets for which bootstrapping is inappropriate (Manly, 1997). A simple example, based on a problem illustrated by Bevington and Robinson (1992) is given as follows.

Consider determining the area of an irregular plane figure (Fig. 37.3).

In a relatively simple figure the area could be measured by planimetry or the trapezoid rule, but random sampling gives the same results. Place the figure inside a square with the base of 2 units and plot on the figure a series of points determined by random sampling from a set of numbers ranging from -1 to $+1$. After many samples, say 1000, there will be a distribution of points resembling the few shown in the figure. Because the distribution is random, the ratio of the number of points in the irregular figure to the total number of

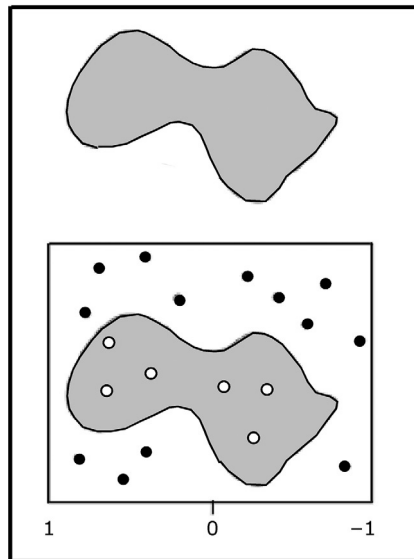


Fig. 37.3 Example of Monte Carlo method.

points will be the same as the ratio of the two areas. Because the area of the square is known, it is possible to determine the area of the irregular figure.

Many problems pertaining to these resampling methods are unsolved, and not every bootstrapped derived value is automatically correct. Outliers and severe heteroscedasticity may distort results. It is a field in which considerable experience is needed to avoid errors.

REFERENCES

- Berger, D., 2015. A gentle introduction to resampling techniques. Available: <http://wise.cgu.edu/wp-content/uploads/2015/04/Introduction-to-Resampling-Techniques-110901.pdf>.
- Bevington, P.R., Robinson, D.K., 1992. *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill, Inc., New York.
- Boos, D., Stefanski, L., 2010. Efron's bootstrap. *Significance* 7, 186–188.
- Curran-Everett, D., 2009. Explorations in statistics: the bootstrap. *Adv. Physiol. Educ.* 33, 286–292.
- Curran-Everett, D., 2012. Explorations in statistics: permutation methods. *Adv. Physiol. Educ.* 36, 181–187.
- Diaconis, P., Efron, B., 1983. Computer-intensive methods in statistics. *Sci. Am.* 248, 116–130.
- Efron, B., 1979. Bootstrap methods; another look at the jackknife. *Ann. Stat.* 7, 1–26.
- Efron, B., 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Capital City Press, Montpelier, VT.
- Efron, B., Tibshirani, R.J., 1986. Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. *Stat. Sci.* 1, 54–77.
- Hesterberg, T., Monaghan, S., Moore, D.S., Clipson, A., Epstein, R., 2003. Bootstrap methods and permutation tests. In: *The Practice of Business Statistics*. W.H. Freeman and Company, New York.
- Ludbrook, J., 1994. Advantages of permutation (randomization) tests in clinical and experimental pharmacology and physiology. *Clin. Exp. Pharmacol. Physiol.* 21, 673–686.

- Ludbrook, J., 1995a. Issues in biomedical statistics: comparing means by computer-intensive tests. *Aust. N. Z. J. Surg.* 65, 812–819.
- Ludbrook, J., 1995b. Issues in biomedical statistics: comparing means under normal distribution theory. *Aust. N. Z. J. Surg.* 65, 267–272.
- Ludbrook, J., 2002. Statistical techniques for comparing measurers and methods of measurement: a critical review. *Clin. Exp. Pharmacol. Physiol.* 29, 527–536.
- Ludbrook, J., Dudley, H., 1994. Issues in biomedical statistics: statistical inference. *Aust. N. Z. J. Surg.* 64, 630–636.
- Manly, B.F.J., 1997. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman and Hall, London.
- Mosteller, F., Rourke, R.E.K., 1973. *Sturdy Statistics*. In: *Nonparametrics and Order Statistics*. Addison-Wesley Publishing Company, London.
- Mosteller, F., Tukey, J.W., 1977. *Data Analysis and Regression. A Second Course in Statistics*. Addison-Wesley, Reading, CA.
- Simon, J.L., 1997. *Resampling: The New Statistics*. Wadsworth, Boston.
- Teknomo, K., 2006. Bootstrap Sampling Tutorial. Available: <http://people.revoledu.com/kardi/tutorial/Bootstrap/examples.htm>.

CHAPTER 38

Design: Sampling, Clinical Trials

SAMPLING PROBLEMS

To design studies efficiently, it is essential to start by asking important questions, identifying the likely important relationships, and using effective methods of assessing them. This is what training programs are for. *In addition, whatever the question and the study design it is essential to know ahead of time how to analyze the results.* Without this knowledge, the designs may be inefficient or even produce unanalyzable results (Hulley et al., 2007). The two main types of study are observational and experimental: the former observes what happens in a sample, the latter perturbs the sample for specific purposes. To determine if smoking cigarettes caused bladder cancer it would be unethical to take two groups of people, make one group smoke cigarettes, and then after 40 years find out if there was more bladder cancer in the smoking than the nonsmoking group. The best alternative is to follow two groups, one consisting of cigarette smokers and one of nonsmokers, and find out if they developed different incidences of bladder cancer.

If the cigarette smokers did indeed have more bladder cancer, what more needs to be done? The first consideration is to decide what population was being investigated. Is it a target or a convenience population? The target population is the theoretical population desired. The target might be a group of men and women with a racial mixture and range of socioeconomic classes equivalent those in the US population, all of whom began smoking cigarettes in their teens and continued to smoke cigarettes throughout the period of study. What we might get is a convenience sample, a selection of people with the time or inclination to join the study, and if a financial reward was offered for joining the study then we might get an excess of poorer people. The results would apply only to the observed sample and hence could be cautiously extended only to the base population from which the sample came. This might not be the same as the target population. Another major problem concerns confounders—hidden differences that might be the true explanation for the observations. For example, could smoking be innocuous, but could smokers be more likely to take a medication or food that is excreted in urine and will eventually cause bladder cancer? How are we going to find out the true cause of bladder cancer?

Cautionary Tales

The history of observational studies is rich with misleading results. One of the most famous is the Literary Digest survey in 1936 of the likely outcome of the presidential

Continued

Cautionary Tales—cont'd

election between Franklin Roosevelt and Alf Landon (Poll, 1936). The Literary Digest, having correctly predicted the outcome of the five preceding Presidential elections, mailed 10,000,000 questionnaires (each ballot contained the annual subscription card) and over 2,300,000 responses showed an overwhelming response of 57% to 43% in favor of the Republican candidate Alf Landon. Even the chairman of the National Democratic party, James Farley, was impressed enough to state:

“Any sane person cannot escape the implication of such a gigantic sampling of popular opinion as is embraced in The Literary Digest straw vote. I consider this conclusive evidence as to the desire of the people of this country for a change in the National Government. The Literary Digest poll is an achievement of no little magnitude. It is a Poll fairly and correctly conducted (Poll, 1936).”

Yet in the actual election it was Roosevelt who won with 62% to 38%, one of the largest majorities in electoral history. A similar poll conducted at the same time with only 50,000 people by George Gallup, a journalist who specialized in assessing public opinion, correctly predicted the Roosevelt victory. Why was there such a discrepancy?

The Literary Digest sent its questionnaire to people whose names were drawn from lists of its subscribers, as well as from lists of automobile and telephone owners, but this was a time when telephones were scarce, cars were relatively expensive, and magazine subscribers were among the minority with money to spare; the United States was just emerging from the Great Depression. Consequently, the people polled were among the more affluent members of society, most of whom detested Roosevelt's policies. Furthermore, even among the affluent, a response to the questionnaire was more likely among those against Roosevelt than in favor of his policies; people who feel strongly about an issue are more likely to respond. Therefore despite the huge sample, the Literary Digest made two fatal errors. They did not sample the target population that should have been all the potential US voters, and they did not allow for the fact that responders and nonresponders may be different. A 5% failure to respond might not have been important, but when 77% do not respond the chances of a biased sample are very high. The convenience sample, even one as large as this one was, was totally unrepresentative of the target population. There may indeed be safety in numbers, but only if they represent the desired sample.

How did Gallup with smaller numbers of people polled produce an accurate estimate of the final results? He used a method called quota sampling, in which he attempted to sample representative proportions of all the voters. This has now been replaced by a more accurate method based on randomization. As a footnote to history, the Gallup Organization made a serious error when in 1948 they predicted that Dewey would defeat Truman in the presidential election. Their error was in stopping polling 3 weeks before the election, thereby missing the big late swing of independent voters to Truman.

Sometimes an observational group may be large enough to be more like a target than a convenience population. In the Framingham Heart Study, in 1948 the National Heart Institute (subsequently the National Heart, Lung, and Blood Institute) enlisted the city of Framingham in Massachusetts to help them follow several thousand of its

inhabitants by a series of clinical and laboratory tests to find out what factors might lead to cardiovascular disease and stroke. The investigators chose a random selection of 2/3 of the households, and then the participants were invited to participate. The majority of subjects approached enrolled. Initially they recruited 5209 people aged 30–62 years and have followed them and two cohorts of their children since then. By maintaining good relations with the community there have been few subjects defecting from the study, and almost all of them return every 2 years for testing and examination. Nevertheless, despite what was close to a random sample for the Framingham population, this was far from being representative of all adults of that age. There are vast cultural and socioeconomic differences between the population of Framingham and comparably sized cities in Louisiana or Bulgaria, and what is true of one city might not be true of another. One study in Great Britain (Brindle et al., 2003) showed that the criteria established in Framingham for several cardiovascular risk factors overestimated the risk in Great Britain.

Another prominent study of this type is the Nurses' Health Study (NHS), started in 1976 with the primary objective of evaluating the long-term consequences of oral contraceptives. Subsequently, other factors such as diet were also investigated. Nurses were selected because they had a high educational standard so that they could respond accurately to questionnaires and also be motivated to join the study. Invitations were sent out to the population of 170,000 nurses aged 30–55 years in the 11 most populous states, and 122,000 responded and have been followed since then.

One important component of the NHS was evaluating the effects of hormone replacement therapy in healthy postmenopausal women with respect to deaths from cardiovascular disease. A major finding was that over 20 years there was about a 50% reduction in deaths from cardiovascular disease among those who took hormone replacement (Manson and Martin, 2001).

The Nurses' Health Study sample was large, with a widely based selection. It is likely but not certain that the 122,000 responders and the 48,000 (28%) nonresponders were similar. On the other hand, this is not the same as a universal target for women because it was restricted to subjects with a specific educational and perhaps socioeconomic criterion. Whether they respond in the same way to diet and contraceptives as do, say, women who are poor, uneducated, and belong to underprivileged minorities is uncertain. In fact, when the results of this study on deaths from cardiovascular disease were tested formally by randomized clinical trials (Grady et al., 2002; Grodstein et al., 2000), hormone replacement therapy was shown to *increase* the risk of death from cardiovascular disease. One possible explanation of the discrepancy between the observational and randomized studies was that in the NHS, hormone replacement therapy was started at an average age of 51 years, whereas in the randomized trials the average age at starting therapy was 63–67 years (Coulter, 2011). Once again, the target population must match the wider population if the results are to be applicable.

HISTORICAL CONTROLS

In some studies, instead of performing a randomized clinical trial, a group of subjects exposed to some potential disease-causing factor is compared to historical controls. Historical controls are not randomized, have many drawbacks, and should be avoided if possible. Diseases change over time, especially infectious diseases where severity waxes and wanes during an epidemic. Furthermore, patient selection changes over time. Current patients are often diagnosed earlier than they used to be, so that the average severity of the disease is less. Specific advances in treatment, on the other hand, may attract the more severe forms of disease disproportionately. Nonspecific therapeutic advances have altered outcomes. For example, [Harrison et al. \(1978, 1986\)](#) assessed the natural history of fetuses diagnosed in utero with congenital diaphragmatic hernia born in Norway between 1969 and 1975 and in California between 1980 and 1982; in both groups about 60%–75% of the patients died. This observation was the basis for carrying out a clinical trial of inflating the collapsed lung in the fetuses between 1999 and 2001 ([Harrison et al., 2003](#)). By the time of the trial, the medical care of these infants had improved survival so much that it was not possible to show any improvement from the fetal surgery. Many more differences are listed in the comprehensive book by [Schwartz et al. \(1980\)](#).

Historical controls may, however, lead to useful hypotheses. In the 1940s the development of incubators for warming and providing high oxygen concentrations to premature infants increased their survival, but an increasing incidence of retrolental fibroplasia (retinopathy of prematurity) began to be reported ([Wheatley et al., 2002](#)). In 1951 in Australia [Campbell \(1951\)](#) suggested that high oxygen concentrations could be the cause of the retinopathy, basing this hypothesis on the high incidence of this retinopathy in Australia compared with the low incidence in Great Britain where incubators and high oxygen concentrations were used far less frequently ([Silverman, 1980](#)). Nevertheless, it was not until [Patz \(1957\)](#) carried out a randomized clinical trial that the hypothesis was established.

RANDOMIZATION

The advantages of randomization were described in [Chapter 1](#), but certain problems need to be checked.

Cautionary Tales

An early example of randomization was the Medical Research Council (MRC) trials of the treatment of tuberculosis. In the first trial they allocated patients at random to receive either streptomycin or the symptomatic therapy in use at the time, and found a marked benefit from streptomycin ([A Medical Research Council Investigation, 1948](#)). The

groups were comparable in all important criteria. They then conducted later trials with patients randomized to receive streptomycin (now the new standard of treatment) or streptomycin and para-amino-salicylic acid (PAS); the combination proved to be superior ([A Medical Research Council Investigation, 1950](#)). Once again, the groups were comparable in all important respects. In 1952 the MRC conducted another trial in which patients were allocated at random to receive either streptomycin plus PAS or isoniazid ([Anon, 1952](#)). This time when they compared the degree of lung cavitation, one of the markers of severity, they noted that the isoniazid group had 55% of the patients with bilateral cavitation but only 40% of those with unilateral cavitation. This discrepancy per se biased the results against isoniazid. Furthermore, when they graded the degree of cavitation from nil to 3+, the isoniazid group had 20% with no cavitation as against 7% in the other group and had 69% with grades 1+ and 2+ as against 84% with these grades in the streptomycin PAS group. This degree of imbalance, not seen in the prior trials, occurred despite a formal randomization procedure. As another example, in the VA Cooperative Study of the treatment of chronic angina pectoris by surgery or medical treatment, patients were randomized based on age (<50, >50) and degree of coronary arterial involvement on angiography. When the investigators examined several other factors (blood pressure, previous myocardial infarction, history of smoking, diabetes, and blood cholesterol, all of these were equally distributed in both groups except for cholesterol, with many more with high blood cholesterol in the medically treated group. Therefore randomization, although eliminating conscious bias, does not necessarily eliminate all bias.

One particularly subtle type of error, that of *stage migration*, was discussed in 1985 by [Feinstein et al. \(1985\)](#) in an article entitled “The Will Rogers phenomenon. Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer.” This was based on the quip by the humorist Will Rogers: “When the Okies left Oklahoma and moved to California, they raised the average intelligence level in both states.” In one report, the authors observed that cancers treated in 1977 appeared to have higher survival rates compared with those treated between 1953 and 1964. This seemed to be due to the use of newer imaging techniques, because when survival based on initial symptoms was studied, there was no difference between the two periods.

A modern example of this was described by [Chee et al. \(2008\)](#). In lung cancer, stage III (locally advanced) has a better outcome than stage IV (metastatic). Until about the year 2000 this distinction was made by X-ray imaging, but in 2000 [Pieterman et al. \(2000\)](#) reported on the value of PET scanning with fluoro¹⁸-deoxyglucose (FDG) for detecting small distant metastases. [Chee et al. \(2008\)](#) showed that after the introduction of PET scanning with FDG, many patients formerly graded as stage III were found to be in stage IV. As a result, the prognosis of stage IV patients was slightly improved by addition of patients with smaller stage III lesions with less obvious metastasis, and the prognosis of stage III patients was improved by removal of patients with concealed metastases.

Continued

Cautionary Tales—cont'd

Randomization may also be important when studying the potential advantages of a screening test. Chapter 21 described issues associated with selection of an appropriate screening test but did not address whether the screening test was effective in reducing morbidity and mortality. To study this there should be two comparable groups, applying the screening test to one and not the other, and determining if indeed screening is beneficial. Randomization and equalization of the groups is essential. If the screened group had more patients with mild disease and the control group had more with severe disease, the screened group would show the better outcomes, but it would be impossible to tell if the better outcome was due to screening or patient selection. The number of possible errors in selecting groups to compare is large and described in detail by Feinstein (1984).

Simple randomization can be done with a table of random numbers (Gore and Altman, 1982). For example, to divide a sample of 60 subjects into 3 equal sized groups, allocate numbers 00–29 to group A, numbers 30–59 to group B, and numbers 60–89 to group C. Then take a table of random numbers (these are usually in pairs), turn pages blindly and select a page, and, while still not looking, put a pin point somewhere on the page. Assume this number is 75. The first subject is then allocated to group C. Then as more subjects come in take the next pair of numbers just below the first number (or just above it, or just to the right of it, but be consistent); if this is 23, that subject goes into group A, and so on. Any number between 90 and 99 is ignored. Random numbers can be generated online at <http://stattrek.com/statistics/random-number-generator.aspx>, <http://graphpad.com/quickcalcs/randomN1.cfm>, and <http://researchbasics.education.uconn.edu/random-number-table/>.

Random numbers do not have to come from a table but can be generated by computer. Because some program has to be used to create them, these numbers are not truly random but it may take enormous effort to disprove randomness. These numbers are often known as pseudorandom numbers, and they suffice for biomedical studies. If truly random outcomes are needed, they can be based on physically unpredictable events such as radioactive decay or atmospheric noise. Such a program is available at <http://www.random.org/integers/>.

If the plan was to have twice as many subjects in group A as in each of the other groups, then assign numbers 00–49 to group A, 50–74 to group B, and 75–99 to group C. There is no guarantee, however, that the groups will end up with the planned numbers. Quite by chance, there might be 65 in group A, 28 in group B, and 7 in group C. This is unlikely, but nothing prevents this imbalance from occurring. If this imbalance cannot be tolerated, then stratified sampling may be done. In a trial to determine if indomethacin closes the patent ductus arteriosus in premature infants, merely allocating patients at random to indomethacin or placebo has a major disadvantage. The chances

of spontaneous closure of the patent ductus arteriosus increase with increased gestational age or birth weight. Therefore if at the end of the study there were more older infants in the indomethacin group and if this group had more closures, it would not be possible to tell if closure was due to age or treatment, even though randomization was intended to avoid this difficulty. To ensure randomization, carry out stratified sampling. Define homogeneous blocks by birth weight or gestational age, and then randomize each of these blocks. With two groups (indomethacin, placebo), designate blocks of 4, 6, or 8, and arrange to have each block have equal numbers of subjects. If block 1 has 4 patients, patient 1 is allocated (at random) to either A or B group. The second patient is also randomized by the same method. If these two patients are randomized to A and B, then the third patient is randomized to A or B, and the fourth patient automatically goes into the other group. If the first two patients are randomized to group A, then the next two are allocated to group B. The possibilities are presented in Table 38.1 with the nonrandom allocations in bold type after a space:

Table 38.1 Randomized blocks of 4

A,	B,	A,	B
A,	A,	B ,	B
A,	B,	B,	A
B,	A,	B,	A
B,	A,	A,	B
B,	B,	A ,	A

Providing that the statistician does the allocation and the investigator is unaware of the allocation, there is no way that the investigator can know how the allocations are done. To avoid any possibility that the investigator may guess which patient has one of the agents, the blocks themselves may be changed at random from 4–6 to 8.

Stratified sampling has other advantages. For example, to determine if body weight and insulin resistance are associated, select at random 150 subjects and look for the association. Because exercise might be a confounding factor, subjects can be classified as no exercise outside normal activity, moderate exercise, and extreme exercise. By sampling randomly within each group, not only is there added information, but because variability will probably be smaller within each exercise group the ability to discern differences will be improved. This is the equivalent of blocking in ANOVA.

