Summary of Some Statistical Methods to Test Hypotheses

Scale of measurement	Type of experiment: Two treatment groups consisting of different individuals	Three or more treatment groups consisting of different individuals	Before and after a single treatment in the same individuals	Multiple treatments in the same individuals	Association between two variables
Interval (and drawn from normally distributed populations*)	Unpaired *t* test (Chapter 4)	Analysis of variance (Chapter 3)	Paired *t* test (Chapter 9)	Repeated-measures analysis of variance (Chapter 9)	Linear regression, Pearson product-moment correlation, or Bland-Altman analysis (Chapter 8)
Nominal	Chi-square analysis-of-contingency table (Chapter 5)	Chi-square analysis-of-contingency table (Chapter 5)	McNemar's test (Chapter 9)	Cochrane Q†	Relative rank or odds ratio (Chapter 5)
Ordinal‡	Mann-Whitney rank-sum test (Chapter 10)	Kruskal-Wallis statistic (Chapter 10)	Wilcoxon signed-rank test (Chapter 10)	Friedman statistic (Chapter 10)	Spearman rank correlation (Chapter 8)
Survival time	Log-rank test or Gehan's test (Chapter 11)				

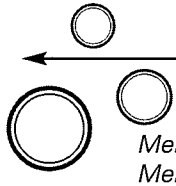
*If the assumption of normally distributed populations is not met, rank the observations and use the methods for data measured on an ordinal scale.

†Not covered in this text.

‡Or interval data that are not necessarily normally distributed.

primer of biostatistics

FIFTH EDITION



Stanton a.
Glantz, PhD
Professor of Medicine
Member, Cardiovascular Research Institute
Member, Institute for Health Policy Studies
Member, Cancer Center
University of California, San Francisco

McGRAW-HILL
Medical Publishing Division

New York Chicago San Francisco Lisbon London Madrid Mexico City
Milan New Delhi San Juan Seoul Singapore Sydney Toronto

McGraw-Hill

A Division of The McGraw-Hill Companies

PRIMER OF BIOSTATISTICS, FIFTH EDITION

1 2 3 4 5 6 7 8 9 0 DOC DOC 0 9 8 7 6 5 4 3 2 1

ISBN 0-07-137946-0

This book was set in Times Roman by Progressive Information Technologies.
The editors were Shelley Reinhardt and Karen Davis.
The production supervisor was Lisa T. Mendez.
R.R. Donnelley & Sons/Crawfordsville was the printer and binder.

This book is printed on acid-free paper.

Library of Congress Cataloging-in-Publication Data

Glantz, Stanton A.
Primer of biostatistics / Stanton A. Glantz. —5th ed.
p. ; cm.
Includes bibliographical references and index.
ISBN 0-07-137946-0
1. Medical statistics. 2. Biometry. I. Title.
[DNLM: 1. Biometry. WA 950 G545p 2001]
RA409.G55 2001
610'.7'27—dc21

2001034232

Hunches and intuitive impressions are essential for getting the work started, but it is only through the quality of the numbers at the end that the truth can be told.*

Lewis Thomas
Memorial Sloan-Kettering Cancer Center

*L. Thomas, "Biostatistics in Medicine," *Science* **198**:675, 1977.

Contents

Location of Tables for Tests of Significance

Preface

I have always thought of myself as something of an outsider and troublemaker, so it is with some humility that I prepare the fifth edition of this book, 20 years after the first edition originally appeared. Then, as now, the book had an unusual perspective: that many papers in the medical literature contained avoidable errors. At the time, the publisher, McGraw-Hill, expressed concern that this "confrontational approach" would put off reader and hurt sales. They also worried that the book was not organized like a traditional statistics text.

Time has shown that the biomedical community was ready for such an approach and the book has achieved remarkable success. Over time, it has evolved to include more topics, including power and sample size, more on multiple comparison procedures, and survival analysis. This fifth edition adds more on multiple comparison testing—the Holm test, which seems superior to the older Bonferroni, SNK, and Tukey tests—and a discussion of relative risks and odds ratios. It also replaces many of the examples that date from the first edition of this book with more current examples. In doing so, however, I have maintained the book's original tone and attitude.

This book has its origins in 1973, when I was a postdoctoral fellow. Many friends and colleagues came to me for advice and explanations about biostatistics. Since most of them had less knowledge of statistics than I did, I tried to learn what I needed to help them. The need to develop quick and intuitive, yet

correct, explanations of the various tests and procedures slowly evolved into a set of stock explanations and a two-hour slide show on common statistical errors in the biomedical literature and how to cope with them. The success of this slide show led many people to suggest that I expand it into an introductory book on biostatistics, which led to the first edition of *Primer of Biostatistics* in 1981.

As a result, this book is oriented as much to the individual reader—be he or she student, postdoctoral research fellow, professor, or practitioner—as to the student attending formal lectures.

This book can be used as a text at many levels. It has been the required text for the biostatistics portion of the epidemiology and biostatistics course required of all medical students at the University of California, San Francisco. This course covered the material in the first eight chapters in eight 1-hour lectures. The material is discussed along with other epidemiology in an additional problem session each week. The book is also used for a more abbreviated set of lectures on biostatistics (covering the first three chapters) given to our dental students. In addition, it has served me (and others) well in a one quarter four-unit course in which we cover the entire book in depth. This course meets for 3 lecture hours and has a 1 hour problem session. It is attended by a wide variety of students, from undergraduates through graduate students and postdoctoral fellows, as well as an occasional faculty member.

Since this book includes the technical material covered in any introductory statistics course, it is suitable as either the primary or supplementary text for a general undergraduate introductory statistics course (which is essentially the level at which this material is taught in medical schools), especially for a teacher seeking a way to make statistics relevant to students majoring in the life sciences.

This book differs from other introductory texts on biostatistics in several ways, and it is these differences which seem to account for the book's popularity.

First, it is based on the premise that much of what is published in the biomedical literature uses dubious statistical practices, so that a reader who takes what he reads at face value may often be absorbing erroneous information. Most of the errors (at least as they relate to statistical inference) center on misuse of the *t* test, probably because the people doing the research were unfamiliar with anything else. The *t* test is usually the first procedure presented in a statistics book that will yield the highly prized *P* value. Analysis of variance, if presented at all, is deferred to the end of the book to be ignored or rushed through at the end of the term. Since so much is published that should probably be analyzed with analysis of variance, and since analysis of variance is really the paradigm of all parametric statistical tests, I present it first, then discuss the *t* test as a special case.

Second, in keeping with the problems I see in the literature, there is a discussion of multiple comparison testing.

Third, the book is organized around hypothesis testing and estimation of the size of treatment effects, as opposed to the more traditional (and logical from a theory of statistics perspective) organization that goes from one-sample to two-sample to general k-sample estimation and hypotheses testing procedures. I believe my approach goes directly to the kinds of problems one most commonly encounters when reading about or doing biomedical research.

The examples are mostly based on interesting studies from the literature and are reasonably true to the original data. I have, however, taken some liberty in recreating the raw data to simplify the statistical problems (for example, making the sample sizes equal) so that I could focus on the important intuitive ideas behind the statistical procedures rather than getting involved in the algebra and arithmetic. When the text only discusses the case of equal sample sizes, the formulas for the more general unequal sample size case are included in an appendix.

It is worth mentioning a few items I have not added. Some people suggested that I add an explicit discussion of probability calculus and expected values, rather than the implicit discussion of them in the existing text. Others suggested that I make the distinction between P and α more precise. (I purposefully blurred this distinction.) I also was tempted to use the platform this book has created within the research community to popularize multivariate statistical methods—in particular multiple regression—within the biomedical community. These methods have been applied widely with good results in the social sciences and I have found them very useful in my work on cardiac function and tobacco control. I decided against making these changes, however, because they would have fundamentally changed the scope and tone of the book, which are the keys to its success.*

As with all books, there are many people who deserve thanks. Julien Hoffman gave me the first really clear and practically oriented course in biostatistics that allowed me to stay one step ahead of the people who came to me for expert help. His continuing interest and discussion of statistical issues has helped me learn enough to even think of writing this book. Philip Wilkinson and Marion Nestle suggested some of the best examples and also offered very useful criticisms of the manuscript. Mary Giammona offered many useful criticisms from a student's point of view and helped develop the original problem sets. Bryan Slinker, Jim Lightwood, and Kristina Thayer helped develop new problems included in the later editions. Virginia Ernster and Susan Sacks not only offered many helpful suggestions, but also unleashed their 300 first- and

*These suggestions did, however, lead to a new book on the subject of multiple regression and analysis of variance, written with the same approach in *Primer of Biostatistics*. It is: *Primer of Applied Regression and Analysis of Variance* (2 ed.), S. A. Glantz and B. K. Slinker, New York: McGraw-Hill, 2000.

second-year medical students on the manuscript for the first edition when they graciously offered to use it as the required text for their course. Bryan Slinker, Ken Resser, B. S. Appleyard, Robin Booth, Kristina Thayer, and others, who served as teaching assistants to me in the course I taught from this book, offered many insightful criticisms and concrete suggestions on how to sharpen the explanations, examples, and problems in the book.

Mary Hurtado typed the manuscript for the original version of this book, with amazing speed and accuracy. Thomas Summer, Sonja Bock, and Mike Matrigali helped with the original text-editing with UNIX. Dale Johnson prepared the original illustrations.

Since the first edition of this book in 1981, many things have changed. There is a much wider appreciation for the need to use appropriate statistical methods in biomedical research than there was in 1981. While the problem persists, many journals have recognized the problems caused by ignorance of statistical issues among many scientists and have explicitly included biostatistical considerations in the review process for manuscripts. Indeed, in a classic example of the fact that the individual who complains the loudest gets put in charge, I now serve as an Associate Editor of the *Journal of the American College of Cardiology* with special responsibility for reviewing tentatively accepted manuscripts for statistical problems prior to publication. About half the papers still have some sort of problem (of varying severity), but we catch them *before* publication now.

Finally, I thank the many others who have used the book, both as students and as teachers of biostatistics, who took the time to write me questions, comments, and suggestions on how to improve it. I have done my best to heed their advice in preparing this fifth edition.

Many of the pictures in this book are direct descendants of my original slides. In fact, as you read this book, you would do best to think of it as a slide show that has been set to print. Most people who attend my slide show leave more critical of what they read in the biomedical literature. After I gave it to the MD-PhD candidates at the University of California, San Francisco, I heard that the candidates gave every subsequent speaker a hard time about misuse of the standard error of the mean as a summary statistic and abuse of t tests. This book has had a similar effect on many others. Nothing could be more flattering or satisfying to me. I hope that this book will continue to make more people more critical and help improve the quality of the biomedical literature and, ultimately, the care of people.

Stanton A. Glantz

Chapter 1

Biostatistics and Clinical Practice

In an ideal world, editors of medical journals would do such an excellent job of ensuring the quality and accuracy of the statistical methods of the papers they publish that readers with no personal interest in this aspect of the research work could simply take it for granted that anything published was correct. If past history is any guide, however, we will probably never reach that ideal. In the meantime, consumers of the medical literature—practicing physicians and nurses, biomedical researchers, and health planners—must be able to assess statistical methods on their own in order to judge the strength of the arguments for or against the specific diagnostic test or therapy under study. As discussed below, these skills will become more important as financial constraints on medical practice grow.

THE CHANGING MEDICAL ENVIRONMENT

The practice of medicine, like the delivery of all medical services, is entering a new era. Until the second quarter of the last century,

medical treatment had little positive effect on when, or even whether, sick people recovered. With the discovery of ways to reverse the biochemical deficiencies that caused some diseases and the development of antibacterial drugs, it became possible to cure sick people. These early successes and the therapeutic optimism they engendered stimulated the medical research community to develop a host of more powerful agents to treat heart disease, cancer, neurologic disorders, and other ailments. These successes led society to continue increasing the amount of resources devoted to the delivery of medical services. In 1997, the United States spent $1.1 trillion (13.5 percent of the gross domestic product) on medical services. In addition, both the absolute amount of money and the fraction of the gross domestic product devoted to the medical sector have grown rapidly (Fig. 1-1). Today, many government and business leaders view this continuing explosion with concern. Containing medical care costs has become a major fo-

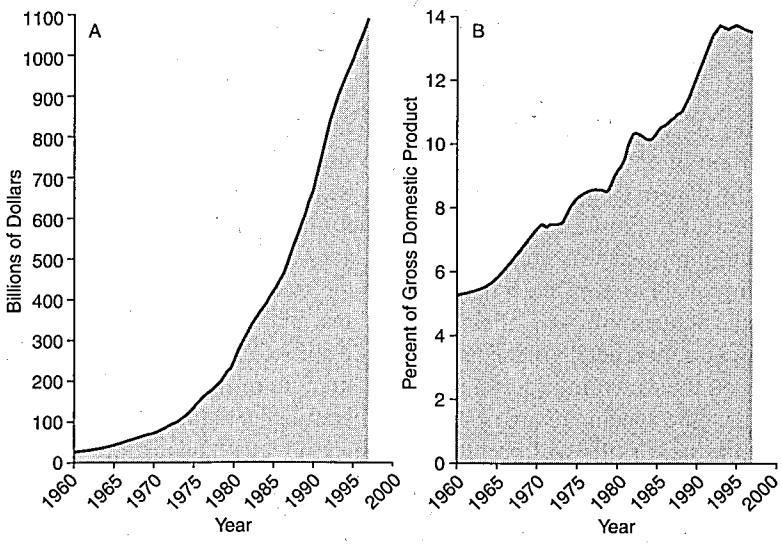


Figure 1-1 (*A*) Total annual expenditures for medical services in the United States between 1960 and 1997. (*B*) Expenditures for medical services as a percentage of the gross domestic product. (Source: *Statistical Abstract of the United States, 1999.* Washington DC: U.S. Department of Commerce, pp. 118, 459.)

cus of activity for both state and federal governments and the private sector.

During the period of rapid growth that is probably ending, there were ample resources to enable physicians and other health care providers to try tests, procedures, and therapies with little or no restriction on their use. As a result, much of what is considered good medical practice developed without firm evidence demonstrating that these practices actually help the patient. Even for effective therapies, there has been relatively little systematic evaluation of precisely which patients these therapies help.* In addition to wasting money, these practices regularly expose people to powerful drugs, surgery, or other interventions with potentially dangerous side effects in cases where such treatment does not do the patient any good.

What does this have to do with biostatistics?

As the resources available to provide medical care grow more slowly, health professionals will have to identify more clearly which tests, procedures, and therapies are of demonstrated value. In addition to assessing whether or not one intervention or another made a difference, it will become important to assess how great the difference was. Such knowledge will play a growing role in decisions on how to allocate medical resources among potential health care providers and their patients. These issues are, at their heart, statistical issues. Because of factors such as the natural biological variability between individual patients and the placebo effect,† one usually cannot conclude that some therapy was beneficial on the basis of simple experience. For example, about one-third of people given placebos in place of pain killers experience relief. Biostatistics provides the tools for turning clinical and laboratory experience into quantitative statements about whether and by how much a treatment or procedure affected a group of patients.

In addition to studies of procedures and therapies, researchers are beginning to study how physicians, nurses, and other health care

*A. L. Cochrane, *Effectiveness and Efficiency: Random Reflections on Health Services,* Nuffield Provincial Hospitals Trust, London, 1972.

†The placebo effect is a response attributable to therapy per se as opposed to the therapy's specific properties. Examples of placebos are an injection of saline, sugar pill, and surgically opening and closing without performing any specific surgical procedure.

professionals go about their work. For example, one study* demonstrated that patients with uncomplicated pyelonephritis, a common kidney infection, who were treated in accordance with the guidelines in the *Physicians' Desk Reference* remained in the hospital an average of 2 days less than those who were not treated appropriately. Since hospitalization costs constitute a sizable element of total medical care expenditures, it would seem desirable to minimize the length of stay when it does not affect the patient's recovery adversely. Traditionally, there have been few restrictions on how individual physicians prescribed drugs. This study suggests that steps making physicians follow recommended prescribing patterns more closely could significantly reduce hospitalization and save money without harming the patient. Such evidence could be used to support efforts to restrict the individual physician's freedom in using prescription drugs.

Hence, evidence collected and analyzed using biostatistical methods can potentially affect not only how physicians choose to practice medicine but what choices are open to them. Intelligent participation in these decisions requires an understanding of biostatistical methods and models that will permit one to assess the quality of the evidence and the analysis of that evidence used to support one position or another.

Clinicians have not, by and large, participated in debates on these quantitative questions, probably because the issues appear too technical and seem to have little impact on their day-to-day activities. As pressure for more effective use of medical resources grows, clinicians will have to be able to make more informed judgments about claims of medical efficacy so that they can participate more intelligently in the debate on how to allocate medical resources. These judgments will be based in large part on statistical reasoning.

WHAT DO STATISTICAL PROCEDURES TELL YOU?

Suppose researchers believe that administering some drug increases urine production in proportion to the dose and to study it they give different doses of the drug to five different people, plotting their urine production against the dose of drug. The resulting data, shown in

*D. E. Knapp, D. A. Knapp, M. K. Speedie, D. M. Yaeger, and C. L. Baker, "Relationship of Inappropriate Drug Prescribing to Increased Length of Hospital Stay," *Am. J. Hosp. Pharm.*, **36**:1334–1337, 1979. This study will be discussed in detail in Chaps. 3 to 5.

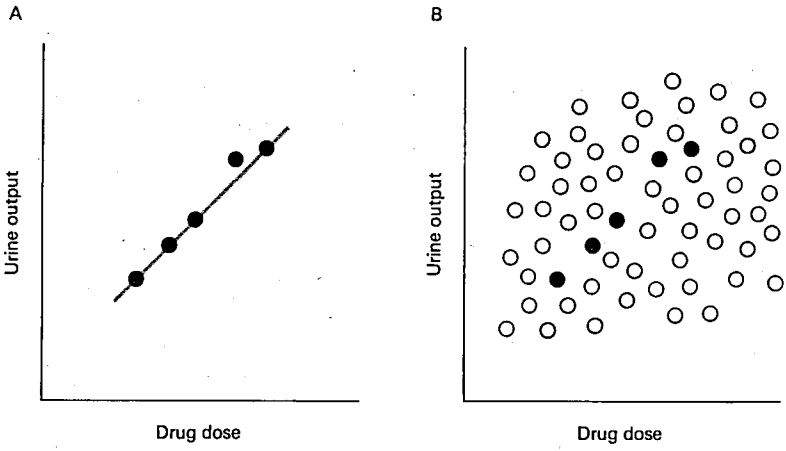


Figure 1-2 (*A*) Results of an experiment in which researchers administered five different doses of a drug to five different people and measured their daily urine production. Output increased as the dose of drug increased in these five people, suggesting that the drug is an effective diuretic in all people similar to those tested. (*B*) If the researchers had been able to administer the drug to all people and measure their daily urine output, it would have been clear that there is no relationship between the dose of drug and urine output. The five specific individuals who happened to be selected for the study in panel *A* are shown as shaded points. It is possible, but not likely, to obtain such an unrepresentative sample that leads one to believe that there is a relationship between the two variables when there is none. A set of statistical procedures called tests of hypotheses permits one to estimate the chance of getting such an unrepresentative sample.

Fig. 1-2*A*, reveal a strong relationship between the drug dose and daily urine production in the five people who were studied. This result would probably lead the investigators to publish a paper stating that the drug was an effective diuretic.

The only statement that can be made with absolute certainty is that as the drug dose increased, so did urine production *in the five people in the study.* The real question of interest, however, is: How is the drug likely to affect *all people who receive it?* The assertion that the drug is effective requires a leap of faith from the limited experience shown in Fig. 1-2*A* to all people. Of course, one cannot know in advance how all people will respond to the drug.

Now, pretend that we knew how every person who would ever receive the drug would respond. Figure 1-2*B* shows this information. There is no systematic relationship between the drug dose and urine production! The drug is not an effective diuretic.

How could we have been led so far astray? The shaded points in Fig. 1-2*B* represent the specific individuals who happened to be studied to obtain the results shown in Fig. 1-2*A*. While they are all members of the population of people we are interested in studying, the five specific individuals we happened to study, taken as a group, were not really representative of how the entire population of people responds to the drug.

Looking at Fig. 1-2*B* should convince you that obtaining such an unrepresentative sample of people, though possible, is not very likely. One set of statistical procedures, called *tests of hypotheses,* permits you to estimate the likelihood of concluding that two things are related as Fig. 1-2*A* suggests when the relationship is really due to bad luck in selecting people for study and not a true effect of the drug investigated. In this example, we will be able to estimate that such a sample of people should turn up in a study of the drug only about 5 times in 1000 when the drug actually has no effect.

Of course it is important to realize that although biostatistics is a branch of mathematics, there can be honest differences of opinion about the best way to analyze a problem. This fact arises because all statistical methods are based on relatively simple mathematical models of reality, so the results of the statistical tests are accurate only to the extent that the reality and the mathematical model underlying the statistical test are in reasonable agreement.

WHY NOT DEPEND ON THE JOURNALS?

Aside from direct personal experience, most health care professionals rely on medical journals to keep them informed about the current concepts on how to diagnose and treat their patients. Since few members of the clinical or biomedical research community are conversant in the use and interpretation of biostatistics, most readers assume that when an article appears in a journal, the reviewers and editors have scrutinized every aspect of the manuscript, including the use of statistics. Unfortunately, this is often not so.

Throughout the 1950s through the 1970s, several critical reviews* of the use of statistics in the general medical literature consistently found that about half the articles used incorrect statistical methods. This situation led many of the larger journals to incorporate formal statistical reviews (by a statistician) into the peer review process. More recent reviews of the use of statistical methods in the medical literature have concentrated on the efficacy of providing these secondary statistical reviews of tentatively accepted papers. These reviews have revealed that about half (or more) of the papers tentatively accepted for publication have statistical problems. For the most part, these errors are resolved before publication, together with substantive issues raised by the other (content) reviewers, and the rate of statistical problems in the final published papers is much lower.

The fact remains, however, that most journals still do not provide a complete secondary statistical review of all papers, so the fraction of published papers containing statistical errors is probably still about 50% for many journals. Indeed, reviews of specialty journals continue to show a high frequency of statistical problems in published papers.†

*O. B. Ross, Jr. ("Use of Controls in Medical Research." *JAMA,* **145:**72–75, 1951) evaluated 100 papers published in *Journal of the American Medical Association, American Journal of Medicine, Annals of Internal Medicine, Archives of Neurology and Psychiatry,* and *American Journal of Medical Sciences* during 1950. R. F. Badgley ("An Assessment of Research Methods Reported in 103 Scientific Articles from Two Canadian Medical Journals," *Can M.A.J.,* **85:**256–260, 1961) evaluated 103 papers published in the *Canadian Medical Association Journal* and *Canadian Journal of Public Health* during 1960. S. Schor and I. Karten ("Statistical Evaluation of Medical Journal Manuscripts," *JAMA,* **195:**1123–1128, 1966) evaluated 295 papers published in *Annals of Internal Medicine, New England Journal of Medicine, Archives of Surgery, American Journal of Medicine, Journal of Clinical Investigation, American Archives of Neurology, Archives of Pathology,* and *Archives of Internal Medicine* during 1964. S. Gore, I. G. Jones, and E. C. Rytter ("Misuses of Statistical Methods: Critical Assessment of Articles in B.M.J. from January to March, 1976," *Br. Med. J.* **1**(6053):85–87, 1977) evaluated 77 papers published in the *British Medical Journal* in 1976.

†More recent reviews, while dealing with a more limited selection of journals, have shown that this problem still persists. See S. J. White, "Statistical Errors in Papers in the *British Journal of Psychiatry,*" *Br. J. Psychiatry,* **135:**336–342, 1979; M. J. Avram, C. A. Shanks, M. H. M. Dykes, A. K. Ronai, W. M. Stiers, "Statistical Methods in Anesthesia Articles: An Evaluation of Two American Journals during Two Six-Month Periods," *Anesth. Analg.* **64:**607–611, 1985; J. Davies, "A Critical Survey of Scientific Methods in Two Psychiatry Journals," *Aust. N.Z. J. Psych.,* **21:**367–373, 1987; D. F. Cruess, "Review of the Use of Statistics in *The American Journal of Tropical Medicine and Hygiene* for January–December 1988," *Am. J. Trop. Med. Hyg.,* **41:**619–626, 1990; C. A. Silagy, D. Jewell, D. Mant, "An Analysis of Randomized Controlled Trials Published in

continued

When confronted with this observation—or the confusion that arises when two seemingly comparable articles arrive at different conclusions—people often conclude that statistical analyses are maneuverable to one's needs, or are meaningless, or are too difficult to understand.

Unfortunately, except when a statistical procedure merely confirms an obvious effect (or the paper includes the raw data), a reader cannot tell whether the data in fact support the author's conclusions or not. Ironically, the errors rarely involve sophisticated issues that provoke debate between professional statisticians but are simple mistakes, such as neglecting to include a control group, not allocating treatments to subjects at random, or misusing elementary tests of hypotheses. These errors generally bias the study on behalf of the treatments.

The existence of errors in experimental design and misuse of elementary statistical techniques in a substantial fraction of published papers is especially important in clinical studies. These errors may lead investigators to report a treatment or diagnostic test to be of statistically demonstrated value when, in fact, the available data fail to support this conclusion. Physicians who believe that a treatment has been proved effective on the basis of publication in a reputable journal may use it for their patients. Because all medical procedures involve some risk, discomfort, or cost, people treated on the basis of erroneous research reports gain no benefit and may be harmed. On the other hand, errors could produce unnecessary delay in the use of helpful treatments. Scientific studies which document the effectiveness of medical procedures will become even more important as efforts grow to control medical costs without sacrificing quality. Such studies must be designed and interpreted correctly.

In addition to indirect costs, there are significant direct costs associated with these errors: money is spent, animals may be sacrificed, and human subjects may be put at risk to collect data that are not interpreted correctly.

(*continued*) the US Family Medicine Literature, 1987–1991," *J. Fam. Pract.* **39:** 236–242, 1994; M. H. Kanter and J. R. Taylor, "Accuracy of Statistical Methods in *Transfusion:* A Review of Articles from July/August 1992 through June 1993," *Transfusion* **34:**697–701, 1994; N. R. Powe, J. M. Tielsch, O. D. Schein, R. Luthra, E. P. Steinberg, "Rigor of Research Methods in Studies of the Effectiveness and Safety of Cataract Extraction with Intraocular Lens Implantation," *Arch. Ophthalmol.* **112:**228–238, 1994.

WHY HAS THE PROBLEM PERSISTED?

Because so many people are making these errors, there is little peer pressure on academic investigators to use statistical techniques carefully. In fact, one rarely hears a word of criticism. Quite the contrary, some investigators fear that their colleagues—and especially reviewers—will view a correct analysis as unnecessarily theoretical and complicated.

The journals are the major force for quality control in scientific work. Some journals have recognized that the regular reviewers often are not competent to review the use of elementary statistics in papers submitted for publication, and these journals have modified their refereeing process accordingly. Generally, they have someone familiar with statistical methods review manuscripts before they are accepted for publication. This secondary statistical review, generally conducted before a decision is made to invite revision of a paper, often leads to substantial revision in the tests used for statistical analysis, the presentation of the results, or the conclusions that are drawn.*

Most editors, however, apparently still assume that the reviewers will examine the statistical methodology in a paper with the same level of care that they examine the clinical protocol or experimental preparation. If this assumption were correct, one would expect all papers to describe, in detail as explicit as the description of the protocol or preparation, how the authors have analyzed their data. Yet, often the statistical procedures used to test hypotheses in medical journals are not even identified. It is hard to believe that the reviewers examined the methods of data analysis with the same diligence with which they evaluated the experiment used to collect the data.

In short, to read the medical literature intelligently, you will have to be able to understand and evaluate the use of the statistical methods used to analyze the experimental results as well as the laboratory methods used to collect the data. Fortunately, the basic ideas needed to be an intelligent reader—and, indeed, to be an intelligent investigator—are quite simple. The next chapter begins our discussion of these ideas and methods.

*For a discussion of the experiences of two journals, see M. J. Gardner and J. Bond, "An Exploratory Study of Statistical Assessment of Papers Published in the *British Medical Journal*," *JAMA* **263:**1355–1357, 1990 and S. A. Glantz, "It Is All in the Numbers," *J. Am. Coll. Cardiol.* **21:**835–837, 1993.

Chapter 2

How to Summarize Data

An investigator collecting data generally has two goals: to obtain descriptive information about the population from which the sample was drawn and to test hypotheses about that population. We focus here on the first goal: to summarize data collected on a single variable in a way that best describes the larger, unobserved population.

When the value of the variable associated with any given individual is more likely to fall near the mean (average) value for all individuals in the population under study than far from it and equally likely to be above the mean and below it, the *mean* and *standard deviation* for the sample observations describe the location and amount of variability among members of the population. When the value of the variable is more likely than not to fall below (or above) the mean, one should report the *median* and values of at least two other percentiles.

To understand these rules, assume that we observe *all* members of the population, not only a limited (ideally representative) sample as in an experiment.

For example, suppose we wish to study the height of Martians and to avoid any guesswork, we visit Mars and measure the entire population—all 200 of them. Figure 2-1 shows the resulting data with each Martian's height rounded to the nearest centimeter and represented by a circle. There is a *distribution* of heights of the Martian population. Most Martians are between about 35 and 45 cm tall, and only a few (10 out of 200) are 30 cm or shorter or 50 cm or taller.

Having successfully completed this project and demonstrated the methodology, we submit a proposal to measure the height of Venusians. Our record of good work assures funding, and we proceed to make the measurements. Following the same conservative approach, we measure the heights of *all* 150 Venusians. Figure 2-2 shows the measured heights for the entire population of Venus, using the same presentation as Fig. 2-1. As on Mars, there is a distribution of heights among members of the population, and all Venusians are around 15 cm tall, almost all of them being taller than 10 cm and shorter than 20 cm.

Comparing Figs. 2-1 and 2-2 demonstrates that Venusians are shorter than Martians and that the variability of heights within the Venusian population is smaller. Whereas almost all (194 of 200) the Martians' heights fell in a range 20 cm wide (30 to 50 cm), the

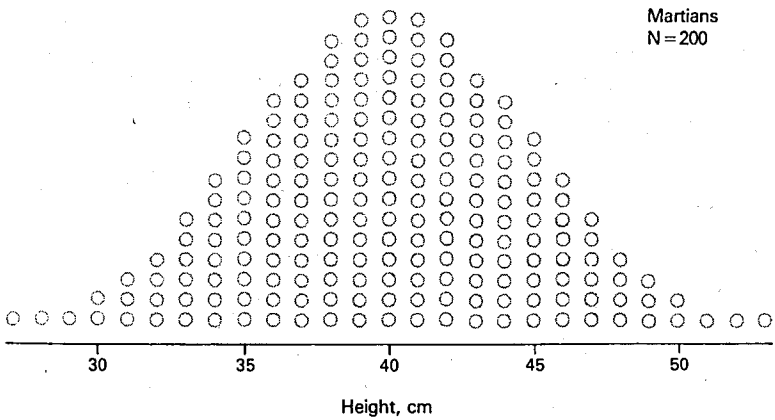


Figure 2-1 Distribution of heights of 200 Martians, with each Martian's height represented by a single circle. Notice that any individual Martian is more likely to have a height near the mean height of the population (40 cm) than far from it and is equally likely to be shorter or taller than average.

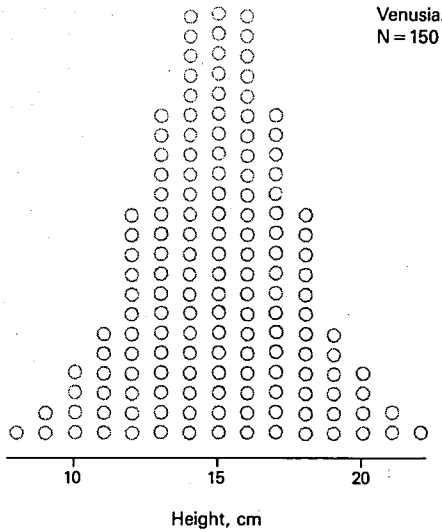


Figure 2-2 Distribution of heights of 150 Venusians. Notice that although the average height and dispersion of heights about the mean differ from those of Martians (Fig. 2-1), they both have a similar bell-shaped appearance.

analogous range for Venusians (144 of 150) is only 10 cm (10 to 20 cm). Despite these differences, there are important similarities between these two populations. In both, any given member is more likely to be near the middle of the population than far from it and equally likely to be shorter or taller than average. In fact, despite the differences in population size, average height, and variability, the *shapes* of the distributions of heights of the inhabitants of both planets are almost identical. A most striking result!

We can now reduce all this information to a few numbers, called *parameters* of the distribution. Since the shapes of the two distributions are similar, we need only describe how they differ; we do this by computing the *mean* height and the *variability* of heights about the mean.

THE MEAN

To indicate the location along the height scale, define the *population mean* to be the average height of all members of the population.

Population means are often denoted by μ, the Greek letter mu. When the population is made up of discrete members,

$$\text{Population mean} = \frac{\text{sum of values, e.g., heights, for each member of population}}{\text{number of population members}}$$

The equivalent mathematical statement is

$$\mu = \frac{\Sigma X}{N}$$

in which Σ, Greek capital sigma, indicates the sum of the value of the variable X for all N members of the population. Applying this definition to the data in Figs. 2-1 and 2-2 yields the result that the mean height of Martians is 40 cm and the mean height of Venusians is 15 cm. These numbers summarize the qualitative conclusion that the distribution of heights of Martians is higher than the distribution of heights of Venusians.

MEASURES OF VARIABILITY

Next, we need a measure of dispersion about the mean. A value an equal distance above or below the mean should contribute the same amount to our index of variability, even though in one case the deviation from the mean is positive and in the other it is negative. Squaring a number makes it positive, so let us describe the variability of a population about the mean by computing the *average squared deviation from the mean.* The average squared deviation from the mean is larger when there is more variability among members of the population (compare the Martians and Venusians). It is called the *population variance* and is denoted by σ^2, the square of the lower case Greek sigma. Its precise definition for populations made up of discrete individuals is

$$\text{Population variance} = \frac{\text{sum of (value associated with member of population} - \text{population mean)}^2}{\text{number of population members}}$$

The equivalent mathematical statement is

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

Note that the units of variance are the square of the units of the variable of interest. In particular, the variance of Martian heights is 25 cm^2 and the variance of Venusian heights is 6.3 cm^2. These numbers summarize the qualitative conclusion that there is more variability in heights of Martians than in heights of Venusians.

Since variances are often hard to visualize, it is more common to present the square root of the variance, which we might call the *square root of the average squared deviation from the mean.* Since that is quite a mouthful, this quantity has been named the *standard deviation* σ. Therefore, by definition,

Population standard deviation

$$= \sqrt{\text{population variance}}$$

$$= \sqrt{\frac{\text{sum of (value associated with member of population} - \text{population mean)}^2}{\text{number of population members}}}$$

or mathematically,

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$$

where the symbols are defined as before. Note that the standard deviation has the same units as the original observations. For example, the standard deviation of Martian heights is 5 cm, and the standard deviation of Venusian heights is 2.5 cm.

THE NORMAL DISTRIBUTION

Table 2–1 summarizes what we found out about Martians and Venusians. The three numbers in the table tell a great deal: the population size, the mean height, and how much the heights vary about the mean.

Table 2-1 Population Parameters for Heights of Martians and Venusians

	Size of population	Population mean, cm	Population standard deviation, cm
Martians	200	40	5.0
Venusians	150	15	2.5

The distributions of heights on both planets have a similar shape, so that *roughly 68 percent of the heights fall within 1 standard deviation from the mean and roughly 95 percent within 2 standard deviations from the mean.* This pattern occurs so often that mathematicians have studied it and found that if the observed measurement is the sum of many independent small random factors, the resulting measurements will take on values that are distributed like the heights we observed on both Mars and Venus. This distribution is called the *normal (or gaussian) distribution.*

Its height at any given value of X is

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2 \right]$$

Note that the distribution is completely defined by the population mean μ and population standard deviation σ. Therefore, the information given in Table 2–1 is not just a good abstract of the data, it is *all* the information one needs to describe the population fully *if the distribution of values follows a normal distribution.*

PERCENTILES

Armed with this theoretical breakthrough, we renew our grant by proposing not only to measure the heights of all Jupiter's inhabitants but also to compute the mean and standard deviation of the heights of all Jovians. The resulting data show the mean height to be 37.6 cm and the standard deviation of heights to be 4.5 cm. By comparison with Table 2–1, Jovians appear quite similar in height to Martians, since these two parameters completely specify a normal distribution.

The raw data, however, tell a different story. Figure 2-3*A* shows that, unlike those living on the other two planets, a given Jovian is not equally likely to have a height above average as below average; the distribution of heights of all population members is no longer symmetric but *skewed.* The few individuals who are much taller than the rest increase the mean and standard deviation in a way that led us to think that

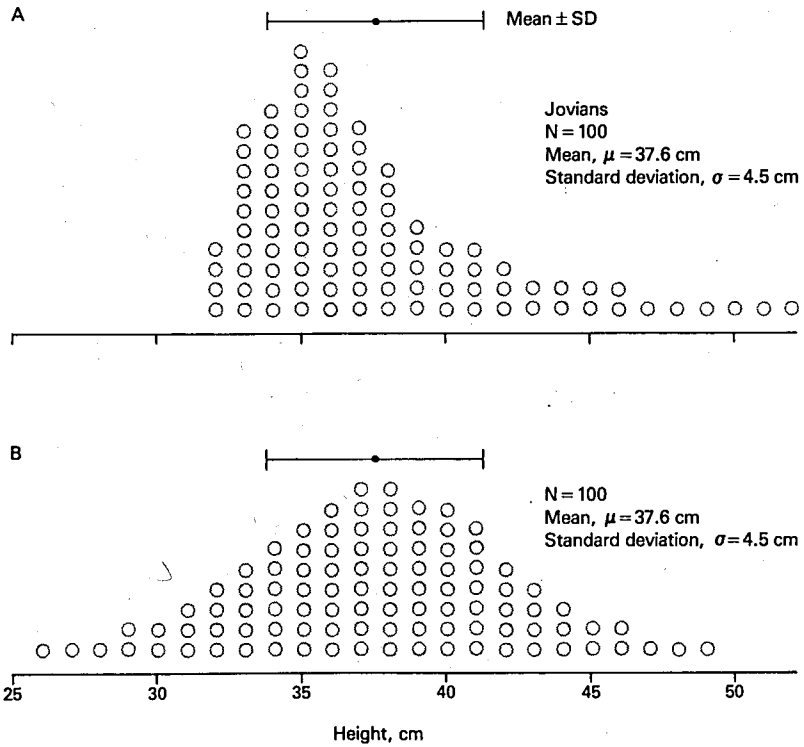


Figure 2-3 When the population values are not distributed symmetrically about the mean, reporting the mean and standard deviation can give the reader an inaccurate impression of the distribution of values in the population. Panel *A* shows the true distribution of the heights of the 100 Jovians (note that it is skeswed toward taller heights). Panel *B* shows a normally distributed population with 100 members and the same mean and standard deviation as in panel *A*. Despite this, the distribution of heights in the two populations is quite different.

most of the heights were higher than they actually are and that the variability of heights was greater than it actually is. Specifically, Fig. 2-3*B* shows a population of 100 individuals whose heights are distributed according to a normal or gaussian distribution with the same mean and standard deviation as the 100 Jovians in Fig. 2-3*A*. It is quite different. So, although we can compute the mean and standard deviation of heights of Jupiter's — or, for that matter, any — population, these two numbers do not summarize the distribution of heights nearly so well as they did when the heights in the population followed a normal distribution.

An alternative approach which better describes such data is to report the *median.* The median is the value that half the members of the population fall below. Figure 2-4*A* shows that half the Jovians are shorter than 36 cm; 36 cm is the median. Since 50 percent of the population values fall below the median, it is also called the *50th percentile.*

Calculation of the *median* and other *percentiles* is simple. First, list the n observations in order. The median, the value that defines the lower half of the observations, is simply the $(n + 1)/2$ observation. When there are an odd number of observations, the median falls on one of the observations. For example, if there are 27 observations, the $(27 + 1)/2 = 14$th observation (listed from smallest to largest) is the median. When there are an even number of observations, the median falls between two observations. For example, if there are 40 observations, the median would be the $(40 + 1)/2 = 20.5$th observation. Since there is no 20.5th observation, we take the average of 20th and 21st observation.

Other percentile points are defined analogously. For example the 25th percentile point, the point that defines the lowest quarter of the observations, is just the $(n + 1)/4$ observation. Again, if the value falls between two observations, take the mean of the two surrounding observations. In general, the pth percentile point is the $(n + 1)/(100/p)$ observation.

To give some indication of the dispersion of heights in the population, report the value which separates the lowest (shortest) 25 percent of the population from the rest and the value which separates the shortest 75 percent of the population from the rest. These two points are called the *25th* and *75th percentile* points, respectively. For the Jovians, Fig. 2-4*B* shows that these percentiles are 34 and 40 cm. While these three numbers (the 25, 50, and 75 percentile points, 34,

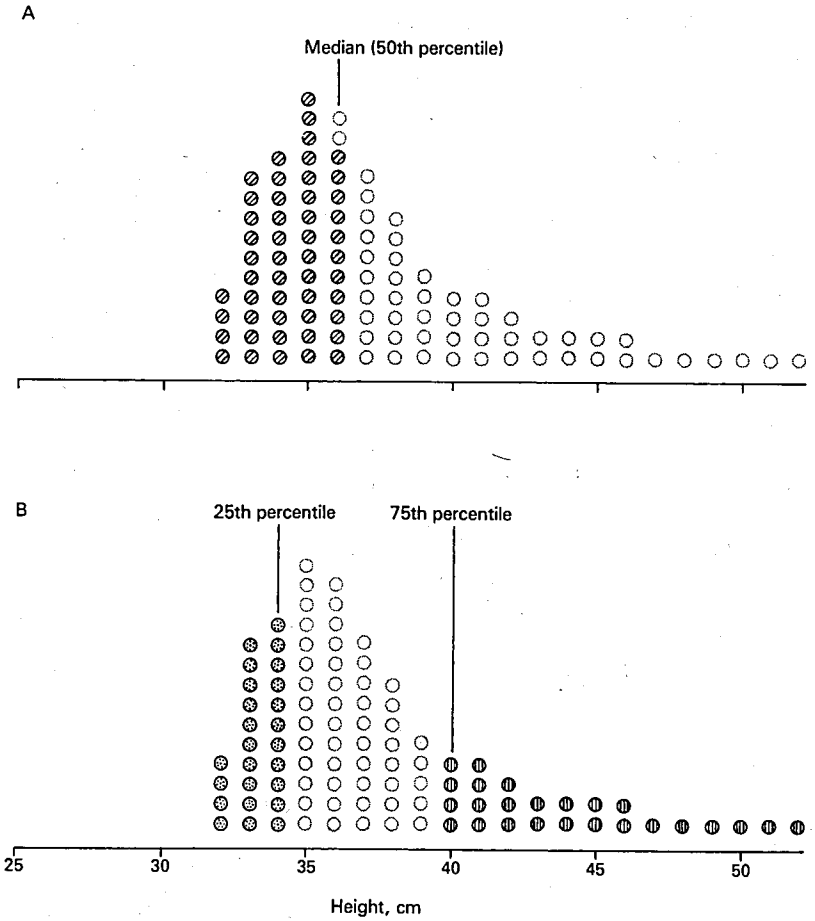


Figure 2-4 One way to describe a skewed distribution is with percentiles. The median is the point which divides the population in half. Panel *A* shows that 36 cm is the median height on Jupiter. Panel *B* shows the 25th and 75th percentiles, the points locating the lowest and highest quarter of the heights, respectively. The fact that the 25th percentile is closer to the median than the 75th percentile indicates that the distribution is skewed toward higher values.

36, and 40 cm) do not precisely describe the distribution of heights, they do indicate what the range of heights is and that there are a few very tall Jovians but not many very short ones.