CLINICAL TRIALS

Firm evidence on which to base treatment is uncommon in Medicine, and instead treatment is often based on anecdotes or on poorly designed and often underpowered trials. This implies that patients often do not get optimal treatment, and instead get treatment that is futile or occasionally harmful. For example, in the early 1900s the groundbreaking

surgeon Sir Arbuthnot Lane decided on the basis of little or no evidence that numerous diseases were the consequence of intestinal stasis. His solution to the problem was either colectomy or colostomy. Fortunately, these procedures were not adopted by other physicians and surgeons, and better ways of treating bowel disorders are now used (Smith, 1982). Are we free of these errors today? Certainly not. After a randomized trial had shown that aspirin taken for 30 days after a myocardial infarction reduced mortality and morbidity, long-term aspirin was used after a myocardial infarction and subsequently in patients with chronic congestive heart failure. The few trials conducted to validate this treatment were either underpowered or poorly controlled, and sometimes reached conflicting conclusions. Nevertheless, long-term aspirin has become the standard of care. Only recently has this been questioned in a large study in Denmark that found that not only did aspirin confer no benefit, but it was sometimes harmful (Madelaine et al., 2018). The editorial that accompanied this publication is well worth reading (Cleland, 2018).

To provide good evidence-based results some type of controlled trial is needed so that sample size is large enough and we do not compare apples and oranges. The most rigorous of these is the randomized controlled clinical trial that will be discussed next, but there are now several alternatives that can be considered.

People have tried various proposed cures from time immemorial, but Dr. James Lind (1716–94) probably performed the first controlled clinical trial. In the 18th century scurvy was a devastating disease of sailors. For example, in Anson's circumnavigation of the globe from 1740 to 1744, reported in 1748, 380 out of 510 sailors died of scurvy, and in the British navy regularly about one-third of sailors died or were disabled by the disease. Lind (1753) selected 12 sailors who could not work because of scurvy. He gave 2 of them cider, 2 took elixir of vitriol, 2 had vinegar, 2 had copious drinks of seawater, 2 were given purgatives, and two were given fresh lemon and orange juice (one of these was the most severely affected of the 12). The 2 given citrus juice recovered rapidly, but the others remained ill. In the 1760s, too, John Hunter described that treating gonorrhea with (inert) bread pills produced the same cure rate as standard treatment (Palmer, 1835). It was not until 1944, however, that the Medical Research Council in Great Britain published the results of the first truly well-designed clinical trial (MRC, 1944).

There are observational studies that because of their unique nature cannot be considered without confounding issues and cannot be duplicated. Abraham Wald studied aviation casualties for the Statistical Research Group that was formed just after the attack on Pearl Harbor in 1941 (Mangel and Samaniego, 1984). Planes that returned from missions had more bullet holes in the fuselage than the engine cover, so the Air Force proposed reinforcing the fuselage. The difficulty was the absence of data from the planes that did not return, often because of damage to the engine. Wald demonstrated that this was likely, and his tentative conclusions were put into practice in the Korean and Vietnam wars. Modern investigation of airline crashes, too, has to work also with incomplete data,

often different for each crash, but the investigators usually manage to form a tentative conclusion and suggest possible safety measures. For example, on July 17, 1996, TWA flight 800 crashed into the sea soon after leaving John F. Kennedy airport in New York. Despite incomplete recovery of plane fragments, the National Transport Safety Board concluded that the center wing fuel tank was empty but filled with gasoline vapor that exploded, perhaps due to overheating by the air-conditioning packs that were just under the tank. This led to reconstruction of fuel systems so that this type of calamity could be avoided in the future.

Intent to treat is a cardinal principle of clinical trials (Peto et al., 1976, 1977). Remember that randomization is an attempt to minimize possible bias. If the subjects are randomized, then not only should factors such as age, body weight, smoking history, presence or absence of diabetes be approximately equalized between the two (or more) groups being compared in the trial, but the hope is that factors not yet known to be important will also be equalized. Without randomizing, one group might have an excess of subjects with a factor later found to have an important influence on the outcome.

Once the groups have been randomized, each should receive its designated treatment, but this is not always possible. For example, in the large VA trial of coronary artery surgery versus medical treatment, a substantial number of patients assigned to medical treatment eventually had surgery because of deterioration of their disease, and a smaller number assigned to surgery declined surgery (Peduzzi et al., 1998). How should the investigators analyze their data? The correct approach is termed the *intent-to-treat* principle, (Peto et al., 1976, 1977) and this requires the investigators to record the end points (acute myocardial infarction, death) based on the original assignments. To make this clear, if a patient assigned to medical treatment had surgery and then died, that would be considered a death due to medical treatment. Now this seems intuitively wrong, because death followed surgery. However, end points based on final treatment cause a problem in that those who crossed over to surgery were not a representative group. They might have had more with diabetes, or more with a high cholesterol level, or more with obesity. If they did, then transferring to surgery might deplete the medical treatment group of its high-risk patients, and the final analysis of those who remained on medical treatment versus those who had surgery would be made on groups that were not comparable. Vickers gave a simple description of the issues involved in the intent-to-treat principle (Vickers, 2009).

Of course, there is nothing to stop the investigators from analyzing subgroups that are matched, but once the randomization scheme has been broken the balance of all the other unknown factors (some of which might in the future turn out to be important) is altered in an unknown manner, and there is no longer a randomized clinical trial.

In planning a clinical trial, thought must be given to the end points. Often there is a single end point such as death or readmission to hospital, but sometimes there may be multiple end points. A trial of a method of reducing the incidence of acute myocardial

infarction might have this as its primary goal, but might also examine deaths, need for coronary revascularization, and congestive heart failure as secondary end points. Examining all of these raises all the issues of multiple comparisons (Chapter 24). If the secondary end points had not been considered a priori, then some form of Bonferroni or equivalent test can be applied to keep the type I error low. These tests are probably too conservative for planned secondary end points, and there is no generally accepted method for evaluating them (Fisher, 1999; Fisher and Moyé, 1999). One approach is to divide the alpha spending function into portions dealing with the primary and secondary end points, much as the Lan and De Mets approach to partitioning alpha spending functions for interim examinations. As an example, make the alpha allocation (α_E) for the whole study 0.05, and then choose $\alpha_P = 0.02$ for the primary end point. This leaves 0.03 to be apportioned among, say, three secondary end points.

There is no generally accepted method, but the issue of multiplicity must always be considered.

Another problem of multiple end points is that of sample size. Because clinical trials are costly in terms of time and money, the sample size is usually based on the primary end point and may be too small for the secondary end points unless they have large effect sizes. Under these circumstances if the primary end point is not achieved and a secondary end point seems to be important but is underpowered, a specific trial with that end point as a goal is needed.

The results of a well-conducted clinical trial must be considered carefully and related to the composition of the sample. At one extreme the sample may be very homogeneous; for example, a trial of the value of a statin in reducing the risk of myocardial infarction in men between the ages of 45 and 55 years, who do not smoke, are not hypertensive or diabetic, and have no family history of coronary heart disease. If the trial shows a reduction of risk, a physician might cautiously use that treatment for subjects who match these characteristics but can have no assurance that the treatment would help others with different characteristics. More often, the sample has subjects with different combinations of the previous variables, and more caution is needed in putting the results into practice. Furthermore, a decrease in relative risk, even if marked, has to be put into context by considering attributable risk, population attributable risk, and number needed to treat (Chapter 20).

An important aspect of all clinical trials is the distinction between internal and external validity, the former referring to the success of the trial in eliminating bias, the latter to the extension of the trial results to a wider population (Rothwell, 2005). Failure of internal and external validity to agree explains why frequently a clear-cut result of a clinical trial cannot be confirmed in a larger population. Statisticians over the years have come to accept this distinction and to be less rigid about the conclusions to be drawn from a trial. Sir Austin Bradford Hill, one of the earliest leaders in the field of Medical Statistics, changed from an unswerving reliance on figures to a more cautious approach (Horton, 2000).

In the 11th edition of his book *Principles of Medical Statistics*, published in 1984 (47 years after the 1st edition), [Hill \(1984\)](#) wrote:

“At its best such a trial shows what can be accomplished with a medicine under careful observation and certain restricted conditions. The same results will not invariably or necessarily be observed when the medicine passes into general use; but the trial has at the least provided background knowledge which the physician can adapt to the individual patient.”

Even if the trial shows, for example, that a specific treatment reduces the risk of some event, for example, a myocardial infarction, the benefit may not apply equally to all members of the group because the outcome will be influenced by sets of variables that distinguish one member of the group from another. As discussed by [Dorresteijn et al. \(2011\)](#) unless the group is unusually homogeneous some subjects will benefit more than others from the new treatment, some may not benefit at all, and some may be harmed. They described methods for predicting optimal response based on the individual effects of known variables. As an example, they used the data from the Justification for the Use of Statins in Prevention (JUPITER) trial to evaluate the use of rosuvastatin in the primary prevention of cardiovascular disease. Based on Framingham and Reynolds risk scores involving gender, smoking history, blood pressure, and family history of premature coronary heart disease, they calculated the risk of cardiovascular events with and without treatment. Such a method is the antithesis of the “one size fits all” approach and would allow physicians to tailor their treatment for specific individuals. This should maximize outcomes and minimize complications and costs.

Recently [Ebrahim et al. \(2014\)](#) reexamined 37 clinical trials that were reevaluated by independent or involved investigators. The reanalyses were often done with different statistical procedures, definitions, or measurements of outcomes of interest. Thirty-five percent of the reanalyses led to interpretations that differed from those in the original article. Because details of the trial are frequently not made public, there is a potential source of error that should make us be careful about accepting the results of these trials.

In addition to all the problems discussed before, clinical trials have their own problems associated with difficulties of organization and assessment. An important issue concerns outside interference with a perfectly good plan, as occurred in the early trials of the effect of oxygen therapy in premature infants. Although giving high oxygen concentrations to breathe was once standard because of the immature lungs of these patients, there was concern about an increase in the incidence of retrolental fibroplasia (retinopathy of prematurity). A trial was therefore planned to compare the effect of breathing high oxygen versus lower oxygen concentrations. (In its initial application, the NIH rejected the application for funding because the referees “knew” that it was dangerous to lower oxygen concentrations for premature infants!) Appropriate randomization was carried out. Unfortunately, some of the nurses had strong opinions about the usefulness of high

oxygen concentrations and deliberately placed infants assigned to low oxygen onto high oxygen concentrations (Patz, 1957). Once again, this made interpretation of the results difficult. Nevertheless, lower oxygen concentrations were shown to be beneficial.

Humans are particularly susceptible to suggestion. An extreme example of psychological effect has been termed the Hawthorne effect. The Hawthorne Works in Cicero, Illinois, were owned by the Western Electric Company and manufactured a variety of consumer products including telephones, electric fans, and refrigerators. Between 1927 and 1932 Elton Mayo studied the effect of environmental influences on worker productivity. A series of changes were made: for example, increased illumination of the factory floor, removal of obstacles, decreased humidity, and increased frequency of rest periods. After each change, productivity increased. Then all the variables were returned to their original states, and once again productivity increased! The conclusion was that the increases in productivity were nonspecific and not due to the individual interventions but rather to the workers' feelings that someone was taking an interest in them. The conclusions have been challenged and debated since then, although the principle that workers perform better when they believe that people take interest in them still applies.

The ideal clinical trial should make sure that everyone involved in the trial—doctors, patients, ancillary personnel, and statisticians analyzing the data—should be unaware of (blinded to) which patients are getting which treatment. Sometimes, as in comparing surgical with medical treatment of coronary artery disease, it is impossible for doctors and patients to be unaware of which group is which, but those analyzing the results should be blinded. If the patients do not know which group they are in, but the doctors know, the trial is termed a single blind. If neither doctors nor patients know which group they are in, it is a double-blind trial. Failure to randomize and conduct blinded trials may lead to bias, usually in favor of the new experimental treatment. This was the conclusion of a study by Chalmers et al. (1983) who studied controlled trials of the treatment of acute myocardial infarction and noted also that at least one prognostic variable was maldistributed in 14.0% of blinded trials, 26.7% of unblinded but randomized studies, and in 58.1% of the non-randomized studies. Other studies have shown that investigators may try to subvert randomization in an attempt to secure what they thought was the better treatment for a given patient (Schulz, 1995; Hoare, 2010). Those trials in which randomizations been inadequately concealed tended to yield larger estimates of treatment effects (Schulz et al., 1995).

PLACEBO EFFECT

The placebo effect refers to treatment with a supposedly inert substance without the patient's knowledge. Many studies have shown that a variable, often high percentage of patients with a variety of diseases and symptoms may improve after being given the

placebo (Beecher, 1955). The effect is real and presumably due to psychosomatic interactions, perhaps release of endorphins. The placebo recipients, if adequately randomized, do serve as a type of control. This is a more specific example of the complexities of human thought and emotions, and the issues were brought into focus by Kaptchuk (2003) who wrote:

Facts do not accumulate on the blank slates of researchers' minds and data simply do not speak for themselves. Good science inevitably embodies a tension between the empiricism of concrete data and the rationalism of deeply held convictions. Unbiased interpretation of data is as important as performing rigorous experiments. This evaluative process is never totally objective or completely independent of scientists' convictions or theoretical apparatus.....'At the cutting edge of scientific progress, where new ideas develop, we will never escape subjectivity.' Interpretation can produce sound judgments or systematic error. Only hindsight will enable us to tell which has occurred.

Placebos can form one arm of a clinical trial providing there is no available therapy. A placebo cannot substitute for a current therapy (Michels and Rothman, 2003).

ALTERNATIVES TO RANDOMIZED CLINICAL TRIALS

Although randomized clinical trials are desirable, they have some disadvantages. They are usually very expensive, take a long time, and often have restricted entry criteria that make it difficult for clinicians to apply those results to patients who do not meet those criteria. Recently Eapen et al. (2013) cited studies to show that the cost to conduct a large phase 3 or phase 4 clinical trial could exceed \$400 million. Over the years the costs of trials have been increasing and the number of trials that can be funded has consequently decreased. These authors discussed ways of reducing the costs of these important clinical trials.

The ALLHAT study (Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial), one of the largest clinical trials to date, followed over 40,000 patients with mild or moderate hypertension between 1994 and 2002 to compare the effectiveness of various treatments to reduce the blood pressure and the incidence of various cardiovascular complications (ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group and The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial, 2002; Salvetti and Ghiadoni, 2004) It cost \$105 million, according to the NIH Public Affairs Department. The trial concluded, among other things, that a low-cost diuretic agent was as effective in lowering blood pressure and reducing cardiovascular complications as newer and more expensive agents. (The estimated savings to the medical system by using the cheaper agent were estimated to be \$3.1 billion dollars over 10 years.) Then the question arose about what second-line drug to add when the primary diuretic agent did not achieve the desired lowering of blood pressure. Another randomized clinical trial of equivalent size, cost, and duration seemed to be impractical. An alternative was provided by Magid et al. (2010) who used

the electronic records of the Kaiser Permanente Health System with over 8.6 million enrolled patients to compare the results of an angiotensin blocker versus a beta-adrenergic blocker as a second blood pressure lowering treatment. Although the individual patient's physician decided treatment without any attempt at randomization, the data pool was so vast that by using sophisticated methods the investigators were able to match for each treatment patients with similar characteristics. Although excluding confounding factors is more difficult to do by this method, it had the advantage of minimizing them by using a huge database, producing the results in a year and a half (a fraction of the time that a prospective study would have taken), covering a much wider range of patient ages and associated conditions than a randomized trial would have covered, and did it all for only \$200,000. With the proper care, an observational study of this type can lead to results as acceptable as those attained by randomized clinical trials.

One of the other disadvantages of the usual RCT is that it adopts a "one size fits all" approach. Rather than worrying primarily about whether a treatment "works" in general, we should ask: For whom (if anyone) is the treatment beneficial and for whom is it harmful? What individual and circumstantial characteristics are conducive to a positive (or negative) response? (Weisberg, 2015; Berry et al., 2015). Performing subgroup analyses may introduce methodological problems, but new approaches may be more helpful (Weisberg and Pontes, 2015). As an alternative, Berry et al. (2015) described the concept of platform trials that had multiple differences from the usual RCT. These differences included having heterogeneous populations and assuming heterogeneous results, having multiple treatments, having treatments vary over time, depending upon initial results, and removing subgroups with demonstrated efficacy or futility of treatment while continuing with the rest of the trial.

Propensity Analysis

One method of allowing for differences in the covariates between two or more samples is to use propensity analysis. As summarized by D'Agostino (2007) "The propensity score is the probability that a participant is in the "treated" group given his/her background (pretreatment) characteristics." As an example, in comparing the risks of getting pancreatic cancer in smokers and nonsmokers, it would be unethical to assign experimental subjects to these groups randomly, so we elect to compare the outcomes in the two groups observed for 15 years. Unfortunately for simple comparisons, people may become smokers because they have underlying covariates that may be the real cause of pancreatic cancer. Perhaps, for example, people who drink a lot of coffee are more likely to be smokers and to get pancreatic cancer (this was a theory once held, but never proven). Therefore smokers and nonsmokers would differ in the proportion with the "true" underlying cause. It is to correct for this that propensity analysis was developed.

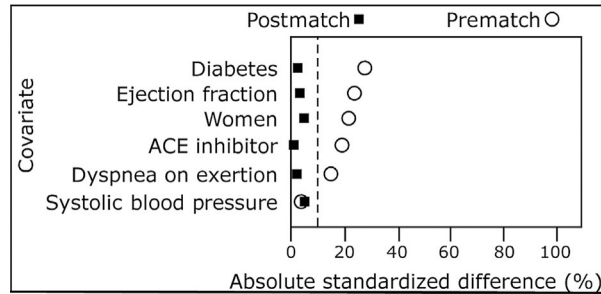


Fig. 38.1 Love plot shown for covariates before and after matching. Before matching there were often quite big differences between individual covariates. These differences almost disappeared after matching.

By using logistic analysis in which the outcome is treated vs nontreated, it is possible to correct for imbalances of covariates between the two groups. A simplified Love plot is shown in [Fig. 38.1](#), based on a study comparing the effect of using or not using diuretics in patients with chronic congestive heart failure ([Ahmed et al., 2006](#)).

It is likely that carefully conducted and analyzed nonrandom studies will be used more frequently. [Corrao et al. \(2011\)](#) studied a cohort of 209,650 patients from Lombardy, Italy, who were treated with antihypertensive drugs between 2000 and 2001 with the goal of determining whether starting with one or two medications was better. They selected 10,688 hospitalized with cardiovascular disease and selected 3 controls at random for each patient. By careful analysis they were able to show improved results by starting with combination therapy. The huge population base allowed them to study a wide range of patient comorbidities that might not have been investigated had a smaller controlled clinical trial been done, and they did so with a fraction of the time and cost that a formal randomized clinical trial would have incurred.

REFERENCES

- A Medical Research Council Investigation, 1948. Streptomycin treatment of pulmonary tuberculosis. *BMJ* 2, 769–782.
- A Medical Research Council Investigation, 1950. Treatment of pulmonary tuberculosis with streptomycin and Para-aminosalicylic acid; a Medical Research Council investigation. *BMJ* 2, 1073–1085.
- Ahmed, A., Husain, A., Love, T.E., Gambassi, G., Dell’Italia, L.J., Francis, G.S., Gheorghiade, M., Allman, R.M., Meleth, S., Bourge, R.C., 2006. Heart failure, chronic diuretic use, and increase in mortality and hospitalization: an observational study using propensity score methods. *Eur. Heart J.* 27, 1431–1439.
- ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group, The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial, 2002. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: The antihypertensive and lipid-lowering treatment to prevent heart attack trial (ALLHAT). *JAMA* 288, 2981–2997.
- Anon, 1952. Isoniazid in pulmonary tuberculosis. *Br. Med. J.* 2, 764–765.

- Beecher, H.K., 1955. The powerful placebo. *JAMA* 159, 1602–1606.
- Berry, S.M., Connor, J.T., Lewis, R.J., 2015. The platform trial: an efficient strategy for evaluating multiple treatments. *JAMA* 313, 1619–1620.
- Brindle, P., Emberson, J., Lampe, F., Walker, M., Whincup, P., Fahey, T., Ebrahim, S., 2003. Predictive accuracy of the Framingham coronary risk score in British men: prospective cohort study. *BMJ* 327, 1267.
- Campbell, K., 1951. Intensive oxygen therapy as a possible cause of retrolental fibroplasia; a clinical approach. *Med. J. Aust.* 2, 48–50.
- Chalmers, T.C., Celano, P., Sacks, H.S., Smith Jr., H., 1983. Bias in treatment assignment in controlled clinical trials. *N. Engl. J. Med.* 309, 1358–1361.
- Chee, K.G., Nguyen, D.V., Brown, M., Gandara, D.R., Wun, T., Lara Jr., P.N., 2008. Positron emission tomography and improved survival in patients with lung cancer: the will Rogers phenomenon revisited. *Arch. Intern. Med.* 168, 1541–1549.
- Cleland, J.G.F., 2018. Physicians addicted to prescribing aspirin—a disorder of cardiologists (PAPA-DOC) syndrome: The Headache of Nonevidence-Based Medicine for Ischemic Heart Disease? *JACC Heart Fail* 6, 168–171.
- Corrao, G., Nicotra, F., Parodi, A., Zambon, A., Heiman, F., Merlino, L., Fortino, I., Cesana, G., Mancia, G., 2011. Cardiovascular protection by initial and subsequent combination of antihypertensive drugs in daily life practice. *Hypertension* 58, 566–572.
- Coulter, S.A., 2011. Heart disease and hormones. *Texas Heart Inst J* 38, 137–141.
- D’Agostino Jr., R.B., 2007. Propensity scores in cardiovascular research. *Circulation* 115, 2340–2343.
- Dorresteyn, J.A., Visseren, F.L., Ridker, P.M., Wassink, A.M., Paynter, N.P., Steyerberg, E.W., Van Der Graaf, Y., Cook, N.R., 2011. Estimating treatment effects for individual patients based on the results of randomised clinical trials. *BMJ* 343, d5888.
- Eapen, Z.J., Vavalle, J.P., Granger, C.B., Harrington, R.A., Peterson, E.D., Califf, R.M., 2013. Rescuing clinical trials in the United States and beyond: A call for action. *Am. Heart J.* 165, 837–847.
- Ebrahim, S., Sohani, Z.N., Montoya, L., Agarwal, A., Thorlund, K., Mills, E.J., Ioannidis, J.P., 2014. Re-analyses of randomized clinical trial data. *JAMA* 312, 1024–1032.
- Feinstein, A.R., 1984. Current problems and future challenges in randomized clinical trials. *Circulation* 70, 767–774.
- Feinstein, A.R., Sosin, D.M., Wells, C.K., 1985. The will Rogers phenomenon: stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. *New Engl J Med* 312, 1604–1608.
- Fisher, L.D., 1999. Carvedilol and the Food and Drug Administration (FDA) approval process: the FDA paradigm and reflections on hypothesis testing. *Control Clin Trials* 20, 16–39.
- Fisher, L.D., Moyé, L.A., 1999. Carvedilol and the Food and Drug Administration approval process: an introduction. *Control. Clin. Trials* 20, 1–15.
- Gore, S.M., Altman, D.G., 1982. *Statistics in Practice*. Devonshire, Torquay.
- Grady, D., Herrington, D., Bittner, V., Blumenthal, R., Davidson, M., Hlatky, M., Hsia, J., Hulley, S., Herd, A., Khan, S., Newby, L.K., Waters, D., Vittinghoff, E., Wenger, N., 2002. Cardiovascular disease outcomes during 6.8 years of hormone therapy: heart and estrogen/progestin replacement study follow-up (HERS II). *JAMA* 288, 49–57.
- Grodstein, F., Manson, J.E., Colditz, G.A., Willett, W.C., Speizer, F.E., Stampfer, M.J., 2000. A prospective, observational study of postmenopausal hormone therapy and primary prevention of cardiovascular disease. *Ann. Intern. Med.* 133, 933–941.
- Harrison, M.R., Bjordal, R.I., Langmark, F., Knutrud, O., 1978. Congenital diaphragmatic hernia: the hidden mortality. *J. Pediatr. Surg.* 13, 227–230.
- Harrison, M.R., Adzick, N.S., Nakayama, D.K., Delorimier, A.A., 1986. Fetal diaphragmatic hernia: pathophysiology, natural history, and outcome. *Clin. Obstet. Gynecol.* 29, 490–501.
- Harrison, M.R., Keller, R.L., Hawgood, S.B., Kitterman, J.A., Sandberg, P.L., Farmer, D.L., Lee, H., Filly, R.A., Farrell, J.A., Albanese, C.T., 2003. A randomized trial of fetal endoscopic tracheal occlusion for severe fetal congenital diaphragmatic hernia. *N. Engl. J. Med.* 349, 1916–1924.
- Hill, A.B., 1984. *Principles of Medical Statistics*. Oxford University Press, London.
- Hoare, Z.S.J., 2010. Randomisation: what, why and how? *Significance* 7, 136–138.
- Horton, R., 2000. Common sense and figures: the rhetoric of validity in medicine (Bradford Hill Memorial Lecture 1999). *Stat. Med.* 19, 3149–3164.

- Hulley, S.B., Cummings, S.R., Browner, W.S., Grady, D.G., Newman, T.B., 2007. *Designing Clinical Research*. Lippincott, Williams & Wilkins, Philadelphia.
- Kaptschuk, T.J., 2003. Effect of interpretive bias on research evidence. *BMJ* 326, 1453–1455.
- Lind, J., 1753. A treatise of the scurvy. In three parts. Containing an inquiry into the nature, causes and cure, of that disease. Together with a critical and chronological view of what has been published on the subject. Sands, Murray and Cochran for A Kincaid and A Donaldson, Edinburgh.
- Madelaire, C., Gislason, G., Kristensen, S.L., Fosbol, E.L., Bjerre, J., D'souza, M., Gustafsson, F., Kober, L., Torp-Pedersen, C., Schou, M., 2018. Low-dose aspirin in heart failure not complicated by atrial fibrillation: a nationwide propensity-matched study. *JACC Heart Fail* 6, 156–167.
- Magid, D.J., Shetterly, S.M., Margolis, K.L., Tavel, H.M., O'connor, P.J., Selby, J.V., Ho, P.M., 2010. Comparative effectiveness of angiotensin-converting enzyme inhibitors versus beta-blockers as second-line therapy for hypertension. *Circ. Cardiovasc Quality Outcomes* 3, 453–458.
- Mangel, M., Samaniego, F.J., 1984. Abraham Wald's work on aircraft survivability. *J. Am. Statist. Assoc.* 79, 259–267.
- Manson, J.E., Martin, K.A., 2001. Clinical practice. Postmenopausal hormone-replacement therapy. *New Engl J Med* 345, 34–40.
- Michels, K.B., Rothman, K.J., 2003. Update on unethical use of placebos in randomised trials. *Bioethics* 17, 188–204.
- Mrc, 1944. Clinical trial of Patulin in the common cold. *Lancet*, 373–375.
- Palmer, J., 1835. *The Works of John Hunter*. Longman, Rees, Orme, Brown and Breen, London.
- Patz, A., 1957. The role of oxygen in retrolental fibroplasia. *Pediatrics* 19, 504–524.
- Peduzzi, P., Kamina, A., Detre, K., 1998. Twenty-two-year follow-up in the VA cooperative study of coronary artery bypass surgery for stable angina. *Am. J. Cardiol.* 81, 1393–1399.
- Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J., Smith, P.G., 1976. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br. J. Cancer* 34, 585–612.
- Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J., Smith, P.G., 1977. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *Br J Cancer* 35, 1–39.
- Pieterman, R.M., Van Putten, J.W., Meuzelaar, J.J., Mooyaart, E.L., Vaalburg, W., Koeter, G.H., Fidler, V., Pruim, J., Groen, H.J., 2000. Preoperative staging of non-small-cell lung cancer with positron-emission tomography. *N. Engl. J. Med.* 343, 254–261.
- Poll, T.L.D., 1936. <https://www.math.upenn.edu/~deturck/m170/wk4/lecture/case1.html>.
- Rothwell, P.M., 2005. External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet* 365, 82–93.
- Salvetti, A., Ghiadoni, L., 2004. Guidelines for antihypertensive treatment: an update after the ALLHAT study. *J. Am. Soc. Nephrol.* 15 (Suppl 1), S51–S54.
- Schulz, K.F., 1995. Subverting randomization in controlled trials. *JAMA* 274, 1456–1458.
- Schulz, K.F., Chalmers, I., Hayes, R.J., Altman, D.G., 1995. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 273, 408–412.
- Schwartz, D., Flamant, R., Lellouch, J., 1980. *Clinical Trials*. Academic Press, London.
- Silverman, W.A., 1980. *Retrolental Fibroplasia: A Modern Parable*. Grune & Stratton, New York.
- Smith, J.L., 1982. Sir Arbuthnot lane, chronic intestinal stasis, and autointoxication. *Ann. Intern. Med.* 96, 365–369.
- Vickers, A.J., 2009. Why Mr. Jones got surgery even If he didn't: intention-to-treat analysis. Available: <http://www.medscape.com/viewarticle/707140?src=mp&spon=2&uac=105072MT>.
- Weisberg, H.I., 2015. What next for randomised clinical trials? *Significance* 22–27.
- Weisberg, H.I., Pontes, V.P., 2015. Post hoc subgroups in clinical trials: anathema or analytics? *Clin Trials* 12, 357–364.
- Wheatley, C.M., Dickinson, J.L., Mackey, D.A., Craig, J.E., Sale, M.M., 2002. Retinopathy of prematurity: recent advances in our understanding. *Br. J. Ophthalmol.* 86, 696–700.

ANSWERS TO PROBLEMS

CHAPTER 3

Problem 3.1

- a. Interval
- b. Ordinal
- c. Nominal
- d. Ordinal
- e. Ordinal

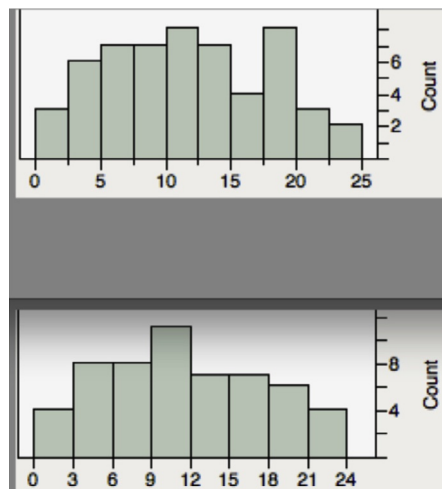
Problem 3.2

It tells us to add up all the X values, starting with the third and ending with value before the last value.

CHAPTER 4

Problem 4.1

The two histograms should look like these: Produced from http://www.wessa.net/rwasp_varia1.wasp#output.



Problem 4.2**Stem and Leaf Plot:**

Stem	Leaf
0	0 0 1 2 2 3 3 3 4 4
0	5 5 5 6 6 6 6 7 7 7 8 8 9 9 9
1	0 0 0 0 1 1 1 1 2 2 3 4 4
1	5 6 6 7 7 7 7 8 8 9 9 9
2	1 1 2 2
2	

Here is one plot that uses the data, produced from calculator at <http://www.calculatorsoup.com/calculators/statistics/stemleaf.php>. Yours may differ if you used different stems or a different calculator.

CHAPTER 5**Problem 5.1**

There are 52 cards and 4 kings. Therefore the probability is $4/52 = 1/13$.

Problem 5.2

The probability of drawing a king is $1/13$. This leaves 51 cards, and the probability of drawing a queen is thus $4/51$. This leaves 50 cards, and the probability of drawing a jack is $4/50$. These probabilities are independent of each other, so the combined probability is

$$\frac{4}{52} \times \frac{4}{51} \times \frac{4}{50} = \frac{64}{132600} = 0.0004826.$$

Problem 5.3

Person 1 can choose any of the 100 numbers. Person 2 can choose a different number in 99/100 ways. Person 3 can choose one of the remaining numbers in 98/100 ways, and so on. Therefore the chance of choosing 20 different numbers is $1 \times 0.99 \times 0.98 \times \dots \times 0.81 = 0.1187$. Then the chances of at least two numbers being the same are $1 - 0.1187 = 0.8813$, or 88.13%. The correct answer is (e).

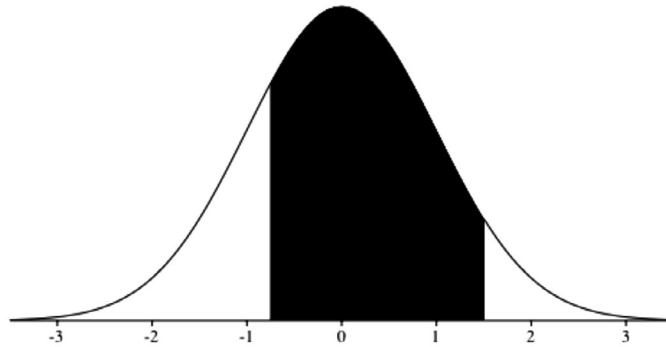
CHAPTER 6**Problem 6.1**

The cumulative area beneath the curve for $z = -0.75$ is 0.2263

The cumulative area beneath the curve for $z = 1.5$ is 0.9332

Therefore the area between these two limits is $0.9332 - 0.2263 = 0.7069$

From http://davidmlane.com/hyperstat/z_table.html we get



- ☒ Area from a value (Use to compute p from Z)
☐ Value from an area (Use to compute Z for confidence intervals)

Specify Parameters:

Mean

SD

☐ Above

☐ Below

☒ Between and

☐ Outside and

Results:

Area (probability)

Problem 6.2

Skewness and Kurtosis Test
<pre>> agostino D'Agostino skewness test data: x skew = -0.4598, z = -0.9738, p-value = 0.3302 alternative hypothesis: data have a skewness</pre>
<pre>> anscombe Anscombe-Glynn kurtosis test data: x kurt = 7.9059, z = 3.6771, p-value = 0.0002359 alternative hypothesis: kurtosis is not equal to 3</pre>
<pre>> jarque Jarque-Bera Normality Test data: x JB = 56.0562, p-value = 6.722e-13 alternative hypothesis: greater</pre>
<pre>> geary [1] 0.6603605</pre>

The JB test shows that this is not normal. Produced from http://www.wessa.net/rwasp_skewness_kurtosis.wasp#output.

Interquartile distance = $1.39 - 1.16 = 0.23$

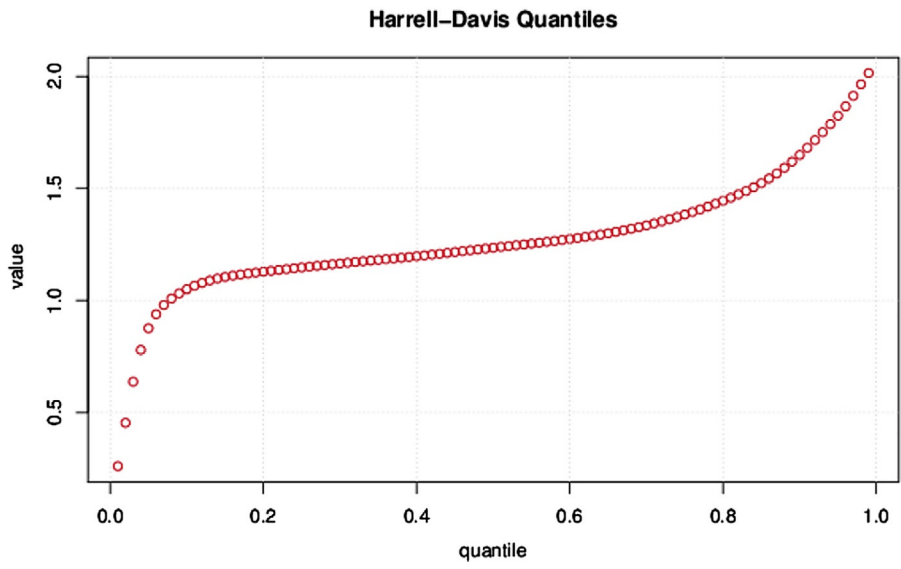
Standard deviation = 0.2333

PSD = $0.23/1.349 = 0.17$

Therefore distribution is not normal.

Problem 6.3

Quantile plot drawn from http://www.wessa.net/rwasp_harrell_davis.wasp#output.



This is not a straight line, so that the data set is not normal. It is straight in the middle 80%, but is abnormal at each end and is asymmetrical.

CHAPTER 7

Problem 7.1

	95%	95% range	99%	99% range
$N=14$	85.29–101.11	15.82	82.18–104.22	22.04
$N=77$	90.14–96.26	6.12	88.17–97.23	9.06

Note that the limits are narrower for the larger sample sizes, and narrower for 95% limits than for 99% limits.

CHAPTER 11

Problems 11.1 and 11.2

Power	$s = 0.4$	$s = 0.64$
0.80	14	36
0.85	16	41
0.90	19	48

Sample sizes needed.

By Lehr equation for paired samples, $s=0.4$ and power of 0.80, $k=8$ and $N = \frac{8 \times 0.4^2}{0.3^2} = 14.2$ subjects tested before and after the intervention. If the tests were unpaired, then $k=16$ and we need 28 subjects in all.

For power of 0.9, with paired samples we need $N = \frac{10.5 \times 0.4^2}{0.3^2} = 18.7$ subjects, and for unpaired subjects we need 37.

CHAPTER 12

Problem 12.1

For any one suit the chance of drawing 5 cards out of 13 is $\frac{13!}{5!8!} = 1287$. For all four suits, the chances are $4 \times 1287 = 5148$.

If this is divided by the total number of 5 card hands (2,598,960), the chances of a flush are $\frac{5148}{2,598,960} = 0.00198$, or about 1/505.

Problem 12.2

The possible sets are 1, 2, 3, 4, 5; 2, 3, 4, 5, 6; 3, 4, 5, 6, 7; 4, 5, 6, 7, 8; 5, 6, 7, 8, 9; 6, 7, 8, 9, 10; 7, 8, 9, 10, J; 8, 9, 10, J, Q; 9, 10, J, Q, K; 10, J, Q, K, Ace; the Ace can rank as either a 1 or above the King. Therefore numerically there are 10 possible sequences. Each sequence represents 5 cards drawn from any of the 4 suits for a total of 4^5 hands. Therefore the total number of straights is $10 \times 4^5 = 10,240$, which represents $\frac{10,240}{2,598,960} = 0.00394$ or about 1/254 hands.

Problem 12.3

For any suit, the number of possible sequences is 10 (remember that an Ace can be either the highest or the lowest card). There are 4 suits, therefore there are 40 straight flushes possible. In 2,598,960 possible hands, this represents $40/2,598,960 = 0.0001539$, or about one in 64,974 hands.

The secret to carrying out these calculations is to set out the problem in words before calculating.

CHAPTER 13

Problem 13.1

Total number of patients = $N=16$

Number of hyperreactors = $X=5$

Number of nonhyperreactors = $N - X=11$

Number chosen = $n=6$

Number of hyperreactors chosen = 2 or 5

Calculated with <http://stattrek.com/online-calculator/hypergeometric.aspx>.

■ Enter a value in each of the first four text boxes (the unshaded boxes).
■ Click the **Calculate** button.

Population size

Number of successes in population

Sample size

Number of successes in sample (x)

Hypergeometric Probability: $P(X = 2)$

Cumulative Probability: $P(X < 2)$

Cumulative Probability: $P(X \leq 2)$

Cumulative Probability: $P(X > 2)$

Cumulative Probability: $P(X \geq 2)$

■ Enter a value in each of the first four text boxes (the unshaded boxes).
■ Click the **Calculate** button.

Note:
One or more outputs use E-notation.

Population size

Number of successes in population

Sample size

Number of successes in sample (x)

Hypergeometric Probability: $P(X = 5)$

Cumulative Probability: $P(X < 5)$

Cumulative Probability: $P(X \leq 5)$

Cumulative Probability: $P(X > 5)$

Cumulative Probability: $P(X \geq 5)$

CHAPTER 14

Problem 14.1

*	Outcome Occurred	Outcome did not Occur	Totals
Risk Factor Present or Dx Test Positive	64 = a	216 = b	280 = r1
Risk Factor Absent or Dx Test Negative	47 = c	273 = d	320 = r2
Totals	111 = c1	489 = c2	600 = t

Confidence Level: %

Compute

Chi-Square Tests

Type of Test	Chi Square	d.f.	p-value
Pearson Uncorrected	6.610	1	0.010
Yates Corrected	6.080	1	0.014

There is a tendency for maternal age and birth weight to be associated. If the issue is important, you should probably acquire further data. Calculations done from <http://statpages.org/ctab2x2.html>.