Although these percentiles are often used, one could equally well report the 5th and 95th percentile points, or, for that matter, report the 5-, 25-, 50-, 75-, and 95-percentile points.

Computing the percentile points of a population is a good way to see how close to a normal distribution it is. Recall that we said that in a population that exhibits a normal distribution of values, about 95 percent of the population members fall within 2 standard deviations of the mean and about 68 percent fall within 1 standard deviation of the mean. Figure 2-5 shows that, for a normal distribution, the values of the associated percentile points are:

2.5th percentile	mean − 2 standard deviation
16th percentile	mean − 1 standard deviation
25th percentile	mean − 0.67 standard deviation
50th percentile (median)	mean
75th percentile	mean + 0.67 standard deviation
84th percentile	mean + 1 standard deviation
97.5th percentile	mean + 2 standard deviation

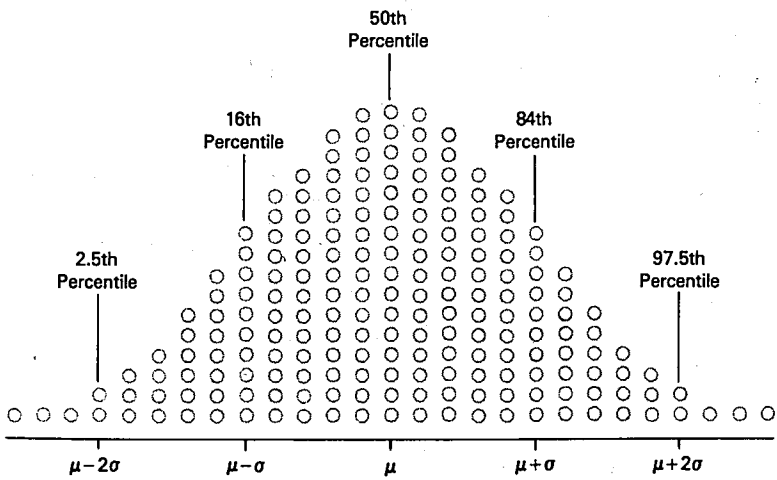


Figure 2-5 Percentile points of the normal distribution.

If the values associated with the percentiles are not too different from what one would expect on the basis of the mean and standard deviation, the normal distribution is a good approximation to the true population and then the mean and standard deviation describe the population adequately.

Why care whether or not the normal distribution is a good approximation? Because many of the statistical procedures used to test hypotheses—including the ones we will develop in Chaps. 2, 4, and 9—require that the population follow a normal distribution at least approximately for the tests to be reliable. (Chapters 10 and 11 present alternative tests that do not require this assumption.)

HOW TO MAKE ESTIMATES FROM A LIMITED SAMPLE

So far, everything we have done has been exact because we followed the conservative course of examining every single member of the population. Usually it is physically or fiscally impossible to do this, and we are limited to examining a *sample* of n individuals drawn from the population in the hope that it is representative of the complete population. Without knowledge of the entire population, we can no longer know the population mean μ and population standard deviation σ. Nevertheless, we can estimate them from the sample. The estimate of the population mean is called the *sample mean* and is defined analogously to the population mean:

$$\text{Sample mean} = \frac{\text{sum of values, e.g., heights, of each observation in sample}}{\text{number of observations in sample}}$$

The equivalent mathematical statement is

$$\bar{X} = \frac{\Sigma X}{n}$$

in which the bar over the X denotes that it is the mean of the n observations of X.

The estimate of the population standard deviation is called the *sample standard deviation* s and is defined by

$$\text{Sample standard deviation} = \sqrt{\frac{\text{sum of (value of observation in the sample} - \text{sample mean)}^2}{\text{number of observations in sample} - 1}}$$

or, mathematically,*

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}}$$

(The standard deviation is also often denoted SD.) This definition differs from the definition of the population standard deviation σ in two ways: (1) the population mean μ has been replaced by our estimate of it, the sample mean $\bar{X}$, and (2) we compute the "average" squared deviation of a sample by dividing by $n - 1$ rather than n. The precise reason for this requires substantial mathematical arguments, but we can present the following intuitive justification. The sample will never show as much variability as the entire population and dividing by $n - 1$ instead of n compensates for the resultant tendency to underestimate the population standard deviation.

In conclusion, when there is no evidence that the sample was not drawn from a normal distribution, summarize data with the sample mean and sample standard deviation, the best estimates of the population mean and population standard deviation, because these two parameters completely define the normal distribution. When there is evidence that the population under study does not follow a normal distribution, summarize data with the median and upper and lower percentiles.

HOW GOOD ARE THESE ESTIMATES?

The mean and standard deviation computed from a random sample are estimates of the mean and standard deviation of the entire population from which the sample was drawn. There is nothing special about the

*All equations in the text will be presented in the form most conducive to understanding statistical concepts. Often there is another, mathematically equivalent, form of the equation which is more suitable for computation. These forms are tabulated in Appendix A.

specific random sample used to compute these statistics, and different random samples will yield slightly different estimates of the true population mean and standard deviation. To quantitate how accurate these estimates are likely to be, we can compute their *standard errors.* It is possible to compute a standard error for any statistic, but here we shall focus on the *standard error of the mean.* This statistic quantifies the certainty with which the mean computed from a random sample estimates the true mean of the population from which the sample was drawn.

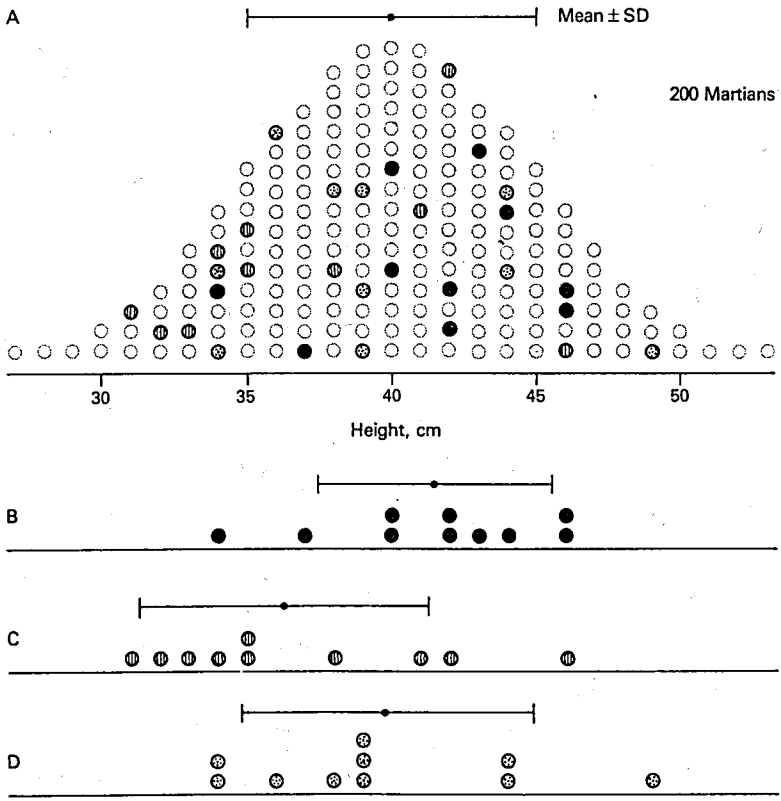


Figure 2-6 If one draws three different samples of 10 members each from a single population, one will obtain three different estimates of the population mean and standard deviation.

What is the standard error of the mean?

Figure 2-6*A* shows the same population of Martian heights we considered before. Since we have complete knowledge of every Martian's height, we will use this example to explore how accurately statistics computed from a random sample describe the entire population. Suppose that we draw a random sample of 10 Martians from the entire population of 200, then compute the sample mean and sample standard deviation. The 10 Martians in the sample are indicated by solid circles in Fig. 2-6*A*. Figure 2-6*B* shows the results of this random sample as it might be reported in a journal article, together with the sample mean ($\bar{X}$= 41.5 cm) and sample standard deviation (s = 3.8 cm). The values are close, but not equal, to the population mean (μ = 40 cm) and standard deviation (σ = 5 cm).

There is nothing special about this sample—after all, it was drawn at random—so let us consider a second random sample of 10 Martians from the same population of 200. Figure 2-6*C* shows the results of this sample, with the specific Martians identified in Fig. 2-6*A* as hatched circles. While the mean and standard deviation, 36 and 5 cm, of this second random sample are also similar to the mean and standard deviation of the whole population, they are not the same. Likewise, they are also similar, but not identical, to those from the first sample.

Figure 2-6*D* shows a third random sample of 10 Martians, identified in Fig. 2-6*A* with circles containing dots. This sample leads to estimates of 40 and 5 cm for the mean and standard deviation.

Now, we make an important change in emphasis. Instead of concentrating on the population of all 200 Martians, let us examine the *means of all possible random samples of 10 Martians.* We have already found three possible values for this mean, 41.5, 36, and 40 cm, and there are many more possibilities. Figure 2-7 shows these three means, plotted as circles, just as we plotted the individual heights, using the same symbols as Fig. 2-6. To better understand the amount of variability in the means of samples of 10 Martians, let us draw another 22 random samples of 10 Martians each and compute the mean of each sample. These additional means are plotted in Fig. 2-7 as open circles.

Now that we have drawn 25 random samples of 10 Martians each, have we exhausted the entire population of 200 Martians? No. There

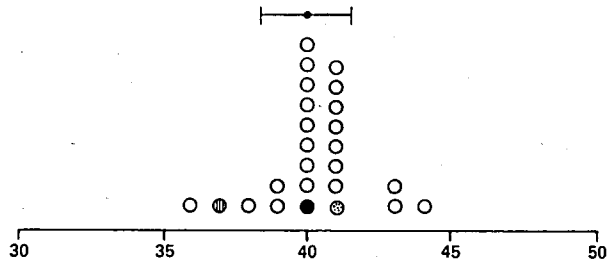


Figure 2–7 If one draws more and more samples—each with 10 members—from a single population, one obtains the population of all possible sample means. This figure illustrates the means of 25 samples of 10 Martians each drawn from the population of 200 Martians shown in Figs. 2–1 and 2–6*A*. The means of the three specific samples shown in Fig. 2–6 are shown using circles filled with corresponding patterns. This new population of all possible sample means will be normally distributed regardless of the nature of the original population; its mean will equal the mean of the original population; its standard deviation is called the standard error of the mean.

are more than 10^{16} different ways to select 10 Martians at random from the population of 200 Martians.

Look at Fig. 2-7. The collection of the means of 25 random samples, each of 10 Martians, has a roughly bell-shaped distribution, which is similar to the normal distribution. When the variable of interest is the sum of many other variables, its distribution will tend to be normal, regardless of the distributions of the variables used to form the sum. Since the sample mean is just such a sum, its distribution will tend to be normal, with the approximation improving as the sample size increases. (If the sample were drawn from a normally distributed population, the distribution of the sample means would have a normal distribution regardless of the sample size.) Therefore, it makes sense to describe the data in Fig. 2-7 by computing their mean and standard deviation. Since the mean value of the 25 points in Fig. 2-7 is the mean of the means of 25 samples, we will denote it $\overline{X}_{\overline{X}}$. The standard deviation is the *standard deviation of the means* of 25 independent random samples of 10 Martians each, and so we will denote it $s_{\overline{X}}$. Using the formulas for mean and standard deviation presented earlier, we compute $\overline{X}_{\overline{X}} = 40$ cm and $s_{\overline{X}} = 1.6$ cm.

The mean of the sample means $\overline{X}_{\overline{X}}$ is (within measurement and rounding error) equal to the mean height μ of the entire population of

200 Martians from which we drew the random samples. This is quite a remarkable result, since $\bar{X}_{\bar{X}}$ is *not* the mean of a sample drawn directly from the original population of 200 Martians; $\bar{X}_{\bar{X}}$ is the mean of 25 random samples of size 10 drawn from the *population consisting of all 10^{16} possible values of the mean of random samples of size 10 drawn from the original population of 200 Martians.*

Is $s_{\bar{X}}$ equal to the standard deviation σ of the population of 200 Martians? No. In fact, it is quite a bit smaller; the standard deviation of the collection of sample means $s_{\bar{X}}$ is 1.6 cm while the standard deviation for the whole populations is 5 cm. Just as the standard deviation of the original sample of 10 Martians s is an estimate of the variability of Martians' heights, $s_{\bar{X}}$ is an estimate of the *variability of possible values of means of samples of 10 Martians.* Since when one computes the mean, extreme values tend to balance each other, there will be less variability in the values of the sample means than in the original population. $s_{\bar{X}}$ is a measure of the precision with which a sample mean $\bar{X}$ estimates the population mean μ. We might name $s_{\bar{X}}$ "standard deviation of means of random samples of size 10 drawn from the original population." To be brief, statisticians have coined a shorter name, the *standard error of the mean* (SEM).

Since the precision with which we can estimate the mean increases as the sample size increases, the standard error of the mean decreases as the sample size increases. Conversely, the more variability in the original population, the most variability will appear in possible mean values of samples; therefore, the standard error of the mean increases as the population standard deviation increases. The true standard error of the mean of samples of size n drawn from a population with standard deviation σ is*

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

The best estimate of $\sigma_{\bar{X}}$ from a single sample is

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

*This equation is derived in Chapter. 4.

Since the possible values of the sample mean tend to follow a normal distribution, the true (and unobserved) mean of the original population will lie within 2 standard errors of the sample mean about 95 percent of the time.

As already noted, mathematicians have shown that the distribution of mean values will always approximately follow a normal distribution *regardless* of how the population from which the original samples were drawn is distributed. We have developed what statisticians call the *central-limit theorem.* It says:

- *The distribution of sample means will be approximately normal regardless of the distribution of values in the original population from which the samples were drawn.*
- *The mean value of the collection of all possible sample means will equal the mean of the original population.*
- *The standard deviation of the collection of all possible means of samples of a given size, called the standard error of the mean, depends on both the standard deviation of the original population and the size of the sample.*

Figure 2-8 illustrates the relationship between the sample mean, the sample standard deviation, and the standard error of the mean and how they vary with sample size as we measure more and more Martians.* As we add more Martians to our sample, the sample mean $\overline{X}$ and standard deviation s estimate the population mean μ and standard deviation σ with increasing precision. This increase in the precision with which the sample mean estimates the population mean is reflected by the smaller standard error of the mean with larger sample sizes. Therefore, the standard error of the mean tells not about variability in the original population, as the standard deviation does, but about the certainty with which a sample mean estimates the true population mean.

The *standard deviation* and *standard error of the mean* measure two very different things and are often confused. Most medical investigators summarize their data with the standard error of the mean

*Figure 2-8 was obtained by selecting two Martians from Fig. 2-1 at random, then computing $\overline{X}$, s, and $s_{\overline{X}}$. Then one more Martian was selected and the computations done again. Then, a fourth, a fifth, and so on, always adding to the sample already drawn. Had we selected different random samples or the same samples in a different order, Fig. 2-8 would have been different.

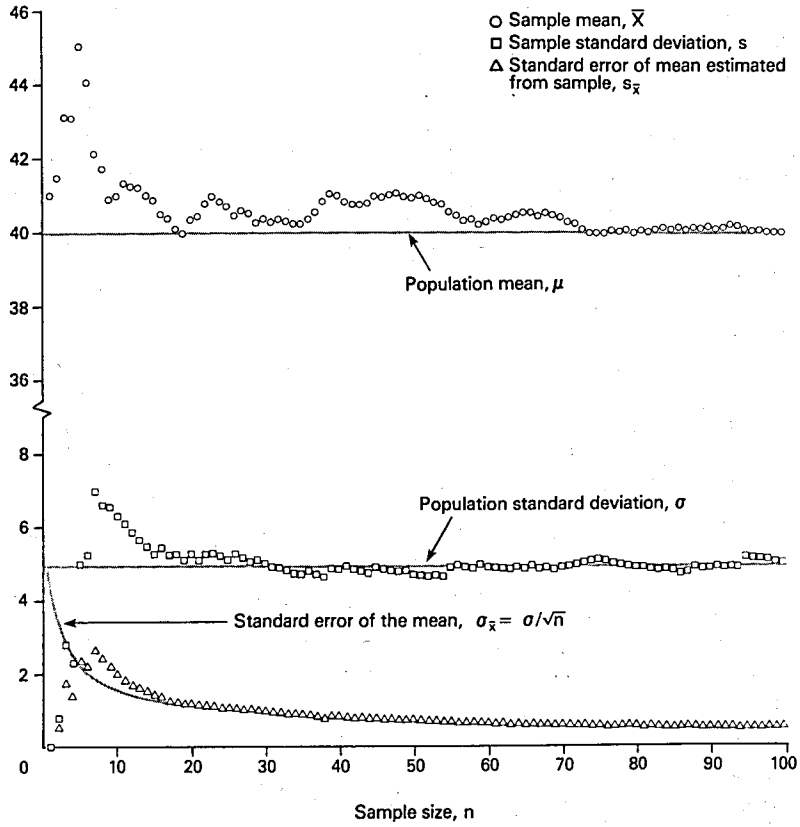


Figure 2-8 As the size of a random sample of Martians drawn from the population depicted in Fig. 2-1 grows, the precision with which the sample mean and sample standard deviation, $\bar{X}$ and s, estimate the true population mean and standard deviation, μ and σ, increases. This increasing precision appears in two ways: (1) the difference between the statistics computed from the sample (the points) moves closer to the true population values (the lines), and (2) the size of the standard error of the mean decreases.

because it is always smaller than the standard deviation. It makes their data look better. However, unlike the standard deviation, which quantifies the *variability in the population,* the standard error of the mean quantifies *uncertainty in the estimate of the mean.* Since readers are generally interested in knowing about the population, data should never be summarized with the standard error of the mean.

To understand the difference between the standard deviation and standard error of the mean and why one ought to summarize data using the standard deviation, suppose that in a sample of 20 patients an investigator reports that the mean cardiac output was 5.0 L/min with a standard deviation of 1 L/min. Since about 95 percent of all population members fall within about 2 standard deviations of the mean, this report would tell you that, assuming that the population of interest followed a normal distribution, it would be unusual to observe a cardiac output below about 3 or above about 7 L/min. Thus, you have a quick summary of the population described in the paper and a range against which to compare specific patients you examine. Unfortunately, it is unlikely that these numbers would be reported, the investigator being more likely to say that the cardiac output was 5.0 ± 0.22 (SEM) L/min. If you confuse the standard error of the mean with the standard deviation, you would believe that the range of most of the population was narrow indeed—4.56 to 5.44 L/min. These values describe the range which, with about 95 percent confidence, contains the mean cardiac output of the entire population from which the sample of 20 patients was drawn. (Chapter 7 discusses these ideas in detail.) In practice, one generally wants to compare a specific patient's cardiac output not only with the population mean but with the spread in the population taken as a whole.

SUMMARY

When a population follows a normal distribution, we can describe its location and variability completely with two parameters, the mean and standard deviation. When the population does not follow a normal distribution at least roughly, it is more informative to describe it with the median and other percentiles. Since one can rarely observe all members of a population, we will estimate these parameters from a sample drawn at random from the population. The standard error quantifies the precision of these estimates. For example, the standard error of the mean quantifies the precision with which the sample mean estimates the population mean.

In addition to being useful for describing a population or sample, these numbers can be used to estimate how compatible measurements

are with clinical or scientific assertions that an intervention affected some variable. We now turn our attention to this problem.

PROBLEMS

2-1 Viral load of HIV-1 is a known risk factor for heterosexual transmission of HIV; people with higher viral loads of HIV-1 are significantly more likely to transmit the virus to their uninfected partners. Thomas Quinn and associates ("Viral Load and Heterosexual Transmission of Human Immunodeficiency Virus Type 1." *N. Engl. J. Med.*, **342**: 921–929, 2000) studied this question by measuring the amount of HIV-1 RNA detected in blood serum. The following data represent HIV-1 RNA levels in the group whose partners seroconverted, which means that an initially uninfected partner became HIV positive during the course of the study; 79725, 12862, 18022, 76712, 256440, 14013, 46083, 6808, 85781, 1251, 6081, 50397, 11020, 13633, 1064, 496433, 25308, 6616, 11210, 13900 RNA copies/mL. Find the mean, median, standard deviation, and 25th and 75th percentiles of these concentrations. Do these data seem to be drawn from a normally distributed population? Why or why not?

2-2 When data are not normally distributed, researchers can sometimes *transform* their data to obtain values that more closely approximate a normal distribution. One approach to this is to take the logarithm of the observations. The following numbers represent the same data described in Prob. 2-1 following log (base 10) transformation: 4.90, 4.11, 4.26, 4.88, 5.41, 4.15, 4.66, 3.83, 4.93, 3.10, 3.78, 4.70, 4.04, 4.13, 3.03, 5.70, 4.40, 3.82, 4.05, 4.14. Find the mean, median, standard deviation, and 25th and 75th percentiles of these concentrations. Do these data seem to be drawn from a normally distributed population? Why or why not?

2-3 Polychlorinated biphenyls (PCBs) are a class of environmental chemicals associated with a variety of adverse health effects, including intellectual impairment in children exposed *in utero* while their mothers were pregnant. PCBs are also one of the most abundant contaminants found in human fat. Tu Binh Minh and colleagues analyzed PCB concentrations in the fat of a group of Japanese adults ("Occurrence of Tris (4-chlorophenyl)methane, Tris (4-chlorophenyl)methanol, and "Some Other Persistent Organochlorines in Japanese Human Adipose Tissue," *Environ. Health Perspect.*, **108**:599–603, 2000). They detected 1800, 1800, 2600, 1300, 520, 3200, 1700, 2500, 560, 930, 2300, 2300, 1700, 720 ng/g lipid weight of PCBs in the people they studied. Find the mean, median standard deviation, and 25th and 75th percentiles of these

concentrations. Do these data seem to be drawn from a normally distributed population? Why or why not?

2-4 Sketch the distribution of all possible values of the number on the upright face of a die. What is the mean of this population of possible values?

2-5 Roll a *pair* of dice and note the numbers on each of the upright faces. These two numbers can be considered a sample of size 2 drawn from the population described in Prob. 2-4. This sample can be averaged. What does this average estimate? Repeat this procedure 20 times and plot the averages observed after each roll. What is this distribution? Compute its mean and standard deviation. What do they represent?

2-6 Robert Fletcher and Suzanne Fletcher ("Clinical Research in General Medical Journals: A 30-Year Perspective," *N. Engl. J. Med.*, **301**:180–183, 1979, used by permission) studied the characteristics of 612 randomly selected articles published in the *Journal of the American Medical Association, New England Journal of Medicine,* and *Lancet* since 1946. One of the attributes they examined was the number of authors; they found:

Year	No. of articles examined	Mean no. of authors	SD
1946	151	2.0	1.4
1956	149	2.3	1.6
1966	157	2.8	1.2
1976	155	4.9	7.3

Sketch the populations of numbers of authors for each of these years. How closely do you expect the normal distribution to approximate the actual population of all authors in each of these years? Why? Estimate the certainty with which these samples permit you to estimate the true mean number of authors for all articles published in comparable journals each year.

Chapter 3

How to Test for Differences between Groups

Statistical methods are used to summarize data and test hypotheses with those data. Chapter 2 discussed how to use the mean, standard deviation, median, and percentiles to summarize data and how to use the standard error of the mean to estimate the precision with which a sample mean estimates the population mean. Now we turn our attention to how to use data to test scientific hypotheses. The statistical techniques used to perform such tests are called *tests of significance;* they yield the highly prized *P value.* We now develop procedures to test the hypothesis that, on the average, different treatments all affect some variable identically. Specifically, we will develop a procedure to test the hypothesis that diet has no effect on the mean cardiac output of people living in a small town. Statisticians call this hypothesis of no effect the *null hypothesis.*

The resulting test can be generalized to analyze data obtained in experiments involving any number of treatments. In addition, it is the archetype for a whole class of related procedures known as *analysis of variance.*

THE GENERAL APPROACH

To begin our experiment, we randomly select four groups of seven people each from a small town with 200 healthy adult inhabitants. All

participants give informed consent. People in the control group continue eating normally; people in the second group eat only spaghetti; people in the third group eat only steak; and people in the fourth group eat only fruit and nuts. After 1 month, each person has a cardiac catheter inserted and his or her cardiac output is measured.

As with most tests of significance, we begin with the hypothesis that all treatments (diets) have the same effect (on cardiac output). Since the study includes a control group (as experiments generally should), this hypothesis is equivalent to the hypothesis that diet has no effect on cardiac output. Figure 3-1 shows the distribution of cardiac outputs for the entire population, with each individual's cardiac output represented by a circle. The specific individuals who were randomly selected for each diet are indicated by shaded circles, with different shading for different diets. Figure 3-1 shows that the hypothesis is, in fact, true. Unfortunately, as investigators we cannot observe the entire population and are left with the problem of deciding whether or not it is true from the limited data shown in Fig. 3-2. There are obviously differences between the samples; the question is: *Are these differences due to*

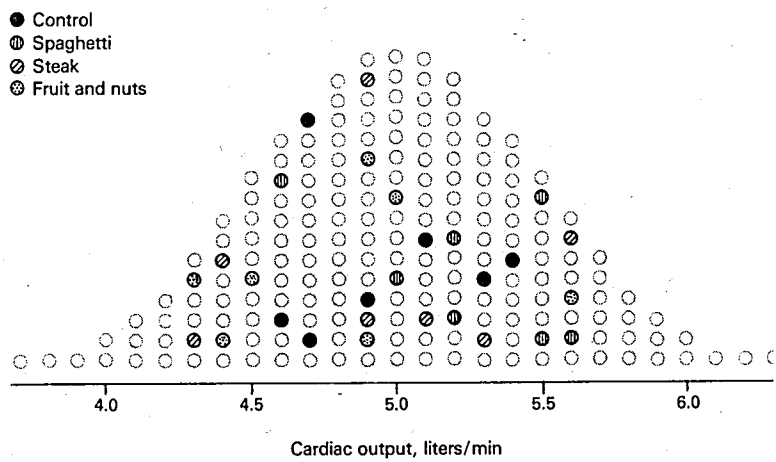


Figure 3-1 The values of cardiac output associated with all 200 members of the population of a small town. Since diet does not affect cardiac output, the four groups of seven people each selected at random to participate in our experiment (control, spaghetti, steak, fruit and nuts) simply represent four random samples drawn from a single population.

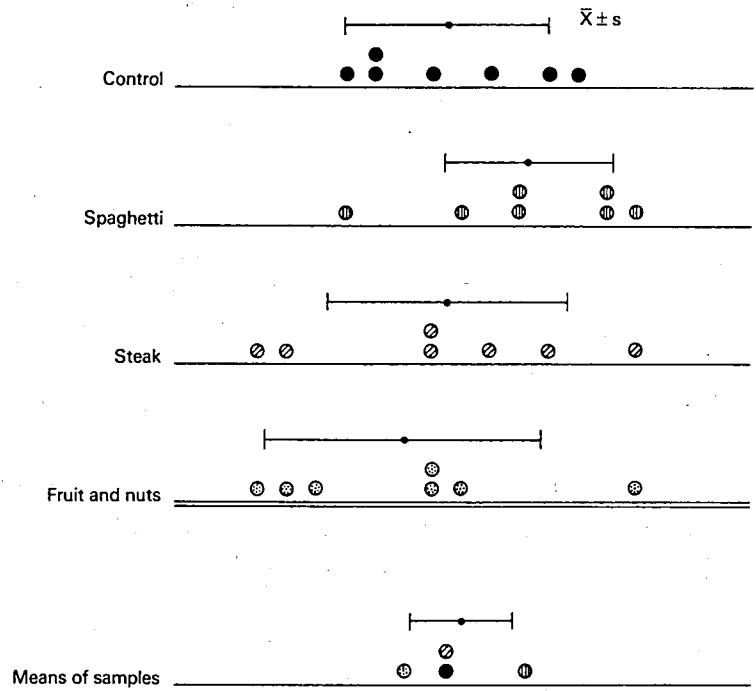


Figure 3-2 An investigator cannot observe the entire population but only the four samples selected at random for treatment. This figure shows the same four groups of individuals as in Fig. 3-1 with their means and standard deviations as they would appear to the investigator. The question facing the investigator is: Are the observed differences due to the different diets or simply random variation? The figure also shows the collection of sample means together with their standard deviation, which is an estimate of the standard error of the mean.

the fact that the different groups of people ate differently or are these differences simply a reflection of the random variation in cardiac output between individuals?

To use the data in Fig. 3-2 to address this question, we proceed under the assumption that the hypothesis that diet has no effect on cardiac output is correct. Since we assume that it does not matter which diet any particular individual ate, we *assume* that the four experimental groups of seven people each are four random samples of size 7 *drawn from a single population* of 200 individuals. Since the samples are

drawn at random from a population with some variance, we expect the samples to have different means and standard deviations, but *if our hypothesis that the diet has no effect on cardiac output is true,* the observed differences are simply due to random sampling.

Forget about statistics for a moment. What is it about different samples that leads you to believe that they are representative samples drawn from different populations? Figures 3-2, 3-3, and 3-4 show three different possible sets of samples of some variable of interest. Simply looking at these pictures makes most people think that the four samples in Fig. 3-2 were all drawn from a single population, while the samples in Figs. 3-3

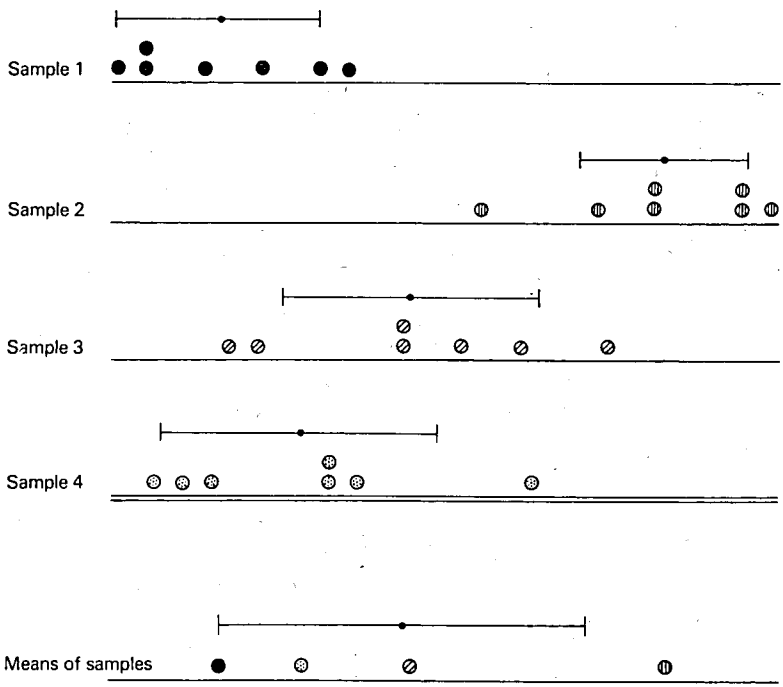


Figure 3-3 The four samples shown are identical to those in Fig. 3-2 except that the variability in the mean values has been increased substantially. The samples now appear to differ from each other because the variability between the sample means is larger than one would expect from the variability within each sample. Compare the relative variability in mean values with the variability within the sample groups with that seen in Fig. 3-2.

and 3-4 were not. Why? The variability within each sample, quantified with the standard deviation, is approximately the same. In Fig. 3-2, the variability in the mean values of the samples is consistent with the variability one observes within the individual samples. In contrast, in Figs. 3-3 and 3-4, the variability among sample means is much larger than one would expect from the variability within each sample. Notice that we reach this conclusion whether all (Fig. 3-3) or only one (Fig. 3-4) of the sample means appear to differ from the others.

Now let us formalize this analysis of variability to analyze our diet experiment. The standard deviation or its square, the variance, is a good measure of variability. We will use the variance to construct a procedure to test the hypothesis that diet does not affect cardiac output.

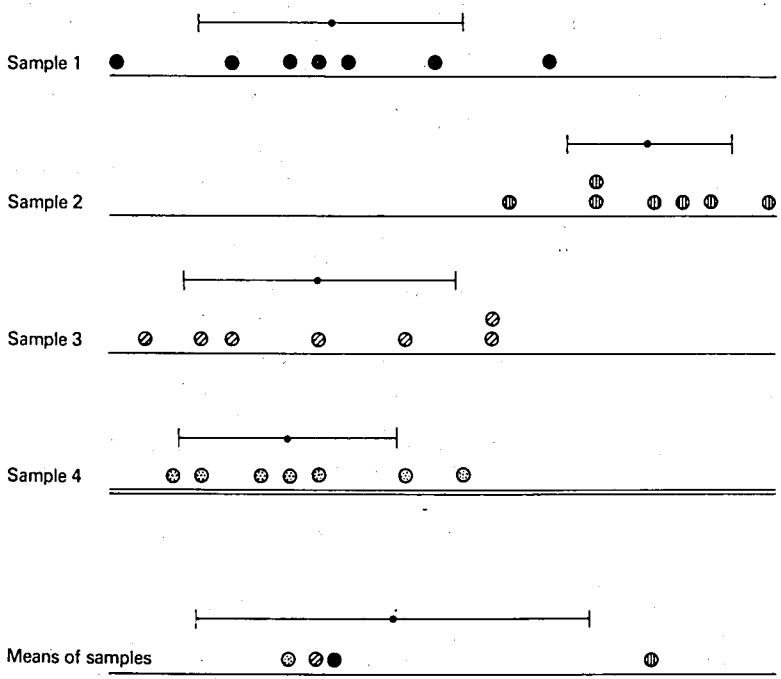


Figure 3-4 When the mean of even one of the samples (sample 2) differs substantially from the other samples, the variability computed from within the means is substantially larger than one would expect from examining the variability within the groups.

Chapter 2 showed that two population parameters—the mean and standard deviation (or, equivalently, the variance)—completely describe a normally distributed population. Therefore, we will use our raw data to compute these parameters and then base our analysis on their values rather than on the raw data directly. Since the procedures we will now develop are based on these parameters, they are called *parametric statistical methods.* Because these methods assume that the population from which the samples were drawn can be completely described by these parameters, they are valid only when the real population approximately follows the normal distribution. Other procedures, called *nonparametric statistical methods,* are based on frequencies, ranks, or percentiles and do not require this assumption.* Parametric methods generally provide more information about the treatment being studied and are more likely to detect a real treatment effect when the underlying population is normally distributed.

We will estimate the parameter population variance in two different ways: (1) The standard deviation or variance computed from each sample is an estimate of the standard deviation or variance of the entire population. Since each of these estimates of the population variance is computed from within each sample group, the estimates will not be affected by any differences in the mean values of different groups. (2) We will use the values of the means of each sample to determine a second estimate of the population variance. In this case, the differences between the means will obviously affect the resulting estimate of the population variance. If all the samples were, in fact, drawn from the same population (i.e., the diet had no effect), these two different ways to estimate the population variance should yield approximately the same number. When they do, we will conclude that the samples were likely to have been drawn from a single population; otherwise, we will reject this hypothesis and conclude that at least one of the samples was drawn from a different population. In our experiment, rejecting the original hypothesis would lead to the conclusion that diet *does* alter cardiac output.

TWO DIFFERENT ESTIMATES OF THE POPULATION VARIANCE

How shall we estimate the population variance from the four sample variances? When the hypothesis that the diet does not affect cardiac output

*We will study these procedures in Chaps. 5, 8, 10, and 11.

is true, the variances of each sample of seven people, regardless of what they ate, are equally good estimates of the population variance, so we simply average our four estimates of *variance within the treatment groups*

> Average variance in cardiac output within treatment groups = 1/4 (variance in cardiac output of controls + variance in cardiac output of spaghetti eaters + variance in cardiac output of steak eaters + variance in cardiac output of fruit and nut eaters)

The mathematical equivalent is

$$s^2_{\text{wit}} = 1/4\,(s^2_{\text{con}} + s^2_{\text{spa}} + s^2_{\text{st}} + s^2_{\text{f}})$$

where s^2 represents variance. The variance of each sample is computed with respect to the mean of that sample. Therefore, the population variance estimated from within the groups, *the within-groups variance* s^2_{wit}, will be the same whether or not diet altered cardiac output.

Next we estimate the population variance from the means of the samples. Since we have hypothesized that all four samples were drawn from a single population, the standard deviation of the four sample means will approximate the standard error of the mean. Recall that the standard error of the mean $\sigma_{\bar{X}}$ is related to the sample size n (in this case 7) and the population standard deviation σ according to

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Therefore, the true population variance σ^2 is related to the sample size and standard error of the mean according to

$$\sigma^2 = n\sigma^2_{\bar{X}}$$

We use this relationship to estimate the population variance from the variability between the sample means using

$$s^2_{\text{bet}} = ns^2_{\bar{X}}$$

where s^2_{bet} is the estimate of the population variance computed from between the sample means and $s_{\bar{X}}$ is the standard deviation of the

means of the four sample groups, the standard error of the mean. This estimate of the population variance computed from between the group means is often called the *between-groups variance.*

If the hypothesis that all four samples were drawn from the same population is true (i.e., that diet does not affect cardiac output), the within-groups variance and between-groups variance are both estimates of the same population variance and so should be about equal. Therefore, we will compute the following ratio, called the F-test statistic,

$$F = \frac{\text{population variance estimated from sample means}}{\text{population variance estimated as average of sample variances}}$$

$$F = \frac{s^2_{\text{bet}}}{s^2_{\text{wit}}}$$

Since both the numerator and the denominator are estimates of the same population variance σ^2, F should be about $\sigma^2/\sigma^2 = 1$. For the four random samples in Fig. 3-2, F is about equal to 1, we conclude that the data in Fig. 3-2 are not inconsistent with the hypothesis that diet does not affect cardiac output and we continue to accept that hypothesis.

Now we have a rule for deciding when to reject the hypothesis that all the samples were drawn from the same population:

> *If* F *is a big number, the variability between the sample means is larger than expected from the variability within the samples, so reject the hypothesis that all the samples were drawn from the same population.*

This quantitative statement formalizes the qualitative logic we used when discussing Figs. 3-2 to 3-4. The F associated with Fig. 3-3 is 68.0, and that associated with Fig. 3-4 is 24.5.

WHAT IS A "BIG" F?

The exact value of F one computes depends on which individuals were drawn for the random samples. For example, Fig. 3-5 shows yet another set of four samples of seven people drawn from the population of 200 people in Fig. 3-1. In this example $F = 0.5$. Suppose we repeated our

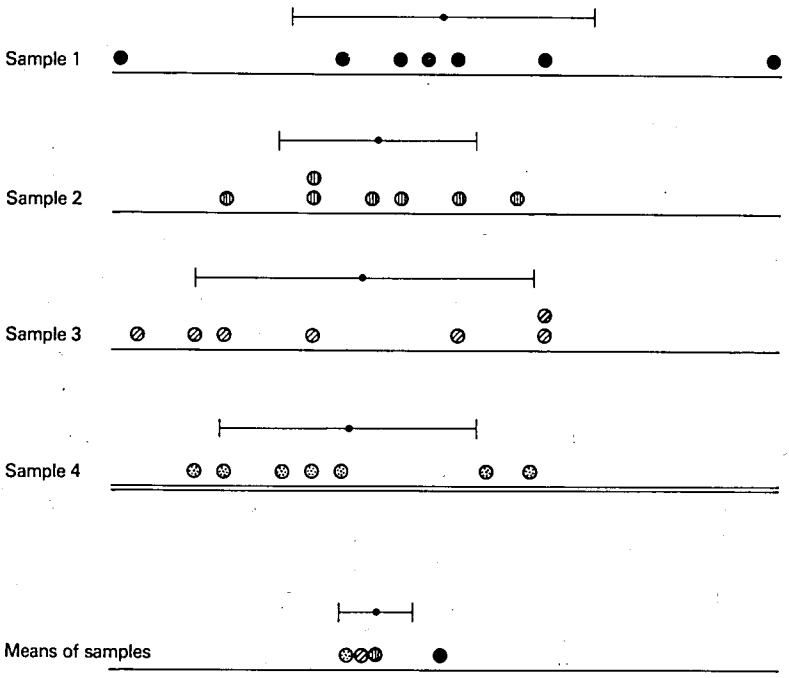


Figure 3-5 Four samples of seven members each drawn from the population shown in Fig. 3-1. Note that the variability in sample means is consistent with the variability within each of the samples, $F = 0.5$.

experiment 200 times on the same population. Each time we would draw four different samples of people and—even if the diet had no effect on cardiac output—get slightly different values for F due to random variation. Figure 3-6*A* shows the result of this procedure, with the resulting F's rounded to one decimal place and represented with a circle; the two dark circles represent the values of F computed from the data in Figs. 3-2 and 3-5. The exact shapes of the distribution of values of F depend on how many samples were drawn, the size of each sample, and the distribution of the population from which the samples were drawn.

As expected, most of the computed F's are around 1 (that is, between 0 and 2), but a few are much larger. Thus, even though most experiments will produce relatively small values of F, it is possible that,

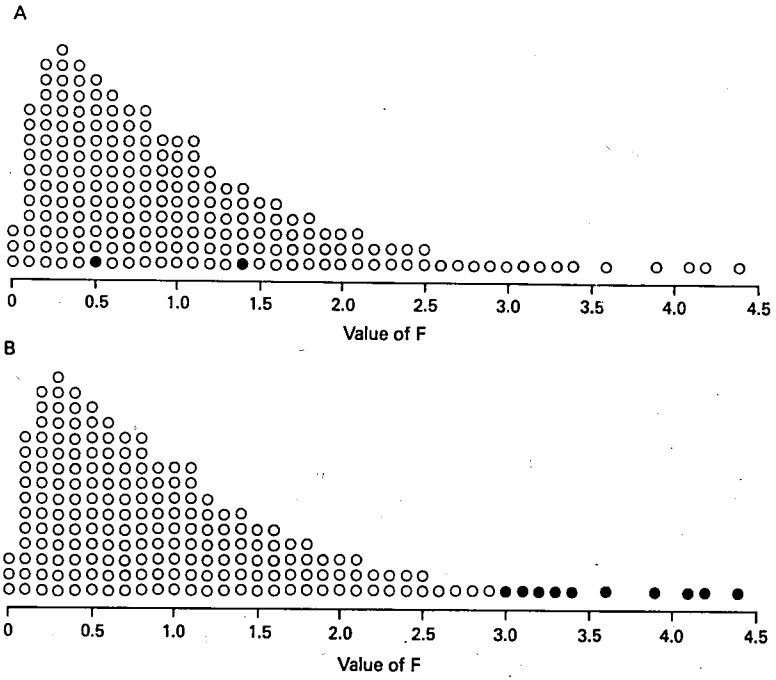


Figure 3-6 (*A*) Values of *F* computed from 200 experiments involving four samples, each of size 7, drawn from the population in Fig. 3-1. (*B*) We expect *F* to exceed 3.0 only 5 percent of the time when all samples were, in fact, drawn from a single population. (*C*) Results of computing the *F* ratio for all possible samples drawn from the original population. The 5 percent of most extreme *F* values are shown darker than the rest. (*D*) The *F* distribution one would obtain when sampling an infinite population. In this case, the cutoff value for considering *F* to be "big" is that value of *F* that subtends the upper 5 percent of the total area under the curve.

by sheer bad luck, one could select random samples that are not good representatives of the whole population. The result is an occasional relatively large value for *F* even though the treatment had no effect. Figure 3-6*B* shows, however, that such values are unlikely. Only 5 percent of the 200 experiments (10 experiments) produced *F* values equal to or greater than 3.0. We now have a tentative estimate of what to call a "big" value for *F*. Since *F* exceeded 3.0 only 10 out of 200 times *when all the samples were drawn from the same population,* we might decide that *F* is big when it exceeds 3.0 and reject the hypothesis that all the samples

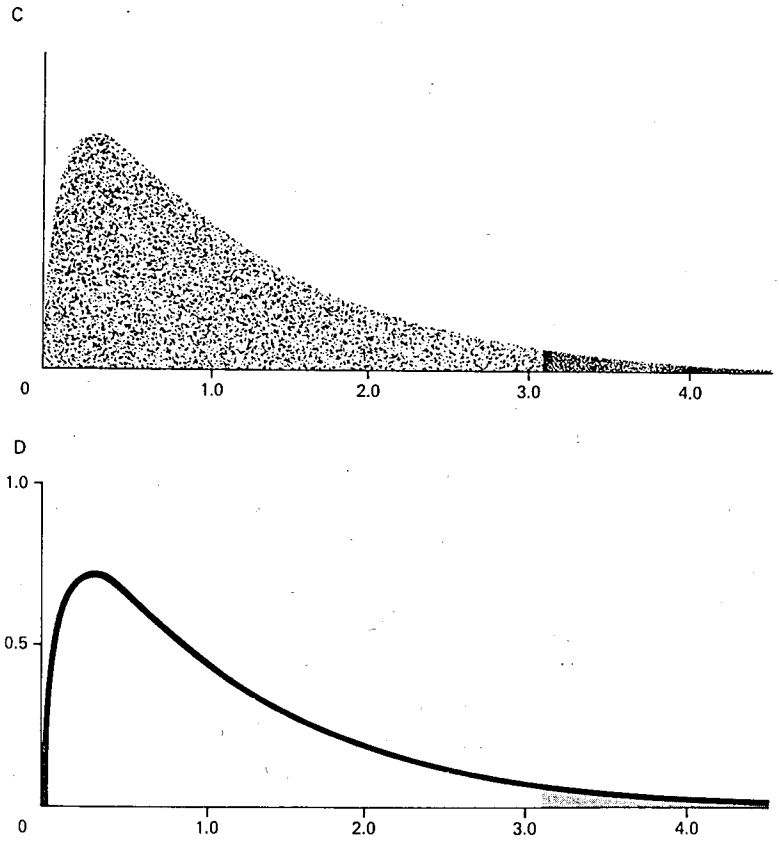


Figure 3-6 *(continued)*

were drawn from the same population (i.e., that the treatment had no effect). In deciding to reject the hypothesis of no effect when F is big, we accept the risk of erroneously rejecting this hypothesis 5 percent of the time because F will be 3.0 or greater about 5 percent of the time, even when the treatment does not alter mean response.