Problem 14.2

The odds ratio is $\frac{64 \times 273}{47 \times 216} = 1.72$. Some calculators, for example, <http://statpages.org/ctab2x2.html> give the confidence limits for the ratio as well.

Problem 14.3

From the calculator, the total chi-square is 20.52, and $P=0.00039$. However, this does not indicate what the association is. To find this information, perform the calculation by hand, as in the text, and examine where there are excessive or very deficient expected values.

Maternal age (year)	Birth weight (g)						Total
	<2500		2500–3000		>3000		
<20	21	12.025	14	13.65	30	39.375	65
	+8.975	80.5506	+0.35	0.1225	−9.325	86.9556	
		6.70		0.01		2.21	
20–25	43	39.775	56	45.15	116	130.075	215
	+3.225	10.4006	+10.85	17.7225	−14.075	198.1056	
		0.26		2.61		1.52	
>25	47	59.2	56	67.2	217	193.6	320
	−12.2	148.84	−11.2	125.44	+23.4	547.46	
		2.51		1.87		2.83	
Total		111		126		363	600

Observed counts in bold, chi-square in italics.

Total chi-square 20.52, 4, df, $P=0.00039$.

As shown, the biggest chi-square value shows an excess of counts in the low birth weight babies born to mothers <20 years of age.

An online calculator that comes closest to providing this information is <http://vassarstats.net/index.html>.

Select the number of rows:	<input type="button" value="2"/>	<input type="button" value="3"/>	<input type="button" value="4"/>	<input type="button" value="5"/>	3
Select the number of columns:	<input type="button" value="2"/>	<input type="button" value="3"/>	<input type="button" value="4"/>	<input type="button" value="5"/>	3

Data Entry

	B ₁	B ₂	B ₃	B ₄	B ₅	Totals
A ₁	21	14	30	-----	-----	65
A ₂	43	56	116	-----	-----	215
A ₃	47	56	217	-----	-----	320
A ₄	-----	-----	-----	-----	-----	-----
A ₅	-----	-----	-----	-----	-----	-----
Totals	111	126	363	-----	-----	600

Chi-Square	df	P	No message for this analysis. -----
20.52	4	0.0004	
Cramer's V = 0.1308			

Percentage Deviations

	B ₁	B ₂	B ₃	B ₄	B ₅
A ₁	+74.6%	+2.6%	-23.7%		
A ₂	+8.1%	+24%	-10.8%		
A ₃	-20.6%	-16.7%	+12.1%		
A ₄					
A ₅					

Standardized Residuals

	B ₁	B ₂	B ₃	B ₄	B ₅
A ₁	+2.59	+0.09	-1.49	-----	-----
A ₂	+0.51	+1.61	-1.23	-----	-----
A ₃	-1.59	-1.37	+1.68	-----	-----

Percentage deviation and standardized residual are both measures of the degree to which an observed chi-square cell frequency differs from the value that would be expected on the basis of the null hypothesis.

For each cell, *percentage deviation* is calculated as

$$\frac{\text{observed} - \text{expected}}{\text{expected}} \times 100$$

Thus, a percentage deviation of +15% within a cell indicates that the observed frequency is 15% greater than the expected, while a percentage deviation of -15% indicates that the observed frequency is 15% smaller than the expected.

Problem 14.4

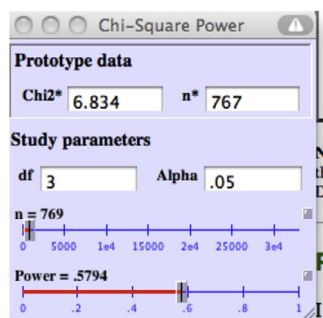
	Outcome 1	Outcome 2	Total
Group 1	3	9	12
Group 2	7	4	11
Total	10	13	23

Fisher's exact test

The two-tailed P value equals 0.0995

The association between rows (groups) and columns (outcomes) is considered to be not quite statistically significant.

Performed by <http://www.graphpad.com/quickcalcs/contingency1.cfm>. A trend, but more data are needed.

Problem 14.5

The power is relatively low. To determine the number needed for a given power, use the slider to set the power, and read the number. Determined from <http://homepage.stat.uiowa.edu/wrlenth/Power/index.html>.

Because of the scale the estimate is approximate, but that does not matter. In practice, you would plan for about 25%–30% greater numbers to allow for dropouts and other mishaps.

CHAPTER 15

Problem 15.1

Data Entry

Category	Observed Frequency	Expected Frequency	Expected Proportion	
A	4	7.13312693	0.02476780	<p>Sums:</p> <p>Observed Frequencies: 288</p> <p>Expected Frequencies: 256.79257</p> <p>Expected Proportions: 0.89164</p>
B	11	14.2662538	0.04953560	
C	19	21.3993808	0.07430340	
D	31	28.5325077	0.09907120	
E	42	35.6656346	0.12383900	
F	50	42.7987616	0.14860681	
G	63	49.9318885	0.17337461	
H	68	57.0650154	0.19814241	
	Reset	Calculate		

Cumulative Proportions

	Observed	Expected	O - E	
A	0.014	0.028	0.014	<p>D_{\max}</p> <p>0.052</p>
B	0.052	0.084	0.032	
C	0.118	0.167	0.049	
D	0.226	0.278	0.052	
E	0.372	0.417	0.045	
F	0.546	0.584	0.038	
G	0.765	0.778	0.013	
H	1.0	1.0	0	

Critical Values of D_{\max} for $n =$ 288

Level of Significance (non-directional)	
.05	.01
0.0801	0.096

The distribution is not uniform, but more numbers would be needed to be sure. Calculated with <http://vassarstats.net/ksm.html>.

Problem 15.2

Frequencies-I	8	17	31	47	52	73	89	98						
Frequencies-II	4	6	19	36	44	68	93	99						
<input type="button" value="CALCULATE"/> <input type="button" value="CLEAR"/>														
Test Statistic 0.078101														
Conclusion														
No real evidence against the null hypothesis														

Calculated with <http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/ks.htm>.

Problem 15.3

Calculated with <http://vassarstats.net/kappa.html>.

Data Entry

		B								
		1	2	3	4	5	6	7	8	Totals
A	1	16	3	1	1	----	----	----	----	21
	2	3	17	2	2	----	----	----	----	24
	3	4	6	14	3	----	----	----	----	27
	4	1	1	5	16	----	----	----	----	23
	5	----	----	----	----	----	----	----	----	----
	6	----	----	----	----	----	----	----	----	----
	7	----	----	----	----	----	----	----	----	----
	8	----	----	----	----	----	----	----	----	----
Totals		24	27	22	22	----	----	----	----	95

Reset

Calculate

Unweighted Kappa

Observed Kappa	.95 Confidence Interval		
	Standard Error	Lower Limit	Upper Limit
0.5512			
Method 1	0.0646	0.4246	0.6778
Method 2	0.0645	0.4247	0.6777

0.9158

maximum possible unweighted kappa, given the observed marginal frequencies

0.6019

observed as proportion of maximum possible

Kappa with Linear Weighting

Observed Kappa	.95 Confidence Interval		
	Standard Error	Lower Limit	Upper Limit
0.6221	0.0631	0.4984	0.7458

0.9141

maximum possible linear-weighted kappa, given the observed marginal frequencies

0.6806

observed as proportion of maximum possible

Kappa with Quadratic Weighting

Observed Kappa	.95 Confidence Interval		
	Standard Error	Lower Limit	Upper Limit
0.6842	0.0903	0.5071	0.8613

0.9211

maximum possible quadratic-weighted kappa, given the observed marginal frequencies

0.7428

observed as proportion of maximum possible

CHAPTER 16

Problem 16.1

By hand:

$$P(X=3) = \frac{50!}{3!47!} (0.07)^3 (0.93)^{47} = 0.2219$$

By calculator (<http://vassarstats.net/binomialX.html>).

n = 100 [the number of opportunities for a head to occur]
k = 60 [the stipulated number of heads]
p = .5 [the probability that a head will occur on any particular toss]
q = .5 [the probability that a head will not occur on any particular toss]

[Show Description of Methods](#)

To proceed, enter the values for **n**, **k**, and **p** into the designated cells below, and then click the «Calculate» button. (The value of **q** will be calculated and entered automatically). The value entered for **p** can be either a decimal fraction such as .25 or a common fraction such as 1/4. Whenever possible, it is better to enter the common fraction rather than a rounded decimal fraction: 1/3 rather than .3333; 1/6 rather than .1667; and so forth.

n	k	p	q
50	3	0.07	0.9299999995

Calculate

Reset

Parameters of binomial sampling distribution:

mean = 3.5

variance = 3.255

standard deviation = 1.8042

binomial z-ratio = (if applicable)

P: exactly 3 out of 50

Method 1. exact binomial calculation 0.22194674155

Method 2. approximation via normal

Method 3. approximation via Poisson

P: 3 or fewer out of 50

Method 1. exact binomial calculation 0.532735302552

Method 2. approximation via normal

Method 3. approximation via Poisson

P: 3 or more out of 50

Method 1. exact binomial calculation 0.689211438998

Method 2. approximation via normal

Method 3. approximation via Poisson

P: 3 or fewer out of 50

For hypothesis testing	One-Tail	Two-Tail
Method 1. exact binomial calculation	0.532735302552	1.0
Method 2. approximation via normal		
Method 3. approximation via Poisson		

By calculator <http://www.stat.tamu.edu/~west/applets/binomialdemo.html>.

Problem 16.2

Using the normal approximation,

$p = 0.37$, $q = 0.63$, $N = 250$. The 95% confidence limits are

$$0.37 \pm \left(1.96 \sqrt{\frac{0.37 \times 0.63}{200} \times \frac{1}{550}} \right) = 0.37 \pm 0.0708 = 0.2992 - 0.4408$$

By calculator <http://www.graphpad.com/quickcalcs/confInterval2/>.

QuickCalcs

[1. Select category](#)

[2. Choose calculator](#)

[3. Enter data](#)

[4. View results](#)

Confidence Intervals

Your data

Numerator = 74

Denominator = 200

Proportion (74/200) = 0.3700

Confidence intervals by modified Wald method

Agresti and Coull (The American Statistician. 52:119-126, 1998) recommend a method they term the modified [Wald method](#). It is easy to compute by hand and is actually more accurate than the so-called "exact" method (below). Here are the results computed by the modified Wald method. (Bug in 90% and 99% CI fixed Feb 2006.)

The 90% confidence interval extends from 0.3159 to 0.4276

The 95% confidence interval extends from 0.3061 to 0.4388

The 99% confidence interval extends from 0.2875 to 0.4609

"Exact" confidence intervals

The confidence intervals below are calculated using the so-called "exact" confidence intervals, computed by the method of Clopper and Pearson (Biometrika 26:404-413, 1934), which is based on a relationship between the F distribution and the binomial distribution. The modified Wald intervals (above) may actually be more exact.

The 90% confidence interval extends from 0.3131 to 0.4298

The 95% confidence interval extends from 0.3030 to 0.4409

The 99% confidence interval extends from 0.2836 to 0.4627

Note small difference between the normal approximation and the more exact calculator method.

Problem 16.3

The Confidence Interval of a Proportion

This unit will calculate the lower and upper limits of the 95% confidence interval for a proportion, according to two methods described by Robert Newcombe, both derived from a procedure outlined by E. B. Wilson in 1927 (references below). The first method uses the Wilson procedure without a correction for continuity; the second uses the Wilson procedure with a correction for continuity.

For the notation used here, n = the total number of observations and k = the number of those n observations that are of particular interest. Thus, if one observes 23 recoveries among 60 patients, $n = 60$, $k = 23$, and the proportion is $23/60 = 0.3833$.

To calculate the lower and upper limits of the confidence interval for a proportion of this sort, enter the values of k and n in the designated places, then click the «Calculate» button.

$k =$	<input type="text" value="74"/>	$\text{Proportion} =$	<input type="text" value="0.37"/>
$n =$	<input type="text" value="200"/>		
<input type="button" value="Reset"/> <input type="button" value="Calculate"/>			
<i>95% confidence interval: no continuity correction</i>			
Lower limit =	<input type="text" value="0.3061"/>	Upper limit =	<input type="text" value="0.4388"/>
<i>95% confidence interval: including continuity correction</i>			
Lower limit =	<input type="text" value="0.3038"/>	Upper limit =	<input type="text" value="0.4413"/>

These limits are slightly different from the others and are regarded as more accurate. Calculated from <http://vassarstats.net/> (under proportions).

Problem 16.4

By hand

$$P(A = 3; B = 26; C = 43; D = 9) = \frac{81!}{3!26!43!9!} 0.06^3 0.37^{26} 0.41^{43} 0.16^9$$

$$= 0.0002153.$$

This is tedious and error prone, so use calculator <http://stattrek.com/online-calculator/multinomial.aspx>.

- First, enter the number of outcomes.
- Then, enter the probability and frequency for each outcome.
- Click the **Calculate** button.

Number of outcomes

Outcome	Probability	Frequency
1	<input style="width: 50px;" type="text" value="0.06"/>	<input style="width: 50px;" type="text" value="3"/>
2	<input style="width: 50px;" type="text" value="0.37"/>	<input style="width: 50px;" type="text" value="26"/>
3	<input style="width: 50px;" type="text" value="0.41"/>	<input style="width: 50px;" type="text" value="43"/>
4	<input style="width: 50px;" type="text" value="0.16"/>	<input style="width: 50px;" type="text" value="9"/>

Multinomial probability 0.00022

CHAPTER 17**Problem 17.1**

By hand

Adjusted p for the Wald limits are:

$$p = \frac{19 + 2}{113 + 4} = 0.1795$$

Therefore limits are $0.1795 \pm 1.96 \sqrt{\frac{0.1795 \times 0.8205}{113}} = 0.1087 - 0.2503$

By calculator <http://www.graphpad.com/quickcalcs/confInterval2/cfm>.

Confidence Intervals

Your data

Numerator = 19

Denominator = 113

Proportion (19/113) = 0.1681

Confidence intervals by modified Wald method

Agresti and Coull (The American Statistician. 52:119-126, 1998) recommend a method they term the modified [Wald method](#). It is easy to compute by hand and is actually more accurate than the so-called "exact" method (below). Here are the results computed by the modified Wald method. (Bug in 90% and 99% CI fixed Feb 2006.)

The 90% confidence interval extends from 0.1177 to 0.2341

The 95% confidence interval extends from 0.1095 to 0.2486

The 99% confidence interval extends from 0.0948 to 0.2783

"Exact" confidence intervals

The confidence intervals below are calculated using the so-called "exact" confidence intervals, computed by the method of Clopper and Pearson (Biometrika 26:404-413, 1934), which is based on a relationship between the F distribution and the binomial distribution. The modified Wald intervals (above) may actually be more exact.

The 90% confidence interval extends from 0.1130 to 0.2369

The 95% confidence interval extends from 0.1044 to 0.2501

The 99% confidence interval extends from 0.0887 to 0.2766

Problem 17.2

By calculator <http://www.cct.cuhk.edu.hk/stat/proportion/Casagrande.htm>.

Input		Results	
α	0.05 <input checked="" type="radio"/> one sided test <input type="radio"/> two sided test	Calculate	
β	0.20	m	1018
P₁	0.06		
P₂	0.09	N	2036
r	1		

Note:

Variables	Descriptions
α	Significance level
$1-\beta$	Power of the test
P_1	Success proportion in arm 1
P_2	Success proportion in arm 2
r	Ratio of arm 2 to arm 1
m	Sample size for arm 1
N	Total sample size for arm 1 and 2

By the Lehr equation, it is $n = \frac{16 \times \frac{0.06 + 0.09}{2} \times \frac{[(1 - 0.06) + (1 - 0.09)]}{2}}{(0.06 - 0.09)^2} = 1233$ for each group.

The calculator allows for different ratios of sample sizes.

CHAPTER 18**Problem 18.1**

Although this can be done by hand, it is much easier to use the calculator <http://vassarstats.net/poissonfit.html>. This gives

k	Observed		Fitted Poisson	
	Frequency	Proportion	Probability	Expected Frequency
0	57	0.0218	0.01964	51.2503
1	203	0.0778	0.0772	201.4138
2	383	0.1468	0.1517	395.7782
3	525	0.2012	0.19872	518.4694
4	532	0.2039	0.19525	509.3962
5	408	0.1564	0.15346	400.3854
6	273	0.1046	0.10052	262.2525
7	139	0.0533	0.05643	147.236
8	45	0.0172	0.02772	72.3297
9	27	0.0103	0.01211	31.584
10	10	0.0038	0.00476	12.4125
11	4	0.0015	0.0017	4.4346
12	0	0	0.00056	1.4523
13	1	0.0004	0.00017	0.4391
14	1	0.0004	0.00005	0.1232
15	1	0.0004	0.00001	0.0323
<div>Reset Calculate</div>				
<div>mean of observed sample = 3.88</div> <div>variance of observed sample = 3.74</div> <div>mean and variance of fitted Poisson distribution = 3.93</div>				

If needed, the observed and expected frequencies can be compared by a chi-square test. However, the fit to a Poisson distribution is close, and in keeping with random distribution. The easiest way to test for a Poisson distribution with the above data is to take the ratio of the variance to the mean, which is $3.88/3.74 = 1.0374$, and multiply this by the degrees of freedom. This gives chi-square = $1.0374 \times 2609 = 2706$. From the tables this gives $P = 0.91$, so that we cannot reject the null hypothesis that the data fit a Poisson distribution.

Problem 18.2

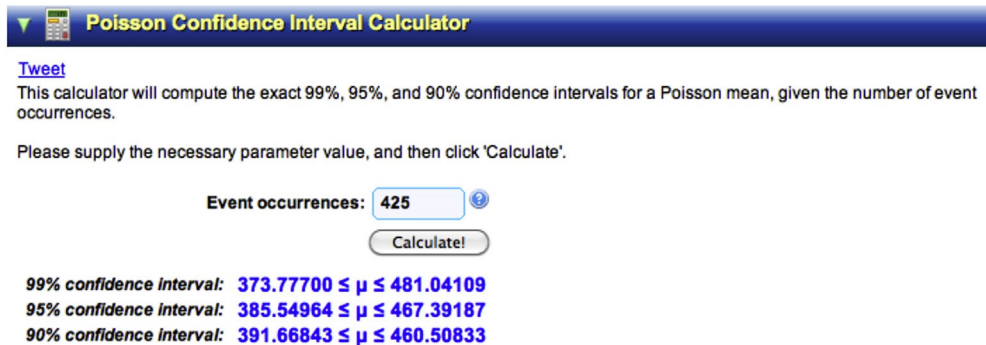
$$d = \sqrt{2 \times 2706} - \sqrt{2 \times 2609 - 1} = 1.3375.$$

This is much less than the critical value of 1.96, so we cannot reject the null hypothesis.

Problem 18.3

By the normal approximation, the 95% limits are

$425 \pm 1.96\sqrt{425} = 384.59 - 465.41$. By using the calculator <http://www.danielsoper.com/statcalc3/calc.aspx?id=86> we get



Poisson Confidence Interval Calculator

[Tweet](#)

This calculator will compute the exact 99%, 95%, and 90% confidence intervals for a Poisson mean, given the number of event occurrences.

Please supply the necessary parameter value, and then click 'Calculate'.

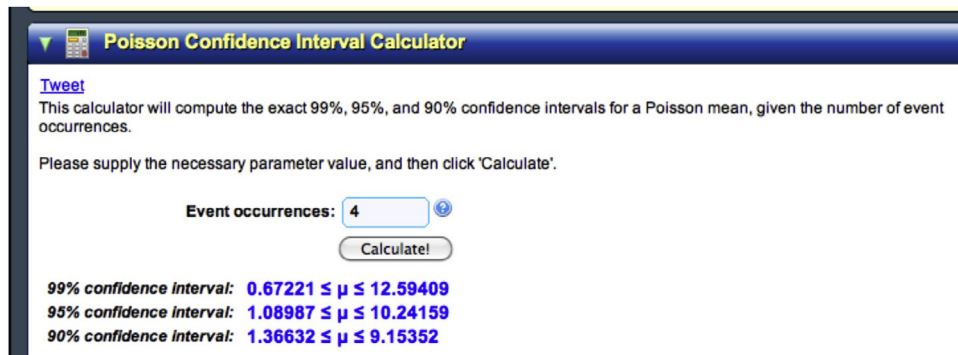
Event occurrences:

99% confidence interval: $373.77700 \leq \mu \leq 481.04109$
 95% confidence interval: $385.54964 \leq \mu \leq 467.39187$
 90% confidence interval: $391.66843 \leq \mu \leq 460.50833$

The results by the exact method are similar.

Problem 18.4

Using the above calculator, the confidence limits are



Poisson Confidence Interval Calculator

[Tweet](#)

This calculator will compute the exact 99%, 95%, and 90% confidence intervals for a Poisson mean, given the number of event occurrences.

Please supply the necessary parameter value, and then click 'Calculate'.

Event occurrences:

99% confidence interval: $0.67221 \leq \mu \leq 12.59409$
 95% confidence interval: $1.08987 \leq \mu \leq 10.24159$
 90% confidence interval: $1.36632 \leq \mu \leq 9.15352$

Problem 18.5

Using the calculator at <http://stattrek.com/Tables/poisson.aspx>, and fill in the number in the top line until an approximate probability is reached.

- Enter a value in BOTH of the first two text boxes.
- Click the **Calculate** button.
- The Calculator will compute the Poisson and Cumulative Probabilities.

Poisson random variable (x)	6
Average rate of success	3
Poisson Probability: $P(X = 6)$	0.0504094067224622
Cumulative Probability: $P(X < 6)$	0.916082057968697
Cumulative Probability: $P(X \leq 6)$	0.966491464691159
Cumulative Probability: $P(X > 6)$	0.033508535308841
Cumulative Probability: $P(X \geq 6)$	0.083917942031303

Calculate

There is a 96.66% probability of getting 6 or more in the next hour.

For 32 patients in 8h, place 24 in the second panel. The calculator gives a 95.3% chance.

CHAPTER 19

Problem 19.1

Calculating from <http://stattrek.com/online-calculator/negative-binomial.aspx> gives

- Enter a value in each of the first three text boxes (the unshaded boxes).
- Click the **Calculate** button.
- The Calculator will compute the Negative Binomial Probability.

Number of trials	30
Number of successes	8
Probability of success on a single trial	0.25
Negative binomial probability: $P(X = 8)$	0.0424824500374265

Calculate

CHAPTER 20

Problem 20.1

Data Entry

	Condition		Totals	Expected Cell Frequencies per Null Hypothesis	
	Absent	Present			
Group 1	445	652	1097	517	580
Group 2	794	738	1532	722	810
Totals	1239	1390	2629		

	Rate	Risk Ratio	Odds	Odds Ratio	Log Odds
Group 1	0.5943	1.2338	1.4652	1.5763	0.4551
Group 2	0.4817		0.9295		

Rate = proportion in group with condition present

Risk Ratio = Rate[1]/Rate[2]

Odds[1] = present[1]/absent[1]

Odds[2] = present[2]/absent[2]

Odds Ratio = Odds[1]/Odds[2]

Log Odds = natural logarithm of Odds Ratio

	Observed	.95 Confidence Intervals	
		Lower Limit	Upper Limit
Risk Ratio	1.2338	1.1489	1.325
Odds Ratio	1.5763	1.3477	1.8438

	Chi-Square	
	Yates	Pearson
Phi	32.09	32.54
P	<.0001	<.0001

Chi-square is calculated only if all expected cell frequencies are equal to or greater than 5. The Yates value is corrected for continuity; the Pearson value is not. Both probability estimates are non-directional.

Fisher Exact Probability Test:

P	one-tailed	Sample size too large
	two-tailed	for the Fisher test.

Derived from <http://vassarstats.net/odds2x2.html>.

The confidence limits and chi-square are also provided. Because these were not samples drawn at random from a population, we cannot be sure about their representation in the population.

Problem 20.2

Using the calculator <http://www.stat.ubc.ca/wrollin/stats/ssize/caco.html> gives 2733 for the first set, 358 for the second set (lower because of the higher risk ratio), and 150 for the last set.

Unmatched Case/Control Studies

(To use this page, your browser must recognize JavaScript.)

Choose which calculation you desire, enter the relevant population values (as decimal fractions) for p_0 (exposure in the controls) and RR (relative risk of disease associated with exposure) and, if calculating power, a sample size (assumed the same for each sample). You may also modify α (type I error rate) and the power, if relevant. After making your entries, hit the **calculate** button at the bottom.

- ☒ Calculate Sample Size (for specified Power)
- ☐ Calculate Power (for specified Sample Size)

Enter a value for p_0 :

Enter a value for RR:

- ☐ 1 Sided Test
- ☒ 2 Sided Test

Enter a value for α (default is .05):

Enter a value for desired power (default is .80):

The sample size (for cases and controls, separately) is:

Problem 20.3

From <http://statpages.org/ctab2x2.html> we get

Observed Contingency Table

*	Outcome Occurred	Outcome did not Occur	Totals
Risk Factor Present or Dx Test Positive	53 = a	34 = b	87 = r1
Risk Factor Absent or Dx Test Negative	148 = c	165 = d	313 = r2
Totals	201 = c1	199 = c2	400 = t

Confidence Level: 95 %

Compute

Chi-Square Tests

Type of Test	Chi Square	d.f.	p-value
Pearson Uncorrected	5.063	1	0.024
Yates Corrected	4.532	1	0.033
Mantel-Haenszel	5.050	1	0.025

Odds Ratio (OR) = (a/b)/(c/d);	1.738	1.042	2.904
Relative Risk (RR) = (a/r1)/(c/r2);	1.288	1.020	1.569

Number Needed to Treat (NNT) = 1 / absolute value of DP; which = 1 / absolute value of ARR;	7.334	3.933	97.510
Absolute Risk Reduction (ARR) = c/r2 - a/r1; which = - DP	-0.136	-0.254	-0.010
Relative Risk Reduction (RRR) = ARR/(c/r2); more info	-0.288	-0.569	-0.020

NNT is calculated from

$$\text{NNT} = \frac{1}{\frac{53}{148} - \frac{34}{165}} = \frac{1}{0.1520} = 6.6$$

CHAPTER 21

Problem 21.1

Quantities Derived from the 2-by-2 Contingency Table	Value	95% Conf. Interval	
Odds Ratio (OR) = $(a/b)/(c/d)$;	18.592	9.099	38.376
Relative Risk (RR) = $(a/r1)/(c/r2)$;	16.518	8.390	33.005
Kappa	0.174	0.121	0.219
Overall Fraction Correct = $(a+d)/t$; (often referred to simply as "Accuracy")	0.908	0.902	0.913
Mis-classification Rate, = $1 - \text{Overall Fraction Correct}$;	0.092	0.087	0.098
Sensitivity = $a/c1$; (use exact Binomial confidence intervals instead of these)	0.641	0.476	0.781
Specificity = $d/c2$; (use exact Binomial confidence intervals instead of these)	0.912	0.909	0.915
Positive Predictive Value (PPV) = $a/r1$; (use exact Binomial confidence intervals instead of these)	0.118	0.087	0.144
Negative Predictive Value (NPV) = $d/r2$; (use exact Binomial confidence intervals instead of these)	0.993	0.990	0.996

A positive test is of little help because of the large number of false positives. On the other hand, a negative test is strongly against the diagnosis. Performed with calculator at <http://statpages.org/ctab2x2.html>.

Problems 21.2 and 21.3

From the same calculator we get

Positive Likelihood Ratio (+LR) = $\text{Sensitivity} / (1 - \text{Specificity})$;	7.315	5.247	9.182
Negative Likelihood Ratio (-LR) = $(1 - \text{Sensitivity}) / \text{Specificity}$;	0.393	0.239	0.577
Diagnostic Odds Ratio = $(\text{Sensitivity}/(1-\text{Sensitivity})) / ((1-\text{Specificity})/\text{Specificity})$;	18.592	9.099	38.376

This suggests that a positive test supports the possibility of bowel cancer.

MedCalc: Bayesian Analysis Model

Enter PREVALENCE, SENSITIVITY, and SPECIFICITY:

PREV: 0.0179 SENS: 0.641 SPEC: 0.9124

Calculate Reset

Or enter TP, FN, TN, and FP:

N = 2173	Disease	No Disease
Positive Test	TP: 25	FP: 187
Negative Test	FN: 14	TN: 1947

Calculate Reset

Positive Predictive Value = 0.1179

Negative Predictive Value = 0.9929

Positive Likelihood Ratio = 7.3152 Pre-test Probability = 0.0179 (prevalence)

Negative Likelihood Ratio = 0.3935 Post-test Probability = 0.1179

Remember:

Sensitivity = $TP / (TP + FN)$
 Specificity = $TN / (FP + TN)$
 Prevalence = $(TP + FN) / (TP + FN + FP + TN)$
 Predictive value positive = $TP / (TP + FP)$
 Predictive value negative = $TN / (FN + TN)$
 Positive Likelihood Ratio = $SENS / (1 - SPEC)$
 Negative Likelihood Ratio = $(1 - SENS) / SPEC$
 Pre-test Probability = Prevalence
 Pre-test Odds = Pre-test Prob / (1 - Pre-test Prob)
 Post-test Odds = Pre-test Odds x Likelihood Ratio
 Post-test Probability = Post-test Odds / (1 + Post-test Odds)

Calculated from <http://www.medcalc.com/bayes.html>.**CHAPTER 22****Problem 22.1** $t = 4.8554$, 8 degrees of freedom. $P = 0.0013$.

Difference between means = 49.78, with standard error of mean 10.252.

95% confidence intervals are 26.14 to 73.42.

Conclusion: There is good reason to reject the null hypothesis, and exercise reduces peak flow rate in these subjects.

Problem 22.2 $t = 1.8259$ with 16 degrees of freedom. $P = 0.0866$.

Difference between means = 49.78 (same as for paired test), with standard error of mean 27.263.

95% confidence intervals are -8.02 to 107.57 .

The peak flow rate has been decreased, but the difference is not strongly supportive of rejecting the null hypothesis. This is due to the big increase in standard error of the mean in the unpaired versus the paired test because intersubject variability is added to the between-subject variability.

Problem 22.3

Wilcoxon Signed-Rank Test Calculator

Subject	Before exercise	After exercise	Sign	Absolute value	Rank	Signed rank
1	320	297	+	23	2	2
2	235	200	+	35	3	3
3	322	220	+	102	9	9
4	376	334	+	42	4	4
5	286	210	+	76	8	8
6	254	255	—	1	1	−1
7	381	338	+	43	5	5
8	397	341	+	56	6	6
9	299	227	+	72	7	7

Obtained at <http://www.socscistatistics.com/tests/signedranks/Default2.aspx>.

Result Details

W value: 1

Mean Difference: 118.89

Sum of positive ranks: 44

Sum of negative ranks: 1

Z value: -2.5471 (nb. N too small)

Sample Size (N): 9

Significance Level: 0.05

Two-tailed

Result 1—Z value

The Z value is -2.5471 . However, the size of N (9) is not large enough for the distribution of the Wilcoxon W statistic to form a normal distribution. Therefore, it is not possible to calculate an accurate p value.

Result 2—W value

The W value is 1. The critical value of W for $N=9$ at $p \leq 0.05$ is 5. Therefore, the result is significant at $p \leq 0.05$.

Explanation of results

We have calculated both a W value and Z value. If the size of N is at least 20—see the Results Details box—then the distribution of the Wilcoxon W statistic tends to form a

normal distribution. This means you can use the Z value to evaluate your hypothesis. If, on the other hand, the size of N is low, and particularly if it is below 10, you should use the W value to evaluate your hypothesis.

You should also note that if a subject's difference score is zero—that is, if a subject has the same score in both treatment conditions—then the test discards the individual from the analysis and reduces the sample size. If you have a lot of ties, this procedure will undermine the reliability of the test (and also suggests that the requirement that the data are continuous has not been met).

Problem 22.4

Z ratio

The Z score is 1.5011. The p value is 0.13362. The result is *not* significant at $p \leq 0.05$.

Before	After	Ranks before	Ranks after
235	200	5	1
254	210	6	2
286	220	8	3
299	227	10	4
320	255	11	7
322	297	12	9
376	334	16	13
381	338	17	14
397	341	18	15
		$\Sigma = 103$	$\Sigma = 68$

Based on <http://www.socscistatistics.com/tests/mannwhitney/Default2.aspx>.

CHAPTER 25

Problem 25.1

I used the online calculator at <http://vassarstats.net/anova1u.html> because it allows you to cut and paste. The results were

Setup

Number of samples in analysis =

Independent Samples Independent Samples k=4

Correlated Samples standard weighted-means analysis

Unweighted Click this button only if you wish to perform an unweighted-means analysis. Advice: do not perform an unweighted-means analysis unless you have a clear reason for doing so.

Weighted Click this button to return to a standard weighted-means analysis

Data Entry

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
1190	1774	524	-40	
907	1854	544	-60	
665	1552	444	20	
544	1431	363	20	
423	1310	423	121	
927	1169	605	-504	
1048	1048	544		
645	1230	181		
464	826	60		
262	604	161		
	665	n		

Data Summary

	Samples					Total
	1	2	3	4	5	
N	27	17	14	6		64
ΣX	37292	31748	6670	-443		75267
Mean	1381.1852	1867.5294	476.4286	-73.8333		1176.0469
ΣX^2	89339498	98628324	4564342	274657		19280682
Variance	1455090.0	2458625.0	106658.72	48389.766		1655385.7
Std.Dev.	1206.2711	1568.0003	326.5865	219.9767		1286.6179
Std.Err.	232.147	380.2959	87.2839	89.8051		160.8272

standard weighted-means analysis					
ANOVA Summary Independent Samples k=4					
Source	SS	df	MS	F	P
Treatment [between groups]	25490448.28	3	8496816.096	6.47	0.000723
Error	78798852.57	60	1313314.209		
Ss/Bl					Graph Maker
Total	104289300.8	63			

Ss/Bl = Subjects or Blocks depending on the design.
Applicable only to correlated-samples ANOVA.

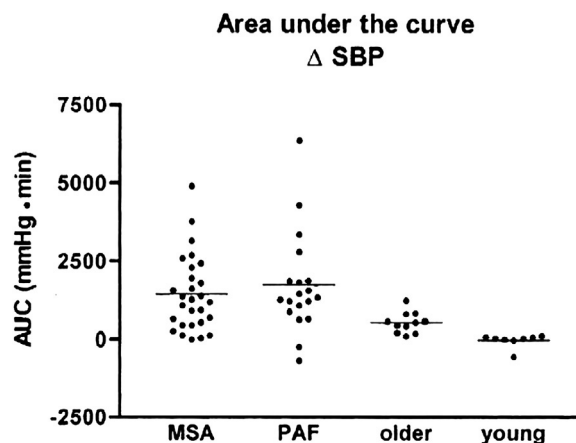
Tukey HSD Test

HSD[.05]=1241.51;
HSD[.01]=1524.63
M1 vs M2 nonsignificant
M1 vs M3 nonsignificant
M1 vs M4 P<.05
M2 vs M3 P<.05
M2 vs M4 P<.01
M3 vs M4 nonsignificant

M1 = mean of Sample 1
M2 = mean of Sample 2
and so forth.

HSD = the absolute [unsigned] difference between any two sample means required for significance at the designated level. HSD[.05] for the .05 level; HSD[.01] for the .01 level.

The figure from which these data came is



Problem 25.2

When the data are ranked, we get

	Ranks for Sample			
count	A	B	C	D
1	3.5	7	3.5	9
2	3.5	8	10	15
3	3.5	14	12	20
4	3.5	18.5	16	21
5	3.5		17	22
6	11		18.5	24
7	13		23	26
8			25	27
Reset	Calculate from Ranks			

The mean ranks of the four samples are, respectively, 5.9, 11.9, 15.6, and 20.5. Then $H=13.23$ with 3 df, and $P=0.0042$. The conclusion is that the null hypothesis of equality of mean ranks might be rejected.

If a standard ANOVA is done on these data, $P=0.0063$. ANOVA is relatively robust.

If we compare groups A and C by Dunn's test, then

$$Q = \frac{15.6 - 5.9}{\sqrt{\frac{27 \times 26}{12} \left(\frac{1}{7} + \frac{1}{8} \right)}} = \frac{9.7}{3.96} = 2.45.$$

The critical value for a family-wise error rate of 0.05 is a z corresponding to $\frac{0.05}{4 \times 3} = \frac{0.05}{6} = 0.0083$. The equivalent value of z is 2.638.

CHAPTER 26**Problem 26.1**

I used the calculator at <http://vassarstats.net/anova2u.html>, and got

Number of rows in analysis = 2
 Number of columns in analysis = 3
 Setup
 2rows x 3columns
 standard weighted-means analysis

Click this button only if you wish to perform an unweighted-means analysis. Advice: do not perform an unweighted-means analysis unless you have a clear reason for doing so.

Unweighted

Click this button to return to a standard weighted-means analysis

Weighted

Data Entry

	Col 1	Col 2	Col 3	Col 4
Row 1	4	7	10	
	5	9	12	
	6	8	11	
	5	12	9	
Row 2	6	13	12	
	6	15	13	
	4	12	10	
	4	12	13	
Row 3				
Row 4				

Reset Calculate

standard weighted-means analysis

ANOVA Summary 2rows x 3columns

Source	SS	df	MS	F	P
Rows	20.17	1	20.17	9.81	0.0058
Columns	200.33	2	100.17	48.73	<.0001
r x c	16.33	2	8.16	3.97	0.0373
Error	37	18	2.06		
Total	273.83	23			

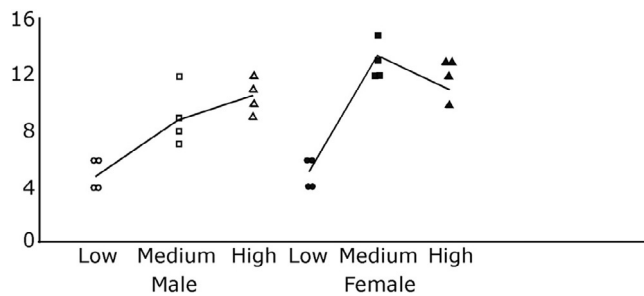
Critical Values for the Tukey HSD Test

		HSD[.05]	HSD[.01]
Rows	2	1.23	1.68
Columns	3	1.83	2.39
Cells	6	3.23	4.02

HSD=the absolute [unsigned] difference between any two means (row means, column means, or cell means) required for significance at the designated level: HSD[.05] for the .05 level; HSD[.01] for the .01 level. The HSD test between row means can be meaningfully performed only if the row effect is significant; between column means, only if the column effect is significant; and between cell means, only if the interaction effect is significant.

The term $r \times c$ indicates the interaction.

It is good practice to put the data on a graph

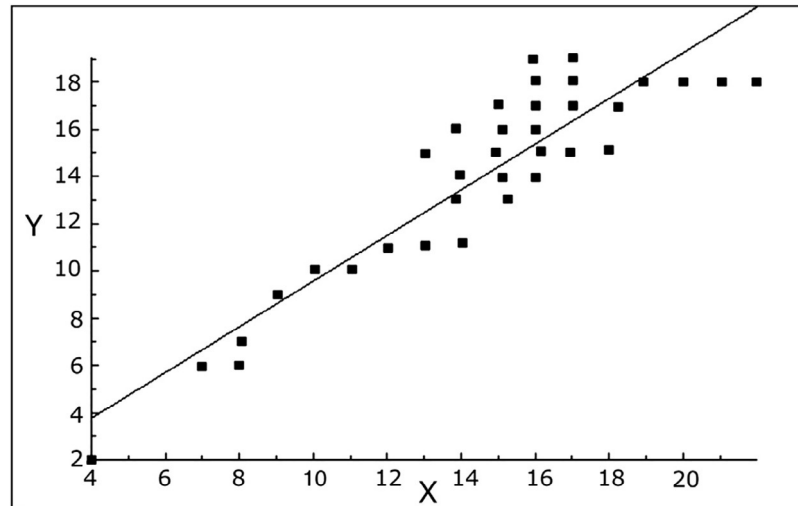


The decrease in weight gain in women on high carbohydrate content indicates the interaction.