When we obtain such a "big" F, we reject the original hypothesis that all the means are the same and report $P < 0.05$. $P < 0.05$ means that there is less than a 5 percent chance of getting a value of F as big or bigger than the computed value if the original hypothesis were true (i.e., diet did not affect cardiac output).

The critical value of F should be selected not on the basis of just 200 experiments but all 10^{42} possible experiments. Suppose we did all 10^{42} experiments and computed the corresponding F values, then plotted the results, just as we did for Fig. 3-6*B*. Figure 3-6*C* shows the results with grains of sand to represent each observed F value. The darker sand indicates the biggest 5 percent of the F values. Notice how similar it is to Fig. 3-6*B*. This similarity should not surprise you, since the results in panel *B* are just a random sample of the population in panel *C*. Finally, recall that everything so far has been based on an original population containing only 200 members. In reality, populations are usually much larger, so that there can be many more than 10^{42} possible values of F. Often, there are essentially an infinite number of possible experiments. In terms of Fig. 3-6*C*, it is as if all the grains of sand melted together to yield the continuous line in Fig. 3-6*D*.

Therefore, *areas under the curve* are analogous to the fractions of total number of circles or grains of sand in panels *B* and *C*. Since the shaded region in Fig. 3-6*D* represents 5 percent of the total area under the curve, it can be used to compute that the cutoff point for a "big" F with the number of samples and sample size in this study is 3.01. This and other cutoff values that correspond to $P < 0.05$ and $P < 0.01$ are listed in Table 3-1.

To construct these tables, mathematicians have assumed four things about the underlying population that must be at least approximately satisfied for the tables to be applicable to real data:

- *Each sample must be independent of the other samples.*
- *Each sample must be randomly selected from the population being studied.*
- *The populations from which the samples were drawn must be normally distributed.**
- *The variances of each population must be equal, even when the means are different, i.e., when the treatment has an effect.*

When the data suggest that these assumptions do not apply, one ought not to use the procedure we just developed, the analysis of variance. Since there is one factor (the diet) that distinguishes the different exper-

*This is another reason parametric statistical methods require data from normally distributed populations.

Table 3-1 Critical Values of F Corresponding to $P < .05$ (Lightface) and $P < .01$ (Boldface)

ν_d	ν_n 1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248	249	250	251	252	253	253	254	254	254
	4052	**4999**	**5403**	**5625**	**5764**	**5859**	**5928**	**5981**	**6022**	**6056**	**6082**	**6106**	**6142**	**6169**	**6208**	**6234**	**6261**	**6286**	**6302**	**6323**	**6334**	**6352**	**6361**	**6366**
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41	19.42	19.43	19.44	19.45	19.46	19.47	19.47	19.48	19.49	19.49	19.50	19.50
	98.49	**99.00**	**99.17**	**99.25**	**99.30**	**99.33**	**99.36**	**99.37**	**99.39**	**99.40**	**99.41**	**99.42**	**99.43**	**99.44**	**99.45**	**99.46**	**99.47**	**99.48**	**99.48**	**99.49**	**99.49**	**99.49**	**99.50**	**99.50**
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71	8.69	8.66	8.64	8.62	8.60	8.58	8.57	8.56	8.54	8.54	8.53
	34.12	**30.82**	**29.46**	**28.71**	**28.24**	**27.91**	**27.67**	**27.49**	**27.34**	**27.23**	**27.13**	**27.05**	**26.92**	**26.83**	**26.69**	**26.60**	**26.50**	**26.41**	**26.35**	**26.27**	**26.23**	**26.18**	**26.14**	**26.12**
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87	5.84	5.80	5.77	5.74	5.71	5.70	5.68	5.66	5.65	5.64	5.63
	21.20	**18.00**	**16.69**	**15.98**	**15.52**	**15.21**	**14.98**	**14.80**	**14.66**	**14.54**	**14.45**	**14.37**	**14.24**	**14.15**	**14.02**	**13.93**	**13.83**	**13.74**	**13.69**	**13.61**	**13.57**	**13.52**	**13.48**	**13.46**
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.64	4.60	4.56	4.53	4.50	4.46	4.44	4.42	4.40	4.38	4.37	4.36
	16.26	**13.27**	**12.06**	**11.39**	**10.97**	**10.67**	**10.45**	**10.29**	**10.15**	**10.05**	**9.96**	**9.89**	**9.77**	**9.68**	**9.55**	**9.47**	**9.38**	**9.29**	**9.24**	**9.17**	**9.13**	**9.07**	**9.04**	**9.02**
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68	3.67
	13.74	**10.92**	**9.78**	**9.15**	**8.75**	**8.47**	**8.26**	**8.10**	**7.98**	**7.87**	**7.79**	**7.72**	**7.60**	**7.52**	**7.39**	**7.31**	**7.23**	**7.14**	**7.09**	**7.02**	**6.99**	**6.94**	**6.90**	**6.88**
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.52	3.49	3.44	3.41	3.38	3.34	3.32	3.29	3.28	3.25	3.24	3.23
	12.25	**9.55**	**8.45**	**7.85**	**7.46**	**7.19**	**7.00**	**6.84**	**6.71**	**6.62**	**6.54**	**6.47**	**6.35**	**6.27**	**6.15**	**6.07**	**5.98**	**5.90**	**5.85**	**5.78**	**5.75**	**5.70**	**5.67**	**5.65**
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.23	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.94	2.93
	11.26	**8.65**	**7.59**	**7.01**	**6.63**	**6.37**	**6.19**	**6.03**	**5.91**	**5.82**	**5.74**	**5.67**	**5.56**	**5.48**	**5.36**	**5.28**	**5.20**	**5.11**	**5.06**	**5.00**	**4.96**	**4.91**	**4.88**	**4.86**
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07	3.02	2.98	2.93	2.90	2.86	2.82	2.80	2.77	2.76	2.73	2.72	2.71
	10.56	**8.02**	**6.99**	**6.42**	**6.06**	**5.80**	**5.62**	**5.47**	**5.35**	**5.26**	**5.18**	**5.11**	**5.00**	**4.92**	**4.80**	**4.73**	**4.64**	**4.56**	**4.51**	**4.45**	**4.41**	**4.36**	**4.33**	**4.31**
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77	2.74	2.70	2.67	2.64	2.61	2.59	2.56	2.55	2.54
	10.04	**7.56**	**6.55**	**5.99**	**5.64**	**5.39**	**5.21**	**5.06**	**4.95**	**4.85**	**4.78**	**4.71**	**4.60**	**4.52**	**4.41**	**4.33**	**4.25**	**4.17**	**4.12**	**4.05**	**4.01**	**3.96**	**3.93**	**3.91**
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79	2.74	2.70	2.65	2.61	2.57	2.53	2.50	2.47	2.45	2.42	2.41	2.40
	9.65	**7.20**	**6.22**	**5.67**	**5.32**	**5.07**	**4.88**	**4.74**	**4.63**	**4.54**	**4.46**	**4.40**	**4.29**	**4.21**	**4.10**	**4.02**	**3.94**	**3.86**	**3.80**	**3.74**	**3.70**	**3.66**	**3.62**	**3.60**
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69	2.64	2.60	2.54	2.50	2.46	2.42	2.40	2.36	2.35	2.32	2.31	2.30
	9.33	**6.93**	**5.95**	**5.41**	**5.06**	**4.82**	**4.65**	**4.50**	**4.39**	**4.30**	**4.22**	**4.16**	**4.05**	**3.98**	**3.86**	**3.78**	**3.70**	**3.61**	**3.56**	**3.49**	**3.46**	**3.41**	**3.38**	**3.36**
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60	2.55	2.51	2.46	2.42	2.38	2.34	2.32	2.28	2.26	2.24	2.22	2.21
	9.07	**6.70**	**5.74**	**5.20**	**4.86**	**4.62**	**4.44**	**4.30**	**4.19**	**4.10**	**4.02**	**3.96**	**3.85**	**3.78**	**3.67**	**3.59**	**3.51**	**3.42**	**3.37**	**3.30**	**3.27**	**3.21**	**3.18**	**3.16**
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13
	8.86	**6.51**	**5.56**	**5.03**	**4.69**	**4.46**	**4.28**	**4.14**	**4.03**	**3.94**	**3.86**	**3.80**	**3.70**	**3.62**	**3.51**	**3.43**	**3.34**	**3.26**	**3.21**	**3.14**	**3.11**	**3.06**	**3.02**	**3.00**
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48	2.43	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.12	2.10	2.08	2.07
	8.68	**6.36**	**5.42**	**4.89**	**4.56**	**4.32**	**4.14**	**4.00**	**3.89**	**3.80**	**3.73**	**3.67**	**3.56**	**3.48**	**3.36**	**3.29**	**3.20**	**3.12**	**3.07**	**3.00**	**2.97**	**2.92**	**2.89**	**2.87**

16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42	2.37	2.33	2.28	2.24	2.20	2.16	2.13	2.09	2.07	2.04	2.02	2.01
	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.61	3.55	3.45	3.37	3.25	3.18	3.10	3.01	2.96	2.98	2.86	2.80	2.77	2.75
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38	2.33	2.29	2.23	2.19	2.15	2.11	2.08	2.04	2.02	1.99	1.97	1.96
	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.45	3.35	3.27	3.16	3.08	3.00	2.92	2.86	2.79	2.76	2.70	2.67	2.65
18	4.41	3.55	3.16	2.93	2.77	3.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29	2.25	2.19	2.15	2.11	2.07	2.04	2.00	1.98	1.95	1.93	1.92
	8.28	6.01	5.09	4.58	4.25	4.01	3.85	3.71	3.60	3.51	3.44	3.37	3.27	3.19	3.07	3.00	2.91	2.83	2.78	2.71	2.68	2.62	2.59	2.57
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31	2.26	2.21	2.15	2.11	2.07	2.02	2.00	1.96	1.94	1.91	1.90	1.88
	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.19	3.12	3.00	2.92	2.84	2.76	2.70	2.63	2.60	2.54	2.51	2.49
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28	2.23	2.18	2.12	2.08	2.04	1.99	1.96	1.92	1.90	1.87	1.85	1.84
	8.10	5.85	4.94	4.43	4.10	3.87	3.71	3.56	3.45	3.37	3.30	3.23	3.13	3.05	2.94	2.86	2.77	2.69	2.63	2.56	2.53	2.47	2.44	2.42
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.20	2.15	2.09	2.05	2.00	1.96	1.93	1.89	1.87	1.84	1.82	1.81
	8.02	5.78	4.87	4.37	4.04	3.81	3.65	3.51	3.40	3.31	3.24	3.17	3.07	2.99	2.88	2.80	2.72	2.63	2.58	2.51	2.47	2.42	2.38	2.36
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26	2.23	2.18	2.13	2.07	2.03	1.98	1.93	1.91	1.87	1.84	1.81	1.80	1.78
	7.94	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	3.02	2.94	2.83	2.75	2.67	2.58	2.53	2.46	2.42	2.37	2.33	2.31
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.20	2.14	2.10	2.04	2.00	1.96	1.91	1.88	1.84	1.82	1.79	1.77	1.76
	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	2.97	2.89	2.78	2.70	2.62	2.53	2.48	2.41	2.37	2.32	2.28	2.26
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18	2.13	2.09	2.02	1.98	1.94	1.89	1.86	1.82	1.80	1.76	1.74	1.73
	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.25	3.17	3.09	3.03	2.93	2.85	2.74	2.66	2.58	2.49	2.44	2.36	2.33	2.27	2.23	2.21
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16	2.11	2.06	2.00	1.96	1.92	1.87	1.84	1.80	1.77	1.74	1.72	1.71
	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.21	3.13	3.05	2.99	2.89	2.81	2.70	2.62	2.54	2.45	2.40	2.32	2.29	2.23	2.19	2.17
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.10	2.05	1.99	1.95	1.90	1.85	1.82	1.78	1.76	1.72	1.70	1.69
	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.17	3.09	3.02	2.96	2.86	2.77	2.66	2.58	2.50	2.41	2.36	2.28	2.25	2.19	2.15	2.13
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13	2.08	2.03	1.97	1.93	1.88	1.84	1.80	1.76	1.74	1.71	1.68	1.67
	7.68	5.49	4.60	4.11	3.79	3.56	3.39	3.26	3.14	3.06	2.98	2.93	2.83	2.74	2.63	2.55	2.47	2.38	2.33	2.25	2.21	2.16	2.12	2.10
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12	2.06	2.02	1.96	1.91	1.87	1.81	1.78	1.75	1.72	1.69	1.67	1.65
	7.64	5.45	4.57	4.07	3.76	3.53	3.36	3.23	3.11	3.03	2.95	2.90	2.80	2.71	2.60	2.52	2.44	2.35	2.30	2.22	2.18	2.13	2.09	2.06
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.05	2.00	1.94	1.90	1.85	1.80	1.77	1.73	1.71	1.68	1.65	1.64
	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.08	3.00	2.92	2.87	2.77	2.68	2.57	2.49	2.41	2.32	2.27	2.19	2.15	2.10	2.06	2.03
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09	2.04	1.99	1.93	1.89	1.84	1.79	1.76	1.72	1.69	1.66	1.64	1.62
	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.06	2.98	2.90	2.84	2.74	2.66	2.55	2.47	2.38	2.29	2.24	2.16	2.13	2.07	2.03	2.01
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07	2.02	1.97	1.91	1.86	1.82	1.76	1.74	1.69	1.67	1.64	1.61	1.59
	7.50	5.34	4.46	3.97	3.66	3.42	3.25	3.12	3.01	2.94	2.86	2.80	2.70	2.62	2.51	2.42	2.34	2.25	2.20	2.12	2.08	2.02	1.98	1.96
34	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.17	2.12	2.08	2.05	2.00	1.95	1.89	1.84	1.80	1.74	1.71	1.67	1.64	1.61	1.59	1.57
	7.44	5.29	4.42	3.93	3.61	3.38	3.21	3.08	2.97	2.89	2.82	2.76	2.66	2.58	2.47	2.38	2.30	2.21	2.15	2.08	2.04	1.98	1.94	1.91

Note: ν_n = degrees of freedom for numerator; ν_d = degrees of freedom for denominator.

(*Continued on next page*)

Table 3-1 Critical Values of F Corresponding to $P < .05$ (Lightface) and $P < .01$ (Boldface) (*Continued*)

ν_d	ν_n 1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.03	1.98	1.93	1.87	1.82	1.78	1.72	1.69	1.65	1.62	1.59	1.56	1.55
	7.39	**5.25**	**4.38**	**3.89**	**3.58**	**3.35**	**3.18**	**3.04**	**2.94**	**2.86**	**2.78**	**2.72**	**2.62**	**2.54**	**2.43**	**2.35**	**2.26**	**2.17**	**2.12**	**2.04**	**2.00**	**1.94**	**1.90**	**1.87**
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.96	1.92	1.85	1.80	1.76	1.71	1.67	1.63	1.60	1.57	1.54	1.53
	7.35	**5.21**	**4.34**	**3.86**	**3.54**	**3.32**	**3.15**	**3.02**	**2.91**	**2.82**	**2.75**	**2.69**	**2.59**	**2.51**	**2.40**	**2.32**	**2.22**	**2.14**	**2.08**	**2.00**	**1.97**	**1.90**	**1.86**	**1.84**
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00	1.95	1.90	1.84	1.79	1.74	1.69	1.66	1.61	1.59	1.55	1.53	1.51
	7.31	**5.18**	**4.31**	**3.83**	**3.51**	**3.29**	**3.12**	**2.99**	**2.88**	**2.80**	**2.73**	**2.66**	**2.56**	**2.49**	**2.37**	**2.29**	**2.20**	**2.11**	**2.05**	**1.97**	**1.94**	**1.88**	**1.84**	**1.81**
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99	1.94	1.89	1.82	1.78	1.73	1.68	1.64	1.60	1.57	1.54	1.51	1.49
	7.27	**5.15**	**4.29**	**3.80**	**3.49**	**3.26**	**3.10**	**2.96**	**2.86**	**2.77**	**2.70**	**2.64**	**2.54**	**2.46**	**2.35**	**2.26**	**2.17**	**2.08**	**2.02**	**1.94**	**1.91**	**1.85**	**1.80**	**1.78**
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98	1.92	1.88	1.81	1.76	1.72	1.66	1.63	1.58	1.56	1.52	1.50	1.48
	7.24	**5.12**	**4.26**	**3.78**	**3.46**	**3.24**	**3.07**	**2.94**	**2.84**	**2.75**	**2.68**	**2.62**	**2.52**	**2.44**	**2.32**	**2.24**	**2.15**	**2.06**	**2.00**	**1.92**	**1.88**	**1.82**	**1.78**	**1.75**
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97	1.91	1.87	1.80	1.75	1.71	1.65	1.62	1.57	1.54	1.51	1.48	1.46
	7.21	**5.10**	**4.24**	**3.76**	**3.44**	**3.22**	**3.05**	**2.92**	**2.82**	**2.73**	**2.66**	**2.60**	**2.50**	**2.42**	**2.30**	**2.22**	**2.13**	**2.04**	**1.98**	**1.90**	**1.86**	**1.80**	**1.76**	**1.72**
48	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	1.99	1.96	1.90	1.86	1.79	1.74	1.70	1.64	1.61	1.56	1.53	1.50	1.47	1.45
	7.19	**5.08**	**4.22**	**3.74**	**3.42**	**3.20**	**3.04**	**2.90**	**2.80**	**2.71**	**2.64**	**2.58**	**2.48**	**2.40**	**2.28**	**2.20**	**2.11**	**2.02**	**1.96**	**1.88**	**1.84**	**1.78**	**1.73**	**1.70**
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95	1.90	1.85	1.78	1.74	1.69	1.63	1.60	1.55	1.52	1.48	1.46	1.44
	7.17	**5.06**	**4.20**	**3.72**	**3.41**	**3.18**	**3.02**	**2.88**	**2.78**	**2.70**	**2.62**	**2.56**	**2.46**	**2.39**	**2.26**	**2.18**	**2.10**	**2.00**	**1.94**	**1.86**	**1.82**	**1.76**	**1.71**	**1.68**
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.86	1.81	1.75	1.70	1.65	1.59	1.56	1.50	1.48	1.44	1.41	1.39
	7.08	**4.98**	**4.13**	**3.65**	**3.34**	**3.12**	**2.95**	**2.82**	**2.72**	**2.63**	**2.56**	**2.50**	**2.40**	**2.32**	**2.20**	**2.12**	**2.03**	**1.93**	**1.87**	**1.79**	**1.74**	**1.68**	**1.63**	**1.60**
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89	1.84	1.79	1.72	1.67	1.62	1.56	1.53	1.47	1.45	1.40	1.37	1.35
	7.01	**4.92**	**4.08**	**3.60**	**3.29**	**3.07**	**2.91**	**2.77**	**2.67**	**2.59**	**2.51**	**2.45**	**2.35**	**2.28**	**2.15**	**2.07**	**1.98**	**1.88**	**1.82**	**1.74**	**1.69**	**1.62**	**1.56**	**1.53**
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88	1.82	1.77	1.70	1.65	1.60	1.54	1.51	1.45	1.42	1.38	1.35	1.32
	6.96	**4.88**	**4.04**	**3.56**	**3.25**	**3.04**	**2.87**	**2.74**	**2.64**	**2.55**	**2.48**	**2.41**	**2.32**	**2.24**	**2.11**	**2.03**	**1.94**	**1.84**	**1.78**	**1.70**	**1.65**	**1.57**	**1.52**	**1.49**
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85	1.79	1.75	1.68	1.63	1.57	1.51	1.48	1.42	1.39	1.34	1.30	1.28
	6.90	**4.82**	**3.98**	**3.51**	**3.20**	**2.99**	**2.82**	**2.69**	**2.59**	**2.51**	**2.43**	**2.36**	**2.26**	**2.19**	**2.06**	**1.98**	**1.89**	**1.79**	**1.73**	**1.64**	**1.59**	**1.51**	**1.46**	**1.43**
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.87	1.84	1.78	1.73	1.66	1.61	1.56	1.50	1.46	1.39	1.37	1.32	1.28	1.25
	6.85	**4.79**	**3.95**	**3.48**	**3.17**	**2.96**	**2.79**	**2.66**	**2.56**	**2.47**	**2.40**	**2.34**	**2.23**	**2.15**	**2.03**	**1.95**	**1.86**	**1.76**	**1.70**	**1.61**	**1.56**	**1.48**	**1.42**	**1.38**
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75	1.69	1.64	1.57	1.52	1.46	1.40	1.35	1.28	1.24	1.17	1.11	1.00
	6.63	**4.60**	**3.78**	**3.32**	**3.02**	**2.80**	**2.64**	**2.51**	**2.41**	**2.32**	**2.24**	**2.18**	**2.07**	**1.99**	**1.87**	**1.79**	**1.69**	**1.59**	**1.52**	**1.41**	**1.36**	**1.25**	**1.15**	**1.00**

Note: ν_n = *degrees of freedom for numerator;* ν_d = *degrees of freedom for denominator.*

Source: Adapted from G. W. Snedecor and W. G. Cochran, *Statistical Methods*, Iowa State University Press, Ames, 1978, pp. 560–563.

imental groups, this is known as a *single-factor* or *one-way analysis of variance.* Other forms of analysis of variance (not discussed here) can be used to analyze experiments in which there is more than one experimental factor.

Since the distribution of possible F values depends on the size of each sample and number of samples under consideration, so does the exact value of F which corresponds to the 5 percent cutoff point. For example, in our diet study, the number of samples was 4 and the size of each sample was 7. This dependence enters into the mathematical formulas used to determine the value at which F gets "big" as two parameters known as *degree-of-freedom* parameters, often denoted ν (Greek nu). For this analysis, the between-groups degrees of freedom (also called the numerator degrees of freedom because the between-groups variance is in the numerator of F) is defined to be the number of samples m minus 1, or $\nu_n = m - 1$. The within-groups (or denominator) degrees of freedom is defined to be the number of samples times 1 less than the size of each sample, $\nu_d = m(n - 1)$. For our diet example, the numerator degrees of freedom are $4 - 1 = 3$, and the denominator degrees of freedom are $4(7 - 1) = 24$. Degrees of freedom often confuse and mystify people who are trying to work with statistics. They simply represent the way *number of samples* and *sample size* enter the mathematical formulas used to construct all statistical tables.

THREE EXAMPLES

We now have the tools needed to form conclusions using statistical reasoning. We will examine examples, all based on results published in the medical literature. I have exercised some literary license with these examples for two reasons: (1) Medical and scientific authors usually summarize their raw data with descriptive statistics (like those developed in Chap. 2) rather than including the raw data. As a result, the "data from the literature" shown in this chapter—and the rest of the book—are usually my guess at what the raw data probably looked like based on the descriptive statistics in the original article.* (2) The analysis of variance as we developed it requires that each sample con-

*Since authors often failed to include a complete set of descriptive statistics, I had to simulate them from the results of their hypothesis tests.

tain the same number of members. This is often not the case in reality, so I adjusted the sample sizes in the original studies to meet this restriction. We later generalize our statistical methods to handle experiments with different numbers of individuals in each sample or treatment group.

Glucose Levels in Children of Parents with Diabetes

Diabetes is a disease caused by abnormal carbohydrate metabolism and is characterized by excessive amounts of sugar in the blood and urine. Type I, or insulin-dependent diabetes mellitus (IDDM), occurs in children and young adults. Type II, or non-insulin-dependent diabetes mellitus (NIDDM), usually occurs in people over 40 years old and is often detected by elevated glucose levels rather than illness. Both types of diabetes tend to run in families, but because type II tends to occur in adults, few studies focus on determining when abnormalities in sugar regulation first appear in children and young adults. Gerald Berenson and colleagues* wanted to investigate whether abnormalities in carbohydrate metabolism could be detected in non-diabetic young adults whose parents had a history of type II diabetes.

They identified parents who had a history of diabetes in Bogalusa, Louisiana in 1987 and 1988 from a survey of school age children. Next, in 1989 and 1991, they recruited children from these families, which they denoted the *cases*. Similarly aged offspring were recruited from families without a history of diabetes to serve as *controls*. Berenson and colleagues then measured many physiological variables that might be related to diabetes, including indicators of carbohydrate tolerance (fasting glucose, insulin, glucagon), blood pressure, cholesterol, weight, and body mass index.

This approach is called an *observational study* because the investigator obtains data by simply observing events without controlling them. Such studies are prone to two potentially serious problems. First, the groups may vary in ways the investigators do not notice or choose to ignore, and these differences, rather than the treatment itself, may

*G. S. Berenson, W. Bao, S. R. Srinivasan, "Abnormal Characteristics in Young Offspring of Parents with Non-Insulin-Dependent Diabetes Mellitus." *Am. J. Epidemiol.*, **144**:962–967, 1996.

account for the differences the investigators find. These effects are called *confounding factors*. For example, smokers are more likely to develop lung cancer than nonsmokers who appear similar in all other aspects. Most people interpret this statistically demonstrated relationship as proving that smoking causes lung cancer. The tobacco industry's "scientific consultants" argue that people with a genetic predisposition to lung cancer also have a genetic predisposition to smoking cigarettes. Nothing in such an observational study alone can definitely prove that this is not the case; one has to apply other information to dispose of this possibility (such as the fact that smoke contains cancer-causing chemicals). Second, such studies can be subject to bias in patient recall, investigator assessment, and selection of the treatment group or the control group.

Observational studies do, however, have advantages. First, they are relatively inexpensive because they are often based on reviews of existing information or information that is already being collected for other purposes (like medical records) and because they generally do not require direct intervention by the investigator. Second, ethical considerations or prevailing medical practice can make it impossible to carry out active manipulation of the variable under study.

Because of the potential difficulties in all observational studies, it is critical that the investigators explicitly specify the criteria they used for classifying each subject in the control or case group. Such specifications help minimize biases when the study is done as well as help you, as the consumer of the resulting information, judge whether the classification rules made sense.

Berenson and colleagues developed explicit criteria for including a person in their study, including the following.

- *Parental history of diabetes was verified by a physician through medical records to exclude possible type I diabetics.*
- *No child had parents who were both diabetic.*
- *The cases and control groups had similar ages (15.3 ± 4.5 SD and 15.1 ± 5.7 SD).*
- *All parents were white.*
- *Control offspring were matched according to age of parents from families with no history of diabetes in parents, grandparents, uncles, or aunts.*

A comparison of controls and cases found that the prevalence of potentially confounding lifestyle factors, such as smoking, drinking alcohol, and using oral contraceptives was similar between the two groups.

Figure 3-7 shows data for fasting glucose levels in the 25 offspring of parents with type II diabetes and 25 control offspring. On the average, the offspring of parents with type II diabetes had fasting glucose levels of 86.1 mg/dL, whereas the control offspring had glucose levels of 82.2 mg/dL. There was not much variability in either group's fasting glucose levels. The standard deviations in glucose levels were 2.09 and 2.49 mg/dL, respectively.

How consistent are these data with the null hypothesis that fasting glucose levels do not differ in children of parents with type II diabetes compared to children of parents without a history of type II diabetes? In

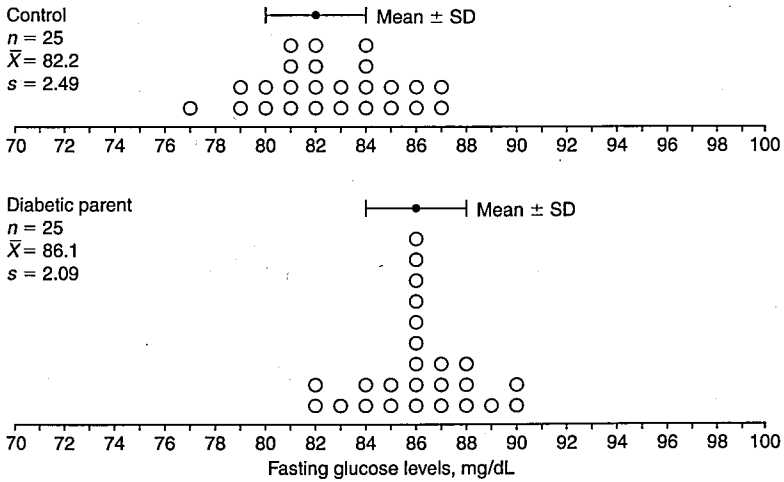


Figure 3-7 Results of a study comparing fasting glucose levels in children of parents who have type II diabetes and children of parents without type II diabetes. Each child's fasting glucose level is indicated by a circle at the appropriate glucose level. The average fasting glucose level for children of diabetic parents is higher than the average glucose level in children whose parents did not have diabetes. The statistical question is to assess whether or not the difference is due simply to random sampling or to an actual effect of the differences in parental history.

other words, how likely are the differences between the two samples of offspring shown in Fig. 3-7 to be due to random sampling rather than the difference based on parental history of diabetes?

To answer this question, we perform an analysis of variance.

We begin by estimating the within-groups variance by averaging the variances of the two groups of children.

$$s^2_{\text{wit}} = \tfrac{1}{2}(s^2_{\text{diabetes}} + s^2_{\text{control}})$$
$$= \tfrac{1}{2}(2.09^2 + 2.49^2) = 5.28\ (\text{mg/dL})^2$$

We then go on to calculate the between-groups variance. The first step is to estimate the standard error of the mean by computing the standard deviation of the two sample means. The mean of the two sample means is

$$\overline{X} = \tfrac{1}{2}(\overline{X}_{\text{diabetes}} + \overline{X}_{\text{control}})$$
$$= \tfrac{1}{2}(86.1 + 82.2) = 84.2\ \text{mg/dL}$$

Therefore the standard deviation of the sample means is

$$s_{\overline{X}} = \sqrt{\frac{(\overline{X}_{\text{diabetes}} - \overline{X})^2 + (\overline{X}_{\text{control}} - \overline{X})^2}{m - 1}}$$
$$= \sqrt{\frac{(86.1 - 84.2)^2 + (82.2 - 84.2)^2}{2 - 1}} = 2.76\ \text{mg/dL}$$

Since the sample size n is 25, the estimation of the population variance from between the groups is

$$s^2_{\text{bet}} = n s^2_{\overline{X}} = 25(2.76^2) = 190.13\ (\text{mg/dL})^2$$

Finally, the ratio of these two different estimates of the population variance is

$$F = \frac{s^2_{\text{bet}}}{s^2_{\text{wit}}} = \frac{190.13}{5.28} = 36.01$$

The degrees of freedom for the numerator are the number of groups minus 1, and so $\nu_n = 2 - 1 = 1$, and the degrees of freedom for the denominator are the number of groups times 1 less than the sample size, or $\nu_d = 2(25 - 1) = 48$. Look in the column headed 1 and the row headed 70 in Table 3-1. The resulting entry indicates that there is less than a 1 percent chance of F exceeding 7.19; we therefore conclude that the value of F associated with our observations is "big" and we reject the hypothesis that there is no difference in the average glucose level in the two groups of children shown in Fig. 3-7.

Hence, we reject the null hypothesis of no difference and conclude that there are higher levels of fasting glucose in children of diabetics than in children of nondiabetics.

Halothane versus Morphine for Open-Heart Surgery

Halothane is a popular drug to induce general anesthesia because it is potent, nonflammable, easy to use, and very safe. Since halothane can be carried with oxygen, it can be vaporized and administered to the patient with the same equipment used to ventilate the patient. The patient absorbs and releases it through the lungs, making it possible to change anesthetic states more rapidly than would be possible with drugs that have to be administered intravenously. It does, however, lessen the heart's ability to pump blood directly by depressing the myocardium itself and indirectly by increasing peripheral venous capacity. Some anesthesiologists believed that these effects could produce complications in people with cardiac problems and suggested using morphine as an anesthetic agent in these patients because it has little effect on cardiac performance in supine individuals. Conahan and colleagues* directly compared these two anesthetic agents in a large number of patients during routine surgery for cardiac valve repair or replacement.

To obtain two similar samples of patients who differed only in the type of anesthesia used, they selected the anesthesia at random for each patient who was suitable for the study.

*T. J. Conahan III, A. J. Ominsky, H. Wollman, and R. A. Stroth, "A Prospective Random Comparison of Halothane and Morphine for Open-Heart Anesthesia: One Year's Experience," *Anesthesiology,* **38**:528–535, 1973.

This procedure, called a *randomized clinical trial,* is the method of choice for evaluating therapies because it avoids the selection biases that can creep into observational studies. The randomized clinical trial is an example of what statisticians call an *experimental study* because the investigator actively manipulates the treatment under study, making it possible to draw much stronger conclusions than are possible from observational studies about whether or not a treatment produced an effect. Experimental studies are the rule in the physical sciences and animal studies in the life sciences but are less common in studies involving human subjects. Randomization reduces biases that can appear in observational studies, and since all clinical trials are *prospective,* no one knows how things will turn out at the beginning. This fact also reduces the opportunity for bias. Perhaps for these reasons, randomized clinical trials often show therapies to be of little or no value, even when observational studies have suggested that they were efficacious.*

Why, then, are not all therapies subjected to randomized clinical trials? Once something has become part of generally accepted medical practice—even if it did so without any objective demonstration of its value—it is extremely difficult to convince patients and their physicians to participate in a study that requires withholding it from some of the patients. Second, randomized clinical trials are always prospective; a person recruited into the study must be followed for some time, often many years. People move, lose interest, or die for reasons unrelated to the study. Simply keeping track of people in a randomized clinical trial is often a major task.

To collect enough patients to have a meaningful sample, it is often necessary to have many groups at different institutions participating. While it is great fun for the people running the study, it is often just one more task for the people at the collaborating institutions. All these factors often combine to make randomized clinical trials expensive and difficult to execute. Nevertheless, when done, they provide the most

*For a readable and classic discussion of the place of randomized clinical trials in providing useful clinical knowledge, together with a sobering discussion of how little of commonly accepted medical practice has ever been actually shown to do any good, see A. K. Cochran, *Effectiveness and Efficiency: Random Reflections on Health Services,* Nuffield Provincial Hospitals Trust, London, 1972.

definitive answers to questions regarding the relative efficacy of different treatments.

Conahan and colleagues were not faced with many of these problems because they were studying the effects of different types of anesthesia on how a patient did during the operation and the immediate postoperative recovery in a single hospital. During the operation, they recorded many hemodynamic variables, such as blood pressures before induction of anesthesia, after anesthesia but before incision, and during other important periods during the operation. They also recorded information relating to length of stay in the postsurgical intensive care unit, total length of hospitalization, and any deaths that occurred during this period. We will analyze these latter data after we have developed the necessary statistical tools in Chap. 5. For now, we will focus on a representative pressure measurement, the lowest mean arterial blood pressure between the start of anesthesia and the time of incision. This variable is thought to be a good measure of depression of the cardiovascular system before any surgical stimulation occurs. Specifically, we will investigate the hypothesis that, on the average, there was no difference in patients anesthetized with halothane or morphine.

Figure 3-8 shows the lowest mean arterial blood pressure observed from the start of anesthesia until the time of incision for 122 patients, half of whom were anesthetized with each agent. Pressures were rounded to the nearest even number, and each patient's pressure is represented by a circle. On the average, patients anesthetized with halothane had pressures 6.3 mmHg below those anesthetized with morphine. There is quite a bit of overlap in the pressures observed in the two different groups because of biological variability in how different people respond to anesthesia. The standard deviations in pressures are 12.2 and 14.4 mmHg for the people anesthetized with halothane and morphine, respectively. Given this, is the 6.3 mmHg difference large enough to assert that halothane produced lower lowest mean arterial pressures?

To answer this question, we perform an analysis of variance exactly as we did to compare length of hospitalizations in the appropriately and inappropriately treated pyelonephritis patients. We estimate the within-groups variance by averaging the estimates of the variance obtained from the two samples:

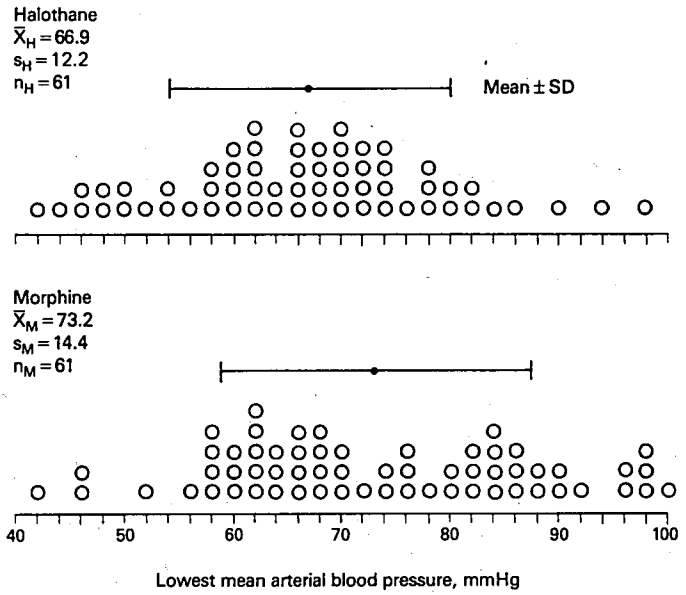


Figure 3-8 Lowest mean arterial blood pressure between the beginning of anesthesia and the incision in patients during open-heart surgery for patients anesthetized with halothane and morphine. Are the observed differences consistent with the hypothesis that, on the average, anesthetic did not affect blood pressure?

$$s^2_{wit} = \tfrac{1}{2}(s^2_{hlo} + s^2_{mor}) = \tfrac{1}{2}(12.2^2 + 14.4^2) = 178.1 \text{ mmHg}^2$$

Since this estimate of the population variance is computed from the variances of the separate samples, it does not depend on whether or not the means are different.

Next, we estimate the population variance by assuming that the hypothesis that halothane and morphine produce the same effect on arterial blood pressure is true. In that case, the two groups of patients in Fig. 3-8 are simply two random samples drawn from a single population. As a result, the standard deviation of the sample means is an estimate of the standard error of the mean. The mean of the two samples means is

$$\bar{\bar{X}} = \tfrac{1}{2}(\bar{X}_{\text{hlo}} + \bar{X}_{\text{mor}}) = \tfrac{1}{2}(66.9 + 73.2) = 70 \text{ mmHg}$$

The standard deviation of the $m = 2$ sample means is

$$s_{\bar{X}} = \sqrt{\frac{(\bar{X}_{\text{hlo}} - \bar{\bar{X}})^2 + (\bar{X}_{\text{mor}} - \bar{\bar{X}})^2}{m - 1}}$$

$$= \sqrt{\frac{(66.9 - 70.0)^2 + (73.2 - 70.0)^2}{2 - 1}} = 4.46 \text{ mmHg}$$

Since the sample size n is 61, the estimate of the population variance computed from the variability in the sample means is

$$s^2_{\text{bet}} = n s^2_{\bar{X}} = 61(4.46^2) = 1213 \text{ mmHg}^2$$

To test whether these two estimates are compatible, we compute

$$F = \frac{s^2_{\text{bet}}}{s^2_{wit}} = \frac{1213}{178.1} = 6.81$$

The degrees of freedom for the numerator are $\nu_n = m - 1 = 2 - 1 = 1$, and the degrees of freedom for the denominator are $\nu_d = m(n - 1) = 2(61 - 1) = 120$. Since $F = 6.81$ is greater than the critical value of 3.92 from (interpolating in) Table 3-1, we conclude that there is less than a 5 percent chance that our data were all drawn from a single population. In other words, we conclude that halothane produced lower lowest mean arterial blood pressures than morphine did, on the average.

Given the variability in response among patients to each drug (quantified by the standard deviations), do you expect this *statistically* significant result to be *clinically significant?* We will return to this question later.

Menstrual Dysfunction in Distance Runners

Infrequent or suspended menstruation can be a symptom of serious metabolic disorders, such as anorexia nervosa (a psychological disorder that leads people to stop eating, then waste away) or tumors of

the pituitary gland. Infrequent or suspended menstruation can also frustrate a woman's wish to have children. It can also be a side effect of birth control pills or indicate that a woman is pregnant or entering menopause. Gynecologists see many women who complain about irregular menstrual cycles and must decide how to diagnose and perhaps treat this possible problem. In addition to these potential explanations, there is some evidence that strenuous exercise, perhaps by changing the percentage of body fat, may affect the ovulation cycle. Since jogging and long-distance running have become popular, Edwin Dale and colleagues* decided to investigate whether there is a relationship between the frequency of menstruation and the amount of jogging young women do, as well as to look for possible effects on body weight, fat, and levels of circulating hormones that play an important role in the menstrual cycle.

They did an observational study of three groups of women. The first two groups were volunteers who regularly engaged in some form of running, and the third, a control group, consisted of women who did not run but were otherwise similar to the other two groups. The runners were divided into *joggers* who jog "slow and easy" 5 to 30 miles per week, and *runners* who run more than 30 miles per week and combine long, slow distance with speed work. The investigators used a survey to show that the three groups were similar in the amount of physical activity (aside from running), distribution of ages, heights, occupations, and type of birth control methods being used.

Figure 3-9 shows the number of menstrual periods per year for the 26 women in each experimental group. The women in the control group averaged 11.5 menses per year, the joggers averaged 10.1 menses per year, and the runners averaged 9.1 menses per year. Are these differences in mean number of menses compatible with what one would expect from the variability within each group?

To answer this question, we first estimate the population variance by averaging the variance from within the groups

*E. Dale, D. H. Gerlach, and A. L. Wilhite, "Menstrual Dysfunction in Distance Runners," *Obstet. Gynecol.*, **54**:47–53, 1979.