CHAPTER 27

Problem 27.1

Draw the graph first. I used and got <http://www.wessa.net/slr.wasp>.



Simple Linear Regression - Ungrouped Data				
Parameter	Value	S.E.	T-STAT	Notes
Constant	2.513420			
Beta	0.850277	0.067717	12.556351	H0: beta = 0
Elasticity	0.826835	0.065850	-2.629690	H0: elast. = 1

Simple Linear Regression - Analysis of Variance			
ANOVA	DF	Sum of Squares	Mean Square
Regression	1.000000	446.250325	446.250325
Residual	33.000000	93.404026	2.830425
Total	34.000000	539.654350	15.872187
F-TEST		157.661949	

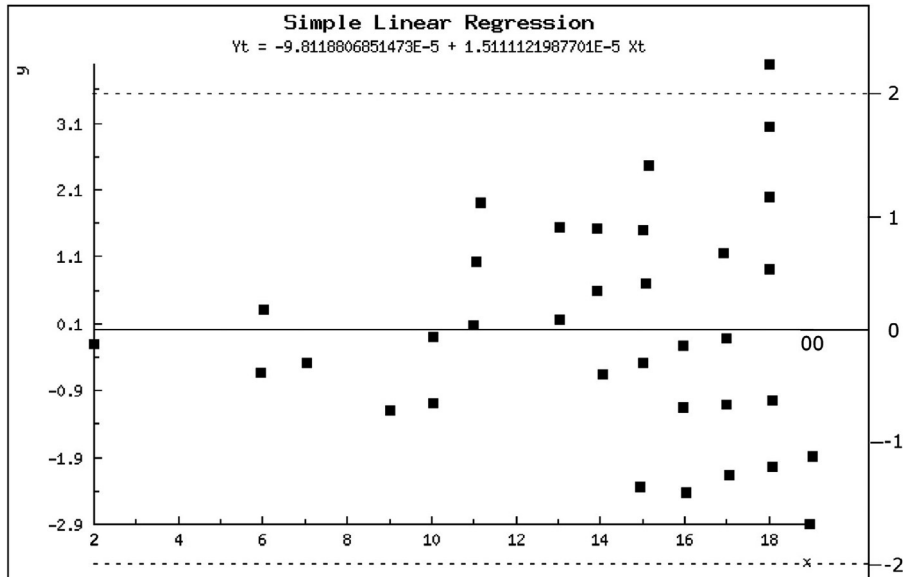
Problem 27.2

I obtained a set of residuals from http://vassarstats.net/corr_stats.html.

Data Entry

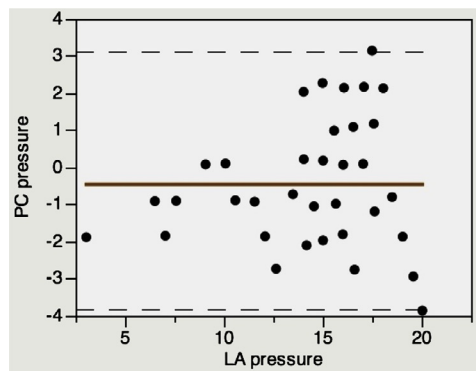
	Data Cells		
Pairs	X	Y	Residuals
1	2.098	3.98	-0.224
2	6.055	6.964	-0.652
3	6.115	7.959	0.292
4	7.134	8.036	-0.51
5	9.113	9.031	-1.221
6	10.132	10.026	-1.105
7	10.132	11.02	-0.111
8	11.091	12.015	0.057
9	11.151	13.01	1
10	11.271	14.005	1.892
11	13.129	13.852	0.137
12	13.129	15.23	1.515
13	14.029	15.995	1.504
14	14.029	15.077	0.586
15	14.149	13.929	-0.666
16	15.048	13.01	-2.36
17	15.108	14.923	-0.499
18	15.168	16.148	0.675
19	15.108	16.913	1.491
20	15.228	17.985	2.46
21	16.067	15.995	-0.253
22	16.067	15.077	-1.171
23	16.127	13.852	-2.448
24	17.146	15	-2.179
25	17.086	15.995	-1.132
26	17.086	16.99	-0.137
27	17.026	18.214	1.139
28	18.106	21.964	3.958
29	18.106	21.046	3.04
30	18.106	19.974	1.968
31	18.106	18.903	0.897
32	18.165	16.99	-1.067
33	18.165	15.995	-2.062
34	19.065	15.918	-2.915
35	19.125	16.99	-1.895

and then plotted the residuals against the original X values with <http://www.wessa.net/slr.wasp> to get



The residuals are scattered evenly above and below the line, and they show some heteroscedasticity. If we divide the Y axis by the standard deviation from regression, which is the square root of the residual mean square of $2.830425 = 1.6823$, we get a Y scale in standard deviations, as shown in the right-hand vertical axis. The range from +2 to -2 standard deviations is shown by dashed lines. As expected, almost all the residuals fall within this range.

Problem 27.3



Note: heteroscedasticity

CHAPTER 29

Problem 29.1

I used <http://www.easycalculation.com/statistics/correlation.php> and got

Results:	
Total Numbers :	35
Correlation :	0.756186190223962

Problem 29.2

Using any of the online calculators gives:

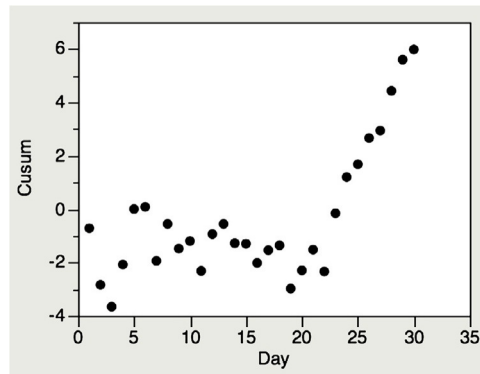
Spearman: $r_s = 0.3935$, $P = 0.023$.

Kendall: $\tau = 0.3738$, $P = 0.13$.

There may be slight differences among programs, depending on whether they correct for ties, and by what method they use. Any differences are likely to be negligible.

CHAPTER 31

Problem 31.1



CHAPTER 35

Problem 35.1

Because the entry of data is all important, I will go over the process step by step, using https://statcom.dk/K-M_plot.php. For batch entry

1. The data must be set out as in the problem, using two separate tables and saving them as text files.
2. Open the online calculator.

Data Format:

Time	Status	Group	Strata	Allowed delimiters
real	integer	optional	optional	comma
decimal point= '.'	event=1	unquoted	unquoted	semicolon
	censored=0	characters[8]	characters[8]	tab

Copy/paste or write Your data below
Column names implicit!!

Type of Confidence Interval

Log

LogLog

Plain

Show CI

Ex 1: Single survival curve. Choose type of confidence interval and check Show CI above to display.

Ex 2: Data includes two groups and tree strata. Click runStratified to apply the stratified LogRank Test.

Ex 3: Tree groups (no strata). Moving cursor to the last section will display variance-covariance matrix.

CAVEAT

Any data in memory, whatever added manually, uploaded or one of the samples, will be included in the analysis. Be sure to remove any data previously loaded using the clear buttons below before entering Your own.

If You get an unexpected result, click [here](#) to check that the format of the current loaded data is correct.

addData

clearAdded

clearUploaded

changeColor

runStratified

clearSample

loadSample1

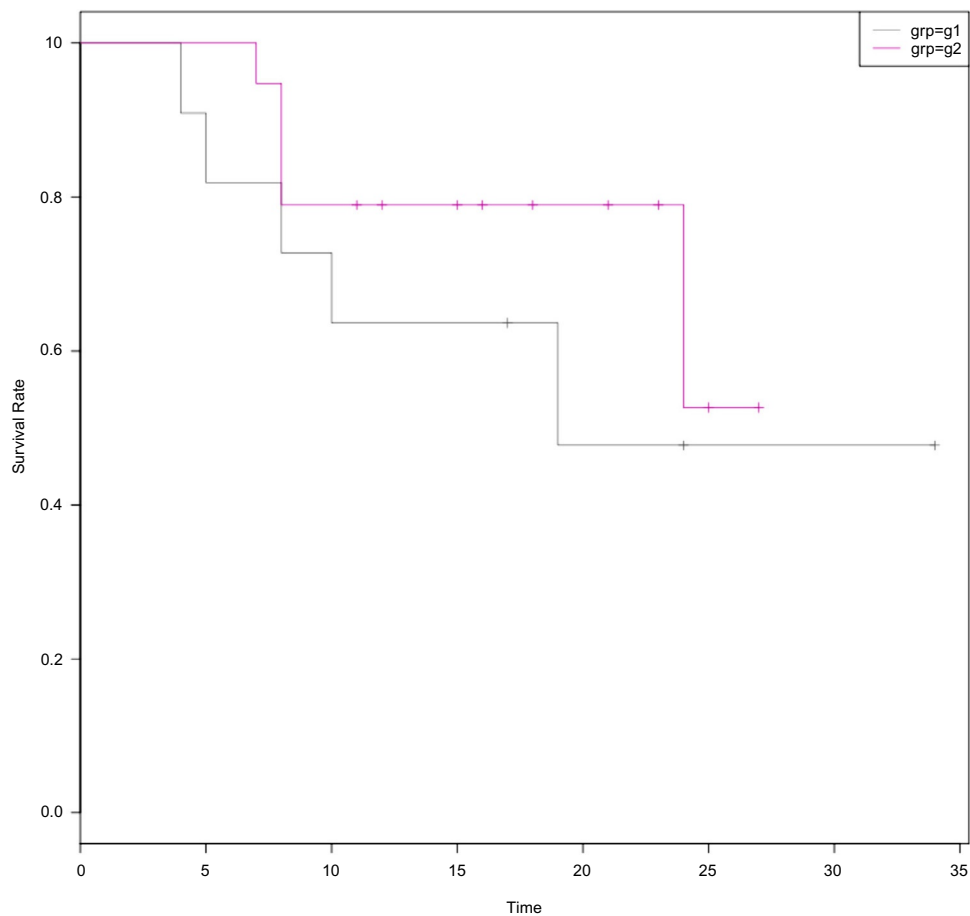
loadSample2

loadSample3

Upload file

Name: Choose File no file selected

3. Select radio button loadSample 1.
4. Select radio button Choose File.
5. Select file from desktop, and select radio button uploadFile.
6. Select radio button loadSample 2.
7. Select radio button Choose File.
8. Select the second file from desktop, and select radio button uploadFile.
9. Your plot should look like this:



10. You may be asked to examine the plot after the first set is entered. Then continue with the second set.
11. No numbers will appear in the box. You can, if you wish, enter data subject by subject by hand.

GLOSSARY

Accuracy This describes how close the measurement is to the true value. For example, if the true cardiac output is 5 L/min (based on some gold standard), then an accurate measurement comes close to that value (e.g., 4.9 L/min), and an inaccurate one will be far from it (e.g., 3.2 L/min).

Alpha The predetermined maximum probability of obtaining a Type I error.

Alternative hypothesis If the null hypothesis (see later) is not accepted, that is, if the two parameters seem to come from different populations, then there are three alternative possibilities:

- Parameter 1 \neq Parameter 2
- Parameter 1 $>$ Parameter 2
- Parameter 1 $<$ Parameter 2

Beta The (predetermined) maximum probability of obtaining a Type II error.

Bias This indicates an outcome differs from the correct answer in a systematic way. Unlike error (see later) the departures from the correct answer are not random.

Biased studies These occur when (intentionally or not) subjects in one group differ from subjects in another group in some meaningful way. For example, if we study the effects of a medication on blood sugar in two groups, one on the medication and one on a placebo, if the placebo group had a large surplus of obese subjects with insulin resistance, the observations will be biased. It is more difficult to make meaningful inferences if samples are biased.

Case-control A study in which groups differing in outcome are identified and compared to determine if there are differences in possible antecedent events (causes). For example, children with and without congenital heart diseases (CHD) are compared to determine if CHD is more common if the mothers took an antidepressant drug during pregnancy.

Cohort A group with a common characteristic, such as being born in the same year, or being exposed to unhealthy atmospheric conditions.

Confidence interval The limits within which the true parameter (such as a mean) is expected to fall with a given probability when observations are sampled from a population.

Confounding If two variables are related to each other by a third factor that might or might not be known, that factor is a confounding factor. For example, an increase in the number of births in a German city parallels the increase in the number of storks nesting there. These two variables are not directly related to each other by cause and effect, but both are due to an increase in the city's population and the number of houses. Population growth is the confounder.

Correlation This describes how closely two (or more) variables are associated with each other.

Dependent variable The outcome of interest (the effect) that might change when an intervention is done.

Effect size The absolute magnitude of the difference between two (or more) parameters. The relative effect size is the absolute effect size related to the variability of the sample.

Error This has a specific meaning in statistics. It refers to variability from a parameter. Thus if the population mean weight of adult females is 50 kg, a particular member of a sample of adult females might weigh 56 kg. This difference from the parameter is termed the error or individual deviation from the parameter. In general, errors might be due to measuring errors, random variation, or to bias. It is the function of a good experimenter to eliminate bias and try to reduce individual variability as much as possible.

Estimation The process of deriving a value for a population parameter, leading to point and interval estimates.

Expectation This is also known as the mathematical expectation, symbolized by E . It can be regarded as the long-run average of the outcomes of many repeated experiments. More formally, it is the weighted

average of all possible values of the random variable. For example, we expect intuitively that in the long run tossing a coin many times will average out as 50% heads and 50% tails. This 1:1 ratio might not apply to any one small series of tosses, but as the number of trials increases the average will approach 50% of each.

Explanatory variable See independent variable.

Independent variable This is what is being manipulated or changed (the cause) to try and produce an effect.

Mathematical expectation See expectation

Null hypothesis The hypothesis that two parameters (means, slopes, etc.) are so similar that they could have been sampled from the same population. Unless the parameters are identical it is never possible to state that they could not have been sampled from two different populations, but the degree of uncertainty can be specified.

Normal An unfortunate word. In ordinary speech it implies health or the usual state. In statistics, it implies that the distribution is compatible with a specific formula. It is a symmetrical distribution with certain properties, and to avoid ambiguity is sometimes called the Gaussian or bell-shaped curve or distribution.

Overdispersion This describes a distribution, usually of counts, with a variance much greater than expected for a typical binomial or Poisson distribution. It may be due to misclassification, outliers, or clumping of data due to extraneous events (such as having high and low malaria attack rates with few intermediate values).

Parameter A quantity that describes or characterizes a population. For example, a measure of central tendency, the mean, is a parameter that gives information about where the center of the distribution is. Parameters are symbolized by Greek letters, such as μ , β , ρ , σ , and so on. This term should not be used when the term variable is meant (see Variable).

Power This is the probability of finding a true difference between parameters and is calculated as $1 - \beta$ where β is the Type II error (see later).

Precision This indicates how reproducible the measurement is. If we repeat the measurement 10 times, are the results all close together, or are they widely scattered? It is important to note that a measurement may be precise but inaccurate. If the cardiac output is 5 L/min, 10 repeated measurements of that output might range from 3.5 to 3.7 L/min; they would be very precise but inaccurate.

Probability density function If there is a random variable of the discrete type, then it is possible to calculate or determine experimentally the probability of $P(X=0)$, $P(X=1)$, ..., $P(X=n)$. The probability $P(X=x)$ is sometimes called the probability density function and symbolized by $f(x)$.

Quantile The fraction (or percent) of points of a distribution below the given value of the quantile. Thus 10% of the distribution is below the 10th quantile.

Random Randomness is a property that pertains to samples drawn from a population. To be random, each member of the sample must have an equal chance of being chosen as a member of that sample, and drawing one member should not influence the drawing of the other members of the sample.

Reexpression Finding a new scale, for example, a logarithmic or square root scale, to simplify the data analysis. There is no necessary physical or physiological connotation in selecting the reexpression.

The fact that a logarithmic scale makes analysis easy does not mean that the process itself is logarithmic, although it might well be.

Regression This indicates how much a dependent variable Y changes for a unit increase in explanatory variables X_1, X_2, \dots, X_n .

Residuals Residuals are what are left over when theoretical expected values are subtracted from the actual observed values. For example, when a straight line is fitted to a series of X, Y points, the vertical differences of the observed Y values from the values of Y that lies on the line for those values of X are the residuals. Therefore

- Residual = Data minus fitted value

Resistant The parameter is little affected by unusually small or big measurements in the data set

Robust This is the quality of a test that implies insensitivity to departures from the assumptions underlying the probabilistic model. For example, the t-test for comparing two groups of measurements is not very robust, and becomes inefficient if the distributions are very asymmetrical, whereas the Mann-Whitney U test is very robust.

Scattergram Also termed scatter plot or scatter diagram. This is a two-dimensional plot that shows how two variables are related.

Significant Another unfortunate word. In ordinary speech, it implies importance. In statistics, however, it has nothing to do with importance, but rather refers to the chances of rejecting the null hypothesis incorrectly (Type I error).

Significant figures These are the digits in a number that denote the accuracy of measurement. If we describe the weight of a small animal as 50.32 g, then we are stating that the weight lies within the range 50.315 and 50.325 g. In this example, we are using 4 significant figures, in which the first three are accurate and the last has potential error.

Statistic This is similar to a parameter, except that it characterizes a sample. Statistics are written as ordinary letters, such as m , b , r , s , and so on. This is not to be confused with other uses of the word “statistic,” which can be taken to mean a given value. For examples, we hear the phrase: “Don’t become a traffic statistic,” or we talk about vital statistics, by which we mean the numbers of births, deaths, and so on, that occur in a community. The context usually distinguishes these two uses.

Transformation See reexpression.

Type I error This occurs if we reject the null hypothesis when in fact it is true. The Type I error rate is usually denoted by α . If $\alpha = 0.05$, then we reject the null hypothesis when $P < 0.05$. Usually we determine what value of α to use; it is sometimes 0.01 or 0.10, or occasionally other values.

Type II error This occurs when we accept the null hypothesis when it is false. The Type II error rate is usually denoted by β . This is determined by the difference between the parameters, the sample size, the sample distribution, and the value of α . It has to be calculated separately for each data set.

Unbiased If many samples are taken with replacement from a population, then any statistic (such as the mean) calculated from those samples will vary. On the average, however, the average sample statistic will equal the population parameter. An estimate in which the average value over all possible samples equals the population parameter is called *unbiased*.

Variable This is a number that can take on one of several values. Thus if we measure the weights of 50 adult females, each weight is a variable. Note that although parameters can also vary (e.g., the mean weight of adult females is not the same as the mean weight of adult males), the parameter is usually a distillate of all the variables and a property of all the variables in the data set.

INDEX

Note: Page numbers followed by “*f*” indicate figures, “*t*” indicate tables, “*b*” indicate boxes, and “*np*” indicate footnotes.

A

- Abnormal distribution, 128–129*f*, 363, 511
- AcaStat program, 31
- Accuracy of statistics measurement, 39–41, 42*f*
 - precision, 41
 - rounding off, 40–41
 - truncation, 40–41
- Acetylcholine, 440
- Achalasia data, 208–209, 209*t*
 - chi-square, 209
 - cross-product ratio, 209
 - hypothetical extension, 209, 209–210*t*
- Actuarial method. *See* Berkson’s actuarial method
- Adaptive designs, 387–388
- Additive *vs.* multiplicative tests, 362*t*
- Additivity, 423–425, 423–424*t*
- Adjusted R^2 , 528–529, 538
- Aerosolized tobramycin, 369
- Affinity, 575–576
- Agresti–Coull adjustment, 264
- Akaike information criterion (AIC), 529
- ALLHAT study, 651–652
- All subsets regression, 528–529, 540–543
- American Statistical Association, 166–167
- American Thoracic Society, 34, 35*t*
- Analysis of variance (ANOVA)
 - all subsets regression, 540–543
 - centered *vs.* noncentered cubic curves, 553*t*
 - cubic *vs.* quartic fit, 550*t*
 - one-way
 - basis of, 391, 391*f*, 394*t*
 - concepts, 391–405
 - with different means, 392, 392*f*
 - effect size, 403–404
 - homogeneity of variance, 399–401
 - independence of observations, 403
 - Kruskal–Wallis test, 401–403, 401*t*
 - linear combinations, 414
 - multiple comparisons, 405–414
 - planned experiments, 415–416
 - requirements, 397–398
 - sample size, 404–405
 - stellate ganglion nerve activity, 398, 398*t*, 398*f*
 - studentized range test, 406–413, 407*t*
 - unplanned experiment, 405–406
 - quadratic *vs.* linear fit, 550*t*
 - two-way, 420–423, 422*t*, 422*f*
 - additivity, 423–425, 423*t*
 - for all anxiety and treatment data, 430*f*
 - Cochrane’s Q-test, 426–427, 426–427*t*
 - Friedman test, 425–426
 - for high *vs.* low anxiety, 430*f*
 - with interactions, 427–432, 428*f*
 - missing data, 433–434
 - model II, 438–439
 - multiple factors, 425
 - nested designs, 434–437, 434*t*
 - with no interaction, 427, 428*f*
 - repeated measures designs, 439–440
 - with replication, 432
 - transformations, 438
 - treated *vs.* untreated, 430*f*
 - unequal cell sizes, 433–434
 - weanling pigs for vitamin B₁₂, 393–394, 393–394*t*
- Angina pectoris, 36
- Anscombe’s data sets, 472, 472*t*, 473*f*
- Anscombe–Tukey transformation, 424
- Antilogarithms, 56, 191
- Apgar score, 36, 36*t*
- Area under the curve (AUC), 331–332
- Argument, 43
- Arithmetic mean, 62–64, 69
- ARL. *See* Average run length (ARL)
- Artificial data set, 60, 60*t*
- Asymmetrical funnel plot, 624–625
- Attributable risk (AR)
 - confidence limits for
 - difference between proportional ratios, 308
 - differences between proportions, 307–308
 - definition, 303–304
 - population attributable risk, 303–304
- Autocorrelation. *See* Serial correlation
- Auxiliary regression, 540, 541*t*
- Average run length (ARL), 570–571

B

Backward regression, 529–530, 548
 Bacterial counts, Poisson distribution, 273–275*b*, 274*t*
 Bacterial infections (OB), 97
 Bar graphs, 77–78, 138, 138–139*f*
 Bartlett's method, 496
 Bayes factor, 167–169, 168*t*
 Bayesian Information Criterion (BIC), 529
 Bayes' theorem, 98–100, 166, 311–314
 neural tube defect diagnosis, 311–314
 sensitivity, 314–336
 specificity, 314–336
 Berkson's actuarial method, 601
 Berkson's fallacy, 212–214, 212–213*f*, 213–214*t*
 Bernoulli coefficients, 251, 251*t*
 Bernoulli formula, 252–253
 Bernoulli trial, 249–250
 negative binomial distribution, 287
 β coefficients, 582
 Between-subjects design, 556–557
 Bias index (BI), 246
 Binomial calculators, 253–254
 Binomial distribution, 270–271, 358*f*
 comparisons, 259
 confidence limits, 256–258
 continuity correction, 256–258, 257*t*, 257*f*
 cumulative probabilities, 256
 mean, 254
 normal approximation, 254, 255*f*
 sample size, estimation, 259
 Tay-Sachs disease, 254, 254*t*
 variance, 254
 Binomial theorem, 108, 233–234
 continuity correction, 256
 proportions and, 263–264
 Bins, 50
 Bivariate correlation, 537, 537*t*
 Body mass index (BMI), 34–35
 Bonferroni correction, 376–382, 377*t*, 414
 control *vs.* runt lambs, 378–379, 379*f*
 Bootstrapping, 631–633, 633*f*
 Bowker's test, 236, 236*t*
 Box plots, 80–81, 80*f*
 Brain abscess (BA), 97
 Break point analysis, linear regression, 496–497, 497*f*
 Breathlessness
 American Thoracic Society, 34, 35*t*
 visual analog scale, 35*f*

Breusch-Pagan test, 476
 BrightStat program, 31, 558
 Bronchiectasis scoring system, 36
 Bubble graph, 78, 79*f*, 80
 Bulging rule, 447–448, 448*f*

C

Calcipotriene, 300
 Cancer biology, 24
 Categorical and cross-classified data
 Berkson's fallacy, 212–214
 Bowker's test, 236, 236*t*
 Cochran-Armitage test, 224–226, 225*t*
 Cochran-Mantel-Haenszel test, 221–224, 222*t*
 concepts, 197–221
 concordance, 241–246
 interobserver agreement, 241*t*
 interrater agreement, 243*t*
 more than two categories, 243
 two categories, 241–243
 weighted kappa, 243–246, 245*t*
 contingency tables, 200–206, 201–203*t*, 205*t*
 continuity correction, 199–200
 degrees of freedom, 200
 extended contingency tables, 214–216, 214*t*
 Fisher's exact test, 216–219, 216–217*t*, 217–218*f*
 goodness of fit, 197–200
 Kolmogorov-Smirnov tests, 236–241, 237*f*
 one-sample test, 237–239, 238*t*
 two-sample test, 236*t*, 240*t*
 McNemar's test, 233–235, 233–235*t*
 multidimensional tables, 227–230, 227*f*, 228*t*
 odds ratio, 206–208
 paired samples, 233–235, 233–234*t*
 pooling data, 224–226
 power and sample size determination, 219–221
 problems with $R \times C$ tables, 226–227
 sample selection, 208–214*b*
 Simpson's paradox, 210–212, 211*t*
 Censoring, 33–34, 613
 Centering, 552, 552*f*
 vs. uncentered regression, 553*t*
 Central limit theorem, 127–131, 375
 Chatillon's balloon trick, 509–510, 509*f*
 Chi-square distribution, 117–118, 145–146, 146*f*
 Chi-square test, 19–20, 203*t*, 256, 272
 of Mendelian experiment, 198–199, 198*t*
 CL. *See* Confidence limits (CL)
 Classical *vs.* robust methods, 81–85
 Clinical trials, 645–650

- Clumping, 290
- Cochran–Armitage test, 224–226, 225*t*
- Cochrane’s Q-test, 426–427, 426–427*t*
- Cochran–Mantel–Haenszel (C–M–H) test, 221–224, 222*t*
- Coefficient of determination, 501
- Coefficient of reliability, 519–520
- Coefficient of variation, 76, 361
- Cohort study, 295–297
- Combinations, 190–191
- Comparison-wise error rate (CWER), 346, 378–380
- Competing risk analysis, 615–617
- Comprehensive meta-analysis, 623
- Conditional probability, 97–98, 161–162, 609
- Confidence interval, 259–260
- Confidence limits (CL), 6–8, 131–134, 135*f*, 139*f*, 256–258, 605–606
 - advantages, 140
 - for attributable risk
 - difference between proportional ratios, 308
 - differences between proportions, 307–308
 - linear regression, 466–469, 466–467*f*, 469*f*
 - for medians, 363
 - for NNT, 309, 309*f*
 - paired *t*-test, 344*f*
 - Poisson distribution
 - exact method, 276–277
 - normal approximation, 276
 - proportions, 264–265
 - Wilson’s method, 260
- Confounders, 7–8, 10–11, 221–224
- Congenital hypothyroid data, 328, 328*f*
- Consecutive measurements, 56, 56*t*
- Constants, 42
- Contagious distribution, 290, 592–593
- Contaminated distribution, 150–152, 151*f*
- Contingency tables
 - extended, 214–216, 214*t*
 - 2×2 table, 200–206, 201–203*t*, 205*t*
- Continuity correction, 256–258, 257*t*, 257*f*, 260
 - categorical and cross-classified data, 199–200
- Continuous distribution, 51, 51*f*
 - chi-square, 145–146
 - exponential, 143–144
 - logarithmic, 144–145
 - uniform, 143
 - variance ratio (*F*), 146–147
- Continuous variable, 33, 36, 38–39, 327
- Control charts, 566–568. *See also* Serial measurements
- Cook’s distance, 484–486
- Coronary blood flow regression
 - Bland–Altman diagram, 470, 471*f*
 - measured *vs.* calculated, 468, 469*f*
 - microsphere method, 470, 471*f*
- Coronary heart disease, 36
- Correctly specified model, 528
- Correlation, 502*f*
 - age *vs.* right ventricular weight, 507*f*
 - Chatillon’s balloon trick, 509–510, 509*f*
 - coefficient, 37
 - confidence limits, 503–504
 - effect of change in slope, 510, 510*f*
 - intraclass, 519–522
 - Kendall’s tau test, 512–514
 - online programs, 502–503
 - partial, 514–515, 515*b*
 - P values, 503–504
 - sample size and power, 504–511
 - Spearman’s test, 511–512
 - spurious, 516–518
- Counting, 49, 49*f*
- Covariance, 454
- Cox regression model, 10–11
 - ablation data, 614*t*
 - in nephrology and hepatology, 615
 - proportional hazards, 612–615
- C_p statistics, 529
- Cross-over designs, 369
- Cross-over trials, 367–371, 368–371*t*
 - acetazolamide in preventing acute mountain sickness, 370, 370–371*t*
 - aerosolized tobramycin in treating patients with cystic fibrosis, 369
 - N of 1 trials, 371
- Cross-product ratio. *See* Odds ratio
- Cubic regression, 552, 552*f*
- Cumming program, 32
- Cumulative binomial probabilities, 256
- Cumulative frequency, 51
- Cumulative poisson probabilities, 278–279, 279*f*
- Cumulative probability density function, 51, 113*f*
- Cumulative sum (Cusum), 568–572
- Cumulative survival function, 609
- Cutting points, 327–329, 329–330*f*
- Cyanotic heart disease, 97

D

D'Agostino's test, 117
 Data, 9
 on blood pressure and strokes, 296*t*
 from children with occult bacteremia
 neutrophil count, 314, 314*t*
 sensitivity, 314, 314*t*
 specificity, 314, 314*t*
 pooling, 224–226
 sets, 38–39, 38*t*
 Degrees of freedom, 73, 200, 354
 Descriptive statistics, 6–7, 19–20
 Deviance, 592
 Diabetes mellitus, 36
 Discrete variable, 33
 Disease incidence, 295
 Distribution-free tests, 355–361
 Distributions, 50–54
 age distribution, hypothetical data, 54*f*
 continuous distribution, 51*f*
 discrete frequency, 50*f*
 frequency polygons, 51–54
 histograms, 51–54
 shapes of, 55–56
 sorting experiment, 50
 transformations, 55–56
 Distribution shape
 kurtosis, 116–117
 skewness, 116
 Dog data, 122*t*
 Dose-response curves, 575–576
 ED50, 575–576, 576*f*, 579
 LD50, 575–576
 quantal response, 575, 577–579
 sigmoid curve, 575, 576*f*, 577
 Double-blind trial, 650
 Dow-Jones index, 561
 Dummy/indicator variable, 555–558
 Duncan's test, 412–413
 Dunnett's test, 416, 426
 Dunn-Sidak equation, 377
 Durbin-Watson test, 527–528, 564–565
 Dyspnea, 37

E

Early Motor Pattern Profile (EMPP), 329
 ED50, 575–576, 576*f*, 579
 Effect cell coding, 556, 556*t*
 Effect size
 analysis of variance, 403–404

hypothesis testing, 175–180
 Efficacy, 575–576, 576*f*
 E-labs, 32
 Emergency medicine, 179
 Empirical probability, 91–92
 Enhancing the QUality And Transparency Of
 health Research (EQUATOR) program,
 24–25
 Equivalence tests, 371–374, 373*f*
 Errors, 5
 independence of, 403
 rates, 378–380
 comparison-wise, 378
 experiment-wise, 378
 Exact method, 276–277
 Expanded Mendelian experiment, 199, 199*t*
 Experiment-wise error rates. *See* Family-wise error
 rates (FWER)
 Explanatory variable
 multiple, 586–588
 single, 583–585
 Exploratory data analysis, 446
 Exploratory descriptive analysis
 advanced and alternative concepts
 applications of variance, 87
 classical *vs.* robust methods, 81–85
 propagation of errors, 85–87
 box plots, 80–81
 counting, 49
 distributions, 50–54
 shapes of, 55–56
 least squares principle, 88
 measures of central tendency (location), 62–70
 measures of variability, 70–76
 stem and leaf diagrams, 56–62
 tables and graphs, 76–80
 Exploratory Software for Confidence
 Intervals (ESCI), 132
 Exponential distribution, 143–144, 144*f*
 Exponents, 191
 Extracorporeal membrane oxygenation (ECMO),
 388
 Extreme values, 153, 154*f*

F

False discovery rate (FDR), 380–382, 381*f*, 382*t*
 Family-wise error rates (FWER), 378–380
F distributions, 146–147, 147*f*
 FDR. *See* False discovery rate (FDR)
 Fever, age and causes, 97*t*

FFR. *See* Fractional flow reserve (FFR)
 Figures, 77–80
 Fisher–Irwin test, 194–195
 Fisher’s test, 161
 exact test, 194–195, 195*t*, 216–219, 216–217*t*,
 217–218*f*
 Z transformation, 504
 Forest graphs, 621–623
 Forward regression, 529–530, 543–547
 Fractional flow reserve (FFR), 321
 Framingham Heart Study, 640–641
 Free graphics programs, 32
 Free online calculators, 116
 Free online tests, 31–33
 Frequency curve, 113*f*
 Frequency polygons, 51–54, 52*f*
 Friedman test, 425–426
 Funnel plots, 623–625
 FWER. *See* Family-wise error rates (FWER)

G

Gaddum’s lambda index, 493, 493*f*
 Galbraith plot, 625, 625*f*
 Gamma regression, 594–595
 Gaussian curve, 107–113, 109–110*f*, 136*f*
 Gaussian distribution, 117
 Geometric mean, 68–69, 68*f*
 Glejser test, 476
 Goldfeld–Quandt test, 475–476
 Graeco–Latin square, 425
 Graphic tests, 118
 Graphs, 76–80
 Grouped frequency distribution, 66*t*
 Grubb’s test, 150

H

Harmonic mean, 69–70, 70*t*
 Hartley’s constant, 567
 Hawthorne effect, 650
 Hazard function, 609–610
 Hazard ratio, 610–612
 Heights, cumulative frequency, 120–121*f*
 Heteroscedasticity, 451, 475–477, 534*f*, 536
 Histograms, 51–54, 53*f*, 57*f*, 110*f*, 117
 Historical controls, 642
 Hochberg test, 413
 Homogeneous variances, 399–401, 400*t*
 Honest significant difference method (HSD), 408,
 410, 411*t*
 Hooke’s law, 168

Hotelling’s T^2 test, 558, 588
 HSD. *See* Honest significant difference
 method (HSD)
 Hypergeometric distribution
 Fisher–Irwin test, 194–195
 Fisher’s exact test, 194–195
 formula, 193–194
 multiple groups, 195
 Hypothesis testing, 7, 159–165, 161*t*
 Bayes factor, 167–169
 effect size, 175–180
 maximum likelihood, 167–168
 null hypothesis significance test, 159, 165–167
 posthoc power analysis, 180–181
 power calculation, 182–184, 182*f*
 statistical power, 173–175
 statistical significance, 163–164
 type I error and α , 169–170
 Hypoxic–ischemic brain injury (HIE), 402

I

Infective endocarditis (IE), 97
 Influence, 484–486
 Inhomogeneity of variance, 400–401
 Intent-to-treat principle, 647
 Interactive statistics, 32
 Interim tests, 383, 384*t*
 International Federation of Clinical Chemistry, 115
 Interobserver agreement, 241*t*
 Interquartile distance (IQD), 74–75, 80–81
 Interrater agreement, 243*t*
 Interval scale, 34
 Intraclass correlation (ICC)
 rater agreement, 520–522
 reliability, 519–520
 Intrinsically linear, 448
 IQD. *See* Interquartile distance (IQD)
 Ischemic heart disease, 36

J

Jackknife, 635
 Jarque–Bera test, 117–118
 Justification for the Use of Statins in Prevention
 (JUPITER) trial, 649

K

Kaplan–Meier method, 602, 602*f*
 Kendall’s tau test, 512–514
 Kolmogorov–Smirnov (K–S) tests, 236–241, 237*f*
 cumulative frequencies, 237, 237*f*

Kolmogorov–Smirnov (K–S) tests (*Continued*)one-sample test, 237–239, 238*t*two-sample test, 236*t*, 240*t*Kruskal–Wallis test, 401–403, 401*t*K–S tests. *See* Kolmogorov–Smirnov (K–S) tests

Kurtosis, 116–117, 123

L

L'Abbé plot, 626–627

Ladder of powers, 447, 447*t*

Lan–De Mets approach, 647–648

Latin squares, 425

LD50, 575–576

Lead-time bias, 335, 336*f*Least significant difference (LSD), 408, 410, 411*t*

Least square method, 14, 88

Left censored data, 33–34

Lehr equation, 177–178, 178*t*, 265–266Leptokurtotic curves, 115*f*

Levene's test, 399, 476

Leverage, 483–484

Life table, 601, 601*t*Likelihood ratio (LR), 167, 167*f*, 323–327nomograms for, 325–327, 326*f*of prolactin, 327, 327*t*

Likert scales, 36

Lilliefors test, 117

Linear combinations, 414

Linear regression

ANCOVA, 480, 480–481*f*, 481*t*ANOVA, 456, 456*t*break point analysis, 496–497, 497*f*

calculations, 459–460

calibration curves, 494, 494*f*

comparison

methods, 470–475, 471*f*

problem, 477–478

two or more lines, 478–482

concepts, 445–475

confidence limits, 466–469, 466–467*f*, 469*f*covariance, 454, 454*t*errors in *X* variate, 495–496

exploratory data analysis, 446

heteroscedasticity, 475–477

inverse prediction, 491–493, 492*f*least squares in, 452*f*

line of best fit passes through zero, 494–495

model I, 450–459

outliers, 482–486, 482*f*product deviation from mean, 454–455, 454–455*t*pulmonary capillary wedge *vs.* left atrial (LAP) pressures, 459ratio measurements and scaling factors, 486–488, 486–488*f*relationship of platelet count to bleeding time, 456, 457–458*t*, 457*f*, 474, 475*f*residuals, 460–465, 460–464*f*

resistant lines, 497–499

slope calculation, 452–453, 453*f*

transforming curves, 446–450

transforming *Y* variate, 491Line diagram, 64*f*

Line graphs, 77–78

Logarithmic distribution, 144–145

Logarithms, 191

Logistic function, 581, 581*f*

Logistic regression, 582, 591

advantages, 588

multicollinearity, 589

multiple explanatory variable, 586–588

sample size and power calculations, 588–589

single explanatory variable, 583–585

Logit transformation, 582

Log linear analysis, 230, 230*t*

Log-rank test, 606, 608–609

Longitudinal regression, 558–559

Love plot, 653, 653*f*

Lower control limit, 567

LOWESS function, 551

LSD. *See* Least significant difference (LSD)

Lyme disease, 318

MMAD. *See* Median absolute deviation (MAD)Mann–Whitney U-test, 359–361, 360*t*, 364, 364*t*Mantel–Haenszel test, 606, 608*t*