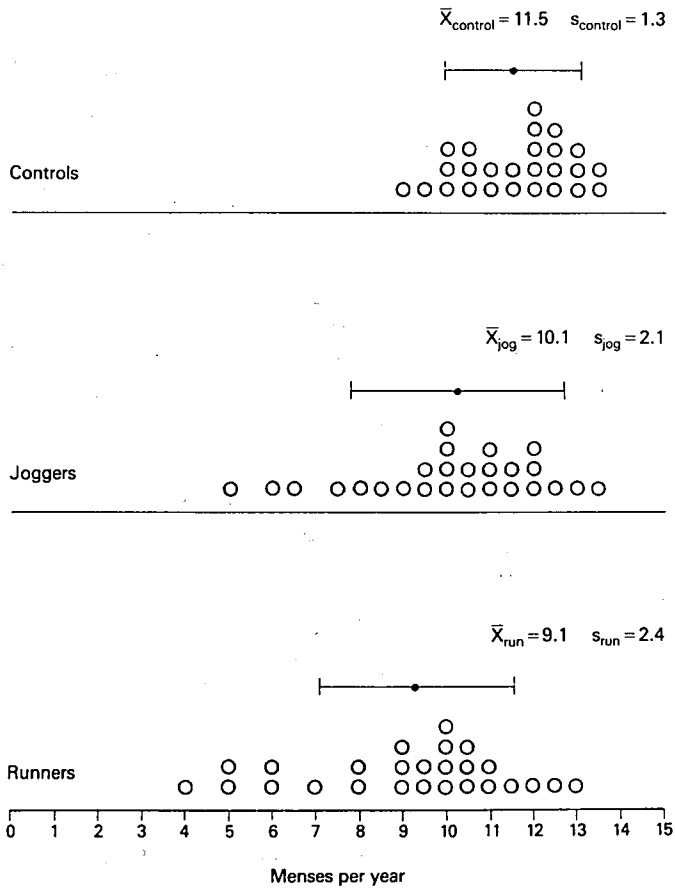


Figure 3-9 The number of menstrual cycles per year in women who were sedentary, joggers, and long-distance runners. The mean values of the three samples of women were different. Is this variation beyond what would be expected from random sampling, i.e., that the amount of running one does has no effect on the number of menstrual cycles, or is it compatible with the view that jogging affects menstruation? Furthermore, if there is an effect, is there a different effect for joggers and long-distance runners?

$$s_{\text{wit}}^2 = \tfrac{1}{3}(s_{\text{con}}^2 + s_{\text{jog}}^2 + s_{\text{run}}^2)$$
$$= \tfrac{1}{3}(1.3^2 + 2.1^2 + 2.4^2) = 3.95 \text{ (menses/year)}^2$$

To estimate the population variance from the variability in the sample means, we must first estimate the standard error of the mean by computing the standard deviation of the means of the three samples. Since the mean of the three means is

$$\overline{X} = \tfrac{1}{3}(\overline{X}_{\text{con}} + \overline{X}_{\text{jog}} + \overline{X}_{\text{run}})$$
$$= \tfrac{1}{3}(11.5 + 10.1 + 9.1) = 10.2 \text{ menses/year}$$

Our estimate of the standard error is

$$s_{\overline{X}} = \sqrt{\frac{(\overline{X}_{\text{con}} - \overline{X})^2 + (\overline{X}_{\text{jog}} - \overline{X})^2 + (\overline{X}_{\text{run}} - \overline{X})^2}{m - 1}}$$
$$= \sqrt{\frac{(11.5 - 10.2)^2 + (10.1 - 10.2)^2 + (9.1 - 10.2)^2}{3 - 1}}$$
$$= 1.2 \text{ menses/year}$$

The sample size n is 26, and so the estimate of the population variance from the variability in the means is

$$s_{\text{bet}}^2 = n s_{\overline{X}}^2 = 26(1.2^2) = 37.44 \text{ (menses/year)}^2$$

Finally,

$$F = \frac{s_{\text{bet}}^2}{s_{\text{wit}}^2} = \frac{37.44}{3.95} = 9.48$$

The numerator has $m - 1 = 3 - 1 = 2$ degrees of freedom and the denominator has $m(n - 1) = 3(26 - 1) = 75$ degrees of freedom. Interpolating in Table 3-1, we find that F will exceed 4.90 only 1 percent of the time when all the groups are drawn from a single population; we conclude that jogging or running has an effect on the frequency of menstruation.

When a woman comes to her gynecologist complaining about irregular or infrequent periods, the physician should not only look for biochemical abnormalities but also ask whether or not she jogs.

One question remains: Which of the three groups differed from the others? Does one have to be a marathon runner to expect menstrual dysfunction, or does it accompany less strenuous jogging? Or is the effect graded, becoming more pronounced with more strenuous exercise? We will have to defer answering these questions until we develop another statistical tool, the *t* test, in Chap. 4.

PROBLEMS

3-1 When labor has to be induced, the mother's cervix can fail to soften and enlarge, prolonging the labor and perhaps requiring delivery by cesarean section. To investigate whether the cervix can be softened and dilated by treating it with a gel containing prostaglandin E_2, C. O'Herlihy and H. MacDonald ("Influence of Preinduction Prostaglandin E_2 Vaginal Gel on Cervical Ripening and Labor," *Obstet. Gynecol.*, **54:**708–710, 1979) applied such a gel to the cervixes of 21 women who were having labor induced and a placebo gel that contained no active ingredients to 21 other women who were having labor induced. The two groups of women were of similar ages, heights, weeks of gestation, and initial extent of cervical dilation before applying the gel. The labor of women treated with prostaglandin E_2 averaged 8.5 h, and the labor of control women averaged 13.9 h. The standard deviations for these two groups were 4.7 and 4.1 h, respectively. Is there evidence that the prostaglandin gel shortens labor?

3-2 It is generally believed that infrequent and short-term exposure to pollutants in tobacco, such as carbon monoxide, nicotine, benzo[*a*]pyrene, and oxides of nitrogen, will not permanently alter lung function in healthy adult nonsmokers. To investigate this hypothesis, James White and Herman Froeb ("Small-Airways Dysfunction in Nonsmokers Chronically Exposed to Tobacco Smoke," *N. Engl. J. Med.*, **302:**720–723, 1980, used by permission) measured lung function in cigarette smokers and nonsmokers during a "physical fitness profile" at the University of California, San Diego. They measured how rapidly a person could force air from the lungs (mean forced midexpiratory flow). Reduced forced midexpiratory flow is associated with small-airways disease of the lungs. For the women they tested White and Froeb found:

Group	No. of subjects	Mean forced midexpiratory flow, L/s	
		Mean	SD
Nonsmokers			
Worked in clean environment	200	3.17	0.74
Worked in smoky environment	200	2.72	0.71
Light smokers	200	2.63	0.73
Moderate smokers	200	2.29	0.70
Heavy smokers	200	2.12	0.72

Is there evidence that the presence of small-airways disease as measured by this test, is any different among the different experimental groups?

3-3 Elevated levels of plasma high-density-lipoprotein (HDL) cholesterol may be associated with a lowered risk of coronary heart disease. Several studies have suggested that vigorous exercise may result in increased levels of HDL. To investigate whether or not jogging is associated with an increase in the plasma HDL concentration, G. Harley Hartung and colleagues ("Relation of Diet to High-Density-Lipoprotein Cholesterol in Middle-Aged Marathon Runners, Joggers, and Inactive Men," *N. Engl. J. Med.*, **302**:357–361, 1980, used by permission) measured HDL concentrations in middle-aged (35 to 66 years old) marathon runners, joggers, and inactive men. The mean HDL concentration observed in the inactive men was 43.3 mg/dL with a standard deviation of 14.2 mg/dL. The mean and standard deviation of the HDL concentration for the joggers and marathon runners were 58.0 and 17.7 mg/dL and 64.8 and 14.3 mg/dL, respectively. If there were 70 men in each group, test the hypothesis that there is no difference in the average HDL concentration between these groups of men.

3-4 If heart muscle is briefly deprived of oxygen—a condition known as ischemia—the muscle stops contracting and, if the ischemia is long enough or severe enough, the muscle dies. When the muscle dies, the person has a myocardial infarction (heart attack). Surprisingly, when the heart muscle is subjected to a brief period of ischemia before a major ischemic episode, the muscle is more able to survive the major ischemic episode. This phenomenon is known as ischemic preconditioning. This protective effect of ischemic preconditioning is known to involve activation of adenosine A_1 receptors, which stimulate protein kinase C (PKC), a

protein involved in many cellular processes including proliferation, migration, secretion, and cell death. Akihito Tsuchida and colleagues ("α_1-Adrenergic Agonist Precondition Rabbit Ischemic Myocardium Independent of Adenosine by Direct Activation of Protein Kinase C," *Circ. Res.*, **75**:576–585, 1994) hypothesized that α_1-adrenergic receptors might have an independent rule in this process. To address this question, Tsuchida and colleagues subjected isolated rabbit hearts to a brief 5-min ischemia or exposed the hearts to a variety of adenosine and α_1-adrenergic agonists and antagonists. In any case, following a 10-min recovery period, the heart was subject to ischemia for 30 min and the size of the resulting infarct measured. The control group was only subjected to 30 min of ischemia. If each group included 7 rabbit hearts, is there evidence that pretreatment with ischemia or a pharmacological agent affected infarct size, measured as the volume of heart muscle that dies?

	Infarct size, cm^3	
Group	Mean	SEM
Control	0.233	0.024
Ischemic preconditioning (PC)	0.069	0.015
α_1-Adrenergic receptor agonist (Phenylephrine)	0.065	0.008
adenosine receptor antagonist (8-p-[sulfophenyl] theophylline)	0.240	0.033
α_1-Adrenergic receptor antagonist (Phenoxybenzamine)	0.180	0.033
Protein kinase C inhibitor (Polymyxin B)	0.184	0.038

3-5 Men and women differ in risk of spinal fracture. Men are at increased risk for all types of bone fractures until approximately 45 years of age, an effect probably due to the higher overall trauma rate in men during this time. However, after age 45, women are at increased risk for spinal fracture, most likely due to age-related increases in osteoporosis, a disease characterized by decreased bone density. S. Kudlacek and colleagues ("Gender Differences in Fracture Risk and Bone Mineral Density," *Maturitas*, **36**:173–180, 2000) wanted to investigate the relationship between gender and bone density in a group of older adults who have had a vertebral bone fracture. Their data are presented below. Are there differences in vertebral bone density between similarly aged men and women who have had a vertebral bone fracture?

	Vertebral bone density (mg/cm^3)		
Group	*n*	Mean	SEM
Women with bone fractures	50	70.3	2.55
Men with bone fractures	50	76.2	3.11

3-6 Burnout is a term that loosely describes a condition of fatigue, frustration, and anger manifested as a lack of enthusiasm for and feeling of entrapment in one's job. It has been argued that professions such as teaching and nursing, which require high levels of commitment, are most subject to burnout. Moreover, since burnout is often linked to stress, it may be that nurses who specialize in so-called high-stress areas, such as intensive care units, experience more burnout and experience it sooner than nurses in less-stressful areas. Anne Keane and her associates ("Stress in ICU and Non-ICU Nurses," *Nurs. Res.*, **34**:231–236, 1985) studied several aspects of nursing burnout, including whether there was a difference in burnout between nurses who worked in intensive care units (ICUs), stepdown units (SDUs) or intermediate care units, and general medical units. They administered a questionnaire that could be used to construct a score called the Staff Burnout Scale for Health Professionals (SBS-HP), in which higher scores indicate more burnout. Do the data below suggest a difference in burnout among the units surveyed?

	ICU		SDU		General	
	Surgical	Medical	Surgical	Medical	Surgical	Medical
Mean score	49.9	51.2	57.3	46.4	43.9	65.2
Standard deviation	14.3	13.4	14.9	14.7	16.5	20.5
Sample size	16	16	16	16	16	16

3-7 High doses of estrogen interfere with male fertility in many animals, including mice. However, there may be significant differences in the response to estrogen in different mouse strains. To compare estrogen responsiveness in different strains of mice, Spearow and colleagues ("Genetic Variation in Susceptibility to Endocrine Disruption by Estro-

gen in Mice," Science, **285**:1259–1261, 1999) implanted capsules containing 1 μg of estrogen into four different strains of juvenile male mice. After 20 days, they measured their testicular weight. They found:

		Testes weight (mg)	
Mouse strain	*n*	Mean	SEM
CD-1	13	142	12
S15/JIs	16	82	3
C17/JIs	17	60	5
B6	15	38	1

Is there sufficient evidence to conclude that any of these strains differ in response to estrogen? (The formulas for analysis of variance with unequal sample sizes are in Appendix A.)

3-8 Several studies suggest that schizophrenic patients have lower IQ scores measured before the onset of schizophrenia (premorbid IQ) than would be expected based on family and environmental variables. These deficits can be detected during childhood and increase with age. Catherine Gilvarry and colleagues ("Premorbid IQ in Patients with Functional Psychosis and Their First-Degree Relatives," *Schizophr. Res.* **41**:417–429, 2000) investigated whether this was also the case with patients diag-

		NART score	
Group	*n*	Mean	SD
Controls	50	112.7	7.8
Psychotic patients (no obstetric complications)	28	111.6	10.3
Relatives of psychotic patients (no obstetric complications)	25	114.3	12.1
Psychotic patients with obstetric complications	13	110.4	10.1
Relatives of psychotic patients with obstetric complications	19	116.4	8.8

nosed with affective psychosis, which encompasses schizoaffective disorder, mania, and major depression. In addition, they also wanted to assess whether any IQ deficits could be detected in first-degree relatives (parents, siblings, and children) of patients with affective psychosis. They administered the National Adult Reading Test (NART), which is an indicator of premorbid IQ, to a set of patients with affective psychosis, their first-degree relatives, and a group of normal subjects without any psychiatric history. Gilvarry and colleagues also considered whether there was an obstetric complication (OC) during the birth of the psychotic patient, which is another risk factor for impaired intellectual development. Is there any evidence that NART scores differ among these groups of people? (The formulas for analysis of variance with unequal sample sizes are in Appendix A.)

Chapter 4

The Special Case of Two Groups: The *t* Test

As we have just seen in Chapter 3, many investigations require comparing only two groups. In addition, as the last example in Chapter 3 illustrated, when there are more than two groups, the analysis of variance allows you to conclude only that the data are not consistent with the hypothesis that all the samples were drawn from a single population. It does not help you decide which one or ones are most likely to differ from the others. To answer these questions, we now develop a procedure that is specifically designed to test for differences in two groups: the *t test* or *Student's t test.* While we will develop the *t* test from scratch, we will eventually show that it is just a different way of doing an analysis of variance. In particular, we will see that $F = t^2$ when there are two groups.

The *t* test is the most common statistical procedure in the medical literature; you can expect it to appear in more than half the papers you read in the general medical literature.* In addition to being used to

*A. R. Feinstein: "Clinical Biostatistics: A Survey of the Statistical Procedures in General Medical Journals," *Clin. Pharmacol. Ther.,* **15**:97–107, 1974.

compare two group means, it is widely applied incorrectly to compare multiple groups, by doing all the pairwise comparisons, for example, by comparing more than one intervention with a control condition or the state of a patient at different times following an intervention. Figure 4-1 shows the results of an analysis of the use of *t* tests for the clinical journal *Circulation;* 54 percent of all the papers used the *t* test, more often than not to analyze experiments for which it is not appropriate. As we will see, this incorrect use increases the chances of rejecting the hypothesis of no effect above the nominal level, say 5 percent, used to select the cutoff value for a "big" value of the test statistic *t*. In practical terms, this boils down to increasing the chances of reporting that some therapy had an effect when the evidence does not support this conclusion.

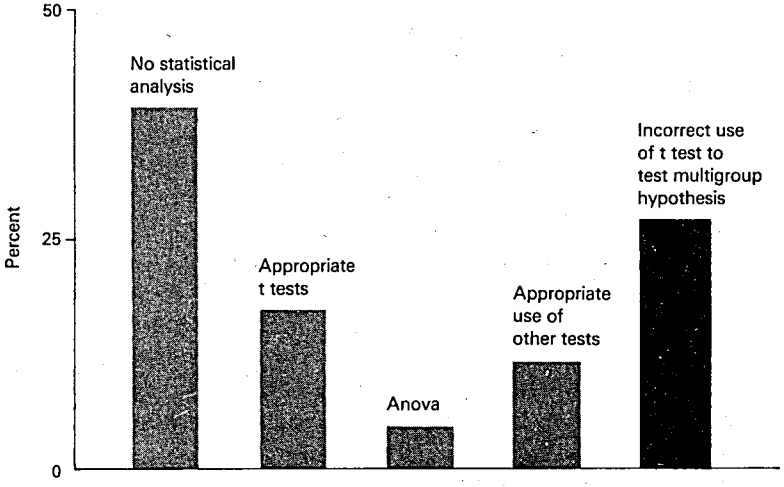


Figure 4-1 Of 142 original articles published in Vol. 56 of *Circulation* (excluding radiology, clinicopathologic, and case reports), 39 percent did not use statistics; 34 percent used a *t* test appropriately to compare two groups, analysis of variance (ANOVA), or other methods; and 27 percent used the *t* test incorrectly to compare more than two groups with each other. *(From S. A. Glantz, "How to Detect, Correct, and Prevent Errors in the Medical Literature," Circulation, 61:1–7, 1980. By permission of the American Heart Association, Inc.)*

THE GENERAL APPROACH

Suppose we wish to test a new drug that may be an effective diuretic. We assemble a group of 10 people and divide them at random into two groups, a control group that receives a placebo and a treatment group that receives the drug; then we measure their urine production for 24 h. Figure 4-2*A* shows the resulting data. The average urine production of the group receiving the diuretic is 240 mL higher than that of the group receiving the placebo. Simply looking at Fig. 4-2*A*, however, does not provide very convincing evidence that this difference is due to anything more than random sampling.

Nevertheless, we pursue the problem and give the placebo or drug to another 30 people to obtain the results shown in Fig. 4-2*B*. The mean responses of the two groups of people, as well as the standard deviations, are almost identical to those observed in the smaller samples shown in Fig. 4-2*A*. Even so, most observers are more confident in claiming that the diuretic increased average urine output from the data in Fig. 4-2*B* than the data in Fig. 4-2*A*, even though the samples in each case are good representatives of the underlying population. Why?

As the sample size increases, most observers become more confident in their estimates of the population means, so they can begin to discern a difference between the people taking the placebo or the drug. Recall that the standard error of the mean quantifies the uncertainty of the estimate of the true population mean based on a sample. Furthermore, as the sample size increases, the standard error of the mean decreases according to

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

where n is the sample size and σ is the standard deviation of the population from which the sample was drawn. As the sample size increases, the uncertainty in the estimate of the difference of the means between the people who received placebo and the patients who received the drug decreases relative to the difference of the means. As a result, we become more confident that the drug actually has an effect. More precisely, we become less confident in the hypothesis that the drug had no effect, in which case, the two samples of patients could be considered two samples drawn from a single population.

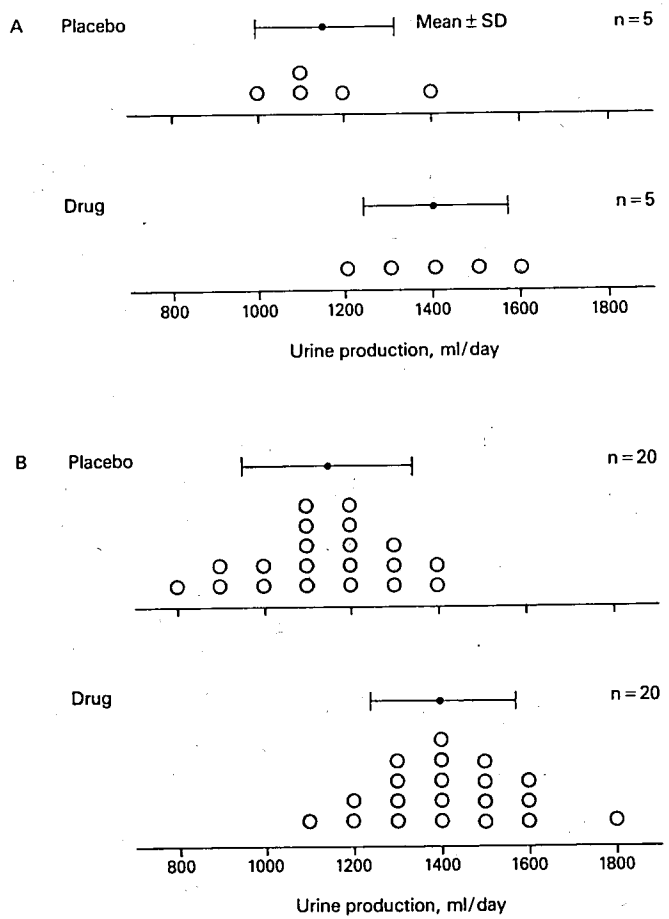


Figure 4-2 (*A*) Results of a study in which five people were treated with a placebo and five people were treated with a drug thought to increase daily urine production. On the average, the five people who received the drug produced more urine than the placebo group. Are these data convincing evidence that the drug is an effective diuretic? (*B*) Results of a similar study with 20 people in each treatment group. The means and standard deviations associated with the two groups are similar to the results in panel *A*. Are these data convincing evidence that the drug is an effective diuretic? If you changed your mind, why did you do it?

To formalize this logic, we will examine the ratio

$$t = \frac{\text{difference in sample means}}{\text{standard error of difference of sample means}}$$

When this ratio is small, we will conclude that the data are compatible with the hypothesis that both samples were drawn from a single population. When this ratio is large, we will conclude that it is unlikely that the samples were drawn from a single population and assert that the treatment (e.g., the diuretic) produced an effect.

This logic, while differing in emphasis from that used to develop the analysis of variance, is essentially the same. In both cases, we are comparing the relative magnitude of the differences in the sample means with the amount of variability that would be expected from looking within the samples.

To compute the *t* ratio we need to know two things: the difference of the sample means and the standard error of this difference. Computing the difference of the sample means is easy; we simply subtract. Computing an estimate for the standard error of this difference is a bit more involved. We begin with a slightly more general problem, that of finding the standard deviation of the difference of two numbers drawn at random from the same population.

THE STANDARD DEVIATION OF A DIFFERENCE OR A SUM

Figure 4-3*A* shows a population with 200 members. The mean is 0, and the standard deviation is 1. Now, suppose we draw two samples at random and compute their difference. Figure 4-3*B* shows this result for the two members indicated by solid circles in panel *A*. Drawing five more pairs of samples (indicated by different shadings in panel *A*) and computing their differences yields the corresponding shaded points in panel *B*. Note that there seems to be more variability in the differences of the samples than in the samples themselves. Figure 4-3*C* shows the results of panel *B*, together with the results of drawing another 50 pairs of numbers at random and computing their differences. The standard deviation of the population of differences is about 40 percent larger than the standard deviation of the population from which the samples were drawn.

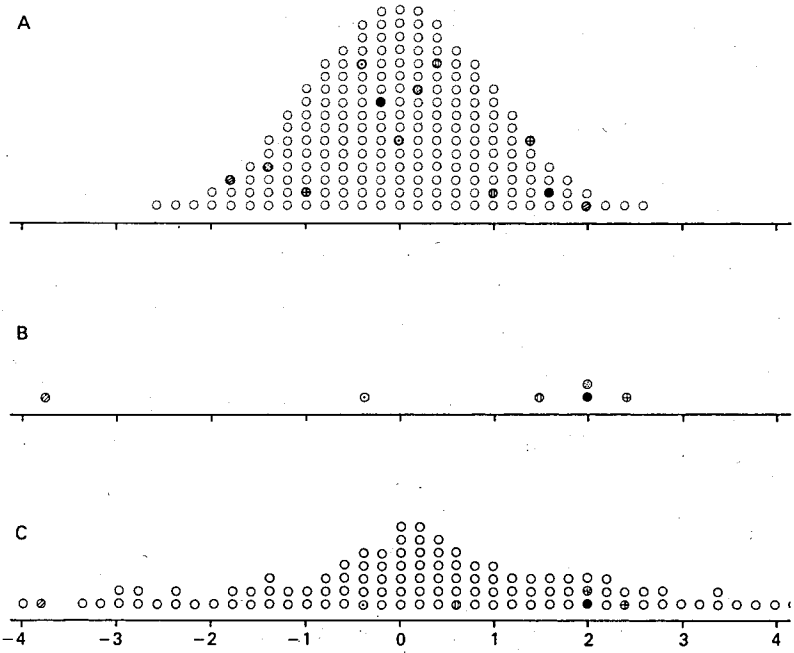


Figure 4-3 If one selects pairs of members of the population in panel *A* at random and computes the difference, the population of differences, shown in panel *B*, has a wider variance than the original population. Panel *C* shows another 100 values for differences of pairs of members selected at random from the population in *A* to make this point again.

In fact, it is possible to demonstrate mathematically that *the variance of the difference (or sum) of two variables selected at random equals the sum of the variances of the two populations from which the samples were drawn.* In other words, if X is drawn from a population with standard deviation σ_X and Y is drawn from a population with standard deviation σ_Y, the distribution of all possible values of $X - Y$ (or $X + Y$) will have variance

$$\sigma^2_{X-Y} = \sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$$

This result should seem reasonable to you because when you select pairs of values that are on opposite (the same) sides of the population

mean and compute their difference (sum), the result will be even farther from the mean.

Returning to the example in Fig. 4-3, we can observe that both the first and second numbers were drawn from the same population, whose variance was 1, and so the variance of the difference should be

$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y = 1 + 1 = 2$$

Since the standard deviation is the square root of the variance, the standard deviation of the population of differences will be $\sqrt{2}$ times the standard deviation of the original population, or about 40 percent bigger, confirming our earlier subjective impression.*

When we wish to estimate the variance in the difference or sum of members of two populations based on the observations, we simply replace the population variances σ^2 in the equation above with the

*The fact that the sum of randomly selected variables has a variance equal to the sum of the variances of the individual numbers explains why the standard error of the mean equals the standard deviation divided by $\sqrt{n}$. Suppose we draw n numbers at random from a population with standard deviation σ. The mean of these numbers will be

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + X_3 + \cdots + X_n)$$

so

$$n\overline{X} = X_1 + X_2 + X_3 + \cdots + X_n$$

Since the variance associated with each of the X_i's is σ^2, the variance of $n\overline{X}$ will be

$$\sigma^2_{n\overline{X}} = \sigma^2 + \sigma^2 + \sigma^2 + \cdots + \sigma^2 = n\sigma^2$$

and the standard deviation will be

$$\sigma_{n\overline{X}} = \sqrt{n}\sigma$$

But we want the standard deviation of $\overline{X}$, which is $n\overline{X}/n$, therefore

$$\sigma_{\overline{X}} = \sqrt{n}\sigma/n = \sigma/\sqrt{n}$$

which is the formula for the standard error of the mean. Note that we made no assumptions about the population from which the sample was drawn. (In particular, we did *not* assume that it had a normal distribution.)

estimates of the variances computed from our samples.

$$s^2_{X-Y} = s^2_X + s^2_Y$$

The standard error of the mean is just the standard deviation of the population of all possible sample means of samples of size n, and so we can find the standard error of the difference of two means using the equation above. Specifically,

$$s^2_{\bar{X}-\bar{Y}} = s^2_{\bar{X}} + s^2_{\bar{Y}}$$

in which case

$$s_{\bar{X}-\bar{Y}} = \sqrt{s^2_{\bar{X}} + s^2_{\bar{Y}}}$$

Now we are ready to construct the t ratio from the definition in the last section.

USE OF t TO TEST HYPOTHESES ABOUT TWO GROUPS

Recall that we decided to examine the ratio

$$t = \frac{\text{difference in sample means}}{\text{standard error of difference of sample means}}$$

We can now use the result of the last section to translate this definition into the equation

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s^2_{\bar{X}_1} + s^2_{\bar{X}_2}}}$$

Alternatively, we can write t in terms of the sample standard deviations rather than the standard errors of the mean

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(s^2_1/n) + (s^2_2/n)}}$$

in which n is the size of each sample.

If the hypothesis that the two samples were drawn from the same population is true, the variances s_1^2 and s_2^2 computed from the two samples are both estimates of the same population variance σ^2. Therefore, we replace the two different estimates of the population variance in the equation above with a single estimate, s^2, that is obtained by averaging these two separate estimates

$$s^2 = \tfrac{1}{2}(s_1^2 + s_2^2)$$

This is called the *pooled-variance estimate* since it is obtained by pooling the two estimates of the population variance to obtain a single estimate. The *t*-test statistic based on the pooled-variance estimate is

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{(s^2/n) + (s^2/n)}}$$

The specific value of t one obtains from any two samples depends not only on whether or not there actually is a difference in the means of the populations from which the samples were drawn but also on which specific individuals happened to be selected for the samples. Thus, as for F, there will be a range of possible values that t can have, even when both samples are drawn from the same population. Since the means computed from the two samples will generally be close to the mean of the population from which they were drawn, the value of t will tend to be small when the two samples are drawn from the same population. Therefore, we will use the same procedure to test hypotheses with t as we did with F in Chapter 3. Specifically, we will compute t from the data, then reject the assertion that the two samples were drawn from the same population if the resulting value of t is "big."

Let us return to the problem of assessing the value of the diuretic we were discussing earlier. Suppose the entire population of interest contains 200 people. In addition, we will assume that the diuretic had no effect, so that the two groups of people being studied can be considered to represent two samples drawn from a single population. Figure 4-4*A* shows this population, together with two samples of 10 people each selected at random for study. The people who received the placebo are shown as dark circles, and the people who received the diuretic are shown as lighter circles. The lower part of panel *A* shows the data as

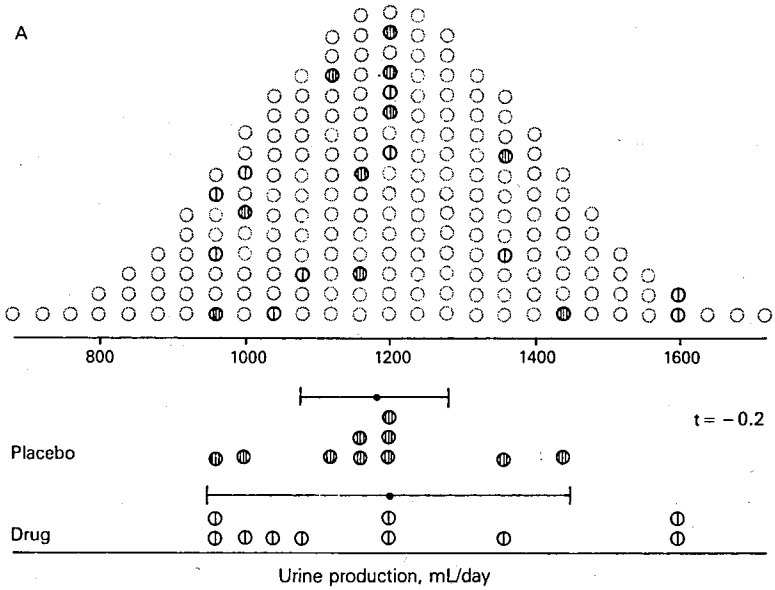


Figure 4-4 A population of 200 individuals and two groups selected at random for study of a drug designed to increase urine production but which is totally ineffective. The people shown as dark circles received the placebo and those with the lighter circles received the drug. An investigator would not see the entire population but just the information as reflected in the lower part of panel *A*; nevertheless, the two samples show very little difference, and it is unlikely that one would have concluded that the drug had an effect on urine production. Of course, there is nothing special about the two random samples shown in panel *A*, and an investigator could just as well have selected the two groups of people in panel *B* for study. There is more difference between these two groups than the two shown in panel *A*, and there is a chance that the investigator would think that this difference is due to the drug's effect on urine production rather than simple random sampling. Panel *C* shows yet another pair of random samples the investigator might have drawn for the study.

they would appear to the investigator, together with the mean and standard deviations computed from each of the two samples. Looking at these data certainly does not suggest that the diuretic had any effect. The value of t associated with these samples is -0.2.

Of course, there is nothing special about these two samples, and we could just as well have selected two different groups of people to study. Figure 4-4*B* shows another collection of people that could have

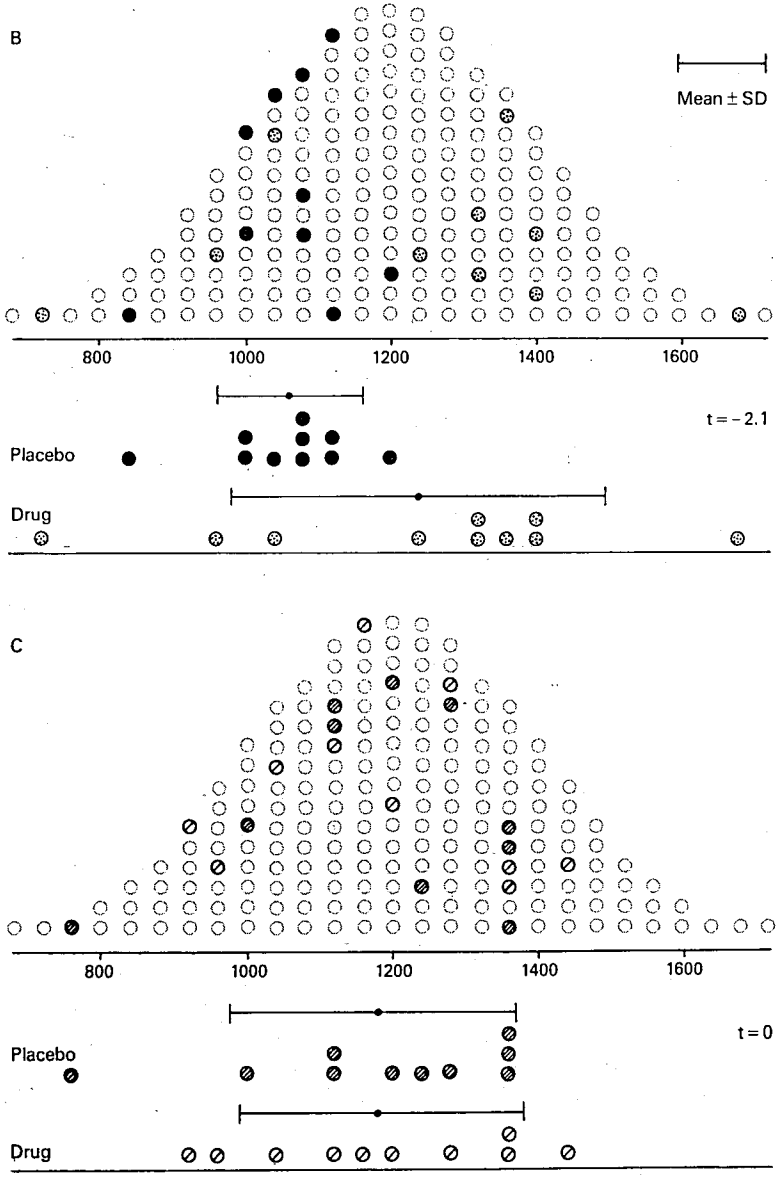


Figure 4-4 (*continued*)

been selected at random to receive the placebo (dark circles) or diuretic (light circles). Not surprisingly, these two samples differ from each other as well as the samples selected in panel *A*. Given only the data in the lower part of panel *B*, we might think that the diuretic increases urine production. The t value associated with these data is -2.1. Panel *C* shows yet another pair of samples. They differ from each other and the other samples considered in panels *A* and *B*. The samples in panel *C* yield a value of 0 for t.

We could continue this process for quite a long time since there are more than 10^{27} different pairs of samples of 10 people each that we could draw from the population of 200 individuals shown in Fig. 4-4*A*. We can compute a value of t for each of these 10^{27} different pairs of samples. Figure 4-5 shows the values of t associated with 200 different pairs of random samples of 10 people each drawn from the original population, including the three specific pairs of samples shown in Fig. 4-4. The distribution of possible t values is symmetrical about $t = 0$ because it does not matter which of the two samples we subtract from the other. As predicted, most of the resulting values of t are close to zero; t rarely is below about -2 or above $+2$.

Figure 4-5 allows us to determine what a big t is. Panel *B* shows that t will be less than -2.1 or greater than $+2.1$ 10 out of 200, or 5 percent of the time. In other words, there is only a 5 percent chance of getting a value of t more extreme than -2.1 or $+2.1$ when the two samples are drawn from the same population. Just as with the F distribution, the number of possible t values rapidly increases beyond 10^{27} as the population size grows, and the distribution of possible t values approaches a smooth curve. Figure 4-5*C* shows the result of this limiting process. We define the cutoff values for t that are large enough to be called "big" on the basis of the total area in the two tails. Panel *C* shows that only 5 percent of the possible values of t will lie beyond -2.1 or $+2.1$ when the two samples are drawn from a single population. When the data are associated with a value of t beyond this range, it is customary to conclude that the data are inconsistent with the hypothesis of no difference between the two samples and report that there was a difference in treatment.

The extreme values of t that lead us to reject the hypothesis of no difference lie in both tails of the distribution. Therefore, the approach we are taking is sometimes called a *two-tailed t test.*

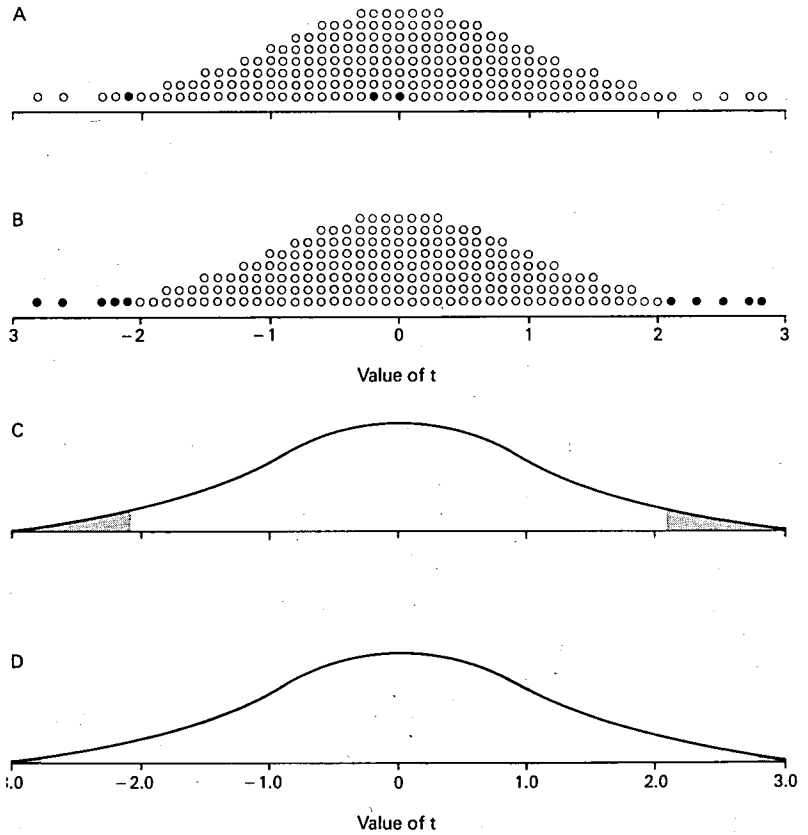


Figure 4-5 The results of 200 studies like that described in Fig. 4-4; the three specific studies from Fig. 4-4 are indicated in panel *A*. Note that most values of the *t* statistic cluster around 0, but it is possible for some values of *t* to be quite large, exceeding 1.5 or 2. Panel *B* shows that there are only 5 chances in 100 of *t* exceeding 2.1 in magnitude if the two samples were drawn from the same population. If one continues examining all possible samples drawn from the population and our pairs of samples drawn from the same population, one obtains a distribution of all possible *t* values which becomes the smooth curve in panel *C*. In this case, one defines the critical value of *t* by saying that it is unlikely that this value of *t* statistic was observed under the hypothesis that the drug had no effect by taking the 5 percent most extreme error areas under the tails of distribution and selecting the *t* value corresponding to the beginning of this region. Panel *D* shows that if one required a more stringent criterion for rejecting the hypothesis for no difference by requiring that *t* be in the most extreme 1 percent of all possible values, the cutoff value of *t* is 2.878.

Occasionally, people use a one-tailed t test, and there are indeed cases where this is appropriate. One should be suspicious of such one-tailed tests, however, because the cutoff value for calling t "big" for a given value of P is smaller. In reality, people are almost always looking for a *difference* between the control and treatment groups, and a two-tailed test is appropriate. This book always assumes a two-tailed test.

Note that the data in Fig. 4-4*B* are associated with a t value of -2.1, which we have decided to consider "big." If all we had were the data shown in Fig. 4-5*B*, we would conclude that the observations were inconsistent with the hypothesis that the diuretic had no effect and report that it *increased* urine production, and even though we did the statistical analysis correctly, *our conclusion about the drug would be wrong.*

Reporting $P < .05$ means that if the treatment had no effect, there is less than a 5 percent chance of getting a value of t from the data as far or farther from 0 as the critical value for t to be called "big." It does not mean it is impossible to get such a large value of t when the treatment has no effect. We could, of course, be more conservative and say that we will reject the hypothesis of no difference between the populations from which the samples were drawn if t is in the most extreme 1 percent of possible values. Figure 4-5*D* shows that this would require t to be beyond -2.88 or $+2.88$ in this case, so we would not erroneously conclude that the drug had an effect on urine output in any of the specific examples shown in Fig. 4-4. In the long run, however, we will make such errors about 1 percent of the time. The price of this conservatism is decreasing the chances of concluding that there is a difference when one really exists. Chapter 6 discusses this trade-off in more detail.

The critical values of t, like F, have been tabulated and depend not only on the level of confidence with which one rejects the hypothesis of no difference — the P value — but also on the sample size. As with the F distribution, this dependence on sample size enters the table as the *degrees of freedom* ν, which is equal to $2(n-1)$ for this t test, where n is the size of each sample. As the sample size increases, the value of t needed to reject the hypothesis of no difference decreases. In other words, as sample size increases, it becomes possible to detect smaller differences with a given level of confidence. Reflecting on Fig. 4-2 should convince you that this is reasonable.

WHAT IF THE TWO SAMPLES ARE NOT THE SAME SIZE?

It is easy to generalize the t test to handle problems in which there are different numbers of members in the two samples being studied. Recall that t is defined by

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s^2_{\overline{X}_1} + s^2_{\overline{X}_2}}}$$

in which $s_{\overline{X}_1}$ and $s_{\overline{X}_2}$ are the standard errors of the means of the two samples. If the first sample is of size n_1 and the second sample contains n_2 members,

$$s^2_{\overline{X}_1} = \frac{s_1^2}{n_1} \quad \text{and} \quad s^2_{\overline{X}_2} = \frac{s_2^2}{n_2}$$

in which s_1 and s_2 are the standard deviations of the two samples. Use these definitions to rewrite the definition of t in terms of the sample standard deviations

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$$

When the two samples are different sizes, the pooled estimate of the variance is given by

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

so that

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{(s^2/n_1) + (s^2/n_2)}}$$

This is the definition of t for comparing two samples of unequal size. There are $v = n_1 + n_2 - 2$ degrees of freedom.

Notice that this result reduces to our earlier results when the two sample sizes are equal, that is, when $n_1 = n_2 = n$.

THE EXAMPLES REVISITED

We can now use the t test to analyze the data from the examples we discussed to illustrate the analysis of variance. The conclusions will be no different from those obtained with analysis of variance because, as already stated, the t test is just a special case of analysis of variance.

Glucose Levels in Children of Parents with Diabetes

From Fig. 3-7, the 25 children of parents with type II diabetes had an average fasting glucose level of 86.1 mg/dL and the 25 children of parents without diabetes had an average fasting glucose level of 82.2 mg/dL. The standard deviations for both these groups were 2.09 and 2.49 mg/dL, respectively. Since the sample sizes are equal, the pooled estimate for the variance is $s^2 = {}^1/_2\,(2.09^2 + 2.49^2) = 5.28\ (\text{mg/dL})^2$.

$$t = \frac{86.1 - 82.2}{\sqrt{(5.28/25) + (5.28/25)}} = 6.001$$

with $\nu = 2(n - 1) = 2(25 - 1) = 48$. Table 4-1 shows that, for 48 degrees of freedom, the magnitude of t will exceed 2.011 only 5 percent of the time, and 2.682 only 1 percent of the time when the two samples are drawn from the same population. Since the magnitude of t associated with our data exceed 2.682, we conclude that the children of parents with type II diabetes have significantly higher fasting glucose levels than children of parents without type II diabetes ($P < .01$).

Halothane versus Morphine for Open-Heart Surgery

Figure 3-8 showed that the lowest mean arterial blood pressure between the start of anesthesia and the beginning of the incision was 66.9 mmHg in the 61 patients anesthetized with halothane and 73.2 in the 61 patients anesthetized with morphine. The standard deviations of the blood pressures in the two groups of patients was 12.2 and 14.4 mmHg, respectively. Thus,

$$s^2 = {}^1/_2(12.2^2 + 14.4^2) = 178.1\ \text{mmHg}^2$$

Table 4-1
Critical Values of t (Two-Tailed)

	Probability of greater value, P								
ν	0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
1	1.000	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
31	0.682	1.309	1.696	2.040	2.453	2.744	3.022	3.375	3.633
32	0.682	1.309	1.694	2.037	2.449	2.738	3.015	3.365	3.622
33	0.682	1.308	1.692	2.035	2.445	2.733	3.008	3.356	3.611
34	0.682	1.307	1.691	2.032	2.441	2.728	3.002	3.348	3.601
35	0.682	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591
36	0.681	1.306	1.688	2.028	2.434	2.719	2.990	3.333	3.582
37	0.681	1.305	1.687	2.026	2.431	2.715	2.985	3.326	3.574
38	0.681	1.304	1.686	2.024	2.429	2.712	2.980	3.319	3.566
39	0.681	1.304	1.685	2.023	2.426	2.708	2.976	3.313	3.558
40	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
42	0.680	1.302	1.682	2.018	2.418	2.698	2.963	3.296	3.538
44	0.680	1.301	1.680	2.015	2.414	2.692	2.956	3.286	3.526
46	0.680	1.300	1.679	2.013	2.410	2.687	2.949	3.277	3.515

Table 4-1
Critical Values of t (Two-Tailed) (*continued*)

	Probability of greater value, P								
ν	0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
48	0.680	1.299	1.677	2.011	2.407	2.682	2.943	3.269	3.505
50	0.679	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
52	0.679	1.298	1.675	2.007	2.400	2.674	2.932	3.255	3.488
54	0.679	1.297	1.674	2.005	2.397	2.670	2.927	3.248	3.480
56	0.679	1.297	1.673	2.003	2.395	2.667	2.923	3.242	3.473
58	0.679	1.296	1.672	2.002	2.392	2.663	2.918	3.237	3.466
60	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
62	0.678	1.295	1.670	1.999	2.388	2.657	2.911	3.227	3.454
64	0.678	1.295	1.669	1.998	2.386	2.655	2.908	3.223	3.449
66	0.678	1.295	1.668	1.997	2.384	2.652	2.904	3.218	3.444
68	0.678	1.294	1.668	1.995	2.382	2.650	2.902	3.214	3.439
70	0.678	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435
72	0.678	1.293	1.666	1.993	2.379	2.646	2.896	3.207	3.431
74	0.678	1.293	1.666	1.993	2.378	2.644	2.894	3.204	3.427
76	0.678	1.293	1.665	1.992	2.376	2.642	2.891	3.201	3.423
78	0.678	1.292	1.665	1.991	2.375	2.640	2.889	3.198	3.420
80	0.678	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
90	0.677	1.291	1.662	1.987	2.368	2.632	2.878	3.183	3.402
100	0.677	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
120	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
140	0.676	1.288	1.656	1.977	2.353	2.611	2.852	3.149	3.361
160	0.676	1.287	1.654	1.975	2.350	2.607	2.846	3.142	3.352
180	0.676	1.286	1.653	1.973	2.347	2.603	2.842	3.136	3.345
200	0.676	1.286	1.653	1.972	2.345	2.601	2.839	3.131	3.340
∞	0.6745	1.2816	1.6449	1.9600	2.3263	2.5758	2.8070	3.0902	3.2905

Source: Adapted from J. H. Zar, *Biostatistical Analysis* (2 ed.), Prentice-Hall, Englewood Cliffs, N.J., 1984, pp. 484–485, table B.3. Used by permission.

and

$$t = \frac{66.9 - 73.2}{\sqrt{(178.1/61) + (178.1/61)}} = -2.607$$

with $\nu = 2(n - 1) = 2(61 - 1) = 120$ degrees of freedom. Table 4-1 shows that the magnitude to t should exceed 2.358 only 2 percent of the time when the two samples are drawn from a single population, as they would be if halothane and morphine both affected patients' blood pressure similarly. Since the value of t associated with the

observations exceeds this value, we conclude that halothane is associated with a lower lowest mean arterial pressure than morphine, on the average.

Conahan and colleagues also measured the amount of blood being pumped by the heart in some of the patients they anesthetized to obtain another measure of how the two anesthetic agents affected cardiac function in people having heart-valve replacement surgery. To normalize the measurements to account for the fact that patients are different sizes and hence have different-sized hearts and blood flows, they computed the cardiac index, which is defined as the rate at which the heart pumps blood (the cardiac output) divided by body surface area. Table 4-2 reproduces some of their results. Morphine seems to produce lower cardiac indexes than halothane, but is this difference large enough to reject the hypothesis that the difference reflects random sampling rather than an actual physiological difference?

From the information in Table 4-2, the pooled estimate of the variance is

$$s^2 = \frac{(9 - 1)(1.05^2) + (16 - 1)(.88^2)}{9 + 16 - 2} = 0.89$$

Table 4-2 Comparison of Anesthetic Effects on the Cardiovascular system

	Halothane ($n = 9$)		Morphine ($n = 16$)	
	Mean	SD	Mean	SD
Best cardiac index, induction to bypass, L/m^2 · min	2.08	1.05	1.75	.88
Mean arterial blood pressure at time of best cardiac index, mmHg	76.8	13.8	91.4	19.6
Total peripheral resistance associated with best cardiac index, dyn · s/cm^5	2210	1200	2830	1130

Source: Adapted from T. J. Conahan et al., "A Prospective Random Comparison of Halothane and Morphine for Open-Heart Anesthesia," *Anesthesiology*, 38:528–535, 1973

and so

$$t = \frac{2.08 - 1.75}{\sqrt{(.89/9) + (.89/16)}} = 0.84$$

which does not exceed the 5 percent critical value of 2.069 for $\nu = n_{\text{hlo}} + n_{\text{mor}} - 2 = 9 + 16 - 2 = 23$ degrees of freedom. Hence, we do not have strong enough evidence to assert that there is really a difference in cardiac index with the two anesthetics. Does this prove that there really was not a difference?

THE *t* TEST IS AN ANALYSIS OF VARIANCE*

The *t* test and analysis of variance we developed in Chapter 3 are really two different ways of doing the same thing. Since few people recognize this, we will prove that when comparing the means of two groups, $F = t^2$. In other words, the *t* test is simply a special case of analysis of variance applied to two groups.

We begin with two samples, each of size n, with means and standard deviations $\overline{X}_1$ and $\overline{X}_2$ and s_1 and s_2, respectively.

To form the F ratio used in analysis of variance, we first estimate the population variance as the average of the variances computed for each group

$$s^2_{\text{wit}} = \tfrac{1}{2}(s_1^2 + s_2^2)$$

Next, we estimate the population variance from the sample means by computing the standard deviation of the sample means with

$$s_{\overline{X}} = \sqrt{\frac{(\overline{X}_1 - \overline{X})^2 + (\overline{X}_2 - \overline{X})^2}{2 - 1}}$$

Therefore

$$s^2_{\overline{X}} = (\overline{X}_1 - \overline{X})^2 + (\overline{X}_2 - \overline{X})^2$$

*This section represents the only mathematical proof in this book and as such is a bit more technical than everything else. The reader can skip this section with no loss of continuity.

in which $\overline{X}$ is the mean of the two sample means

$$\overline{X} = \tfrac{1}{2}(\overline{X}_1 + \overline{X}_2)$$

Eliminate $\overline{X}$ from the equation for $s^2_{\overline{X}}$ to obtain

$$\begin{aligned} s^2_{\overline{X}} &= [\overline{X}_1 - \tfrac{1}{2}(\overline{X}_1 + \overline{X}_2)]^2 + [\overline{X}_2 - \tfrac{1}{2}(\overline{X}_1 + \overline{X}_2)]^2 \\ &= (\tfrac{1}{2}\overline{X}_1 - \tfrac{1}{2}\overline{X}_2)^2 + (\tfrac{1}{2}\overline{X}_2 - \tfrac{1}{2}\overline{X}_1)^2 \end{aligned}$$

Since the square of a number is always positive, $(a - b)^2 = (b - a)^2$ and the equation above becomes

$$\begin{aligned} s^2_{\overline{X}} &= (\tfrac{1}{2}\overline{X}_1 - \tfrac{1}{2}\overline{X}_2)^2 + (\tfrac{1}{2}\overline{X}_1 - \tfrac{1}{2}\overline{X}_2)^2 \\ &= 2[\tfrac{1}{2}(\overline{X}_1 - \overline{X}_2)]^2 = \tfrac{1}{2}(\overline{X}_1 - \overline{X}_2)^2 \end{aligned}$$

Therefore, the estimate of the population variance from between the groups is

$$s^2_{\text{bet}} = n s^2_{\overline{X}} = (n/2)(\overline{X}_1 - \overline{X}_2)^2$$

Finally, F is the ratio of these two estimates of the population variance

$$\begin{aligned} F &= \frac{s^2_{\text{bet}}}{s^2_{\text{wit}}} = \frac{(n/2)(\overline{X}_1 - \overline{X}_2)^2}{\tfrac{1}{2}(s_1^2 + s_2^2)} = \frac{(\overline{X}_1 - \overline{X}_2)^2}{(s_1^2/n) + (s_2^2/n)} \\ &= \left[\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{(s_1^2/n) + (s_2^2/n)}}\right]^2 \end{aligned}$$

The quantity in the brackets is t, hence

$$F = t^2$$

The degrees of freedom for the numerator of F equals the number of groups minus 1, that is, $2 - 1 = 1$ for all comparisons of two groups. The degrees of freedom for the denominator equals the number of groups times 1 less than the sample size of each group, $2(n - 1)$, which is the same as the degrees of freedom associated with the t test.

In sum, the t test and analysis of variance are just two different ways of looking at the same test for two groups. Of course, if there are more than two groups, one cannot use the t-test form of analysis of variance but must use the more general form we developed in Chapter 3.

COMMON ERRORS IN THE USE OF THE t TEST AND HOW TO COMPENSATE FOR THEM

The t test is used to compute the probability of being wrong, the P value, when asserting that the mean values of *two* treatment groups are different, when, in fact, they were drawn from the same population. We have already seen (Fig. 4-1) that it is also used widely but erroneously to test for differences between more than two groups by comparing all possible pairs of means with t tests.

For example, suppose an investigator measured blood sugar under control conditions, in the presence of drug A, and in the presence of drug B. It is common to perform three t tests on these data: one to compare control versus drug A, one to compare control versus drug B, and one to compare drug A versus drug B. This practice is incorrect because the true probability of erroneously concluding that the drug affected blood sugar is actually higher than the nominal level, say 5 percent, used when looking up the "big" cutoff value of the t statistic in a table.

To understand why, reconsider the experiment described in the last paragraph. Suppose that if the value of the t statistic computed in one of the three comparisons just described is in the most extreme 5 percent of the values that would occur if the drugs really had no effect, we will reject that assumption and assert that the drugs changed blood sugar. We will be satisfied if $P < .05$; in other words, in the long run we are willing to accept the fact that 1 statement in 20 will be wrong. Therefore, when we test control versus drug A, we can expect erroneously to assert a difference 5 percent of the time. Similarly, when testing control versus drug B, we expect erroneously to assert a difference 5 percent of the time, and when testing drug A versus drug B, we expect erroneously to assert a difference 5 percent of the time. Therefore, when considering the three tests together, we expect to conclude that at least one pair of groups differs about 5 percent + 5 percent + 5 percent = 15 percent of the time, even if in reality the drugs did not

affect blood sugar (*P* actually equals 14 percent). If there are not too many comparisons, simply adding the *P* values obtained in multiple tests produces a realistic and conservative estimate of the true *P* value for the set of comparisons.

In the example above, there were three *t* tests, so the effective *P* value was about $3(.05) = .15$, or 15 percent. When comparing four groups, there are six possible *t* tests (1 versus 2, 1 versus 3, 1 versus 4, 2 versus 3, 2 versus 4, 3 versus 4); so if the author concludes that there is a difference and reports $P < .05$, the effective *P* value is about $6(.05) = .30$; there is about a 30 percent chance of at least one incorrect statement if the author concludes that the treatments had an effect!

In Chapter 2, we discussed random samples of Martians to illustrate the fact that different samples from the same population yield different estimates of the population mean and standard deviation. Figure 2-6 showed three such samples of the heights of Martians, all drawn from a single population. Suppose we chose to study how these Martians respond to human hormones. We draw three samples at random, give one group a placebo, one group testosterone, and one group estrogen. Suppose that these hormones have no effect on the Martians' heights. Thus, the three groups shown in Fig. 2-6 represent three samples drawn at random from the same population.

Figure 4-6 shows how these data would probably appear in a typical medical journal. The large vertical bars denote the value of the mean responses, and the small vertical bars denote 1 standard error of the mean above or below the sample means. Showing 1 standard deviation would be the appropriate way to describe variability in the samples. Most authors would analyze these data by performing three *t* tests: placebo against testosterone, placebo against estrogen, and testosterone against estrogen. These three tests yield *t* values of 2.39, 0.93, and 1.34, respectively. Since each test is based on 2 samples of 10 Martians each, there are $2(10 - 1) = 18$ degrees of freedom. From Table 4-1, the critical value of *t* with a 5 percent chance of erroneously concluding that a difference exists is 2.101. Thus, the author would conclude that testosterone produced shorter Martians than placebo, whereas estrogen did not differ significantly from placebo, and that the two hormones did not produce significantly different results.

Think about this result for a moment. What is wrong with it?

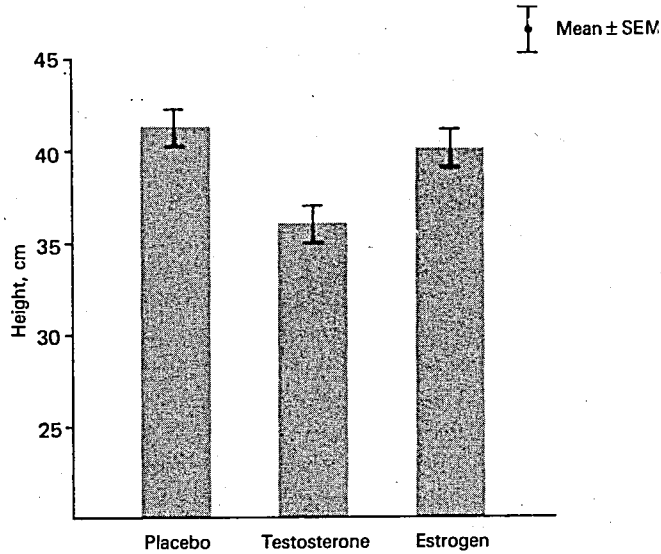


Figure 4-6 Results of a study of human hormones on Martians as it would be commonly presented in the medical literature. Each large bar has a height equal to the mean of the group; the small vertical bars indicate 1 standard error of the mean on either side of the mean (not 1 standard deviation).

If testosterone produced results not detectably different from those of estrogen and estrogen produced results not detectably different from those of placebo, how can testosterone have produced results different from placebo? Far from alerting medical researchers that there is something wrong with their analysis, this illogical result usually leads to a very creatively written "Discussion" section in their paper.

An analysis of variance of these data yields $F = 2.74$ [with numerator degrees of freedom $= m - 1 = 3 - 1 = 2$ and denominator degrees of freedom $m(n - 1) = 3(10 - 1) = 27$], which is below the critical value of 3.35 we have decided is required to assert that the data are incompatible with the hypothesis that all three treatments acted as placebos.

Of course, performing the analysis of variance does not ensure that we will not reach a conclusion that is actually wrong, but it will make it less likely.

We end our discussion of common errors in the use of the *t* test with three rules of thumb.

- *The* t *test can be used to test the hypothesis that two group means are not different.*
- *When the experimental design involves multiple groups, analysis of variance should be used.*
- *When* t *tests are used to test for differences between multiple groups, you can estimate the true* P *value by multiplying the reported* P *value times the number of possible* t *tests.*

HOW TO USE *t* TESTS TO ISOLATE DIFFERENCES BETWEEN GROUPS IN ANALYSIS OF VARIANCE

The last section demonstrated that when presented with data from experiments with more than two groups of subjects, one must do an analysis of variance to determine how inconsistent the observations are with the hypothesis that all the treatments had the same effect. Doing pairwise comparisons with *t* tests increases the chances of erroneously reporting an effect above the nominal value, say 5 percent, used to determine the value of a "big" *t*. The analysis of variance, however, only tests the global hypothesis that *all* the samples were drawn from a single population. In particular, it does not provide any information on which sample or samples differed from the others.

There are a variety of methods, called *multiple-comparison procedures,* that can be used to provide information on this point. All are essentially based on the *t* test but include appropriate corrections for the fact that we are comparing more than one pair of means. We will develop several approaches, beginning with the *Bonferroni t test.* The general approach we take is first to perform an analysis of variance to see whether *anything* appears different, then use a multiple-comparison procedure to isolate the treatment or treatments producing the different results.*

*Some statisticians believe that this approach is too conservative and that one should skip the analysis of variance and proceed directly to the multiple comparisons of interest. For an introductory treatment from this perspective, see Byron W. Brown, Jr., and Myles Hollander, *Statistics: A Biomedical Introduction,* Wiley, New York, 1977, chap. 10, "Analysis of *k*-Sample Problems."

The Bonferroni *t* Test

In the last section, we saw that if one analyzes a set of data with three t tests, each using the 5 percent critical value for concluding that there is a difference, there is about a 3(5) = 15 percent chance of finding it. This result is a special case of a formula called the *Bonferroni inequality,* which states that if k statistical tests are performed with the cutoff value for the test statistics, for example, t or F, at the α level, the likelihood of observing a value of the test statistic exceeding the cutoff value at least once when the treatments did not produce an effect is no greater than k times α. Mathematically, the Bonferroni inequality states

$$\alpha_T < k\alpha$$

where α_T is the true probability of erroneously concluding a difference exists at least once. α_T is the error rate we want to control. From the equation above,

$$\frac{\alpha_T}{k} < \alpha$$

Thus, if we do *each* of the t tests using the critical value of t corresponding to α_T/k, the error rate for *all* the comparisons taken as a group will be at most α_T. For example, if we wish to do three comparisons with t tests while keeping the probability of making at least one mistake to less than 5 percent, we use the t value corresponding to .05/3 = 1.6 percent for each of the individual comparisons. This procedure is called the Bonferroni t test because it is based on the Bonferroni inequality.

This procedure works reasonably well when there are only a few groups to compare, but as the number of comparisons k increases above 8 to 10, the value of t required to conclude that a difference exists becomes much larger than it really needs to be and the method becomes overconservative. Other multiple-comparison procedures, such as the Holm test (discussed in the next section), are less conservative. All, however, are similar to the Bonferroni t test in that they are essentially modifications of the t test to account for the fact that we are making multiple comparisons.

One way to make the Bonferroni t test less conservative is to use the estimate of the population variance computed from within the groups in the analysis of variance. Specifically, recall that we defined t as

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(s^2/n_1) + (s^2/n_2)}}$$

where s^2 is an estimate of the population variance. We will replace this estimate with the population variance estimated from within the groups as part of the analysis of variance, s^2_{wit}, to obtain

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(s^2_{\text{wit}}/n_1) + (s^2_{\text{wit}}/n_2)}}$$

When the sample sizes are equal, the equation becomes

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{2s^2_{\text{wit}}/n}}$$

The degrees of freedom for this test are the same as the denominator degrees of freedom for the analysis of variance and will be higher than for a simple t test based on the two samples being compared.* Since the critical value of t decreases as the degrees of freedom increase, it will be possible to detect a difference with a given confidence with smaller absolute differences in the means.

More on Menstruation and Jogging

In Chapter 3 we analyzed the data in Fig. 3-9 and concluded that they were inconsistent with the hypothesis that a control group, a group of joggers, and a group of runners had on the average the same number of menstrual periods per year. At the time, however, we were unable to isolate where the difference came from. Now we can use our Bonferroni t test to compare the three groups pairwise.

Recall that our best estimate of the within-groups variance s^2_{wit} is 3.95 (menses/year)^2. There are $m = 3$ samples, each containing $n = 26$ women. Therefore, there are $m(n - 1) = 3(26 - 1) = 75$ degrees of

*The number of degrees of freedom is the same if there are only two groups.

freedom associated with the estimate of the within-groups variance. [Note that if we just used the pooled variance from the two samples, there would be only $2(n-1) = 2(26-1) = 50$ degrees of freedom.] Therefore, we can compare the three different groups by computing three values of t. To compare control with the joggers, we compute

$$t = \frac{\overline{X}_{\text{jog}} - \overline{X}_{\text{con}}}{\sqrt{2s^2_{\text{wit}}/n}} = \frac{10.1 - 11.5}{\sqrt{2(3.95)/26}} = -2.54$$

To compare the control group with the runners, we compute

$$t = \frac{\overline{X}_{\text{run}} - \overline{X}_{\text{con}}}{\sqrt{2s^2_{\text{wit}}/n}} = \frac{9.1 - 11.5}{\sqrt{2(3.95)/26}} = -4.35$$

To compare the joggers with the runners, we compute

$$t = \frac{\overline{X}_{\text{jog}} - \overline{X}_{\text{run}}}{\sqrt{2s^2_{\text{wit}}/n}} = \frac{10.1 - 9.1}{\sqrt{2(3.95)/26}} = 1.81$$

There are three comparisons, so to have an overall error rate of less than 5 percent we compare each of these values of t with the critical value of t associated with the $.05/3 = 1.6$ percent level and 75 degrees of freedom. Interpolating* in Table 4-1 shows this value to be about 2.45.

Thus, we have sufficient evidence to conclude that both jogging and running decrease the frequency of menstruation, but we do not have evidence that running decreases menstruation any more than simply jogging.

A Better Approach to Multiple Comparisons: The Holm t Test

There have been several refinements of the Bonferroni t test designed to maintain the computational simplicity while avoiding the low power that the Bonferroni correction brings, beginning with the *Holm t test.*† The Holm test is nearly as easy to compute as the Bonferroni t test, but is more powerful.‡ The Holm test is a so-called sequentially

*Appendix A describes how to interpolate.

†S. Holm, "A Simple Sequentially Rejective Multiple Test Procedure," *Scand. J. Stat.*, **6**:65–70, 1979.

‡J. Ludbrook, "Multiple Comparison Procedures Updated," *Clin. Exp. Pharmacol. Physiol.*, **25**:1032–1037 1998; M. Aickin and H. Gensler, "Adjusting for Multiple Testing When Reporting Research Results: The Bonferroni vs. Holm Methods" *Am. J. Public Health,* **86**:726–728, 1996; B. Levin, "Annotation: On the Holm, Simes, and

rejective, or step-down, procedure because it applies an accept/reject criterion to a set of ordered null hypotheses, starting with the smallest P value, and proceeding until it fails to reject a null hypothesis.

To perform the Holm t test, we compute the family of pairwise comparisons of interest (with t test using the pooled variance estimate from the analysis of variance as we did with the Bonferroni t test) and determine the *unadjusted* P value for each test in the family. We then compare these P values (or the corresponding t values) to critical values that have been adjusted to allow for the fact that we are doing multiple comparisons. In contrast to the Bonferroni correction, however, we take into account how many tests we have already done and become less conservative with each subsequent comparison. We begin with a correction just as conservative as the Bonferroni correction, then take advantage of the conservatism of the earlier tests and become less cautious with each subsequent comparison.

Suppose we wish to make k pairwise comparisons.* Order these k uncorrected P values from smallest to largest, with the smallest uncorrected P value considered first in the sequential step-down test procedure. P_1 is the smallest P value in the sequence and P_k is the largest. For the j^{th} hypothesis test in this ordered sequence, Holm's original test applies the Bonferroni criterion in a step-down manner that depends on k and j, beginning with $j = 1$, and proceeding until we fail to reject the null hypothesis or run out of comparisons to do. Specifically, the uncorrected P value for the jth test is compared to $\alpha_j = \alpha_T/(k - j + 1)$. For the first comparison, $j = 1$, and the uncorrected P value needs to be smaller than $\alpha_1 = \alpha_T/(k - 1 + 1) = \alpha_T/k$, the same as the Bonferroni correction. If this smallest observed P value is less than α_1, we reject that null hypothesis and then compare the next smallest uncorrected P value with $\alpha_2 = \alpha_T/(k - 2 + 1) = \alpha_T/(k - 1)$, which is a larger cutoff than we would obtain just using the Bonferroni correction. Because this critical value is larger, the test is less conservative and has higher power.

In the example of the relationship between menstruation and jogging that we have been studying, the t values for control versus joggers,

---

Hochberg Multiple Test Procedures," *Am. J. Public Health,* **86**:628–629, 1996; B. W. Brown and K. Russel, "Methods for Correcting for Multiple Testing: Operating Characteristics," *Stat. Med.* **16**:2511–2528, 1997. T. Morikawa, A. Terao, and M. Iwasaki, "Power Evaluation of Various Modified Bonferroni Procedures by a Monte Carlo Study," *J. Biopharm. Stat.,* **6**:343–359, 1996.

*Like the Bonferroni correction, the Holm procedure can be applied to any family of hypothesis tests, not just multiple pairwise comparisons.

control versus runners, and joggers versus runners were −2.54, −4.35, and 1.81, respectively, each with 75 degrees of freedom. The corresponding uncorrected P values are .013, .001, and .074. The ordered P values, from smallest to largest are:

.001	.013	.074
control vs. runners	control vs. joggers	joggers vs. runners
$j = 1$	$j = 2$	$j = 3$

We have $k = 3$ null hypothesis tests of interest, which led to these three P values. The rejection criterion for the test of the first of these ordered hypotheses ($j = 1$) is $P \leq \alpha_1 = 0.05/(3 - 1 + 1) = 0.05/3 = .0167$, which is identical to the Bonferroni critical level we applied previously to each of the members of this family of three tests. The computed P, .001, is less than this critical α, and so we reject the null hypothesis that there is no difference between runners and controls. Because the null hypothesis was rejected at this step, we proceed to the next step, $j = 2$, using as a rejection criterion for this second test $P \leq \alpha_2 = 0.05/(3 - 2 + 1) = 0.05/2 = .025$. Note that this is a less restrictive criterion than in the traditional Bonferroni procedure we applied previously. The computed P, .013, is less than this critical value, and so we reject the null hypothesis that there is no difference in cortisol level between joggers and controls. Because the null hypothesis was rejected at this step, we proceed to the next and, in this example, final step, $j = 3$, using as a rejection criterion for this third test $P \leq \alpha_3 = 0.05/(3 - 3 + 1) = 0.05/1 = .05$. Note that this is further less restrictive than in the traditional Bonferroni procedure we applied previously, and is, in fact, equal to the criterion for an unadjusted t test. The computed P, .074, is greater than this critical value, and so we do not reject the null hypothesis that there is no difference in cortisol level between joggers and runners.

An alternative, and equivalent, approach is to compute the critical values of t corresponding to 0.0167, 0.025, and 0.05, and compare the observed values of the t test statistic for these comparisons with these critical t values. For 75 degrees of freedom, the corresponding critical values of the t test statistic are 2.45, 2.29, and 1.99. Therefore, like the Bonferroni t test, the Holm test requires the test statistic to exceed 2.45 for the comparison showing the largest difference (smallest P value), but this value drops to 2.29 for the second of the ordered comparisons,

and finally to 1.99 for the last of the three ordered comparisons (largest P value, corresponding to smallest mean difference).

In this example, we reached the same conclusion that we did when using the regular, single-step Bonferroni, procedure. However, you can see from the progressively less conservative α at each step in the sequence that this collection of tests will have more power than the traditional single-step Bonferroni procedure.* Because of the improved power of sequential stepping versions of the traditional Bonferroni tests while controlling the Type I error for the family of comparisons at the desired level, we recommend the Holm test over the Bonferroni test.†

OTHER APPROACHES TO MULTIPLE COMPARISON TESTING: THE STUDENT-NEWMAN-KEULS TEST‡

As noted in the previous section, the Bonferroni t test is overly conservative when there are more than a few group means to compare. This

*There are several other sequential tests, all of which operate in this manner, but apply different criteria. Some of these other tests are computationally more difficult, and less well understood, than Holm's test. Some, like Hochberg's test (Y. Hochberg, "A Sharper Bonferroni Procedure for Multiple Tests of Significance," *Biometrika* **75:**800–802, 1988), are step-up rather than step-down procedures in that they use a reverse stepping logic, starting with the kth (i.e., largest) P value in the ordered list of k P values, and proceeding until the first rejection of a null hypothesis, after which no more testing is done and all smaller P values are considered significant. Hochberg's test is exactly like the Holm test except that it applies the sequentially stepping Bonferroni criterion in this reverse, step-up order. Although Hochberg's test is claimed to be slightly more powerful than Holm's test, it is less well studied and so for the time being it is probably best to use one of the Holm tests, it is less well studied and so for the time being it is probably best to use one of the Holm tests (B. Levin, "Annotation: On the Holm, Simes, and Hochberg Multiple Test Procedures," *Am. J. Public Health,* **86:**628–629, 1996.

†As discussed earlier, the Bonferroni inequality, which forms the basis for Bonferroni corrections in multiple comparison procedures, is that $\alpha_T \leq k\,\alpha$. As noted when we introduced the topic of multiple comparisons, the Bonferroni inequality is more conservative compared with the exact risk of at least one false positive conclusion, which is $\alpha_T = 1 - (1 - \alpha)^k$. Sidak ("Rectangular Confidence Regions for the Means of Multivariate Normal Distributions," *J. Am. Stat. Assoc.,* **62:** 626–633, 1967) proposed adjusting the P values using this exact probability function rather than the inexact Bonferroni inequality. Likewise, the Holm procedure can be based on this more precise formula for the family error rate. This so-called *Holm–Sidak procedure,* works like the Holm test, except that the criterion is $1 - (1 - \alpha_T)^{1/(k - j + 1)}$ for the jth hypothesis test in the ordered sequence of k tests, rather than $\alpha_T/(k - j + 1)$.

‡This material is important for people who are using this book as a guide for analysis of their data; it can be skipped in a course on introductory biostatistics without interfering

continued

Table 4-3
Critical Values of q

	$\alpha_T = 0.05$								
ν_d	$p = 2$	3	4	5	6	7	8	9	10
1	17.97	26.98	32.82	37.08	40.41	43.12	45.40	47.36	49.07
2	6.085	8.331	9.798	10.88	11.74	12.44	13.03	13.54	13.99
3	4.501	5.910	6.825	7.502	8.037	8.478	8.853	9.177	9.462
4	3.927	5.040	5.757	6.287	6.707	7.053	7.347	7.602	7.826
5	3.635	4.602	5.218	5.673	6.033	6.330	6.582	6.802	6.995
6	3.461	4.339	4.896	5.305	5.628	5.895	6.122	6.319	6.493
7	3.344	4.165	4.681	5.060	5.359	5.606	5.815	5.998	6.158
8	3.261	4.041	4.529	4.886	5.167	5.399	5.597	5.767	5.918
9	3.199	3.949	4.415	4.756	5.024	5.244	5.432	5.595	5.739
10	3.151	3.877	4.327	4.654	4.912	5.124	5.305	5.461	5.599
11	3.113	3.820	4.256	4.574	4.823	5.028	5.202	5.353	5.487
12	3.082	3.773	4.199	4.508	4.751	4.950	5.119	5.265	5.395
13	3.055	3.735	4.151	4.453	4.690	4.885	5.049	5.192	5.318
14	3.033	3.702	4.111	4.407	4.639	4.829	4.990	5.131	5.254
15	3.014	3.674	4.076	4.367	4.595	4.782	4.940	5.077	5.198
16	2.998	3.649	4.046	4.333	4.557	4.741	4.897	5.031	5.150
17	2.984	3.628	4.020	4.303	4.524	4.705	4.858	4.991	5.108
18	2.971	3.609	3.997	4.277	4.495	4.673	4.824	4.956	5.071
19	2.960	3.593	3.977	4.253	4.469	4.645	4.794	4.924	5.038
20	2.950	3.578	3.958	4.232	4.445	4.620	4.768	4.896	5.008
24	2.919	3.532	3.901	4.166	4.373	4.541	4.684	4.807	4.915
30	2.888	3.486	3.845	4.102	4.302	4.464	4.602	4.720	4.824
40	2.858	3.442	3.791	4.039	4.232	4.389	4.521	4.635	4.735
60	2.829	3.399	3.737	3.977	4.163	4.314	4.441	4.550	4.646
120	2.800	3.356	3.685	3.917	4.096	4.241	4.363	4.468	4.560
∞	2.772	3.314	3.633	3.858	4.030	4.170	4.286	4.387	4.474

section presents the *Student-Newman-Keuls (SNK) test.* The SNK test statistic q is constructed similarly to the t test statistic, but the sampling distribution used to determine the critical values reflects a more sophisticated mathematical model of the multiple-comparison problem than does the simple Bonferroni inequality. This more sophisticated model gives

(continued) with the presentation of the rest of the material in this book. For a complete discussion of these (and other) multiple comparison procedures, see S. E. Maxwell and H. D. Delaney, *Designing Experiments and Analyzing Data: A Model Comparison Perspective,* Wadsworth, Belmont CA, 1990, chapter 5. "Testing Several Contrasts: The Multiple Comparison Problem."

Table 4-3
Critical Values of *q* (*Continued*)

$\alpha_T = 0.01$

ν_d	$p = 2$	3	4	5	6	7	8	9	10
1	90.03	135.0	164.3	185.6	202.2	215.8	227.2	237.0	245.6
2	14.04	19.02	22.29	24.72	26.63	28.20	29.53	30.68	31.69
3	8.261	10.62	12.17	13.33	14.24	15.00	15.64	16.20	16.69
4	6.512	8.120	9.173	9.958	10.58	11.10	11.55	11.93	12.27
5	5.702	6.976	7.804	8.421	8.913	9.321	9.669	9.972	10.24
6	5.243	6.331	7.033	7.556	7.973	8.318	8.613	8.869	9.097
7	4.949	5.919	6.543	7.005	7.373	7.679	7.939	8.166	8.368
8	4.746	5.635	6.204	6.625	6.960	7.237	7.474	7.681	7.863
9	4.596	5.428	5.957	6.348	6.658	6.915	7.134	7.325	7.495
10	4.482	5.270	5.769	6.136	6.428	6.669	6.875	7.055	7.213
11	4.392	5.146	5.621	5.970	6.247	6.476	6.672	6.842	6.992
12	4.320	5.046	5.502	5.836	6.101	6.321	6.507	6.670	6.814
13	4.260	4.964	5.404	5.727	5.981	6.192	6.372	6.528	6.667
14	4.210	4.895	5.322	5.634	5.881	6.085	6.258	6.409	6.543
15	4.168	4.836	5.252	5.556	5.796	5.994	6.162	6.309	6.439
16	4.131	4.786	5.192	5.489	5.722	5.915	6.079	6.222	6.349
17	4.099	4.742	5.140	5.430	5.659	5.847	6.007	6.147	6.270
18	4.071	4.703	5.094	5.379	5.603	5.788	5.944	6.081	6.201
19	4.046	4.670	5.054	5.334	5.554	5.735	5.889	6.022	6.141
20	4.024	4.639	5.018	5.294	5.510	5.688	5.839	5.970	6.087
24	3.956	4.546	4.907	5.168	5.374	5.542	5.685	5.809	5.919
30	3.889	4.455	4.799	5.048	5.242	5.401	5.536	5.653	5.756
40	3.825	4.367	4.696	4.931	5.114	5.265	5.392	5.502	5.559
60	3.762	4.282	4.595	4.818	4.991	5.133	5.253	5.356	5.447
120	3.702	4.200	4.497	4.709	4.872	5.005	5.118	5.214	5.299
∞	3.643	4.120	4.403	4.603	4.757	4.882	4.987	5.078	5.157

Source: H. L. Harter, *Order Statistics and Their Use in Testing and Estimation*, Vol. I: *Tests Based on Range and Studentized Range of Samples from a Normal Population*, U.S. Government Printing Office, Washington, D.C., 1970.

rise to a more realistic estimate of the total true probability of erroneously concluding a difference exists, α_T, than does the Bonferroni *t* test.

The first step in the analysis is to complete an analysis of variance on all the data to test the global hypothesis that all the samples were drawn from a single population. If this test yields a significant value of *F*, arrange all the means in ~~increasing~~ descending order and compute the SNK test

statistic q according to

$$q = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{\frac{s^2_{\text{wit}}}{2}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

where $\overline{X}_A$ and $\overline{X}_B$ are the two means being compared, s^2_{wit} is the variance within the treatment groups estimated from the analysis of variance, and n_A and n_B are the sample sizes of the two samples being compared.

This value of q is then compared with the table of critical values (Table 4-3). This critical value depends on α_T, the total risk of erroneously asserting a difference for all comparisons combined, ν_d, the denominator degrees of freedom from the analysis of variance, and a parameter p, which is the number of means being tested. For example, when comparing the largest and smallest of four means, $p = 4$; when comparing the second smallest and smallest means, $p = 2$.

The conclusions reached by multiple-comparisons testing depend on the order that the pairwise comparisons are made. The proper procedure is to compare first the largest mean with the smallest, then the largest with the second smallest, and so on, until the largest has been compared with the second largest. Next, compare the second largest with the smallest, the second largest with the next smallest, and so forth. For example, after ranking four means in ascending order, the sequence of comparisons should be: 4 versus 1, 4 versus 2, 4 versus 3, 3 versus 1, 3 versus 2, 2 versus 1.

Another important procedural rule is that if no significant difference exists between two means, then conclude that no difference exists between any means enclosed by the two without testing for them. Thus, in the preceding example, if we failed to find a significant difference between means 3 and 1, we would not test for a difference between means 3 and 2 or means 2 and 1.

Still More on Menstruation and Jogging

To illustrate the procedure, we once again analyze the data in Fig. 3-9, which presents the number of menses per year in women runners, joggers, and sedentary controls. The women in the control group had an average of 11.5 menses per year, the joggers had an average of 10.1

menses per year, and the runners had an average of 9.1 menses per year. We begin by ordering these means in descending order (which is how they happen to be listed). Next, we compute the change in means between the largest and smallest (control versus runners), the largest and the next smallest (control versus joggers), and the second largest and the smallest (joggers versus runners). Finally, we use the estimate of the variance from within the groups in the analysis of variance, $s^2_{wit} = 3.95$ (menses/year)^2 with $\nu_d = 75$ degrees of freedom, and the fact that each test group contained 26 women to complete the computation of each value of q.

To compare the controls with the runners, we compute

$$q = \frac{\overline{X}_{con} - \overline{X}_{run}}{\sqrt{\frac{s^2_{wit}}{2}\left(\frac{1}{n_{con}} + \frac{1}{n_{run}}\right)}} = \frac{11.5 - 9.1}{\sqrt{\frac{3.95}{2}\left(\frac{1}{26} + \frac{1}{26}\right)}} = 6.157$$

This comparison spans three means, so $p = 3$. From Table 4-3, the critical value of q for $\alpha_T = .05$, $\nu_d = 75$ (from the analysis of variance), and $p = 3$ is 3.385. Since the value of q associated with this comparison, 6.157, exceeds this critical value, we conclude that there is a significant difference between the controls and the runners. Since this result is significant, we go on to the next comparison.

To compare the controls with the joggers, we compute

$$q = \frac{\overline{X}_{con} - \overline{X}_{jog}}{\sqrt{\frac{s^2_{wit}}{2}\left(\frac{1}{n_{con}} + \frac{1}{n_{jog}}\right)}} = \frac{11.5 - 10.1}{\sqrt{\frac{3.95}{2}\left(\frac{1}{26} + \frac{1}{26}\right)}} = 3.592$$

For this comparison, α_T and ν_d are the same as before, but $p = 2$. From Table 4-3, the critical value of q is 2.822. The value of 3.592 associated with this comparison also exceeds the critical value, so we conclude that controls are also significantly different from joggers.

To compare the joggers with the runners, we compute

$$q = \frac{\overline{X}_{con} - \overline{X}_{run}}{\sqrt{\frac{s^2_{wit}}{2}\left(\frac{1}{n_{jog}} + \frac{1}{n_{run}}\right)}} = \frac{10.1 - 9.1}{\sqrt{\frac{3.95}{2}\left(\frac{1}{26} + \frac{1}{26}\right)}} = 2.566$$

The value of q associated with this comparison, 2.566, is less than the critical value of 2.822 required to assert that there is a difference between joggers and runners. (The values of v_d and p are the same as before, so the critical value of q is too.)

Tukey Test

The *Tukey test* is computed identically to the SNK test; the only difference is the critical value used to test whether a given difference is significant. (In fact, the SNK test is actually derived from the Tukey test.) In the SNK test, the value of the parameter p used to determine the critical value of q is the number of means spanned in the comparison being considered. As a result, completing a family of comparisons with the SNK test involves changing values of the critical value of q, depending on which comparison is being made. In the Tukey test, the parameter p is set to m, the number of groups in the study for all comparisons.

Had we used the Tukey test for multiple comparisons in the example of the effect of jogging on menstruation discussed previously, we would have used $m = 3$ for p, and so compared the observed values of q with a critical value of 3.385 for *all* the comparisons. Despite the fact that the critical value for the last two comparisons in the example would be larger than the critical values used in the SNK test, we would draw the same conclusions from the Tukey test as the SNK test; that is, the joggers and runners are not significantly different from each other, and both are significantly different from the control group.

The Tukey and SNK tests, however, do not always yield the same results. The Tukey test controls the error rate for *all* comparisons simultaneously, whereas the SNK test controls the error rate for all comparisons that involve spanning p means. As a result, the Tukey test is more conservative (i.e., less likely to declare a difference significant) than SNK. People who like the Tukey test use it because it controls the overall error rate for all multiple comparisons. People who like the SNK test observe that the test is done *after* doing an analysis of variance, and they depend on the analysis of variance to control the overall error rate. They argue that because the SNK test is done only after the analysis of variance finds a significant difference, they need not worry about excess false positives from the SNK test that are the price of the increased power. Some believe that the Tukey test is overly conservative because it requires that all groups be tested

as though they were separated by the maximum number of steps, whereas the SNK procedure allows each comparison to be made with reference to the exact number of steps that separate the two means actually being compared.

WHICH MULTIPLE COMPARISON PROCEDURE SHOULD YOU USE?

There is no strong consensus among statisticians about which multiple comparison test is preferred, and part of the choice is philosophic. For example, some would choose to be more conservative so that they only follow up avenues of inquiry that are more strongly suggested by their data.

Unadjusted *t* tests (also known as Fisher's Protected Least Significant Difference Test) are too liberal and the Bonferroni *t* test is too conservative for all possible comparisons. The SNK test tends to over-detect significant differences between means because it controls the error rate among all comparisons spanning a fixed number of means rather than all pairwise comparisons. The Tukey test tends to under-detect significant differences. Between these two tests, we prefer the SNK test for all pairwise comparisons as a good compromise between sensitivity and caution. The Tukey test is usually used by people who prefer to be more conservative. The Holm test is less conservative than Tukey or Bonferroni, while at the same time controlling the overall risk of a false-positive conclusion at the nominal level for the entire family of pairwise tests (not just all tests spanning a given number of means). We recommend the Holm test as the first line procedure for most multiple comparison testing.