Mathematical model, 9

Maximum likelihood method, 167–168

McNemar's test, 233–235, 233–235*t*Mean, 68, 68*f*, 134*f*, 137–138

dividing, 87

multiplying, 86–87

square, 71–72

Measurement scales, 34–38

Measures of variability, 70–76

Median, 64–68, 68*f*

Median absolute deviation (MAD), 84–85, 152

- Medical Research Council (MRC), 642–643
 - Mendelian experiment
 - chi-square analysis, 198–199, 198*t*
 - expanded, 199, 199*t*
 - hypothetical, 197–198, 197*t*
 - Meta-analysis, 621
 - forest graph, 621–623
 - funnel plot, 623–625
 - L'Abbé plot, 626–627
 - radial plots, 625
 - Michaelis–Menten kinetics, 554
 - Microspheres, 284
 - Midmean, 83
 - Minimum significant difference (MSD), 407
 - Missing at random (MAR), 433
 - Missing completely at random (MCAR), 433
 - Missing not at random (MNAR), 433
 - Mixed regression, 529–530
 - Mode, 68, 68*f*
 - Monotonic function, 37
 - Monte Carlo methods, 635–636
 - Multicollinearity, 527
 - correcting for, 552–554
 - detection, 538–540
 - logistic regression, 589
 - Multidimensional tables, 227–230, 227*f*, 228*t*
 - Multinomial distribution, 261–262
 - Multiple comparisons, 375–376, 375*f*
 - adaptive methods, 387–388
 - analysis of variance, 405–414
 - linear combinations, 414
 - planned experiments, 415–416
 - studentized range test, 406–413, 407*t*
 - unplanned experiment, 405–406
 - Bonferroni correction and equivalent tests, 376–382, 377*t*
 - error rates, 378–380
 - extreme multiplicity, 380–382
 - false discovery rates, 380–382
 - group sequential boundaries, 382–385
 - sequential analysis, 385–387
 - Multiple explanatory variable, 586–588
 - Multiple groups, 195
 - Multiple regression techniques, 432–433
 - dummy/indicator variable, 555–558
 - with many independent *X* variates, 533–538
 - requirements, 527–528
 - with two *X* variables, 525–527
 - Multiplication rule, 94–95
 - Multiplicity, 380–382
 - Multiplying means, 86–87
 - Multivariable graphs, 78
 - Multivariate analysis/MANOVA, 558
 - Multivariate statistics, 5
- N**
- Nausea scale, 34–35, 35*f*
 - NED. *See* Normal equivalent distribution (NED)
 - Negative autocorrelation, residuals, 465, 465*f*
 - Negative binomial distribution
 - Bernoulli trial, 287
 - overdispersed distribution, 289–290
 - probability of success, 287–289, 288*t*
 - uses, 290
 - Negative likelihood ratio (LR–), 323
 - Negative predictive value, 316
 - Nelson's rules, 567, 568*f*
 - Nested designs, 434–437, 434*t*, 434*f*
 - ANOVA, 435, 435*t*, 435–436*f*, 437, 437*t*
 - Neural tube defect diagnosis, Bayes' theorem, 311–314
 - Neuroscience, 179
 - New York Heart Association class (NYHA), 38–39
 - Neyman–Pearson approach, 164–165
 - NHST. *See* Null hypothesis significance test (NHST)
 - Noise, 5
 - Nominal scale, 37
 - Nomograms, 325–327, 326*f*
 - Noncohort study, 297–302
 - Noninferiority tests, 371–374
 - Nonlinear regression, 531–532
 - dose-response analysis, 577
 - principles, 554
 - Nonparametric tests, 124, 355–361
 - Normal chart, 107
 - Normal distribution curves
 - areas under, 112*f*
 - components, 111*f*
 - determining normality, 117–121
 - distribution shape
 - kurtosis, 116–117
 - skewness, 116
 - Gaussian curve, 107–113
 - populations and samples, 114
 - properties of, 111–113
 - quincunx, 110–111
 - ungrouped data, 122–124

Normal equivalent distribution (NED), 119, 119*t*, 122–123*f*, 578
 Normal Gaussian curve, 129*f*
 Normal probability density, 108*f*
 Normal range, 115
 Notation, 42–43
 Null hypothesis, 159, 173, 263–264, 266
 degree of evidence, 163*f*
 Null hypothesis significance
 test (NHST), 159, 165–167
 Number needed to treat (NNT), 305–307
 confidence limits for, 309, 309*f*
 Nurses' Health Study (NHS), 641

O

Objective probability, 91
 O'Brien-Fleming method, 383
 Odds ratio (OR), 206–208, 295
 One size fits all approach, 649
 One-way ANOVA
 basis of, 391, 391*f*, 394*t*
 concepts, 391–405
 with different means, 392, 392*f*
 effect size, 403–404
 homogeneity of variance, 399–401
 independence of observations, 403
 Kruskal-Wallis test, 401–403, 401*t*
 linear combinations, 414
 multiple comparisons, 405–414
 planned experiments, 415–416
 requirements, 397–398
 sample size, 404–405
 stellate ganglion nerve activity, 398, 398*t*, 398*f*
 studentized range test, 406–413, 407*t*
 unplanned experiment, 405–406
 Operators, 43–45
 Ordered measurements, 56, 56*t*
 Order statistics, 84–85
 Ordinal numbers, 511
 Ordinal scales, 34, 36
 Outliers, 482–486, 482*f*
 contamination, 150–152
 cranial capacities, 151*f*
 dealing with, 152
 definition, 149
 Grubb's test, 150
 robust tests, 152
 slippage, 150–152
 Overdispersion

 correcting for, 594–595
 definition, 592–593
 detection, 594
 distribution, 289–290, 289*t*
 Overspecified model, 528

P

Paired patent ductus arteriosus study, 234, 234*t*
 Paired *t*-test, 341–350
 additive model, 361–363, 362*t*
 confidence limits, 344*f*
 histogram of differences, 344*f*
 peak flow rates in asthmatic patients before
 and after exertion, 343
 sample size for, 350
 type I error, 346
 weight gain, 342, 342*t*
 Parameters, 5–6
 Park test, 476
 Partial correlation, 514–515, 515*b*
 Pascal distribution, 287
 Pascal's triangle, 251
 p-capture, 135*f*
 Pearson's correlation coefficient, 19–20, 512, 522
 Permutation tests, 189, 634, 635*f*
 Peto approximation, 606
 P hacking, 205
 Phocomelia, 284
 Piecewise linear regression, 496
 Placebo effect, 650–651
 Platykurtotic curves, 115*f*
 Point estimation, 6
 Poisson distribution, 144, 270*f*, 383–384, 438
 chi-square test, 272
 comparisons
 counts based on same units, 280
 counts not based on same units, 280–282
 ratio of two Poisson variates, 282–283
 confidence limits
 exact method, 276–277
 normal approximation, 276
 cumulative Poisson probabilities, 278–279, 279*f*
 deaths from mule kicks in cavalry corps, 271, 271*t*
 differences between means of, 279–283
 discrete events, 269
 ratio of variance to mean, 274–275
 sample size determination, 283–285
 square root transformation, 277–278
 Poisson regression, 591
 suitability, 592–594

uses, 592
 Poisson variates, 282–283
 Polonium, radioactive decay
 counts of, 273, 275
 Polynomial regression, 531–532
 Population attributable risk (PAR), 304–305
 Populations, 3–4
 Positive autocorrelation, residuals, 464, 464*f*
 Positive likelihood ratio (LR+), 323
 Positive predictive value, 316
 Posteriori probability, 91
 Posthoc power analysis, 180–181
 Posttest probability, 325
 Potency, 575–576, 576*f*
 Power, 446–447, 447*f*, 505
 Powers of Y (POY) test, 534
 Precision, 41, 42*f*
 Prediction limits, 134–137
 Predictive accuracy, 316
 Predictive value, 316
 Pressure–diameter relations, 448–450, 449*f*
 Pretest probability, 324
 Prevalence index (PI), 246
 Principle of least squares, 63, 63*np*
 Principles of Medical Statistics, 648–649
 Probability, 14, 93*f*, 108, 193
 Bayes' theorem, 98–100
 conditional, 97–98
 definition, 92–96, 92*f*
 density distribution, 51
 principles, 92–96, 92*f*
 types, 91–92
 Probits, 578, 578*f*
 Process, 566
 Product-limit method, 601
 Programs, 31–33
 Propagation of errors, 85–87
 Propensity scoring, 652–653
 Proportions, 219–220
 and binomial theorem, 263–264
 comparisons, 266–267
 confidence limits, 264–265
 pooling samples, 267
 sample and population, 265
 sample size, 265–266
 Prostate cancer, 315, 315*t*
 Prostate specific antigen (PSA), 335
 Prostatic acid phosphatase (PAP) test, 315
 Protein kinase inhibitor, 159
 PSD. *See* Pseudostandard deviation (PSD)

Pseudorandom numbers, 644
 Pseudoreplication, 11–13
 Pseudostandard deviation (PSD), 117
 Psychological effect. *See* Hawthorne effect
 Pulmonary vascular resistance (PVR), 491–493, 492*f*

Q

Quadratic fit, 462*f*
 Quality control, 566
 Quantal dose–response curves, 575, 577–579
 Quantile–quantile (Q–Q) plot, 122*f*
 Quantiles, 64–68
 Quincunx, 110–111, 110*f*
 Quota sampling, 640

R

Radial plots, 625
 Ramsey's RESET test, 534, 536
 Random, 269–270
 Random errors, 110
 Randomization, 11, 642–645
 Randomized clinical trial (RCT), 179, 641, 646, 652
 Range, 71
 Rank von Neumann ratio, 564–565
 Rater agreement, 520–522
 Receiver operating characteristic (ROC)
 curves, 330–335, 330*f*
 advantages, 333
 C-reactive protein, 332, 332*f*
 cutting points, 330–331, 330*f*
 for hypothyroidism, 331, 331*f*
 pediatric appendicitis score, 334, 334*f*
 procalcitonin, 332, 332*f*
 white blood cell count, 332, 332*f*
 Reexpression, 447–448, 448*f*
 Reference range, 115
 Regression, 582
 centered *vs.* uncentered, 553*t*
 longitudinal, 558–559
 to mean, 518
 Relative risk ratio (RR), 295
 below 1, 305–307
 cardiovascular death incidence, 296–297, 296*t*
 sample size and power, 302–303, 302*f*
 Repeated measures regression, 439–440, 556–557, 557*f*
 Replication, 11–13

Resampling, 631
 bootstrap, 631–633, 633*f*
 jackknife, 635
 Monte Carlo methods, 635–636
 permutations, 634, 635*f*
 Residual analysis, 531
 linear regression, 460–465, 460–464*f*
 Resistant lines, 497–499
 Respiratory Medicine, 35
 Reynolds risk scores, 649
 Right censored data, 33–34, 601, 601*t*
 Robust tests, 152
 Rounding off, 40–41
 R program, 32
 Ryan–Einot–Gabriel–Welsch (R–E–G–W) test, 413

S

Samples, 3
 Sample size, 176*f*, 177, 588–589
 analysis of variance, 404–405
 for paired *t*-test, 350
 for unpaired *t*-test, 355
 Sampling, 639–641, 639–641*b*
 Scattergram, 45, 46*f*, 446, 446*f*, 475
 Scheffé test, 409
 Scholastic aptitude test (SAT), 212
 Screening tests, 318–321, 335–336
 SegReg, 497
 Sensitivity, 314–336, 315*t*
 nomograms for, 325–327, 326*f*
 of procalcitonin, 327, 327*t*
 Sequential analysis, 385–387, 385*f*
 paired measurements, 386, 387*f*
 trimeprazine *vs.* amylobarbitone, 386
 Serial correlation
 monotonic curves, 573
 peaked response, 572, 572*f*
 rank von Neumann ratio, 564–565
 ratio measurements, 565–566
 up-and-down procedure, 563–564
 Wald–Wolfowitz runs test, 562–563
 Serial measurements, 424, 572–573
 Shapiro and Wilk's test, 117
 Shrinkage, 530–531
 SIDS. *See* Sudden infant death syndrome (SIDS)
 Sigmoid dose–response curve, 575, 576*f*, 577
 Significance test, 160
 Sign test, 357–359, 358*t*
 Simple multiplication (scale factor), 86

Simple random sampling, 11
 Simpson's paradox, 210–212, 211*t*
 Single sample mean, 175
 SISA interactive statistical
 analysis, 32
 Skewing, 117
 Skewness, 116, 123
 Slippage, 150–152, 151*f*
 Social Science Statistics, 32
 Spearman's rank correlation, 511–512, 522
 Specificity, 314–336, 315*t*
 nomograms for, 325–327, 326*f*
 of procalcitonin, 327, 327*t*
 Spectrum bias, 321–323, 323*t*
 Spline, 551
 Spurious correlation, 516–518
 Square of correlation coefficient, 509
 Square root transformation, 277–278
 S-shaped curve. *See* Sigmoid dose–response
 curve
 Standard chart, 107
 Standard deviation (SD), 71–74, 72*t*, 109, 131, 134*f*,
 137–138, 139*f*, 177
 Standard error (SE), 139*f*, 263, 266
 Standard normal deviates, 119
 Statist, 13
 Statistical analysis, 19
 Statistical inference, 7–8, 159
 central limit theorem, 127–131
 generalizations based on single
 sample, 131–134
 graphic representation, 138–140
 prediction limits, 134–137
 reporting results, 137–138
 setting confidence limits, 131–134
 tolerance limits, 134–137
 Statistical power, 173–175
 Statistical practice, 21
 Statistical significance testing, 163–164, 169
 Statisticophobia, 25
 Statistics
 accuracy of measurement, 39–41
 precision, 41
 rounding off, 40–41
 truncation, 40–41
 current-tests, 20
 data, 9
 data sets, 38–39
 errors in biomedical publications, 21–22*t*
 ethical consequences, 23

- guidelines, 24–25
 - history, 13–16
 - measurement scales, 34–38
 - misuse of, 20–24
 - models, 9
 - notation, 42–43
 - operators, 43–45
 - parameters, 5–6
 - populations, 3
 - procedures, list of, 20*t*
 - programs, 31–33
 - pseudoreplication, 11–13
 - replication, 11–13
 - samples, 3
 - statistical inference, 7–8
 - study design, 9–13
 - unethical behavior, 23
 - uses, 6–8, 19–20
 - variability effects, 4–5
 - variables, 5–6, 33–34
 - weights, 45–46
 - Statistics Open For All (SOFA), 32
 - StatPlus:mac LE 2009, 32
 - StatsToDo, 32
 - Stellate ganglion nerve activity
 - analysis of variance, 398, 398*t*, 398*f*, 410, 410*t*
 - pair-wise differences, 410, 410*f*
 - Stem and leaf diagrams, 56–62, 58*f*, 58*t*, 117
 - Stevens' classification, 37
 - Stratified sampling, 645
 - Stroke, second-hand smoke exposure, 297, 297*t*
 - odd ratio, 297–302, 298*t*
 - relative risk ratio, 297–302, 298*t*
 - Studentized range test, 406–413, 407*t*
 - multiple step test, 409–413
 - recommendations, 413
 - single step test, 408–409
 - Student–Newman–Keuls (SNK)
 - method, 409, 412*t*, 426
 - vs.* Ryan–Einot–Gabriel–Welsch test, 413, 413*t*
 - Subjective probability, 91–92
 - Subsets, 94*f*
 - Sudden infant death syndrome (SIDS), 312–313
 - Sum of squares (SS), 391–392
 - Surgical mortality, 229, 229*t*
 - Survey method, 7–8
 - Survival analysis, 599
 - comparative study, 606–607
 - competing risks, 615–617
 - confidence limits, 605–606
 - Cox regression model, 612–615
 - hazard function, 609–610
 - hazard ratio, 610–612
 - partial data set, 603, 603*t*
 - sample size, 607
 - Symmetrical funnel plot, 624–625
- ## T
- Tables, 76–80
 - Target population, 639–642
 - Tay-Sachs disease, binomial
 - distribution, 254, 254*t*
 - Thallium uptake test, 322*t*
 - Threshold, 576*f*, 577
 - Thyroid hormone, 328
 - Thyroxine (T4), 328
 - Tolerance, 134–137, 539
 - Transformations, 55–56
 - TrendAnalyzer, 80
 - Trim and fill technique, 625
 - Trimean, 82
 - Trimming, 82–84, 83*t*
 - Truncation, 40–41
 - t*-test, 19–20, 37, 160, 160*np*, 341
 - cross-over trials, 367–371, 368–371*t*
 - equivalence tests, 371–374, 373*f*
 - noninferiority tests, 371–374
 - Tukey's test, 149, 408, 409*t*, 424
 - Two one-sided test (TOST), 372
 - Two-way ANOVA, 420–423, 422*t*, 422*f*
 - additivity, 423–425, 423*t*
 - for all anxiety and treatment data, 430*f*, 431–432*t*
 - Cochrane's Q-test, 426–427, 426–427*t*
 - Friedman test, 425–426
 - for high *vs.* low anxiety, 430*f*
 - with interactions, 427–432, 428*f*
 - missing data, 433–434
 - model II, 438–439
 - multiple factors, 425
 - nested designs, 434–437, 434*t*
 - with no interaction, 427, 428*f*
 - repeated measures designs, 439–440
 - with replication, 432
 - transformations, 438
 - treated *vs.* untreated, 430*f*
 - unequal cell sizes, 433–434
 - Two-way ANOVA, 521, 522*t*

Type I error, 169–170, 173, 173*f*, 178, 180*f*, 346, 380–381, 381*t*
 Type II error, 173–174*f*, 174, 176, 176*f*, 180*f*, 353–354, 381, 381*t*

U

Uncorrelated *XY* data set, 508, 508*t*
 Underspecified model, 528
 Unequal variances, 353–355
 Ungrouped data, 122–124, 122*t*
 Uniform distribution, 143
 Unit mass, 269
 Unit space, 269
 Unit time, 269
 Unpaired patent ductus arteriosus study, 234–235, 235*t*
 Unpaired *t*-test, 350–353
 data distribution, 352, 352*f*
 homogeneity of variances, 399
 requirements for, 352
 sample size for, 355
 Up-and-down runs test, 563–564
 Upper control limit (UCL), 567

V

Variability, 4–5, 69, 70*f*
 Variables, 5–6, 33–34, 42
 Variance, 71–72, 87
 homogeneity, 399–401, 400*t*
 inhomogeneity, 400–401
 Variance inflation factor (VIF), 539–540
 Variance ratio (*F*) distributions, 146–147, 147*f*
 Vassar College VassarStats program, 31
 Venn diagrams, 94–95, 95*f*, 212, 212*f*
 Verification bias, 323
 Visual Analog Scale (VAS), 35, 35*f*

Vitamin B₁₂ regression, 465, 466*f*
 V mask, 570–571, 571*f*

W

Wald method, 264
 Wald–Wolfowitz runs test, 562–563
 Walter method, 308
 Web-Enabled Scientific Services & Applications, 31
 Weighted kappa, 243–246, 245*t*
 Weighted mean, 45
 Weights, 45–46, 46*f*
 Welch–Satterthwaite test, 353–354
 Wheat varietal test, 420
 additivity, 423–425, 423–424*t*
 with different locations, 420*t*
 one-way ANOVA, 420, 421*t*, 421*f*
 partitioning within group SS_W, 421, 421*t*
 two-way ANOVA, 422, 422*t*, 422*f*
 White’s test, 536
 Wilcoxon signed rank test, 356–357, 356*t*, 607
 Will Rogers phenomenon, 643
 Wilson’s method, 260, 308
 Winsorization, 37, 82–84
 Winsorized standard deviation, 133
 Within-subjects design, 556–557

X

XY plots, 78

Y

Yardstick, 532
 Yates’ correction, 199, 205–206, 234–235

Z

Zero-inflated negative binomial regression, 594–595
 Z (zeta) transformation, 109, 109*f*, 503

Basic Biostatistics *for Medical and Biomedical Practitioners*

Second Edition

Julien I. E. Hoffman, MD, FRCP

Professor of Pediatrics, Emeritus, Senior Member of the Cardiovascular Research Institute, University of California, San Francisco, CA

Basic Biostatistics for Medical and Biomedical Practitioners, second edition, makes it easier to plan experiments, with emphasis on sample size. It also shows what choices are available when simple tests are unsuitable, and offers investigators an overview of complex tests that they will not do on their own, but need to know how they work.

The second edition presents a new revised and enhanced version of the chapters, taking into consideration new developments and tools available recently, and discusses topics such as basic aspects of statistics, continuous distributions, hypothesis testing, discrete distributions, probability in epidemiology and medical diagnosis, comparing means, regression, and correlation.

Basic Biostatistics for Medical and Biomedical Practitioners, second edition is a valuable source for students and researchers looking to expand or refresh the understanding of statistics as it applies to the biomedical and research fields.

Key Features

- Introduces procedures such as multiple regression, Poisson distribution, binomial and multinomial distributions, variance analysis, as well as designing and sampling clinical trials
- Presents a new section on ANCOVA
- Gives references to free online tests
- Each statistical inferential test is followed immediately by a problem, to give readers the opportunity to perform the test and interpret it
- Includes over 200 diagrams, which enables the reader to visualize the results
- Discusses NHST testing in detail, its disadvantages and how to think about probability; the unsafety of accepting $P < .05$ is stressed

About the Author

Dr. Julien I. E. Hoffman has over 40 years of experience teaching biostatistics to fellows in various medical fields. For 15 years, Dr. Hoffman was a member of the Biostatistics group for approving and coordinating statistics at UCSF as well as a consultant for the journal *Circulation Research*, and was intermittently statistical consultant to several other medical journals.

BIOMEDICAL SCIENCE AND MEDICINE



ACADEMIC PRESS

An imprint of Elsevier
elsevier.com/books-and-journals