MULTIPLE COMPARISONS AGAINST A SINGLE CONTROL*

In addition to all pairwise comparisons, the need sometimes arises to compare the values of multiple treatment groups to a single control group. One alternative would be to use Bonferroni *t*, SNK, or Tukey tests to do all pairwise comparisons, then only consider the ones that involve the control group. The problem with this approach is that it

*This material is important for people who are using this book as a guide for analysis of their data; it can be skipped in a course on introductory biostatistics without interfering with the presentation of the rest of the material in this book.

requires many more comparisons than are actually necessary, with the result that each individual comparison is done much more conservatively than is necessary based on the actual number of comparisons of interest. We now present three techniques specifically designed for the situation of multiple comparisons against a single control: additional *Bonferroni* 2nd Holm *t tests* and *Dunnett's test.* As with all pairwise multiple comparisons, use these tests *after* finding significant differences among all the groups with an analysis of variance.

Bonferroni *t* Test

The Bonferroni *t* test can be used for multiple comparisons against a single control group. The *t* test statistic and adjustment of the critical value to control the total error, α_T, proceed as before. The only difference is that the number of comparisons, k, is smaller because comparisons are being done against the control group only.

Suppose that we had only wanted to compare menstruation patterns in joggers and runners with the controls, but not with each other. Because we are only making comparisons against control, there are a total of $k = 2$ comparisons (as opposed to 3 when making all pairwise comparisons). To keep the total error rate at or below $\alpha_T = .05$ with these two comparisons, we do *each* of the *t* tests using the critical value of *t* corresponding to $\alpha_T/k = .05/2 = .025$. There are 75 degrees of freedom associated with the within-groups variance, so, interpolating* in Table 4-1, the critical value of *t* for each of the comparisons is 2.29. (This value compares with 2.45 for all possible comparisons. The lower critical value of *t* for comparisons against control means that it is easier to identify a difference against control than when making all possible comparisons.) From the previous section, the observed values of *t* for the comparisons of joggers with controls and runners with controls are -2.54 and -4.35, respectively. The magnitudes of both these values exceed the critical value of 2.29, so we conclude that both joggers and runners differ significantly from control. *No statement can be made about the comparison of joggers with runners.*

Holm *t* Test

Just as it is possible to use Bonferroni *t* tests for multiple comparisons against a single control group, it is possible to use Holm *t* tests. In

*Appendix A includes the formulas for interpolating.

the menstruation example, there are $k = 2$ comparisons, so the critical value of t for the first comparison is that corresponding to $\alpha_1 = \alpha_T/(k - j + 1) = .05/(2 - 1 + 1) = .025$, 2.29. From the previous section, the observed value of t for the comparison of runners with controls is -4.35, which exceeds the 2.29 critical value, so we reject the null hypothesis of no difference. For the second comparison, $\alpha_2 = \alpha_T/(k - j + 1) = .05/(2 - 2 + 1) = .05$, 1.99. The value of t for the comparison of joggers with controls is -2.54, which exceeds this value. Therefore, we again conclude that both joggers and runners are significantly different from controls.

Dunnett's Test

The analog of the SNK test for multiple comparisons against a single control group is *Dunnett's test.* Like the SNK test statistic, the Dunnett q' test statistic is defined analogously to the t test statistic:

$$q' = \frac{\overline{X}_{\text{con}} - \overline{X}_A}{\sqrt{s^2_{\text{wit}}\left(\frac{1}{n_{\text{con}}} + \frac{1}{n_A}\right)}}$$

The smaller number of comparisons in multiple comparisons against a single control group compared with all possible comparisons is reflected in the sampling distribution of the q' test statistic, which is in turn reflected in the table of critical values (Table 4-4). As with the SNK test, first order the means, then do the comparisons from the largest to smallest difference. In contrast to the SNK test, the parameter p is the same for all comparisons, equal the number of means in the study. The number of degrees of freedom is the number of degrees of freedom associated with the denominator in the analysis of variance F test statistic.

To repeat the analysis of the effect of running on menstruation using Dunnett's test, we first compare the runners with controls (the largest difference) by computing

$$q' = \frac{\overline{X}_{\text{con}} - \overline{X}_{\text{run}}}{\sqrt{s^2_{\text{wit}}\left(\frac{1}{n_{\text{con}}} + \frac{1}{n_{\text{run}}}\right)}} = \frac{11.5 - 9.1}{\sqrt{3.95\left(\frac{1}{26} + \frac{1}{26}\right)}} = 4.35$$

Table 4-4
Critical Values of q'

	$\alpha_T = 0.05$													
ν_d	$p = 2$	3	4	5	6	7	8	9	10	11	12	13	16	21
5	2.57	3.03	3.29	3.48	3.62	3.73	3.82	3.90	3.97	4.03	4.09	4.14	4.26	4.42
6	2.45	2.86	3.10	3.26	3.39	3.49	3.57	3.64	3.71	3.76	3.81	3.86	3.97	4.11
7	2.36	2.75	2.97	3.12	3.24	3.33	3.41	3.47	3.53	3.58	3.63	3.67	3.78	3.91
8	2.31	2.67	2.88	3.02	3.13	3.22	3.29	3.35	3.41	3.46	3.50	3.54	3.64	3.76
9	2.26	2.61	2.81	2.95	3.05	3.14	3.20	3.26	3.32	3.36	3.40	3.44	3.53	3.65
10	2.23	2.57	2.76	2.89	2.99	3.07	3.14	3.19	3.24	3.29	3.33	3.36	3.45	3.57
11	2.20	2.53	2.72	2.84	2.94	3.02	3.08	3.14	3.19	3.23	3.27	3.30	3.39	3.50
12	2.18	2.50	2.68	2.81	2.90	2.98	3.04	3.09	3.14	3.18	3.22	3.25	3.34	3.45
13	2.16	2.48	2.65	2.78	2.87	2.94	3.00	3.06	3.10	3.14	3.18	3.21	3.29	3.40
14	2.14	2.46	2.63	2.75	2.84	2.91	2.97	3.02	3.07	3.11	3.14	3.18	3.26	3.36
15	2.13	2.44	2.61	2.73	2.82	2.89	2.95	3.00	3.04	3.08	3.12	3.15	3.23	3.33
16	2.12	2.42	2.59	2.71	2.80	2.87	2.92	2.97	3.02	3.06	3.09	3.12	3.20	3.30
17	2.11	2.41	2.58	2.69	2.78	2.85	2.90	2.95	3.00	3.03	3.07	3.10	3.18	3.27
18	2.10	2.40	2.56	2.68	2.76	2.83	2.89	2.94	2.98	3.01	3.05	3.08	3.16	3.25
19	2.09	2.39	2.55	2.66	2.75	2.81	2.87	2.92	2.96	3.00	3.03	3.06	3.14	3.23
20	2.09	2.38	2.54	2.65	2.73	2.80	2.86	2.90	2.95	2.98	3.02	3.05	3.12	3.22
24	2.06	2.35	2.51	2.61	2.70	2.76	2.81	2.86	2.90	2.94	2.97	3.00	3.07	3.16
30	2.04	2.32	2.47	2.58	2.66	2.72	2.77	2.82	2.86	2.89	2.92	2.95	3.02	3.11
40	2.02	2.29	2.44	2.54	2.62	2.68	2.73	2.77	2.81	2.85	2.87	2.90	2.97	3.06
60	2.00	2.27	2.41	2.51	2.58	2.64	2.69	2.73	2.77	2.80	2.83	2.86	2.92	3.00
120	1.98	2.24	2.38	2.47	2.55	2.60	2.65	2.69	2.73	2.76	2.79	2.81	2.87	2.95
∞	1.96	2.21	2.35	2.44	2.51	2.57	2.61	2.65	2.69	2.72	2.74	2.77	2.83	2.91

$\alpha_T = 0.01$

ν_d	$p = 2$	3	4	5	6	7	8	9	10	11	12	13	16	21
5	4.03	4.63	4.98	5.22	5.41	5.56	5.69	5.80	5.89	5.98	6.05	6.12	6.30	6.52
6	3.71	4.21	4.51	4.71	4.87	5.00	5.10	5.20	5.28	5.35	5.41	5.47	5.62	5.81
7	3.50	3.95	4.21	4.39	4.53	4.64	4.74	4.82	4.89	4.95	5.01	5.06	5.19	5.36
8	3.36	3.77	4.00	4.17	4.29	4.40	4.48	4.56	4.62	4.68	4.73	4.78	4.90	5.05
9	3.25	3.63	3.85	4.01	4.12	4.22	4.30	4.37	4.43	4.48	4.53	4.57	4.68	4.82
10	3.17	3.53	3.74	3.88	3.99	4.08	4.16	4.22	4.28	4.33	4.37	4.42	4.52	4.65
11	3.11	3.45	3.65	3.79	3.89	3.98	4.05	4.11	4.16	4.21	4.25	4.29	4.30	4.52
12	3.05	3.39	3.58	3.71	3.81	3.89	3.96	4.02	4.07	4.12	4.16	4.19	4.29	4.41
13	3.01	3.33	3.52	3.65	3.74	3.82	3.89	3.94	3.99	4.04	4.08	4.11	4.20	4.32
14	2.98	3.29	3.47	3.59	3.69	3.76	3.83	3.88	3.93	3.97	4.01	4.05	4.13	4.24
15	2.95	3.25	3.43	3.55	3.64	3.71	3.78	3.83	3.88	3.92	3.95	3.99	4.07	4.18
16	2.92	3.22	3.39	3.51	3.60	3.67	3.73	3.78	3.83	3.87	3.91	3.94	4.02	4.13
17	2.90	3.19	3.36	3.47	3.56	3.63	3.69	3.74	3.79	3.83	3.86	3.90	3.98	4.08
18	2.88	3.17	3.33	3.44	3.53	3.60	3.66	3.71	3.75	3.79	3.83	3.86	3.94	4.04
19	2.86	3.15	3.31	3.42	3.50	3.57	3.63	3.68	3.72	3.76	3.79	3.83	3.90	4.00
20	2.85	3.13	3.29	3.40	3.48	3.55	3.60	3.65	3.69	3.73	3.77	3.80	3.87	3.97
24	2.80	3.07	3.22	3.32	3.40	3.47	3.52	3.57	3.61	3.64	3.68	3.70	3.78	3.87
30	2.75	3.01	3.15	3.25	3.33	3.39	3.44	3.49	3.52	3.56	3.59	3.62	3.69	3.78
40	2.70	2.95	3.09	3.19	3.26	3.32	3.37	3.41	3.44	3.48	3.51	3.53	3.60	3.68
60	2.66	2.90	3.03	3.12	3.19	3.25	3.29	3.33	3.37	3.40	3.42	3.45	3.51	3.59
120	2.62	2.85	2.97	3.06	3.12	3.18	3.22	3.26	3.29	3.32	3.35	3.37	3.43	3.51
∞	2.58	2.79	2.92	3.00	3.06	3.11	3.15	3.19	3.22	3.25	3.27	3.29	3.35	2.42

Source: Reprinted from C. W. Dunnett, "New Tables for Multiple Comparisons with a Control," *Biometrics*, **20**:482–491, 1964.

There are three means, so $p = 3$, and there are 75 degrees of freedom associated with the within-groups variance estimate. From Table 4-4 the critical value of q' for $\alpha_T = .05$ is 2.28, so we conclude that there is a difference between the runners and controls ($P < .05$). Next, we compare the joggers with controls by computing

$$q' = \frac{\overline{X}_{\text{con}} - \overline{X}_{\text{jog}}}{\sqrt{s^2_{\text{wit}}\left(\frac{1}{n_{\text{con}}} + \frac{1}{n_{\text{jog}}}\right)}} = \frac{11.5 - 10.1}{\sqrt{3.95\left(\frac{1}{26} + \frac{1}{26}\right)}} = 2.54$$

As before, there are three means, so $p = 3$; from Table 4-4, the critical value of q' remains 2.26, so we conclude that there is a significant difference between the joggers and controls ($P < .05$). Our overall conclusion is that there is a significant difference in menstruation patterns between both the runners and the joggers and the controls. No statement can be made concerning the differences between the runners and the joggers.

In sum, we conclude that runners and joggers have significantly fewer menses per year than women in the control group, but that there is not a significant difference between the runners and the joggers. Since we only did a small number of comparisons (three), this is the same conclusion we drew using the Bonferroni and Holm t tests to conduct the multiple comparisons. Had we had an experiment with more test groups (and hence many more comparisons), we would see that Dunnett's test was capable of detecting differences that the Bonferroni t test missed because of the large values of t (i.e., small values of P) required to assert a statistically significant difference in any individual pairwise comparison. Dunnett's test is more sensitive than the Bonferroni t test because it uses a more sophisticated mathematical model to estimate the probability of erroneously concluding a difference.

It is less clear whether to recommend the Holm test over Dunnett's test for multiple comparisons against a control group. Theoretically, the sequentially rejective step-down Holm test should be more powerful than the single-step Dunnett's test, but there have been no comprehensive studies of the relative power of Holm's versus Dunnett's tests. You can get a simplified idea of the relative power by considering the running example that we have been considering. The critical value of Dunnett's q' (for $p = 3$, DF = 75, and $\alpha = 0.05$) for each of the two

groups of runners versus controls (runners versus controls and joggers versus controls) is 2.26. The Holm test applied to this family of two comparisons would require critical values of 2.29 for the first test and 1.99 for the second test. Thus, it would be slightly less sensitive for the first comparison and slightly more sensitive for the second.

THE MEANING OF *P*

Understanding what *P* means requires understanding the logic of statistical hypothesis testing. For example, suppose an investigator wants to test whether or not a drug alters body temperature. The obvious experiment is to select two similar groups of people, administer a placebo to one and the drug to the other, measure body temperature in both groups, then compute the mean and standard deviation of the temperatures measured in each group. The mean responses of the two groups will probably be different, regardless of whether the drug has an effect or not for the same reason that different random samples drawn from the same population yield different estimates for the mean. Therefore, the question becomes: Is the observed difference in mean temperature of the two groups likely to be due to random variation associated with the allocation of individuals to the two experimental groups or due to the drug?

To answer this question, statisticians first quantify the observed difference between the two samples with a single number, called *a test statistic,* such as *F, t, q* or *q'*. These statistics, like most test statistics, have the property that the greater the difference between the samples, the greater their value. If the drug has no effect, the test statistic will be a small number. But what is "small"?

To find the boundary between "small" and "big" values of the test statistic, statisticians assume that the drug does *not* affect temperature (the *null hypothesis*). If this assumption is correct, the two groups of people are simply random samples from a single population, all of whom received a placebo (because the drug is, in effect, a placebo). Now, in theory, the statistician repeats the experiment using all possible samples of people and computes the test statistic for each hypothetical experiment. Just as random variation produced different values for means of different samples, this procedure will yield a range of values for the test statistic. Most of these values will be relatively small, but

sheer bad luck requires that there be a few samples that are not representative of the entire population. These samples will yield relatively large values of the test statistic *even if the drug had no effect.* This exercise produces only a few of the possible values of the test statistic, say 5 percent of them, above some cutoff point. The test statistic is "big" if it is larger than this cutoff point.

Having determined this cutoff point, we execute an experiment on a drug with unknown properties and compute the test statistic. It is "big." Therefore, we conclude that *there is less than a 5 percent chance of observing data which led to the computed value of the test statistic if the assumption that the drug has had no effect was true.* Traditionally, if the chances of observing the computed test statistic when the intervention has no effect are below 5 percent, one rejects the working assumption that the drug has no effect and asserts that the drug *does* have an effect. There is, of course, a chance that this assertion is wrong: about 5 percent. This 5 percent is known as the *P value* or *significance level.*

Precisely,

The P value is the probability of obtaining a value of the test statistic as large as or larger than the one computed from the data when in reality there is no difference between the different treatments.

Or, in other words,

The P value is the probability of being wrong when asserting that a true difference exists.

If we are willing to assert a difference when $P < .05$, we are tacitly agreeing to accept the fact that, over the long run, we expect 1 assertion of a difference in 20 to be wrong.*

The convention of considering a difference "statistically significant" when $P < .05$ is widely accepted. In fact, it came from an

*The interpretation of P values that we adopt in this book is the *frequentist* approach, in which the P value is an estimate of the frequency of false positives based only on the analysis of the data in the experiment at hand. An alternative approach, which involves bringing other prior knowledge to bear, is known as the *Bayesian* approach. For a discussion of Bayesian interpretation of P values, with a comparison to the frequentist approach and several clinical examples, see W. S. Browner and T. B. Newman, "Are All Significant P Values Created Equal? The Analogy between Diagnostic Tests and Clinical Research," *JAMA* **257**:2459–2463, 1987. Also J. Brophy and L. Joseph, "Placing Trials in Context Using Bayesian Analysis: GUSTO Revisited by Reverend Bayes," *JAMA* **273**:871–875, 1995.

arbitrary decision by one person, Ronald A. Fisher, who invented much of modern parametric statistics (including the F statistic, which is named for him). In 1926, Fisher published a paper* describing how to assess whether adding manure to a field would increase crop yields which introduced the idea of statistical significance and established the 5 percent standard. He said:

> To an acre of ground the manure is applied; a second acre, sown with similar seed and treated in all other ways like the first, receives none of the manure. When the produce is weighed, it is found that the acre which received the manure has yielded a crop larger indeed by, say, 10 percent. The manure has scored a success, but the confidence with which such a result should be received by the purchasing public depends wholly on the manner in which the experiment was carried out.
>
> First, if the experimenter could say that in twenty years of experience with uniform treatment the difference in favour of the acre treated with manure had never before touched 10 percent., the evidence would have reached a point which may be called the verge of significance; for it is convenient to draw the line at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than one in twenty trials." This level, which we may call the 5 percent point, would be indicated, though very roughly, by the greatest chance deviation observed in twenty successive trials. To locate the 5 percent point with any accuracy we should need about 500 years' experience, for we could then, supposing no progressive changes in fertility were in progress, count out the 25 largest deviations and draw the line between the 25th and 26th largest deviation. If the difference between the two acres in our experimental year exceeded this value, we should have reasonable grounds for calling this value significant.
>
> If one in 20 does not seem high enough odds, we may, if we prefer it, draw the line at 1 in 50 (the 2 percent point) or 1 in 100 (the 1 percent point.) *Personally, the writer prefers to set a low standard of significance at the 5 percent point, and ignore entirely all results which fails to reach this level.* [emphasis added]

*R. A. Fisher. "The Arrangement of Field Experiments," *J. Ministry Ag.* **33**:503–513, 1926. for a discussion of this paper in its historical context, icluding evidence that the logic of hypothesis testing dates back to Blaise Pascal and Pierre Fermat, in 1964, see M. Cowles and C. Davis, "On the Origins of the .05 Level of Statistical Significance, *Am. Psychol.* **37**:533–558, 1982.

Although $P < .05$ is widely accepted, and you will certainly not generate controversy if you use it, a more sensible approach is to consider the P value in making decisions about how to interpret your results without slavishly considering 5 percent a rigid criterion for "truth."

It is commonly believed that the *P* value is the probability of making a mistake. There are obviously two ways an investigator can reach a mistaken conclusion based on the data, reporting that the treatment had an effect when in reality it did not or reporting that the treatment did not have an effect when in reality it did. As noted above, the *P* value only quantifies the probability of making the first kind of error (called a *Type I* or α *error*), that of erroneously concluding that the treatment had an effect when in reality it did not. It gives no information about the probability of making the second kind of error (called a *Type II* or β *error*), that of concluding that the treatment had no effect when in reality it did. Chapter 6 discusses how to estimate the probability of making Type II errors.

PROBLEMS

4-1 Conahan and associates also measured the mean arterial pressure and total peripheral resistance (a measure of how hard it is to produce a given flow through the arterial bed) in 9 patients who were anesthetized with halothane and 16 patients who were anesthetized with morphine. The results are summarized in Table 4-2. Is there evidence that these two anesthetic agents are associated with differences in either of these two variables?

4-2 Cocaine has many adverse effects on the heart, to the point that when people under 40 years of age appear in an emergency room with a heart attack, it is a good guess that it was precipitated by cocaine. In experiments, cocaine has been shown to constrict coronary arteries and reduce blood flow to the heart muscle as well as depress the overall mechanical function of the heart. A class of drugs know as calcium channel blockers has been used to treat problems associated with coronary artery vasoconstriction in other contexts, so Sharon Hale and colleagues ("Nifedipine Protects the Heart from the Acute Deleterious Effects of Cocaine if Administered Before but Not After Cocaine," *Circulation,* **83**:1437–1443, 1991) hypothesized that the calcium channel blocker nifedipine could prevent coronary artery vasoconstriction and the attendant reduction in blood flow to the heart and mechanical function. If true, nifedipine might be useful for treating people who had heart problems brought on by cocaine use. They measured mean arterial pressure in two groups of dogs after administering

cocaine, one of whom was treated with nifedipine and the other of which received a placebo.

Mean Arterial Pressure (mmHg) after Receiving Cocaine

Placebo	Nifedipine
156	73
171	81
133	103
102	88
129	130
150	106
120	106
110	111
112	122
130	108
105	99

Does treatment with nifedipine after administering cocaine affect mean arterial pressure?

4-3 Hale and her colleagues also directly measured the diameter of coronary arteries in dogs after receiving cocaine, and then being treated with a placebo or nifedipine. Based on the following data, did the nifedipine affect the diameters of the coronary arteries?

Diameter of Coronary Artery (mm)

Placebo	Nifedipine
2.5	2.5
2.2	1.7
2.6	1.5
2.0	2.5
2.1	1.4
1.8	1.9
2.4	2.3
2.3	2.0
2.7	2.6
2.7	2.3
1.9	2.2

Does treatment with nifedipine affect the diameter of the coronary arteries in dogs who have received cocaine?

4-4 Rework Probs. 3-1 and 3-5 using the t test. What is the relationship between the value of t computed here and the value of F computed for these data in Chap. 3?

4-5 Problem 3-2 presented the data that White and Froeb collected on the lung function of nonsmokers working in smoke-free environments, nonsmokers working in smoky environments, and smokers of various intensity. Analysis of variance revealed that these data were inconsistent with the hypothesis that the lung function was the same in all these groups. Isolate the various subgroups with similar lung function. What does this result mean in terms of the original question they posed: Does chronic exposure to other people's smoke affect the health of healthy adult nonsmokers?

4-6 Directly test the limited hypothesis that exposure to other people's smoke affects the health of healthy nonsmokers by comparing each group of involuntary smokers and active smokers with the nonsmokers working in a clean environment as the control group. Use the data from Prob. 3-2 and Dunnett's test.

4-7 Problem 3-3 led to the conclusion that HDL concentration is not the same in inactive men, joggers, and marathon runners. Use Bonferroni t tests to compare each of these groups pairwise.

4-8 Suppose that we were just interested in comparisons of the joggers and the marathon men with the inactive adults (as the control group). Use the data in Prob. 3-3 and make these comparisons with Bonferroni t tests.

4-9 Use the data from Prob. 3-4 to determine which interventions have protective effects on the heart during a prolonged ischemic attack. Can a pharmacological agent offer the same benefit as a brief ischemic preconditioning?

4-10 Use the Bonferroni t test to isolate which strains of mice discussed in Prob. 3-7 differ in testicular response to estrogen treatment.

4-11 Repeat Prob. 4-10 using the SNK and Holm tests. Compare the results with those of Prob. 4-10 and explain any differences.

4-12 In Prob. 3-6 you determined there was a difference in burnout among nursing staffs of different patient care units. Isolate these differences and discuss them.

4-13 In a test of significance, the P value of the test statistic is .063. Are the data statistically significant at

a. both the $\alpha = .05$ and $\alpha = .01$ levels?

b. the $\alpha = .05$ level but not at the $\alpha = .01$ level?

c. the $\alpha = .01$ level but not at the $\alpha = .05$ level?

d. neither the $\alpha = .05$ nor the $\alpha = .01$ levels?

Chapter 5

How to Analyze Rates and Proportions

The statistical procedures developed in Chapters 2 to 4 are appropriate for analyzing the results of experiments in which the variable of interest takes on a continuous range of values, such as blood pressure, urine production, or length of hospital stay. These, and similar variables, are measured on an *interval scale* because they are measured on a scale with constant intervals, e.g., millimeters of mercury, milliliters, or days. Much of the information physicians, nurses, and medical scientists use cannot be measured on interval scales. For example, an individual may be male or female, dead or alive, or Caucasian, African American, Mexican American, or Asian. These variables are measured on a *nominal scale,* in which there is no arithmetic relationship between the different classifications. We now develop the statistical tools necessary to describe and analyze such information.*

*There is a third class of variables in which responses can be *ordered* without an arithmetic relationship between the different possible states. Ordinal scales often appear in clinical practice; Chaps. 8 and 10 develop statistical procedures to analyze variables measured on ordinal scales.

It is easy to describe things measured on a nominal scale; simply count the number of patients or experimental subjects with each condition and (perhaps) compute the corresponding percentages.

Let us continue our discussion of the use of halothane versus morphine in open-heart surgery.* We have already seen that these two anesthetic agents produce differences in blood pressure that are unlikely to be due to random sampling effects. This finding is interesting, but the important clinical question is: Was there any difference in mortality? Of the patients anesthetized with halothane, 8 of 61 (13.1 percent) died compared with 10 of the 67 anesthetized with morphine (14.9 percent). This study showed that halothane was associated with a 1.8 percent lower mortality rate *in the 128 patients who were studied.* Is this difference due to a real clinical effect or simply to random variation?

To answer this and other questions about nominal data, we must first invent a way to estimate the precision with which percentages based on limited samples approximate the true rates that would be observed if we could examine the entire population, in this case, *all* people who will be anesthetized for open-heart surgery. We will use these estimates to construct statistical procedures to test hypotheses.

BACK TO MARS

Before we can quantify the certainty of our descriptions of a population on the basis of a limited sample, we need to know how to describe the population itself. Since we have already visited Mars and met all 200 Martians (in Chapter 2), we will continue to use them to develop ways to describe populations. In addition to measuring the Martians' heights, we noted that 50 of them were left-footed and the remaining 150 were right-footed. Figure 5-1 shows the entire population of Mars divided according to footedness. The first way in which we can describe this population is by giving the *proportion p* of Martians who are in each class. In this case $p_{\text{left}} = {}^{50}\!/_{200} = 0.25$ and $p_{\text{right}} = {}^{150}\!/_{200} = 0.75$. Since there are only two possible classes, notice that $p_{\text{right}} = 1 - p_{\text{left}}$. Thus,

*When this study was discussed in Chap. 4, we assumed the same number of patients in each treatment group to simplify the computation. In this chapter we use the actual number of patients in the study.

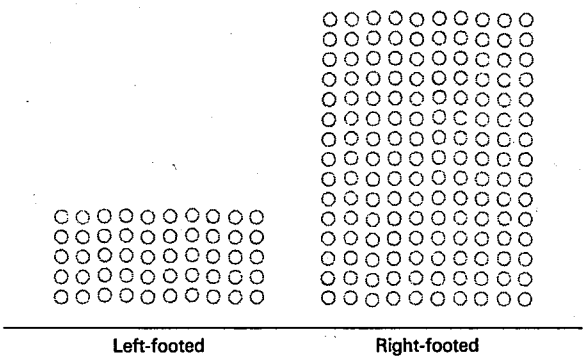


Figure 5-1 Of the 200 Martians 50 are left-footed, and the remaining 150 are right-footed. Therefore, if we select one Martian at random from this population, there is a $p_{\text{left}} = {}^{50}/_{200} = 0.25 = 25$ percent chance it will be left-footed.

whenever there are only two possible classes and they are mutually exclusive, we can completely describe the division in the population with the single parameter p, the proportion of members with one of the attributes. The proportion of the population with the other attribute is *always* $1 - p$.

Note that p also is the *probability* of drawing a left-footed Martian if one selects one member of the population at random.

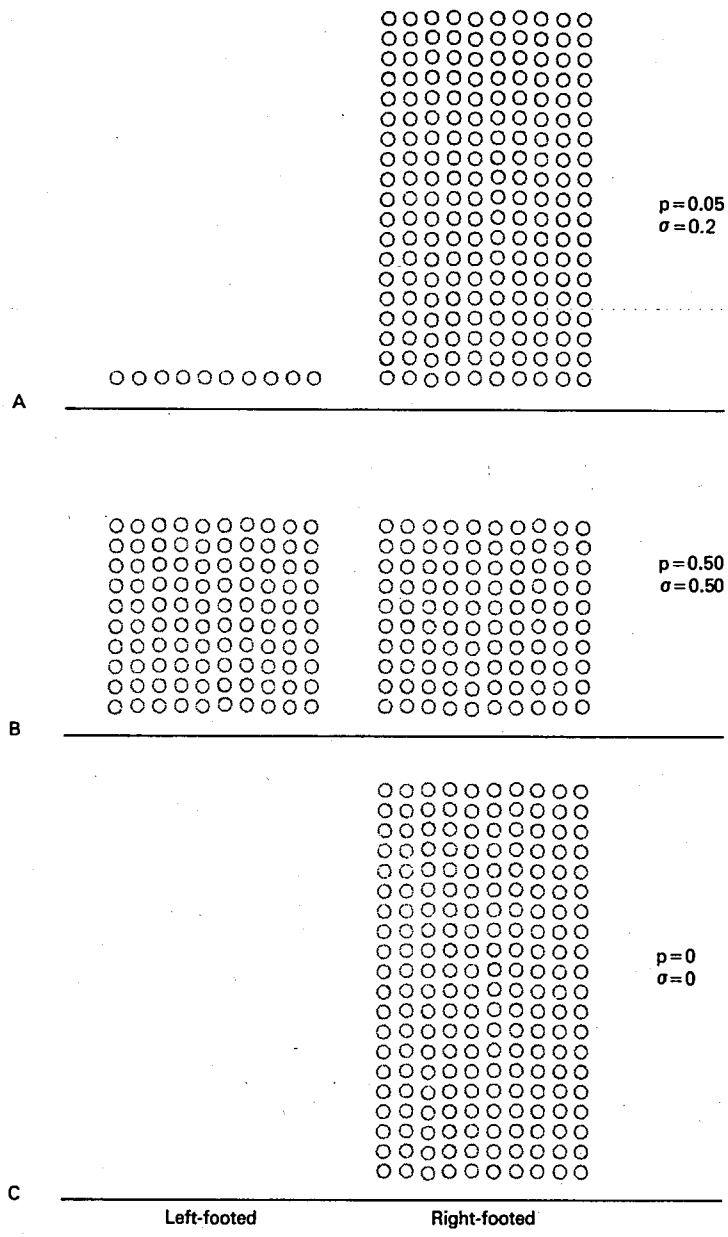
Thus p plays a role exactly analogous to that played by the population mean μ in Chapter 2. To see why, suppose we associate the value $X = 1$ with each left-footed Martian and a value of $X = 0$ with each right-footed Martian. The mean value of X for the population is

$$\mu = \frac{\Sigma X}{N} = \frac{1 + 1 + \cdots + 1 + 0 + 0 + \cdots + 0}{200}$$

$$= \frac{50(1) + 150(0)}{200} = \frac{50}{200} = 0.25$$

which is p_{left}.

This idea can be generalized quite easily using a few equations. Suppose M members of a population of N individuals have some


p=0.05
σ=0.2
A
p=0.50
σ=0.50
B
p=0
σ=0
C
Left-footed
Right-footed

attribute and the remaining $N - M$ members of the population do not. Associate a value of $X = 1$ with the population members having the attribute and a value of $X = 0$ with the others. The mean of the resulting collection of numbers is

$$\mu = \frac{\Sigma X}{N} = \frac{M(1) + (N - M)(0)}{N} = \frac{M}{N} = p$$

the proportion of the population having the attribute.

Since we can compute a mean in this manner, why not compute a standard deviation in order to describe variability in the population? Even though there are only two possibilities, $X = 1$, and $X=0$, the amount of variability will differ, depending on the value of p. Figure 5-2 shows three more populations of 200 individuals each. In Fig. 5-2*A* only 10 of the individuals are left-footed; it exhibits less variability than the population shown in Fig. 5-1. Figure 5-2*B* shows the extreme case in which half the members of the population fall into each of the two classes; the variability is greatest. Figure 5-2*C* shows the other extreme; all the members fall into one of the two classes, and there is no variability at all.

To quantify this subjective impression, we compute the standard deviation of the 1s and 0s associated with each member of the population when we computed the mean. By definition, the population standard deviation is

$$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$$

$X = 1$ for M members of the population and 0 for the remaining $N-M$ members, and $\mu=p$; therefore

---

Figure 5–2 This figure illustrates three different populations, each containing 200 members but with different proportions of left-footed members. The standard deviation, $\sigma = \sqrt{p(1 - p)}$ quantifies the variability in the population. (*A*) When most of the members fall in one class, σ is a small value, 0.2, indicating relatively little variability. (*B*) In contrast, if half the members fall into each class, σ reaches its maximum value of .5, indicating the maximum possible variability. (*C*) At the other extreme, if all members fall into the same class, there is no variability at all and $\sigma = 0$.

$$\sigma = \sqrt{\frac{(1-p)^2+(1-p)^2+\cdots+(1-p)^2+(0-p)^2+(0-p)^2+\cdots+(0-p)^2}{N}}$$

$$= \sqrt{\frac{M(1-p)^2 + (N-M)p^2}{N}} = \sqrt{\frac{M}{N}(1-p)^2 + \left(1-\frac{M}{N}\right)p^2}$$

But since $M/N = p$ is the proportion of population members with the attribute,

$$\sigma = \sqrt{p(1-p)^2 + (1-p)p^2} = \sqrt{[p(1-p) + p^2](1-p)}$$

which simplifies to

$$\sigma = \sqrt{p(1-p)}$$

This equation for the population standard deviation produces quantitative results that agree with the qualitative impressions we developed from Figs. 5-1 and 5-2. As Fig. 5-3 shows, $\sigma = 0$ when $p = 0$ or $p = 1$, that is, when all members of the population either do or do not have the attribute, and σ is maximized when $p = .5$, that is, when any given member of the population is as likely to have the attribute as not.

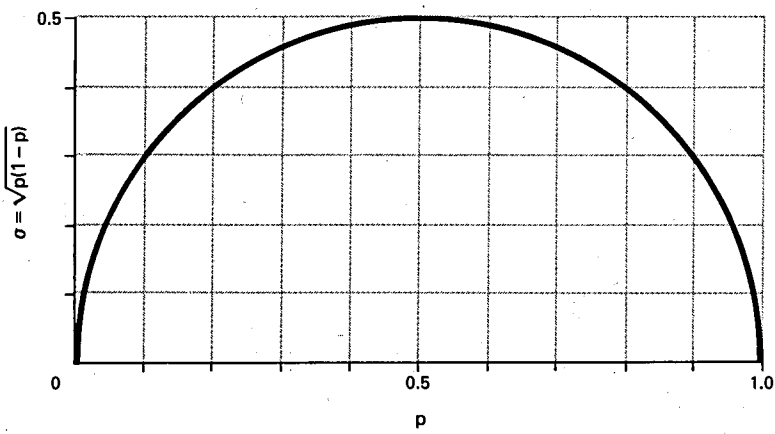


Figure 5–3 The relationship between the standard deviation of a population divided into two categories varies with p, the proportion of members in one of the categories. There is no variation if all members are in one category or the other (so $\sigma = 0$ when $p = 0$ or 1) and maximum variability when a given member is equally likely to fall in one class or the other ($\sigma = 0.5$ when $p = 0.5$).

Since σ depends only on p, it really does not contain any additional information (in contrast to the mean and standard deviation of a normally distributed variable, where μ and σ provide two independent pieces of information). It will be most useful in computing a standard error associated with estimates of p based on samples drawn at random from populations like those shown in Figs. 5-1 or 5-2.

ESTIMATING PROPORTIONS FROM SAMPLES

Of course, if we could observe all members of a population, there would not be any statistical question. In fact, all we ever see is a limited, hopefully representative, sample drawn from that population. How accurately does the proportion of members of a sample with an attribute reflect the proportion of individuals in the population with that attribute? To answer this question, we do a sampling experiment, just as we did in Chapter 2 when we asked how well the sample mean estimated the population mean.

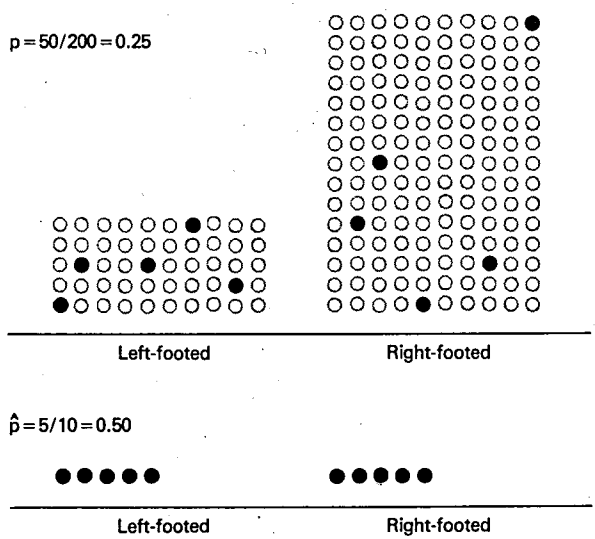


Figure 5-4 The top panel shows one random sample of 10 Martians selected from the population in Fig. 5–1; the bottom panel shows what the investigator would see. Since this sample included 5 left-footed Martians and 5 right-footed Martians, the investigator would estimate the proportion of left-footed Martians to be $\hat{p}_{\text{left}} = {}^{5}/_{10} = .5$, where the circumflex denotes an estimate.

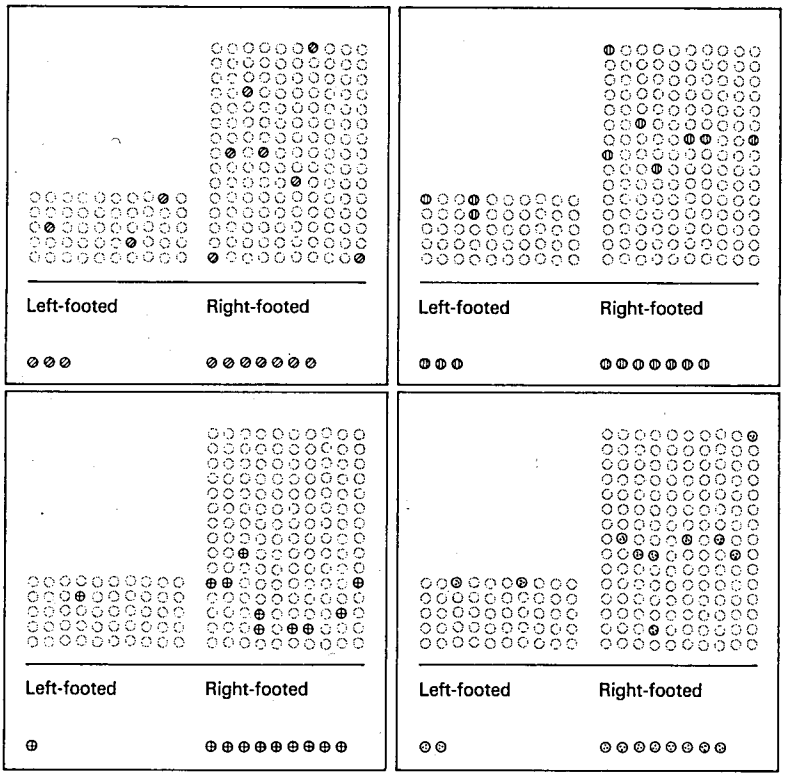


Figure 5-5 Four more random samples of 10 Martians each, together with the sample as it would appear to the investigator. Depending which sample happened to be drawn, the investigator would estimate the proportion of left-footed Martians to be 30, 30, 10, or 20 percent.

Suppose we select 10 Martians at random from the entire population of 200 Martians. Figure 5-4 (top) shows which Martians were drawn; Fig. 5-4 (bottom) shows all the information the investigators who drew the sample would have. Half the Martians in the sample are left-footed and half are right-footed. Given only this information, one would probably report that the proportion of left-footed Martians is 0.5, or 50 percent.

Of course, there is nothing special about this sample, and one of the four other random samples shown in Fig. 5-5 could just as well have been drawn, in which case the investigator would have reported

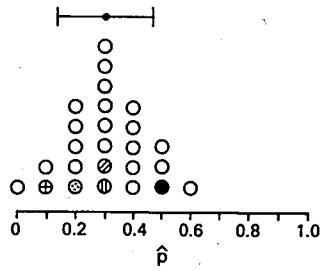


Figure 5-6 There will be a distribution of estimates of the proportion of left-footed Martians $\hat{p}_{left}$ depending on which random sample the investigator happens to draw. This figure shows the 5 specific random samples drawn in Figs. 5–4 and 5–5 together with 20 more random samples of 10 Martians each. The mean of the 25 estimates of p and the standard deviation of these estimates are also shown. The standard deviation of this distribution is the standard error of the estimate of the proportion $\sigma_{\hat{p}}$; it quantifies the precision with which $\hat{p}$ estimates p.

that the proportion of left-footed Martians were 30, 30, 10, or 20 percent, depending on which random sample happened to be drawn. In each case we have computed an estimate of the population proportion p based on a sample. Denote this estimate $\hat{p}$. Like the sample mean, the possible values of $\hat{p}$ depend on both the nature of the underlying population and the specific sample that is drawn. Figure 5-6 shows the five values of $\hat{p}$ computed from the specific samples in Figs. 5-4 and 5-5 together with the results of drawing another 20 random samples of 10 Martians each. Now we change our focus from the population of Martians to the population of all values of $\hat{p}$ computed from random samples of 10 Martians each. There are more than 10^{16} such samples with their corresponding estimates $\hat{p}$ of the value of p for the population of Martians.

The mean estimate of p for the 25 samples of 10 Martians each shown in Fig. 5-6 is 30 percent, which is remarkably close to the true proportion of left-footed Martians in the population (25 percent or 0.25). There is some variation in the estimates. To quantify the variability in the possible values of $\hat{p}$, we compute the *standard deviation* of values of $\hat{p}$ computed from random samples of 10 Martians each. In this case, it is about 14 percent or 0.14. This number describes the variability in the population of all possible values of the proportion of left-footed Martians computed from random samples of 10 Martians each.

Does this sound familiar? It should. It is just like the standard error of the mean. Therefore, we define the *standard error of the estimate of a proportion* to be the standard deviation of the population of all possible values of the proportion computed from samples of a given size.

Just as with the standard error of the mean

$$\sigma_{\hat{p}} = \frac{\sigma}{\sqrt{n}}$$

in which $\sigma_{\hat{p}}$ is the standard error of the proportion, σ is the standard deviation of the population from which the sample was drawn, and n is the sample size. Since $\sigma = \sqrt{p(1-p)}$,

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

We estimate the standard error from a sample by replacing the true value of p in this equation with our estimate $\hat{p}$ obtained from the random sample. Thus,

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The standard error is a very useful way to describe the uncertainty in the estimate of the proportion of a population with a given attribute because the central-limit theorem (Chapter 2) also leads to the conclusion that the distribution of $\hat{p}$ is approximately normal, with mean p and standard deviation $\sigma_{\hat{p}}$ for large enough sample sizes. On the other hand, this approximation fails for values of p near 0 or 1 or when the sample size n is small. When can you use the normal distribution? Statisticians have shown that it is adequate when $n\hat{p}$ and $n(1-\hat{p})$ both exceed about 5.* Recall that about 95 percent of all members of a normally distributed population fall within 2 standard deviations of the mean. When the distribution of $\hat{p}$ approximates the normal distribution, we can assert, with about 95 percent confidence, that the true proportion of population members with the attribute of interest p lies within $2s_{\hat{p}}$ of $\hat{p}$.

These results provide a framework within which to consider the question we posed earlier in the chapter regarding the mortality

*When the sample size is too small to use the normal approximation, you need to solve the problem exactly using the binomial distribution. For a discussion of the binomial distribution, see J. H. Zar, *Biostatistical Analysis,* 2d ed, Prentice-Hall, Englewood Cliffs, N.J., 1984, chap. 22 "The Binomial Distribution."

rates associated with halothane and morphine anesthesia; 13.1 percent of 61 patients anesthetized with halothane and 14.9 percent of 67 patients anesthetized with morphine died following open-heart surgery. The standard errors of the estimates of these percentages are

$$s_{\hat{p}_{\text{hlo}}} = \sqrt{\frac{.131(1 - .131)}{61}} = .043 = 4.3\%$$

for halothane and

$$s_{\hat{p}_{\text{mor}}} = \sqrt{\frac{.149(1 - .149)}{67}} = .044 = 4.4\%$$

for morphine. Given that there was only a 1.8 percent difference in the observed mortality rate, it does not seem likely that the difference in observed mortality rate is due to anything beyond random sampling.

Before moving on, we should pause to list explicitly the assumptions that underlie this approach. We have been analyzing what statisticians call *independent Bernoulli trials,* in which

- *Each individual trial has two mutually exclusive outcomes.*
- *The probability* p *of a given outcome remains constant.*
- *All the trials are independent.*

In terms of a population, we can phrase these assumptions as follows:

- *Each member of the population belongs to one of two classes.*
- *The proportion of members of the population in one of the classes* p *remains constant.*
- *Each member of the sample is selected independently of all other members.*

HYPOTHESIS TESTS FOR PROPORTIONS

In Chapter 4 the sample mean and standard error of the mean provided the basis for constructing the t test to quantify how compatible observations

were with the null hypothesis. We defined the t statistic as

$$t = \frac{\text{difference of sample means}}{\text{standard error of difference of sample means}}$$

The role of $\hat{p}$ is analogous to that of the sample mean in Chapters 2 and 4, and we have also derived an expression for the standard error of $\hat{p}$. We now use the observed proportion of individuals with a given attribute and its standard error to construct a test statistic analogous to t to test the hypothesis that the two samples were drawn from populations containing the same proportion of individuals with a given attribute.

The test statistic analogous to t is

$$z = \frac{\text{difference of sample proportions}}{\text{standard error of difference of sample proportions}}$$

Let $\hat{p}_1$ and $\hat{p}_2$ be the observed proportions of individuals with the attribute of interest in the two samples. The standard error is the standard deviation of the population of all possible values of $\hat{p}$ associated with samples of a given size, and since variances of differences add, the standard error of the difference in proportions is

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{s_{\hat{p}_1}^2 + s_{\hat{p}_2}^2}$$

Therefore

$$z = \frac{\hat{p}_1 - \hat{p}_2}{s_{\hat{p}_1 - \hat{p}_2}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{s_{\hat{p}_1}^2 + s_{\hat{p}_1}^2}}$$

If n_1 and n_2 are the sizes of the two samples,

$$s_{\hat{p}_1} = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1}} \quad \text{and} \quad s_{\hat{p}_2} = \sqrt{\frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

then

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{[\hat{p}_1(1 - \hat{p}_1)/n_1] + [\hat{p}_2(1 - \hat{p}_2)/n_2]}}$$

is our test statistic.

z replaces t because this ratio is approximately normally distributed for large enough sample sizes,* and it is customary to denote a normally distributed variable with the letter z.

Just as it was possible to improve the sensitivity of the t test by pooling the observations in the two sample groups to estimate the population variance, it is possible to increase the sensitivity of the z test for proportions by pooling the information from the two samples to obtain a single estimate of the population standard deviation $s_{\hat{p}}$. Specifically, if the hypothesis that the two samples were drawn from the same population is true, $\hat{p}_1 = m_1/n_1$ and $\hat{p}_2 = m_2/n_2$, in which m_1 and m_2 are the number of individuals in each sample with the attribute of interest, are both estimates of the same population proportion p. In this case, we would consider all the individuals drawn as a single sample of size $n_1 + n_2$ containing a total of $m_1 + m_2$ individuals with the attribute and use this single pooled sample to estimate p:

$$\hat{p} = \frac{m_1 + m_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

$n\hat{p}_1$ — # who died in group 1

in which case

$$s = \sqrt{\hat{p}(1 - \hat{p})}$$

and we can estimate

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}} = \sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Therefore, our test statistic, based on a pooled estimate of the uncertainty in the population proportion, is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})(1/n_1 + 1/n_2)}}$$

Check table 4.1 with ∞ degrees of freedom

Like the t statistic, z will have a range of possible values depending on which random samples happen to be drawn to compute $\hat{p}_1$ and $\hat{p}_2$, even if both samples were drawn from the same population. If z is sufficiently "big," however, we will conclude that the data are inconsistent

*The criterion for a large sample is the same as in the last section, namely that $n\hat{p}$ and $n(1 - \hat{p})$ both exceed about 5 for both samples. When this is not the case, one should use the *Fisher exact test* discussed later in this chapter.

with this hypothesis and assert that there is a difference in the proportions. This argument is exactly analogous to that used to define the critical values of the t for rejecting the hypothesis of no difference. The only change is that in this case we use the standard normal distribution (Fig. 2-5) to define the cutoff values. In fact, the standard normal distribution and the t distribution with an infinite number of degrees of freedom are identical, so we can get the critical values for 5 or 1 percent confidence levels from Table 4-1. This table shows that there is less than a 5 percent chance of z being beyond -1.96 or $+1.96$ and less than a 1 percent chance of z being beyond -2.58 or $+2.58$ when, in fact, the two samples were drawn from the same population.

The Yates Correction for Continuity

The standard normal distribution only approximates the actual distribution of the z test statistic in a way that yields P values that are always smaller than they should be. Thus the results are biased toward concluding that the treatment had an effect when the evidence does not support such a conclusion. The mathematical reason for this problem has to do with the fact that the z test statistic can only take on discrete values, whereas the theoretical standard normal distribution is continuous. To obtain values of the z test statistic which are more compatible with the theoretical standard normal distribution, statisticians have introduced the *Yates correction* (or *continuity correction*), in which the expression for z is modified to become

$$z = \frac{|\hat{p}_1 - \hat{p}_2| - {}^1\!/_2(1/n_1 + 1/n_2)}{\sqrt{\hat{p}(1 - \hat{p})(1/n_1 + 1/n_2)}}$$

This adjustment slightly reduces the value of z associated with the data and compensates for the mathematical problem just described.

Mortality Associated with Anesthesia for Open-Heart Surgery with Halothane or Morphine

We can now formally test the hypothesis that halothane and morphine are associated with the same mortality rate when used as anesthetic agents in open-heart surgery. Recall that the logic of the experiment was that halothane depressed cardiac function whereas morphine did not, so in patients with cardiac problems it ought to be better to use

morphine anesthesia. Indeed, Chapters 3 and 4 showed that halothane produces lower mean arterial blood pressures during the operation than morphine; so the supposed physiological effect is present.

Nevertheless, the important question is: Does either anesthetic agent lead to a detectable improvement in mortality associated with this operation in the period immediately following the operation? Since 8 of the 61 patients anesthetized with halothane (13.1 percent) and 10 of the 67 patients anesthetized with morphine (14.9 percent) died,

$$\hat{p} = \frac{8 + 10}{61 + 67} = 0.141$$

$n\hat{p}$ for the two samples is $0.141(61) = 8.6$ and $0.141(67) = 9.4$. Since both exceed 5, we can use the test described in the last section.* Our test statistic is therefore

$$z = \frac{|\hat{p}_{\text{hlo}} - \hat{p}_{\text{mor}}| - \frac{1}{2}(1/n_{\text{hlo}} + 1/n_{\text{mor}})}{\sqrt{\hat{p}(1 - \hat{p})(1/n_{\text{hlo}} + 1/n_{\text{mor}})}}$$

$$= \frac{|0.131 - 0.149| - \frac{1}{2}\left(\frac{1}{61} + \frac{1}{67}\right)}{\sqrt{(0.141)(1 - 0.141)\left(\frac{1}{61} + \frac{1}{67}\right)}} = 0.04$$

which is quite small. Specifically, it comes nowhere near 1.96, the z value that defines the most extreme 5 percent of all possible values of z when the two samples were drawn from the same population. Hence, we do not have evidence that there is any difference in the mortality associated with these two anesthetic agents, despite the fact that they do seem to have different physiological effects on the patient during surgery.

This study illustrates the importance of looking at *outcomes* in clinical trials. The human body has tremendous capacity to adapt not only to trauma but also in medical manipulation. Therefore, simply showing that some intervention (like a difference in anesthesia)

*$n(1 - \hat{p})$ also exceeds 5 in both cases. We did not need to check this because $\hat{p} < 0.5$, so $n\hat{p} < n(1 - \hat{p})$.

changed a patient's physiological state (by producing different blood pressure) does not mean that in the long run it will make any difference in the clinical outcome. Focusing on these intermediate variables, often called *process variables,* rather than the more important outcome variables may lead you to think something made a clinical difference when it did not. For example, in this study there was the expected change in the process variable, blood pressure, but not the outcome variable, mortality. If we had stopped with the process variables, we might have concluded that morphine anesthesia was superior to halothane in patients with cardiac problems, even though the choice of anesthesia does not appear to have affected the most important variable, whether or not the patient survived.

Keep this distinction in mind when reading medical journals and listening to proponents argue for their tests, procedures, and therapies. It is much easier to show that something affects process variables than the more important outcome variables. In addition to being easier to produce a demonstrable change in process variables than outcome variables, process variables are generally easier to measure. Observing outcomes may require following the patients for some time and often present difficult subjective problems of measurement, especially when one tries to measure "quality of life" variables. Nevertheless, when assessing whether or not some new procedure deserves to be adopted in an era of limited medical resources, you should seek evidence that something affects the patient's outcome. The patient and the patient's family care about outcome, not process.

Prevention of Thrombosis in People Receiving Hemodialysis

People with chronic kidney disease can be kept alive by dialysis; their blood is passed through a machine that does the work of their kidneys and removes metabolic products and other chemicals from their blood. The dialysis machine must be connected to one of the patient's arteries and veins to allow the blood to pass through the machine. Since patients must be connected to the dialysis machine on a regular basis, it is necessary to create surgically a more or less permanent connection that can be used to attach the person's body to the machine. One way of doing this is to attach a small Teflon tube containing a coupling fitting, called a *shunt,* between an artery and vein in the wrist or arm. When the

patient is to be connected to the dialysis machine, the tubing is connected to these fittings on the Teflon tube; otherwise, the two fittings are simply connected together so that the blood just flows directly from the small artery to the vein. For a variety of reasons, including the surgical technique used to place the shunt, disease of the artery or vein, local infection, or a reaction to the Teflon adapter, blood clots (thromboses) tend to form in these shunts. These clots have to be removed regularly to permit dialysis and can be severe enough to require tying off the shunt and creating a new one. The clots can spread down the artery or vein, making it necessary to pass a catheter into the artery or vein to remove the clot. In addition, these clots may break loose and lodge elsewhere in the body, where they may cause problems. Herschel Harter and colleagues* knew that aspirin tends to inhibit blood clotting and wondered whether thrombosis could be reduced in people who were receiving chronic dialysis by giving them a low dose of aspirin (160 mg, one-half a common aspirin tablet) every day to inhibit the blood's tendency to clot.

They completed a randomized clinical trial in which all people being dialyzed at their institution who agreed to participate in the study and who had no reason for not taking aspirin (like an allergy) were randomly assigned to a group that received either a placebo or aspirin. To avoid bias on either the investigators' or patients' parts, the study was *double-blind.* Neither the physician administering the drug nor the patient receiving it knew whether the tablet was placebo or aspirin. This procedure adjusts for the placebo effect in the patients and prevents the investigators from looking harder for clots in one group or the other. The double-blind randomized clinical trial is the best way to test a new therapy.

They continued the study until 24 patients developed thrombi, because they assumed that with a total of 24 patients with thrombi any differences between the placebo and aspirin-treated groups would be detectable. Once they reached this point, they broke the code on the bottles of the pills and analyzed their results: 19 people had received aspirin and 25 people had received placebo (Table 5-1). There did not seem to be any clinically important difference in these two groups in

*H. R. Harter, J. W. Burch, P. W. Majerus, N. Stanford, J. A. Delmez, C. B. Anderson, and C. A. Weerts, "Prevention of Thrombosis in Patients in Hemodialysis by Low-Dose Aspirin," *N. Engl. J. Med.*, **301**:577–579, 1979.

terms of age distribution, sex, time on dialysis at entry into the study, or other variables.

Of the 19 people receiving aspirin, 6 developed thrombi; of the 25 people receiving placebo 18 developed thrombi. Is this difference beyond what we would expect if aspirin had no effect and acted like a placebo, so the two groups of patients could be considered as having been drawn from the same population in which a constant proportion p of patients were destined to develop thrombi?

We first estimate p for the two groups:

$$\hat{p}_{\text{asp}} = \frac{6}{19} = 0.32$$

for the people who received aspirin and

$$\hat{p}_{\text{pla}} = \frac{18}{25} = 0.72$$

for the people who received placebo.

Next, we make sure that $n\hat{p}$ and $n(1-\hat{p})$ are greater than about 5 for both groups, to be certain that the sample sizes are large enough for the normal distribution to reasonably approximate the distribution of our test statistic z if the hypothesis that aspirin had no effect is true. For the people who received aspirin

$$n_{\text{asp}}\hat{p}_{\text{asp}} = 6$$
$$n_{\text{asp}}(1 - \hat{p}_{\text{asp}}) = 13$$

Table 5-1 Thrombus Formation in People Receiving Dialysis and Treated with Placebo or Aspirin

	Number of patients		
Sample group	Developed thrombi	Free of thrombi	Treated
Placebo	18	7	25
Aspirin	6	13	19
Total	24	20	44

Source: H. R. Harter, J. W. Burch, P. W. Majerus, N. Stanford, J. A. Delmez, C. B. Anderson, and C. A. Weerts "Prevention of Thrombosis in Patients on Hemodialysis by Low-Dose Aspirin," *N. Engl. J. Med.*, 301:577–579, 1979. Reprinted by permission of the *New England Journal of Medicine.*

and for the people who received placebo

$$n_{\text{pla}}\hat{p}_{\text{pla}} = 18$$
$$n_{\text{pla}}(1 - \hat{p}_{\text{pla}}) = 7$$

We can use the methods we have developed.

The proportion of all patients who developed thromboses was

$$\hat{p} = \frac{6 + 18}{19 + 25} = 0.55$$

and so

$$s_{\hat{p}_{\text{asp}} - \hat{p}_{\text{pla}}} = \sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_{\text{asp}}} + \frac{1}{n_{\text{pla}}}\right)}$$
$$= \sqrt{0.55(1 - 0.55)\left(\frac{1}{19} + \frac{1}{25}\right)} = 0.15$$

Finally, we compute z according to

$$z = \frac{|\hat{p}_{\text{asp}} - \hat{p}_{\text{pla}}| - \frac{1}{2}\left(\frac{1}{19} + \frac{1}{25}\right)}{s_{\hat{p}_{\text{asp}} - \hat{p}_{\text{pla}}}}$$
$$= \frac{|0.32 - 0.72| - 0.05}{0.15} = 2.33$$

Table 4-1 indicates that z will exceed 2.3263 in magnitude less than 2 percent of the time if the two samples are drawn from the same population. Since the value of z associated with our experiment is more extreme than 2.3263, it is very unlikely that the two samples were drawn from a single population. Therefore, we conclude that they were not, with $P < .02$.* In other words, we will conclude that giving patients low doses of aspirin while they are receiving chronic kidney dialysis decreases the likelihood that they will develop thrombosis in the shunt used to connect them to the dialysis machine.

*The value of z associated with these data, 2.33, is so close to the critical value of 2.3263 associated with $P < .02$ that it would be prudent to report $P < .05$ (corresponding to a critical value of 1.960) because the mathematical models that are used to compute the table of critical values are only approximations of reality.

ANOTHER APPROACH TO TESTING NOMINAL DATA: ANALYSIS OF CONTINGENCY TABLES

The methods we just developed based on the z statistic are perfectly adequate for testing hypotheses when there are only two possible attributes or outcomes of interest. The z statistic plays a role analogous to the t test for data measured on an interval scale. There are many situations, however, where there are more than two samples to be compared or more than two possible outcomes. To do this, we need to develop a testing procedure, analogous to analysis of variance, that is more flexible than the z test just described. While the following approach may seem quite different from the one we just used to design the z test for proportions, it is essentially the same.

To keep things simple, we begin with the problem we just solved, assessing the efficacy of low-dose aspirin in preventing thrombosis. In the last section we analyzed the *proportion* of people in each of the two treatment groups (aspirin and placebo) who developed thromboses. Now we change our emphasis slightly and analyze the *number* of people in each group who developed thrombi. Since the procedure we will develop does not require assuming anything about the nature of parameters of the population from which the samples were drawn, it is called a *nonparametric* method.

Table 5-1 shows the results of placebo and aspirin in the experiment, with the number of people in each treatment group who did and did not develop thromboses. This table is called a 2×2 *contingency table.* Most of the patients in the study fell along the diagonal in this table, suggesting an association between the presence of thrombi and the absence of aspirin treatment. Table 5-2 shows what the experimental results might have looked like *if the aspirin had no effect on thrombus formation.* It also shows the total number of patients who received each treatment as well as the total number who did and did not develop thrombi. These numbers are obtained by summing the rows and columns, respectively, in the table; these sums are the same as Table 5-1. More patients developed thrombi under each treatment; the differences in absolute numbers of patients are due to the fact that more patients received the placebo than aspirin. In contrast to Table 5-1, there does not seem to be a pattern relating treatment to thrombus formation.

To understand better why most people have this subjective impression, let us examine where the numbers in Table 5-2 came from. Of the

Table 5-2 Expected Thrombus Formation If Aspirin Had No Effect

	Number of patients		
Sample group	Developed thrombi	Free of thrombi	Treated
Aspirin	10.36	8.64	19
Placebo	13.64	11.36	25
Total	24	20	44

44 people in the study 25, or $^{25}/_{44} = 57$ percent, received placebo and 19, or $^{19}/_{44} = 43$ percent, received aspirin. Of the people in the study 24, or $^{25}/_{44} = 55$ percent, developed thrombi and 20, or $^{20}/_{44} = 45$ percent, did not. Now, let us hypothesize that the treatment did *not* affect the likelihood that someone would develop a thrombus. In this case, we would expect 55 percent of the 25 patients treated with placebo (13.64 patients) to develop thrombi and 55 percent of the 19 patients treated with aspirin (10.36 patients) to develop thrombi. The remaining patients should be free of thrombi. Note that we compute the expected frequencies to two decimal places (i.e., to the hundredth of a patient); this procedure is necessary to ensure accurate results in the computation of the χ^2 test below. Thus, Table 5-2 shows how we would *expect* the data to look if 25 patients were given placebo and 19 patients were given aspirin and 24 of them were destined to develop thrombi *regardless of how they were treated.* Compare Tables 5-1 and 5-2. Do they seem similar? Not really; the actual pattern of observations seems quite different from what we expected if the treatment had no effect.

The next step in designing a statistical procedure to test the hypothesis that the pattern of observations is due to random sampling rather than a treatment effect is to reduce this subjective impression to a single number, a test statistic, like *F, t,* or *z,* so that we can reject the hypothesis of no effect when this statistic is "big."

Before constructing this test statistic, however, let us return to another example, the relationship between type of anesthesia and mortality following open-heart surgery. Table 5-3 shows the results of our investigation, presented in the same format as Table 5-1. Table 5-4 presents what the table might look like if the type of anesthesia had no effect on mortality. Out of 128 people, 110, or $^{110}/_{128} = 86$ percent, lived.

Table 5-3 Mortality Associated with Open-Heart Surgery

Anesthesia	Lived	Died	Total no. of cases
Halothane	53	8	61
Morphine	57	10	67
Total	110	18	128

If the type of anesthesia had no effect on mortality rate, 86 percent of the 61 people anesthetized with halothane (52.42 people) and 86 percent of the 67 people anesthetized with morphine (57.58 people) would be expected to live, the rest dying in each case. Compare Tables 5-3 and 5-4; there is little difference between the expected and observed frequencies in each cell in the table. The observations are compatible with the assumption that there is no relationship between type of anesthesia and mortality.

The Chi-Square Test Statistic

Now we are ready to design our test statistic. It should describe, with a single number, how much the observed frequencies in each cell in the table differ from the frequencies we would expect if there is no relationship between the treatments and the outcomes that define the rows and columns of the table. In addition, it should allow for the fact that if we expect a large number of people to fall in a given cell, a difference of one person between the expected and observed frequencies is less important than in cases where we expect only a few people to fall in the cell.

Table 5-4 Expected Mortality with Open-Heart Surgery If Anesthesia Did Not Matter

Anesthesia	Lived	Died	Total no. of cases
Halothane	52.42	8.58	61
Morphine	57.58	9.42	67
Total	110	18	128

We define the test statistic χ^2 (the square of the Greek letter chi) as

$$\chi^2 = \text{sum of } \frac{(\text{observed} - \text{expected number of individuals in cell})^2}{\text{expected number of individuals in cell}}$$

The sum is calculated by adding the results for all cells in the contingency table. The equivalent mathematical statement is

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

in which O is the observed number of individuals (frequency) in a given cell, E is the expected number of individuals (frequency) in that cell, and the sum is over all the cells in the contingency table. Note that if the observed frequencies are similar to the expected frequencies, χ^2 will be a small number and if the observed and expected frequencies differ, χ^2 will be a big number.

We can now use the information in Tables 5-1 and 5-2 to compute the χ^2 statistic associated with the data on the use of low-dose aspirin to prevent thrombosis in people undergoing chronic dialysis. Table 5-1 gives the observed frequencies, and Table 5-2 gives the expected frequencies. Thus,

$$\chi^2 = \sum \frac{(O - E)^2}{E} = \frac{(18 - 13.64)^2}{13.64} + \frac{(7 - 11.36)^2}{11.36} + \frac{(6 - 10.36)^2}{10.36} + \frac{(13 - 8.64)^2}{8.64} = 7.10$$

To begin getting a feeling for whether or not 7.10 is "big," let us compute χ^2 for the data on mortality associated with halothane and morphine anesthesia given in Table 5-3. Table 5-4 gives the expected frequencies, so

$$\chi^2 = \frac{(53 - 52.42)^2}{52.42} + \frac{(8 - 8.58)^2}{8.58} + \frac{(57 - 57.58)^2}{57.58} + \frac{(10 - 9.42)^2}{9.42} = 0.09$$

which is pretty small, in agreement with our intuitive impression that the observed and expected frequencies are quite similar. (Of course, it is also in agreement with our earlier analysis of the same data using the

z statistic in the last section.) In fact, it is possible to show that $\chi^2 = z^2$ when there are only two samples and two possible outcomes.

Like all test statistics, χ^2 can take on a range of values even when there is no relationship between the treatments and outcomes because of the effects of random sampling. Figure 5-7 shows the distribution of possible values for χ^2 computed from data in 2×2 contingency tables like those in Tables 5-1 or 5-3. It shows that when the hypothesis of no relationship between the rows and columns of the table is true, χ^2 would be expected to exceed 6.635 only 1 percent of the time. Because the observed value of χ^2, 7.10 exceeds this critical value of 6.635, we can conclude that the data in Table 5-1 are unlikely to occur when the hypothesis that aspirin and placebo have the same effect on thrombus formation is true. We report that aspirin is associated with lower rates of thrombus formation ($P < .01$).

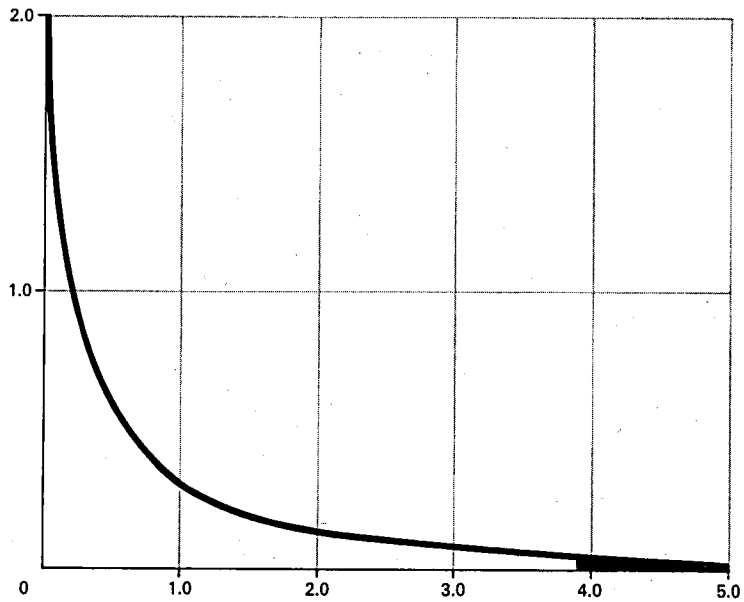


Figure 5-7 The chi-square distribution with 1 degree of freedom. The shaded area denotes the biggest 5 percent of possible values of the χ^2 test statistic when there is no relationship between the treatments and observations.

In contrast, the data in Table 5-3 seem very compatible with the hypothesis that halothane and morphine produce the same mortality rates in patients being operated on for repair of heart valves.

Of course, neither of these cases *proves* that aspirin did or did not have an effect, or that halothane and morphine did or did not produce the same mortality rates. What they show is that in one case the pattern of the observations is unlikely to arise if the aspirin acts like a placebo, whereas on the other hand the pattern of observations are very likely to arise if halothane and morphine produce similar mortality rates. Like all the other procedures we have been using to test hypotheses, however, when we reject the hypothesis of no association at the 5 percent level, we are implicitly willing to accept the fact that, in the long run, about 1 reported effect in 20 will be due to random variation rather than a real treatment effect.

As with all theoretical distributions of test statistics used for testing hypotheses, there are assumptions built into the use of χ^2. For the resulting theoretical distribution to be reasonably accurate, *the expected number of individuals in all the cells must be at least 5.** (This is essentially the same as the restriction on the z test in the last section.)

Like most test statistics, the distribution of χ^2 depends on the number of treatments being compared. It also depends on the number of possible outcomes. This dependency is quantified in a *degrees of freedom* parameter v equal to the number of rows in the table minus 1 times the number of columns in the table minus 1

$$v = (r - 1)(c - 1)$$

where r is the number of rows and c is the number of columns in the table. For the 2×2 tables we have been dealing with so far, $v = (2 - 1)(2 - 1) = 1$.

As with the z test statistic discussed earlier in this chapter, when analyzing 2×2 contingency tables ($v = 1$), the value of χ^2 computed using the formula above and the theoretical χ^2 distribution leads to P values that are smaller than they ought to be. Thus, the results are biased toward concluding that the treatment had an effect when the

*When the data do not meet this requirement, one should use the Fisher exact test.

evidence does not support such a conclusion. The mathematical reason for this problem has to do with the fact that the theoretical χ^2 distribution is continuous whereas the set of all possible values that the χ^2 test statistics can take on is not. To obtain values of the test statistic that are more compatible with the critical values computed from the theoretical χ^2 distribution when $v = 1$, apply the *Yates correction* (or *continuity correction*) to compute a corrected χ^2 test statistic according to

$$\chi^2 = \sum \frac{\left(|O - E| - \frac{1}{2}\right)^2}{E}$$

This correction slightly reduces the value of χ^2 associated with the contingency table and compensates for the mathematical problem just described. The Yates correction is used only when $v = 1$, that is, for 2×2 tables.

To illustrate the use and effect of the continuity correction, let us recompute the value of χ^2 associated with the data on the use of low-dose aspirin to prevent thrombosis in people undergoing chronic dialysis. From the observed and expected frequencies in Tables 5-1 and 5-2, respectively

$$\chi^2 = \frac{\left(|18 - 13.64| - \frac{1}{2}\right)^2}{13.64} + \frac{\left(|7 - 11.36| - \frac{1}{2}\right)^2}{11.36} + \frac{\left(|6 - 10.36| - \frac{1}{2}\right)^2}{10.36} + \frac{\left(|13 - 8.64| - \frac{1}{2}\right)^2}{8.64} = 5.57$$

Note that this value of χ^2, 5.57, is smaller than the uncorrected value of χ^2, 7.10, we obtained before. The corrected value of χ^2 no longer exceeds the critical value of 6.635 associated with the greatest 1 percent of possible χ^2 values (i.e., for $P < .01$). After applying the continuity correction, χ^2 now only exceeds 5.024, the critical value that defines the greatest 2.5 percent of possible values (i.e., for $P < .025$).

CHI-SQUARE APPLICATIONS TO EXPERIMENTS WITH MORE THAN TWO TREATMENTS OR OUTCOMES

It is easy to generalize what we have just done to analyze the results of experiments with more than two treatments or outcomes. The z test we developed earlier in this chapter will not work for such experiments.

Recall that in Chapter 3 we demonstrated that women who jog regularly or engage in long-distance running have fewer menstrual periods on the average than women who do not participate in this sport.* Does this physiological change lead women to consult their physician about menstrual problems? Table 5-5 shows the results of a survey of the same women discussed in conjunction with Fig. 3-9. Are these data consistent with the hypothesis that running does not increase the likelihood that a woman will consult her physician for a menstrual problem?

Of the 165 women in the study 69, or $^{69}/_{165} = 42$ percent, consulted their physicians for a menstrual problem, while the remaining 96, or $^{96}/_{165} = 58$ percent, did not. If the extent of running did not affect the likelihood that a woman would consult her physician, we would expect 42 percent of the 54 controls (22.58 women) to have visited their physicians, and 42 percent of the 23 joggers (9.62 women) to have consulted their physicians, and 42 percent of the 88 distance runners (36.80 women) to have consulted their physicians. Table 5-6 shows these expected frequencies, together with the expected frequencies of

Table 5-5 Consult Physician for Menstrual Problem

Group	Yes	No	Total
Controls	14	40	54
Joggers	9	14	23
Runners	46	42	88
Total	69	96	165

Source: E. Dale, D. H. Gerlach, and A. L. Wilhite, "Menstrual Dysfunction in Distance Runners," *Obstet. Gynecol.*, **54**:47–53, 1979.

*When this study was discussed in Chapter 3, we assumed the same number of patients in each treatment group to simplify the computation. In this chapter we use the actual number of patients in the study.

Table 5-6 Expected Frequencies of Physician Consultation If Running Did Not Matter

Group	Yes	No	Total
Controls	22.58	31.42	54
Joggers	9.62	13.38	23
Runners	36.80	51.20	88
Total	69	96	165

women who did not consult their physicians. Are the differences between the observed and expected frequencies "big?"

To answer this question, we compute the χ^2 statistic

$$\chi^2 = \frac{(14 - 22.58)^2}{22.58} + \frac{(40 - 31.42)^2}{31.42} + \frac{(9 - 9.62)^2}{9.62} + \frac{(14 - 13.38)^2}{13.38} + \frac{(46 - 36.80)^2}{36.80} + \frac{(42 - 51.20)^2}{51.20} = 9.63$$

The contingency table in Table 5-5 has three rows and two columns, so the χ^2 statistic has

$$v = (r - 1)(c - 1) = (3 - 1)(2 - 1) = 2 \text{ degrees of freedom}$$

associated with it. Table 5-7 shows that χ^2 will exceed 9.21 less than 1 percent of the time when the difference between the observed and expected frequencies is due to random variation rather than an effect of the treatment (in this case, running). Thus, there is a relationship between running and the chances that a woman will consult her physician about a menstrual problem ($P < .01$). Note, however, that we do not yet know which group or groups of women account for this difference.

Let us now sum up how to use the χ^2 statistic.

- *Tabulate the data in a contingency table.*
- *Sum the number of individuals in each row and each column and figure the percentage of all individuals who fall in each row and column, independent of the column or row in which they fall.*

- *Use these percentages to compute the number of people that would be expected in each cell of the table if the treatment had no effect.*
- *Summarize the differences between these expected frequencies and the observed frequencies by computing* χ^2. *If the data form a* 2×2 *table, include the Yates correction.*
- *Compute the number of degrees of freedom associated with the contingency table and use Table 5-7 to see whether the observed value of* χ^2 *exceeds what would be expected from random variation.*

Recall that when the data fell into a 2×2 contingency table, all the expected frequencies had to exceed about 5 for the χ^2 test to be accurate. In larger tables, most statisticians recommend that the expected number of individuals in each cell never be less than 1 and that no more than 20 percent of them be less than 5. When this is not the case, the χ^2 test can be quite inaccurate. The problem can be remedied by collecting more data to increase the cell numbers or by reducing the number of categories to increase the numbers in each cell of the table.

Subdividing Contingency Tables

Our analysis of Table 5-6 revealed that there is probably a difference in the likelihood that the different groups of women will consult their physicians regarding a menstrual problem, but our analysis did not isolate *which* groups of women accounted for this effect. This situation is analogous to the multiple-comparison problem in analysis of variance. The analysis of variance will help decide whether *something* is different, but you need to go on to the multiple-comparison procedure to define *which group it was.* You can do the same thing with a contingency table.

Looking at the numbers in Table 5-5 suggests that joggers and runners are more likely to consult their physicians than women in the control group, but they seem similar to each other.

To test this latter hypothesis, we *subdivide* the contingency table to look only at the joggers and runners. Table 5-8 shows the data for the joggers and runners. The numbers in parentheses are the expected number of women in each cell. The observed and expected number

Table 5-7
Critical Values for the χ^2 Distribution

	Probability of greater value P							
ν	.50	.25	.10	.05	.025	.01	.005	.001
1	.455	1.323	2.706	3.841	5.024	6.635	7.879	10.828
2	1.386	2.773	4.605	5.991	7.378	9.210	10.597	13.816
3	2.366	4.108	6.251	7.815	9.348	11.345	12.838	16.266
4	3.357	5.385	7.779	9.488	11.143	13.277	14.860	18.467
5	4.351	6.626	9.236	11.070	12.833	15.086	16.750	20.515
6	5.348	7.841	10.645	12.592	14.449	16.812	18.548	22.458
7	6.346	9.037	12.017	14.067	16.013	18.475	20.278	24.322
8	7.344	10.219	13.362	15.507	17.535	20.090	21.955	26.124
9	8.343	11.389	14.684	16.919	19.023	21.666	23.589	27.877
10	9.342	12.549	15.987	18.307	20.483	23.209	25.188	29.588
11	10.341	13.701	17.275	19.675	21.920	24.725	26.757	31.264
12	11.340	14.845	18.549	21.026	23.337	26.217	28.300	32.909
13	12.340	15.984	19.812	22.362	24.736	27.688	29.819	34.528
14	13.339	17.117	21.064	23.685	26.119	29.141	31.319	36.123
15	14.339	18.245	22.307	24.996	27.488	30.578	32.801	37.697
16	15.338	19.369	23.542	26.296	28.845	32.000	34.267	39.252
17	16.338	20.489	24.769	27.587	30.191	33.409	35.718	40.790
18	17.338	21.605	25.989	28.869	31.526	34.805	37.156	42.312
19	18.338	22.718	27.204	30.144	32.852	36.191	38.582	43.820
20	19.337	23.828	28.412	31.410	34.170	37.566	39.997	45.315
21	20.337	24.935	29.615	32.671	35.479	38.932	41.401	46.797
22	21.337	26.039	30.813	33.924	36.781	40.289	42.796	48.268
23	22.337	27.141	32.007	35.172	38.076	41.638	44.181	49.728
24	23.337	28.241	33.196	36.415	39.364	42.980	45.559	51.179
25	24.337	29.339	34.382	37.652	40.646	44.314	46.928	52.620
26	25.336	30.435	35.563	38.885	41.923	45.642	48.290	54.052
27	26.336	31.528	36.741	40.113	43.195	46.963	49.645	55.476
28	27.336	32.020	37.916	41.337	44.461	48.278	50.993	56.892
29	28.336	33.711	39.087	42.557	45.722	49.588	52.336	58.301
30	29.336	34.800	40.256	43.773	46.979	50.892	53.672	59.703
31	30.336	35.887	41.422	44.985	48.232	52.191	55.003	61.098
32	31.336	36.973	42.585	46.194	49.480	53.486	56.328	62.487
33	32.336	38.058	43.745	47.400	50.725	54.776	57.648	63.870
34	33.336	39.141	44.903	48.602	51.966	56.061	58.964	65.247
35	34.336	40.223	46.059	49.802	53.203	57.342	60.275	66.619
36	35.336	41.304	47.212	50.998	54.437	58.619	61.581	67.985
37	36.336	42.383	48.363	52.192	55.668	59.893	62.883	69.346
38	37.335	43.462	49.513	53.384	56.896	61.162	64.181	70.703

Table 5-7
Critical Values for the χ^2 Distribution (*continued*)

	Probability of greater value P							
ν	.50	.25	.10	.05	.025	.01	.005	.001
39	38.335	44.539	50.660	54.572	58.120	62.428	65.476	72.055
40	39.335	45.616	51.805	55.758	59.342	63.691	66.766	73.402
41	40.335	46.692	52.949	56.942	60.561	64.950	68.053	74.745
42	41.335	47.766	54.090	58.124	61.777	66.206	69.336	76.084
43	42.335	48.840	55.230	59.304	62.990	67.459	70.616	77.419
44	43.335	49.913	56.369	60.481	64.201	68.710	71.893	78.750
45	44.335	50.985	57.505	61.656	65.410	69.957	73.166	80.077
46	45.335	52.056	58.641	62.830	66.617	71.201	74.437	81.400
47	46.335	53.127	59.774	64.001	67.821	72.443	75.704	82.720
48	47.335	54.196	60.907	65.171	69.023	73.683	76.969	84.037
49	48.335	55.265	62.038	66.339	70.222	74.919	78.231	85.351
50	49.335	56.334	63.167	67.505	71.420	76.154	79.490	86.661

Source: Adapted from J. H. Zar, *Biostatistical Analysis* (2nd ed). Prentice-Hall, Englewood Cliffs, N.J., 1984, pp. 479–482, table B.1. Used by permission.

of women in each cell appear quite similar; since it is a 2×2 contingency table, we compute χ^2 with the Yates correction

$$\chi^2 = \sum \frac{\left(|O - E| - \frac{1}{2}\right)^2}{E}$$

$$= \frac{\left(|9 - 11.40| - \frac{1}{2}\right)^2}{11.40} + \frac{\left(|14 - 11.60| - \frac{1}{2}\right)^2}{11.60}$$

$$+ \frac{\left(|46 - 43.60| - \frac{1}{2}\right)^2}{43.60} + \frac{\left(|42 - 44.40| - \frac{1}{2}\right)^2}{44.40} = .79$$

which is small enough for us to conclude that the joggers and runners are equally likely to visit their physicians. Since they are so similar, we combine the two groups and compare this combined group with the control group. Table 5-9 shows the resulting 2×2 contingency table, together with the expected frequencies in parentheses. χ^2 for this

Table 5-8 Physician Consultation among Women Joggers and Runners*

Group	Yes	No	Total
Joggers	9 (11.40)	14 (11.60)	23
Runners	46 (43.60)	42 (44.40)	88
Total	55	56	111

*Numbers in parentheses are expected frequencies if the amount of running does not affect physician consultation.

contingency table is 7.39, which exceeds 6.63, the critical value that defines the upper 1 percent of probable values of χ^2 when there is no relationship between the rows and columns in a 2×2 table.

Note, however, that because we have done *two* tests on the same data, we must use the Bonferroni inequality to adjust the P values to account for the fact that we are doing multiple tests. Since we did two tests, we multiply the nominal 1 percent P value obtained from Table 5-7 by 2 to obtain $2(1) = 2$ percent.* Therefore, we conclude that the joggers and runners did not differ in their medical consultations from each other but did differ from the women in the control group ($P < .02$).

THE FISHER EXACT TEST

The χ^2 test can be used to analyze 2×2 contingency tables when each cell has an expected frequency of at least 5. In small studies, when the expected frequency is smaller than 5, *the Fisher exact test* is the appro-

Table 5-9 Physician Consultation among Women Who Did and Did Not Run*

Group	Yes	No	Total
Controls	14 (22.58)	40 (31.42)	54
Joggers and runners	55 (46.42)	56 (64.58)	111
Total	69	96	165

*Numbers in parentheses are expected frequencies of physician consultation if whether a woman ran or did not affect the likelihood of her consulting a physician for a menstrual problem.

*We could also use a Holm procedure to account for multiple comparisons.

priate procedure. This test turns the liability of small sample sizes into a benefit. When the sample sizes are small, it is possible to simply *list* all the possible arrangements of the observations, then compute the exact probabilities associated with each possible arrangement of the data. The total (two-tailed) probability of obtaining the observed data or more extreme patterns in the data is the P value associated with the hypothesis that the rows and columns in the data are independent.

The Fisher exact test begins with the fact that the probability of observing any given pattern in the 2×2 contingency table with the observed row and column totals in Table 5-10 is

$$P = \frac{\dfrac{R_1!R_2!C_1!C_2!}{N!}}{O_{11}!O_{12}!O_{21}!O_{22}!}$$

where O_{11}, O_{12}, O_{21}, and O_{22} are the observed frequencies in the four cells of the contingency table, C_1 and C_2 are the sums of the two columns, R_1 and R_2 are the sums of the two rows, N is the total number of observations, and the exclamation mark "!" indicates the factorial operator.*

Unlike the χ^2 test statistic, there are one- and two-tailed versions of the Fisher exact test. Unfortunately, most descriptions of the Fisher exact test simply describe the one-tailed version and many computer programs compute the one-tailed version without clearly identifying it as such. Because many researchers do not recognize this issue, results (i.e., P values) may be reported for a single tail without the researchers realizing it. To determine whether or not investigators recognized whether they were using one- or two-tailed Fisher exact tests, W. Paul McKinney and colleagues† examined the use of the Fisher exact test in

Table 5-10 Notation for the Fisher Exact Test

			Row Totals
	O_{11}	O_{12}	R_1
	O_{21}	O_{22}	R_2
Column Totals	C_1	C_2	N

*The definition $n!$ is $n! = (n)(n-1)(n-2) \cdots (2)(1)$; e.g., $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$.

†W. P. McKinney, M. J. Young, A. Harta, and M. B. Lee, "The Inexact Use of Fisher's Exact Test in Six Major Medical Journals," *JAMA*, **261**:3430–3433, 1989.

Table 5-11 Reporting of Use of Fisher Exact Test in the *New England Journal of Medicine* and *The Lancet*

	Test Identified?		
Group	Yes	No	Totals
New England Journal of Medicine	1	8	9
The Lancet	10	4	14
Totals	11	12	23

papers published in the medical literature to see whether or not the authors noted the type of Fisher exact test that was used. Table 5-11 shows the data for the two journals, *New England Journal of Medicine* and *The Lancet.* Because the numbers are small, χ^2 is not an appropriate test statistic. From the equation above, the probability of obtaining the pattern of observations in Table 5-11 for the given row and column totals is

$$P = \frac{\frac{9!14!11!12!}{23!}}{1!8!10!4!} = .00666$$

Thus, it is very unlikely that *this particular* table would be observed. To obtain the probability of observing a pattern in the data this extreme *or more extreme* in the direction of the table, reduce the smallest observation by 1, and recompute the other cells in the table to maintain the row and column totals constant.

In this case, there is one more extreme table, given in Table 5-12. This table has a probability of occurring of

$$P = \frac{\frac{9!14!11!12!}{23!}}{9!0!3!11!} = .00027$$

(Note that the numerator only depends on the row and column totals associated with the table, which does not change, and so only needs to be computed once.) Thus, the one-tailed Fisher exact test yields a P value of $P = .00666 + .00027 = .00695$. This probability represents the probability of obtaining a pattern of observations as extreme or more extreme in one direction as the actual observations in Table 5-11.

Table 5-12 More Extreme Pattern of Observations in Table 5-11, Using Smallest Observed Frequency (in This Case, 1)

	Test Identified?		
Group	Yes	No	Totals
New England Journal of Medicine	0	9	9
The Lancet	11	3	14
Totals	11	12	23

To find the other tail, we list all the remaining possible patterns in the data that would give the same row and column totals. These possibilities, together with the associated probabilities, appear in Table 5-13. These tables are obtained by taking each of the remaining three elements in Table 5-11 one at a time and progressively making it smaller by one, then eliminating the duplicate tables. Two of these tables have probabilities at or below the probability of obtaining the original observations, .00666: the ones with probabilities of .00242 and .00007. These two tables constitute the "other" tail of the Fisher exact test. There is a total probability of being in this table of .00242 + .00007 = .00249.* Thus, the total probability of obtaining a pattern of observations as extreme or more extreme than that observed is $P = .00695 + .00249 = .00944$, and we conclude there is a significant difference in the correct presentation of the Fisher exact test in the *New England Journal of Medicine* and *The Lancet* ($P = .009$). Indeed, it is important when reading papers that use the Fisher exact test to make sure the authors know what they are doing and report the results appropriately.

Let us now sum up how to do the Fisher exact test.

- *Compute the probability associated with the observed data.*
- *Identify the cell in the contingency table with the smallest frequency.*
- *Reduce the smallest element in the table by 1, then compute the elements for the other three cells so that the row and column sums remain constant.*

*Note that the two tails have different probabilities; this is generally the case. The one exception is when either the two rows or two columns have the same sums, in which case the two-tail probability is simply twice the one-tail probability. Some books say that the two-tail value of P is always simply twice the one-tail value. This is not correct unless the row or column sums are equal.

Table 5-13 Other Patterns of Observations in Table 5-11 with the Same Row and Column Totals

			Totals				Totals
	2	7	9		6	3	9
	9	5	14		5	9	14
Totals	11	12	23	Totals	11	12	23
	$P = .05330$				$P = .12438$		
			Totals				Totals
	3	6	9		7	2	9
	8	6	14		4	10	14
Totals	11	12	23	Totals	11	12	23
	$P = .18657$				$P = .02665$		
			Totals				Totals
	4	5	9		8	1	9
	7	7	14		3	11	14
Totals	11	12	23	Totals	11	12	23
	$P = .31983$				$P = .00242$		
			Totals				Totals
	5	4	9		9	0	9
	6	8	14		2	12	14
Totals	11	12	23	Totals	11	12	23
	$P = .27985$				$P = .00007$		

- *Compute the probability associated with the new table.*
- *Repeat this process until the smallest element is zero.*
- *List the remaining tables by repeating this process for the other three elements.* List each pattern of observations only once.*
- *Compute the probabilities associated with each of these tables.*
- *Add all the probabilities together that are equal to or smaller than the probability associated with the observed data.*

This probability is the *two-tail* probability of observing a pattern in the data as extreme or more extreme than observed. Many computer

*Many of these computations can be avoided: see Appendix A.

programs show *P* values for the Fisher exact test, without clearly indicating whether they are one- or two-tail values. Make sure that you know which value is being reported before you use it in your work; the two-tailed *P* value is generally the one you want.

MEASURES OF ASSOCIATION BETWEEN TWO NOMINAL VARIABLES*

In addition to testing whether there are significant differences between two rates or proportions, people often want a measure of the strength of association between some event and different treatments or conditions, particularly in *clinical trials* and *epidemiological studies.* In a *prospective* clinical trial, such as the study of thrombus formation in people treated with aspirin or placebo discussed earlier in this chapter (Table 5-1), investigators randomly assign people to treatment (aspirin) or control (placebo), then follow them to see whether they develop a thrombus or not. In that example, 32 percent (6 out of 19) of the people receiving aspirin developed thrombi and 72 percent (18 out of 25) receiving placebo developed thrombi. These proportions are estimates of the probability of developing a thrombus associated with each of these treatments; these results indicate that the probablity of developing a thrombus was cut by more than half by treatment with aspirin. We will now examine different ways to quantify this effect, *relative risk* and the *odds ratio.*†

Prospective Studies and Relative Risk

We quantify the size of the association between treatment and outcome with the *relative risk,* RR, which is defined as

$$\text{RR} = \frac{\text{Probability of event in } \textit{treatment} \text{ group}}{\text{Probability of event in } \textit{control} \text{ group}}$$

*In an introductory course, this section can be skipped without loss of continuity.

†Another way to quantify this difference is to present the *absolute risk reduction,* which is simply the difference of the probability of an event (in this case, a thrombus) without and with the treatment, .72 − .32 = .40. Treatment with aspirin reduces the probablity of a thrombus by .40. An alternative approach is to present the *number needed to treat,* which is the number of people that would have to be treated to avoid one event. The number needed to treat is simply 1 divided by the absolute risk reduction, in this case 1/.40 = 2.5. Thus, one would expect to avoid one thrombotic event for about every 2.5 people treated (or, if you prefer dealing with whole people, 2 events for every 5 people treated).

For the aspirin trial,

$$\text{RR} = \frac{\hat{p}_{asp}}{\hat{p}_{pla}} = \frac{0.32}{0.72} = 0.44$$

The fact that the relative risk is less than 1 indicates that aspirin reduces the risk of a thrombus. In clinical trials evaluating treatments against placebo (or standard treatment, when it would be unethical to administer a placebo), a relative risk less than 1 indicates that the treatment leads to better outcomes.

In an *epidemiological study,* the probability of an event among people *exposed* to some potential toxin or risk factor is compared to people who are *not exposed.* The calculations are the same as for clinical trials.* Relative risks greater than 1 indicate that exposure to the toxin *increases* the risk of disease. For example, being married to a smoker is associated with a relative risk of heart disease in nonsmokers of 1.3,† indicating that nonsmokers married to smokers are 1.3 times more likely to die from heart disease as nonsmokers married to nonsmokers (and so not breathing secondhand smoke at home).

Table 5-14 shows the general layout for a calculation of relative risk; it is simply a 2×2 contingency table. The probability of an event in the treatment group (also called the *experimental event rate*) is $a/(a + b)$ and the probability of an event in the treatment group (also called the *control event rate*) is $c/(c + d)$. Therefore, the formula for relative risk is

$$\text{RR} = \frac{a/(a + b)}{c/(c + d)}$$

*In clinical trials and epidemological studies one often wants to adjust for other so-called *confounding variables* that could be affecting the probability of an event. It is possible to account for such variables using multivariate techniques using *logistic regression* or the *Cox proportional hazards model.* For a discussion of these issues, see S. A. Glantz and B. K. Slinker, *Primer of Applied Regression and Analysis of Variance,* 2 ed, New York, McGraw-Hill, 2000, chap. 12, "Regression with a Qualitative Dependent Variable."

†S. A. Glantz and W. W. Parmley. "Passive Smoking and Heart Disease: Epidemiology, Physiology, and Biochemistry," *Circulation,* **83**:1–12, 1991. S. Glantz and W. Parmley. Passive Smoking and Heart Disease: Mechanisms and Risk, *JAMA,* **273:**1047–1053, 1995.

Table 5-14 Arrangement of Data to Compute Relative Risk

	Number of people		
Sample group	Disease	No disease	Total
Treated (or exposed to risk factor)	a	b	a + b
Control (or not exposed to risk factor)	c	d	c + d
Total	a + c	b + d	

Using the results of the aspirin trial in Table 5-1, we would compute

$$\text{RR} = \frac{6/(6 + 18)}{13/(13 + 7)} = \frac{0.32}{0.72} = 0.44$$

This formula is simply a restatement of the definition presented above.

The most common null hypothesis that people wish to test related to relative risks is that the relative risk equals 1 (i.e., that the treatment or risk factor does not affect event rate). Although it is possible to test this hypothesis using the standard error of the relative risk, most people simply apply a χ^2 test to the contingency table used to compute the relative risk.*

To compute a relative risk, the data must be collected as part of a *prospective study* in which people are randomized to treatment or control or subjects in an epidemiological study† are followed forward in time after they are exposed to the toxin or risk factor of interest. It is necessary to conduct the study prospectively to estimate the absolute event rates in people in the treatment (or exposed) and control groups.

Such prospective studies are often difficult and expensive to do, particularly if it takes several years for events to occur after treatment or exposure. It is, however, possible to conduct a similar analysis *retrospectively* based on so-called *case-control studies.*

*Traditionally direct hypothesis testing of relative risks is done by examining confidence intervals; see Chap. 7.

†Prospective epidemiological studies are also called *cohort studies.*

Case-Control Studies and the Odds Ratio

Unlike prospective studies, case-control studies are done after the fact. In a case-control study, people who experienced the outcome of interest are identified and the number exposed to the risk factor of interest are counted. These people are the *cases.* You then identify people who did not experience the outcome of interest, but are similar to the cases in all relevant ways and count the number that were exposed to the risk factor. These people are the *controls.* (Often investigators include more than one control per case in order to increase the sample size.) Table 5-15 shows the layout for data from a case-control study.

This information can be used to compute a statistic similar to the relative risk known as the *odds ratio.* The odds ratio, OR, is defined as

$$\text{OR} = \frac{\text{Odds of exposure in } \textit{cases}}{\text{Odds of exposure in } \textit{controls}}$$

The percentage of cases (people with the disease) exposed to the risk factor is $a/(a + c)$ and the percentage of cases not exposed to the risk factor is $c/(a + c)$. (Note that the denominators in each case are appropriate for the numerators; this situation would not exist if one was using case-control data to compute a relative risk.) The odds of disease in the cases is the ratio of these two percentages.

$$\text{Odds of disease in } \textit{cases} = \frac{a/(a + c)}{c/(a + c)} = \frac{a}{c}$$

Table 5-15 Arrangement of Data to Compute Odds Ratio

	Number of People	
Sample group	Disease "cases"	No disease "controls"
Exposed to risk factor (or treatment)	a	b
Not exposed to risk factor (or treatment)	c	d
Total	a + c	b + d

Likewise, the odds of disease in the controls is

$$\text{Odds of disease in } \textit{controls} = \frac{b/(b+d)}{d/(b+d)} = \frac{b}{d}$$

Finally, the odds ratio is

$$\text{OR} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

Because the number of controls depends on how the investigator designs the study (and so b and d in Table 5-15), you cannot use data from a case-control study to compute a relative risk. In a case-control study the investigator decides how many subjects with and without the disease will be studied. This is the opposite of the situation in prospective studies (clinical trials and epidemiological cohort studies), when the investigator decides how many subjects with and without the risk factor will be included in the study. The odds ratio may be used in both case-control and prospective studies, but *must* be used in case-control studies.

While the odds ratio is distinct from the relative risk, the odds ratio is a reasonable estimate of the relative risk when the number of people with the disease is small compared to the number of people without the disease.*

As with the relative risk, the most common null hypothesis that people wish to test related to relative risks is that the odds ratio equals 1 (i.e., that the treatment or risk factor does not affect the event rate). While it is possible to test this hypothesis using the standard error of the odds ratio, most people simply apply a χ^2 test to the contingency table used to compute the odds ratio.†

*In this case, the number of people who have the disease, a and c, is much smaller than the number of people without the disease, b and d, so $a + b \approx b$ and $c + d \approx d$. As a result

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

†Direct hypothesis testing regarding odds ratios is usually done with confidence intervals; see Chapter 7.

Passive Smoking and Breast Cancer

Breast cancer is the second leading cause of cancer death among women (behind lung cancer). Smoking causes lung cancer because of the cancer-causing chemicals in the smoke that enter the body and some of these chemicals appear in breast milk, indicating that they reach the breast. To examine whether exposure to secondhand tobacco smoke increased the risk of breast cancer in lifelong nonsmokers, Johnson and colleagues* conducted a case-control study using cancer registries to identify premenopausal women with histologically confirmed invasive primary breast cancer. They contacted the women and interviewed them about their smoking habits and exposure to secondhand smoke at home and at work. They obtained a group of controls who did not have breast cancer, matched by age group, from a mailing to women using lists obtained from the provincial health insurance authorities. Table 5-16 shows the resulting data.

The fraction of women with breast cancer (cases) who were exposed to secondhand smoke is $50/(50 + 14) = 0.781$ and the fraction of women with breast cancer not exposed to secondhand smoke was $14/(50 + 14) = 0.218$, so the odds of the women with breast cancer having been exposed to secondhand smoke is $0.781/0.218 = 3.58$. Similarly, the fraction of controls exposed to secondhand smoke is $43/(43 + 35) = 0.551$ and the fraction not exposed to secondhand smoke is $35/(43 + 35) = 0.449$, so the odds of the women without

Table 5-16 Passive Smoking and Breast Cancer

	Number of People	
Sample group	Cases (breast cancer)	Controls
Exposed to secondhand smoke	50	43
Not exposed to secondhand smoke	14	35
Total	64	78

*K. C. Johnson, J. Hu, Y. Mao, and the Canadian Cancer Registries Epidemiology Research Group, "Passive and Active Smoking and Breast Cancer Risk in Canada, 1994–1997," *Cancer Causes Control,* **11**:211–221, 2000.

breast cancer having been exposed to secondhand smoke is 0.551/0.449 = 1.23. Finally, the odds ratio of breast cancer associated with secondhand smoke exposure is

$$\text{OR} = \frac{\begin{array}{c}\text{Odds of secondhand smoke exposure} \\ \text{in women with breast cancer}\end{array}}{\begin{array}{c}\text{Odds of secondhand smoke} \\ \text{exposure in controls}\end{array}} = \frac{3.58}{1.23} = 2.91$$

Alternatively, we could use the direct formula for odds ratio and compute

$$\text{OR} = \frac{ad}{bc} = \frac{50 \cdot 35}{14 \cdot 43} = 2.91$$

Based on this study, we conclude that exposure to secondhand smoke increases the odds of having breast cancer by 2.91 times among this population. A χ^2 analysis of the data in Table 15-16 shows that this difference is statistically significant ($P = .007$).

We now have the tools to analyze data measured on a nominal scale. So far we have been focusing on how to demonstrate a difference and quantify the certainty with which we can assert this difference or effect with the P value. Now we turn to the other side of the coin: What does it mean if the test statistic is *not* big enough to reject the hypothesis of no difference?

PROBLEMS

5-1 Because local dental anesthetic administration is often accompanied by patient anxiety and other adverse sequelae, Timothy Bishop ("High Frequency Neural Modulation in Dentistry," *J. Am. Dent. Assoc.*, **112**:176–177, 1986) studied the effectiveness of high frequency neural modulation (similar to that used in physical therapy to control chronic pain) to prevent pain during

	Treatment received	
	Active	Placebo
Effective analgesia	24	3
Ineffective analgesia	6	17

a variety of dental procedures, including restorations and tooth extractions. People were given either the electrical stimulation or simply had the inactive device attached as though it were being used (a placebo control). Neither the dentist nor the patient knew whether or not the neural modulator was turned on. Do the following data suggest that high-frequency neural modulation is an effective analgesic? Use both χ^2 and z statistics.

5-2 Adolescent suicide is commonly associated with alcohol misuse. In a retrospective study involving Finish adolescents who committed suicide, Sami Pirkola and colleagues ("Alcohol-Related Problems Among Adolescent Suicides in Finland," *Alcohol Alcohol.* **34**:320–328, 1999) compared situational factors and family background between victims who abused alcohol and those who did not. Alcohol use was determined by family interview several months following the suicide. Adolescents with alcohol problems, ranging from mild to severe, were classified together in a group called SDAM (Subthreshold or Diagnosable Alcohol Misuse) and compared to victims with no such reported alcohol problems. Some of Pirkola's findings appear below. Use these data to identify the characteristics of SDAM suicides. Are these factors specific enough to be of predictive value in a specific adolescent? Why or why not?

Factor	SDAM group (n = 44)	Not in SDAM group (n = 62)
Violent death (shooting, hanging, jumping, traffic)	32	51
Suicide under influence of alcohol	36	25
Blood alcohol concentration (BAC) $\geq$ 150 mg/dL	17	3
Suicide during weekend	28	26
Parental divorce	20	15
Parental violence	14	5
Parental alcohol abuse	17	12
Paternal alcohol abuse	15	9
Parental suicidal behavior	5	3
Institutional rearing	6	2

5-3 The 106 suicides analyzed in Prob. 5-2 were selected from 116 suicides that occurred between April 1987 and March 1988. Eight of the ten suicides not included in the study were due to lack of family interviews. Discuss the potential problems, if any, associated with these exclusions.

5-4 Major depression can be treated with medication, psychotherapy or a combination of the two. M. Keller and colleagues ("A Comparison of Nefazodone, the Cognitive Behavioral-Analysis System of Psychotherapy, and Their Combination for the Treatment of Chronic Depression," *N. Engl. J. Med.*, **342**:1462–1470, 2000) compared the efficacy of these approaches in outpatients diagnosed with a chronic major depressive disorder. Depression was diagnosed using the 24-item Hamilton Rating Scale for Depression, where a higher score indicates more severe depression. All subjects began the study with a score of at least 20. The investigators randomly assigned patients who met study criteria to the three groups—medication (nefazodone), psychotherapy, or both—for 12 weeks then measured remission, defined as having a follow-up score of 8 or less after 10 weeks of treatment. The responses of the people they studied fell into the following categories.

Treatment	Remission	No remission
Nefazodone	36	131
Psychotherapy	41	132
Nefazodone and psychotherapy	75	104

Is there any evidence that the different treatments produced different responses? If so, which one seems to work best?

5-5 Public health officials often investigate the source of widespread outbreaks of disease. Agnes O'Neil and coworkers ("A Waterborne Epidemic of Acute Infectious Non-Bacterial Gastroenteritis in Alberta, Canada." *Can. J. Public Health*, **76**:199–203, 1985) reported on an outbreak of gastroenteritis in a small Canadian town. They hypothesized that the source of contamination was the municipal water supply. They examined the association between amount of water consumed and the rate at which people got sick. What do these data suggest?

Water consumption, glasses per day	Number ill	Number not ill
Less than 1	39	121
1 to 4	265	258
5 or more	265	146

5-6 In general, the quality of a research project is higher and the applicability of data to a specific question is higher if the data are collected *after* the research is planned. Robert Fletcher and Suzanne Fletcher ("Clinical Research in General Medical Journals: A 30-Year Perspective," *N. Engl. J. Med.*, **301**:180–183, 1979, used by permission) studied 612 articles randomly selected from the *Journal of the American Medical Association, The Lancet,* and *New England Journal of Medicine* to see whether the authors collected their data before or after planning the research. They found:

	1946	1956	1966	1976
No. of articles examined	151	149	157	155
Data collection, %:				
After research planned	76	71	49	44
Before research planned	24	29	51	56

Estimate the certainty with which these percentages estimate the true percentage of articles in which the data were collected before the research was planned. Have things changed over time? If so, when? Was the change (if any) for the better or the worse?

5-7 Dioxin is one of the most toxic synthetic environmental contaminants. An explosion at a herbicide plant in Sevaso, Italy in 1976 released large amounts of this long-lasting contaminant into the environment. Because exposure to dioxin during development is known to be dangerous, researchers have been carefully following the health status of exposed people and their children in Sevaso and surrounding areas. Peter Mocarelli and colleagues ("Paternal Concentrations of Dioxin and Sex Ratio of Offspring," *The Lancet*, **355**:1858–1863, 2000) measured the serum concentration of dioxin in potentially exposed parents and analyzed the number of male and female babies born after 1976. They found that when both parents were exposed to greater that 15 parts per trillion (ppt) of dioxin the proportion of girl babies born was significantly increased compared to couples not exposed to this amount of dioxin. Mocarelli and colleagues also investigated whether there were differences in the proportion of female babies born if only one parent was exposed to greater than 15 ppt of dioxin and whether the sex of the parent (mother or father) made a difference. Based on numbers presented below, are there differences in the proportion of

female babies born when only one parent is exposed to greater than 15 ppt of dioxin?

Parental exposure to dioxin	Female babies	Male babies
Father exposed; mother unexposed	105	81
Father unexposed; mother exposed	100	120

5-8 Fabio Lattanzi and coworkers ("Inhibition of Dipyridamole-Induced Ischemia by Antianginal Therapy in Humans: Correlation with Exercise Electrocardiography," *Circulation,* **83**:1256–1262, 1992) wished to compare the ability of electrocardiography (in which the electrical signals produced by the heart) and echocardiography (in which pictures of the heart are obtained with sound waves) to detect inadequate oxygen supply (ischemia) to the heart of people with heart disease. To obtain the electrocardiographic test (EET), they had the people exercise to increase their heart rate until they developed chest pain or had abnormalities on their electrocardiogram indicating ischemia. To obtain the echocardiographic test (DET), they watched the heartbeat after increasing heart rate with the drug dipyridamole. They compared people receiving therapy for their heart disease in separate experiments. The results they obtained were:

	On therapy	
	Positive DET test	Negative DET test
Positive EET Test	38	2
Negative EET Test	14	3

	Off therapy	
	Positive DET test	Negative DET test
Positive EET Test	21	6
Negative EET Test	16	14

Was there a different response between the two tests in either group of patients?

5-9 David Sackett and Michael Gent ("Controversy in Counting and Attributing Events in Clinical Trials," *N. Engl. J. Med.*, **301**: 1410–1412, 1979, used by permission) took note of two important points with regard to the study described in Prob. 5-5: (1) "available for follow-up" patients had to be discharged alive and free of stroke after their hospitalization; (2) this procedure excluded 15 surgically treated patients (5 who died and 10 who had strokes during or shortly after their operations) but only 1 medically treated patient. Including these 16 patients in the data from the previous problem yields the following result.

	Recurrent ischemia, stroke, or death, no. of patients	
Therapy	Yes	No
Surgical	58	36
Medical	54	19

Does including these patients change the conclusions of the trial? If so, should the trial be analyzed excluding them (as in Prob. 5-5) or including them (as in this problem)? Why?

5-10 The chance of contracting disease X is 10 percent, regardless of whether or not a given individual has disease A or disease B. Assume that you can diagnose all three diseases with perfect accuracy and that in the entire population 1000 people have disease A and 1000 have disease B. People with X, A, and B have different chances of being hospitalized. Specifically, 50 percent of the people with A, 20 percent of the people with B, and 40 percent of the people with X are hospitalized. Then:

- Out of the 1000 people with A, 10 percent (100 people) also have X; 50 percent (50 people) are hospitalized because they have A. Of the remaining 50 (who also have X), 40 percent (20 people) are hospitalized because of X. Therefore, 70 people will be hospitalized with both A and X.
- Out of the 900 people with A but not X, 50 percent are hospitalized for disease A (450 people).
- Out of the 1000 with B, 10 percent (100 people) also have X; 20 percent (20 people) are hospitalized because of B, and of the remaining 80, 40 percent (32 patients) are hospitalized because they have X. Thus, 52 people with B and X are in the hospital.

- Of the 900 with B but not X, 20 percent (180 people) are hospitalized because they have disease B.

A hospital-based investigator will encounter these patients in the hospital and observe the following relationship:

	Disease X	No disease X
Disease A	70	450
Disease B	52	180

Is there a statistically significant difference in the chances that an individual has X depending on whether or not he has A or B in the sample of patients the hospital-based investigator will encounter? Would the investigator reach the same conclusion if she could observe the entire population? If not, explain why. (This example is from D. Mainland, "The Risk of Fallacious Conclusions from Autopsy Data on the Incidence of Diseases with Applications to Heart Disease," *Am. Heart J.*, **45**:644–654, 1953.)

5-11 Cigarette smoking is associated with increased incidence of many types of cancers. Jian-Min Yuan and colleagues ("Tobacco Use in Relation to Renal Cell Carcinoma." *Cancer Epidemiol. Biomarkers Prev.* **7**:429–433, 1998) wanted to investigate whether cigarette smoking was also associated with increased risk of renal cell cancer. They recruited patients with renal cell cancer from the Los Angeles County Cancer Surveillance Program to serve as cases in a retrospective case-control study. Control subjects without renal cell cancer were matched on sex, age (within 5 years), race, and neighborhood of residence to each case subject. After recruiting a total of 2314 subjects for the study, Yuan and colleagues visited subjects in their homes and interviewed them about their smoking habits, both past and present. What effect does smoking cigarettes have on the risk of developing renal cell cancer?

	Number of people	
	Renal cell cancer	No cancer
Ever smoked cigarettes	800	713
Never smoked cigarettes	357	444

5-12 Yuan and colleagues also collected information from subjects who had quit smoking. Is there any evidence that stopping smoking reduces risk of developing renal cell cancer compared to current smokers?

	Number of people	
	Renal cell cancer	No cancer
More than 20 years since quitting	169	177
Current smokers	337	262

5-13 Many postmenopausal women are faced with the decision of whether they want to take hormone replacement therapy or not. Benefits of hormone replacement include decreased risk of cardiovascular disease and osteoporosis. However, hormone replacement therapy has also been associated with increased risk of breast cancer and endometrial cancer. Francine Grodstein and colleagues ("Postmenopausal Hormone Therapy and Mortality." *N. Engl. J. Med.*, **336**: 1769–1775, 1997) investigated the relationship between hormone replacement therapy and overall mortality in a large group of postmenopausal women. The women used in this study were selected from a sample of registered nurses participating in the Nurses' Health Study. This prospective study has been tracking the health status of a large group of registered nurses since 1976, updating information every 2 years. Women became eligible for Grodstein's study when they became menopausal and were included as long as they did not report a history of cardiovascular disease or cancer on the original 1976 questionnaire. Cases were those women who died between 1976–1994, and for each case, up to 10 control women were selected. Control women were those alive at the time of subjects, death, matched on age and age at menopause. Is there any evidence that the risk of death differs in women who were identified as currently using hormone replacement therapy?

	Number of people	
	Decreased	Alive
Currently using hormone replacement therapy	574	8483
Ever used hormone replacement therapy	2051	17,520

5-14 Is there an increase in risk of death in women who reported past hormone replacement therapy use compared to women who never used it?

	Number of people	
	Deceased	Alive
Past use of hormone replacement therapy	1012	8621
Never used hormone replacement therapy	2051	17,520

Chapter 6

What Does "Not Significant" Really Mean?

Thus far, we have used statistical methods to reach conclusions by seeing how compatible the observations were with the hypothesis that the treatment had no effect (the null hypothesis). When the data were unlikely to occur if this hypothesis were true, we rejected it and concluded that the treatment had an effect. We used a test statistic (F, t, q, q', z, or χ^2) to quantify the difference between the actual observations and those we would expect if the hypothesis of no effect were true. We concluded that the treatment had an effect if the value of this test statistic was bigger than 95 percent of the values that would occur if the treatment had no effect. When this is so, it is common for medical investigators to report a *statistically significant* effect. On the other hand, when the test statistic is not big enough to reject the hypothesis of no treatment effect, investigators often report *no statistically significant difference* and then discuss their results as if they had proved that the treatment had no effect. *All they really did was fail to demonstrate that it did have an effect.* The distinction between positively demonstrating that a

treatment had no effect and failing to demonstrate that it does have an effect is subtle but very important, especially in the light of the small numbers of subjects included in most clinical studies.*

As already mentioned in our discussion of the *t* test, the ability to detect a treatment effect with a given level of confidence depends on the size of the treatment effect, the variability within the population, and the size of the samples used in the study. Just as bigger samples make it more likely that you will be able to detect an effect, smaller sample sizes make it harder. In practical terms, this means that studies of therapies that involve only a few subjects and fail to reject the hypothesis of no treatment effect may arrive at this result because the statistical procedures lacked the *power* to detect the effect because of a too small sample size, even though the treatment did have an effect. Conversely, considerations of the power of a test permit you to compute the sample size needed to detect a treatment effect of given size that you believe is present.

AN EFFECTIVE DIURETIC

Now, we make a radical departure from everything that has preceded: we assume that the treatment *does* have an effect.

Figure 6-1 shows the same population of people we studied in Fig. 4-4 except that this time the drug given to increase daily urine production works. It increases the average urine production for members of this population from 1200 to 1400 mL/day. Figure 6-1*A* shows the distribution of values of daily urine production for all 200 members of the population under control conditions, and Fig. 6-1*B* shows how much urine every member of the population on the diuretic would produce.

Of course, an investigator cannot observe all members of the population, so he or she selects two groups of 10 people at random, gives one group the diuretic and the other a placebo, and measures their daily urine production. Figure 6-1*C* shows what the investigator would see.

*This problem is particularly encountered in small clinical studies in which there are no "failures" in the treatment group. This situation often leads to overly optimistic assessments of therapeutic efficacy. See J. A. Hanley and A. Lippman-Hand, "If Nothing Goes Wrong, Is Everything All Right? Interpreting Zero Numerators," *JAMA* **249:** 1743–1745, 1983.

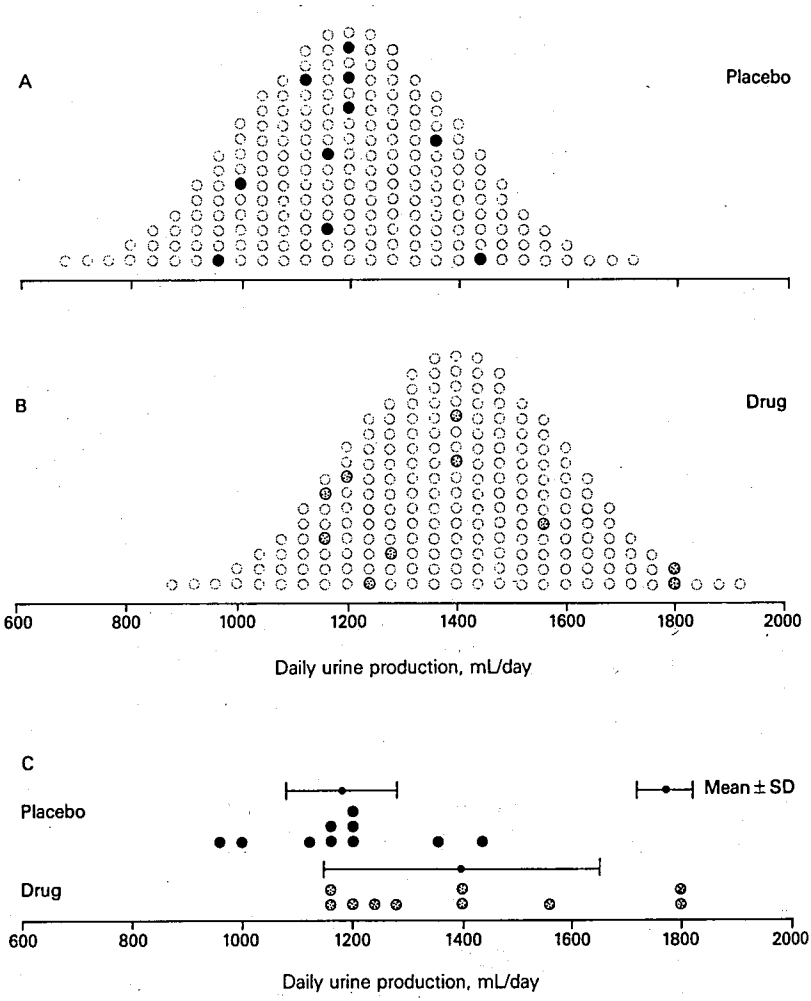


Figure 6-1 Daily urine production in a population of 200 people while they are taking a placebo and while they are taking an effective diuretic that increases urine production by 200 mL/day on the average. Panels *A* and *B* show the specific individuals selected at random for study. Panel *C* shows the results as they would appear to the investigator. $t = 2.447$ for these observations. Since the critical value of t for $P < .05$ with $2(10 - 1) = 18$ degrees of freedom is 2.101, the investigator would probably report that the diuretic was effective.

The people receiving a placebo produced an average of 1180 mL/day, and those receiving the drug produced an average of 1400 mL/day. The standard deviations of these two samples are 144 and 245 mL/day, respectively. The pooled estimate of the population variance is

$$s^2 = \tfrac{1}{2}(s_{dr}^2 + s_{pla}^2) = \tfrac{1}{2}(245^2 + 144^2) = 40{,}381 = 201^2$$

The value of t associated with these observations is

$$t = \frac{\overline{X}_{dr} - \overline{X}_{pla}}{\sqrt{(s^2/n_{dr}) + (s^2/n_{pla})}} = \frac{1400 - 1180}{\sqrt{(201^2/10) + (201^2/10)}} = 2.447$$

which exceeds 2.101, the value that defines the most extreme 5 percent of possible values of the t statistic when the two samples are drawn from the same population. [There are $2(n - 1) = 18$ degrees of freedom.] The investigator would conclude that the observations are not consistent with the assumption that two samples came from the same population and report that the drug increased urine production. And he/she would be right.

Of course, there is nothing special about the two random samples of people selected for the experiment. Figure 6-2 shows two more groups of people selected at random to test the drug, together with the results as they would appear to the investigator. In this case, the mean urine production was 1216 mL/day for the people given the placebo and 1368 mL/day for the people taking the drug. The standard deviation of urine production in the two groups was 97 and 263 mL/day, respectively, so the pooled estimate of the variance is $\tfrac{1}{2}(97^2 + 236^2) = 198^2$. The value of t associated with these observations is

$$t = \frac{1368 - 1216}{\sqrt{(198^2/10) + (198^2/10)}} = 1.71$$

which is less than 2.101. Had the investigator selected these two groups of people for testing, he/she would not have obtained a value of t large enough to reject the hypothesis that the drug had no effect; it would probably be reported "no significant difference." If the investigator went on to conclude that the drug had no effect, he/she would be wrong.

Notice that this is a different type of error from that discussed in Chapters 3 to 5. In the earlier chapters we were concerned with

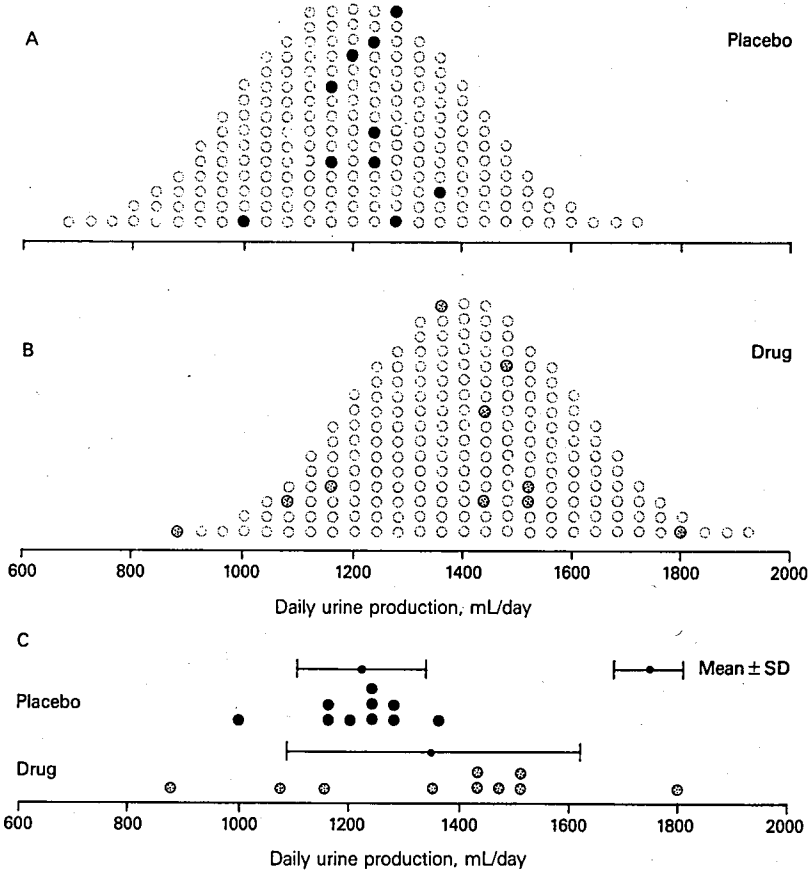


Figure 6-2 There is nothing special about the two random samples shown in Fig. 6-1. This illustration shows another random sample of two groups of 10 people each selected at random to test the diuretic and the results as they would appear to the investigator. The value of t associated with these observations is only 1.71, not great enough to reject the hypothesis of no drug effect with $P < 0.05$, that is, $\alpha = 0.05$. If the investigator reported the drug had no effect, he/she would be wrong.

rejecting the hypothesis of no effect when it was true. Now we are concerned with *not rejecting it when it is not true.*

What are the chances of making this second kind of error?

Just as we could repeat this experiment more than 10^{27} times when the drug had no effect to obtain the distribution of possible values of t

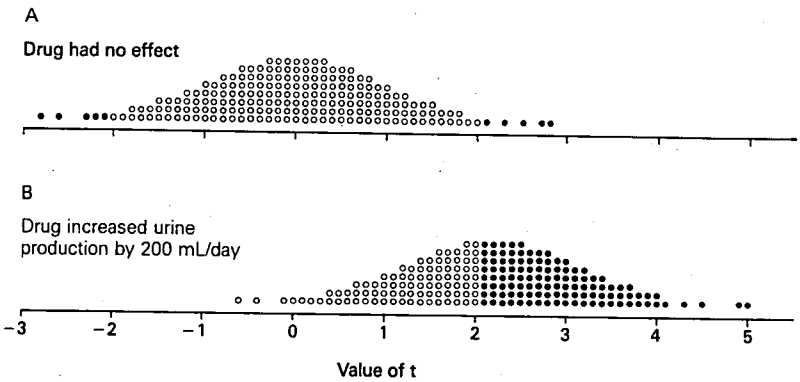


Figure 6-3 (*A*) The distribution of values of the *t* statistic computed from 200 experiments that consisted of drawing two samples of size 10 each from a single population; this is the distribution we would expect if the diuretic had no effect on urine production. (Compare with Fig. 4-5*A*.) (*B*) The distribution of *t* values from 200 experiments in which the drug increased average urine production by 200 mL/day. $t = 2.1$ defines the most extreme 5 percent of the possible values of *t* when the drug has no effect; 111 of the 200 values of *t* we would expect to observe from our data fall above this point when the drug increases urine production by 200 mL/day. Therefore, there is a 55 percent chance that we will conclude that the drug actually increases urine production from our experiment.

(compare the discussion of Fig. 4-5), we can do the same thing when the drug does have an effect. Figure 6-3 shows the results of 200 such experiments; 111 out of the resulting values of *t* fall at or above 2.101, the value we used to define a "big" *t*. Put another way, if we wish to keep the *P* value at or below 5 percent, there is a $^{111}/_{200} = 55$ percent chance of concluding that the diuretic increases urine output when average urine output actually increases by 200 mL/day. We say the *power* of the test is .55. The power quantifies the chances of detecting a real difference of a given size.

Alternatively, we could concentrate on the 89 of the 200 experiments that produced *t* values below 2.101, in which case we would fail to reject the hypothesis that the treatment had no effect and be wrong. Thus, there is a $^{89}/_{200} \doteq 45$ percent $= .45$ chance of continuing to accept the hypothesis of no effect when the drug really increased urine production by 200 mL/day on the average.

TWO TYPES OF ERRORS

Now we have isolated the two different ways the random-sampling process can lead to erroneous conclusions. These two types of errors

are analogous to the false-positive and false-negative results one obtains from diagnostic tests. Before this chapter we concentrated on controlling the likelihood of making a false-positive error, that is, concluding that a treatment has an effect when it really does not. In keeping with tradition, we have generally sought to keep the chances of making such an error below 5 percent; of course, we could arbitrarily select any cutoff value we wanted at which to declare the test statistic "big." Statisticians denote the maximum acceptable risk of this error by α, the Greek letter alpha. If we reject the hypothesis of no effect whenever $P < .05$, $\alpha = .05$, or 5 percent. If we actually obtain data that lead us to reject the hypothesis of no effect when this hypothesis is true, statisticians say that we have made a *Type I error.* They denote the chances of making such an error with α. All this logic is relatively straightforward because we have specified how much we believe the treatment affects the variable of interest, that is, not at all.

What about the other side of the coin, the chances of making a false-negative conclusion and not reporting an effect when one exists? Statisticians denote the chance of erroneously accepting the hypothesis of no effect by β, the Greek letter beta. The chance of detecting a true-positive, that is, reporting a statistically significant difference when the treatment really produces an effect, is $1 - \beta$. The *power* of the test that we discussed earlier is equal to $1 - \beta$. For example, if a test has power equal to .55, there is a 55 percent chance of actually reporting a statistically significant effect when one is really present. Table 6-1 summarizes these definitions.

Table 6-1 Types of Erroneous Conclusions in Statistical Hypothesis Testing

	Actual situation	
Conclude from observations	Treatment has an effect	Treatment has no effect
Treatment has an effect	True positive Correct conclusion $1 - \beta$	False positive Type I error α
Treatment has no effect	False negative Type II error β	True negative Correct conclusion $1 - \alpha$

WHAT DETERMINES A TEST'S POWER?

So far we have developed procedures for estimating and controlling the Type I, or α, error; now we turn our attention to keeping the Type II, or β, error as small as possible. In other words, we want the power to be as high as possible. In theory, this problem is not very different from the one we already solved with one important exception. Since the treatment has an effect, *the size of this effect influences how easy it is to detect.* Large effects are easier to detect than small ones. To estimate the power of a test, you need to specify how small an effect is worth detecting.

Just as with false-positives and false-negatives in diagnostic testing, the Type I and Type II errors are intertwined. As you require stronger evidence before reporting that a treatment has an effect, i.e., make α smaller, you also increase the chance of missing a true effect, i.e., make β bigger or power smaller. The only way to reduce both α and β simultaneously is to increase the sample size, because with a larger sample you can be more confident in your decision, whatever it is.

In other words, the power of a given statistical test depends on three interacting factors:

- *The risk of error you will tolerate when rejecting the hypothesis of no treatment effect.*
- *The size of the difference you wish to detect relative to the amount of variability in the populations.*
- *The sample size.*

To keep things simple, we will examine each of these factors separately.

The Size of the Type I Error α

Figure 6-3 showed the complementary nature of the maximum size of the Type I error α and the power of the test. The acceptable risk of erroneously rejecting the hypothesis of no effect, α, determines the critical value of the test statistic above which you will report that the treatment had an effect, $P < \alpha$. (We have usually taken $\alpha = .05$.) This critical value is defined from the distribution of the test statistic for all possible experiments with a specific sample size *given that the treatment had no effect.* The power is the proportion of possible values of the test statistic that fall above this cutoff value *given that the treatment had a specified effect* (here a 200 mL/day increase in urine production). Changing α, or

the P value required to reject the hypothesis of no difference, moves this cutoff point, affecting the power of the test.

Figure 6-4 illustrates this point further. Figure 6-4*A* essentially reproduces Fig. 6-3 except that it depicts the distribution of t values for all 10^{27} possible experiments involving two groups of 10 people as a continuous distribution. The top part, copied from Fig. 4-5*D*, shows the distribution of possible t values that would occur if the drug did not affect urine production. Suppose we require $P < .05$ before we are willing to assert that the observations were unlikely to have arisen from random sampling rather than the effect of the drug. In other words, we make $\alpha = .05$, in which case -2.101 and $+2.101$ delimit the most extreme 5 percent of all possible t values we would expect to observe if the diuretic did not affect urine production.

We know, however, that the drug actually increased average urine production by 200 mL/day. Therefore, we do not expect the distribution of possible t values associated with our experiment to be given by the distribution at the top of the figure. It will be centered not on zero but above zero (because we expect $\overline{X}_{dr} - \overline{X}_{pla}$ to average around 200 mL/day). The lower distribution in Fig. 6-4*A* shows the actual distribution of possible t values associated with our experiment; 55 percent of these possible values of t, that is, 55 percent of the area under the curve, fall above the 2.101 cutoff, so we say the power of the test is .55. In other words, if the drug increases average urine production by 200 mL/day in this population and we do an experiment using two samples of 10 people each to test the drug, there is a 55 percent chance that we will conclude that the drug is effective ($P < .05$). Conversely, we can say that β, the likelihood that we will make the false-negative, or Type II, error and accept the hypothesis of no effect when it is not true, is $1 - .55 = .45 = 45$ percent.

Now look at Fig. 6-4*B*. The two distributions of t values are identical to those in Fig. 6-4*A*. (After all, the drug's true effect is still the same.) This time, however, we will insist on stronger evidence before concluding that the drug actually increased urine production. We will require that the test statistic fall in the most extreme 1 percent of possible values before concluding that the data are inconsistent with the null hypothesis that the drug has no effect. Thus, $\alpha = .01$ and t must exceed -2.878 or $+2.878$ to fall in the most extreme 1 percent of values. The top part of panel *B* shows this cutoff point. From the actual distribution of t values in the lower part of Fig. 6-4*B*, we see that only 45 percent of

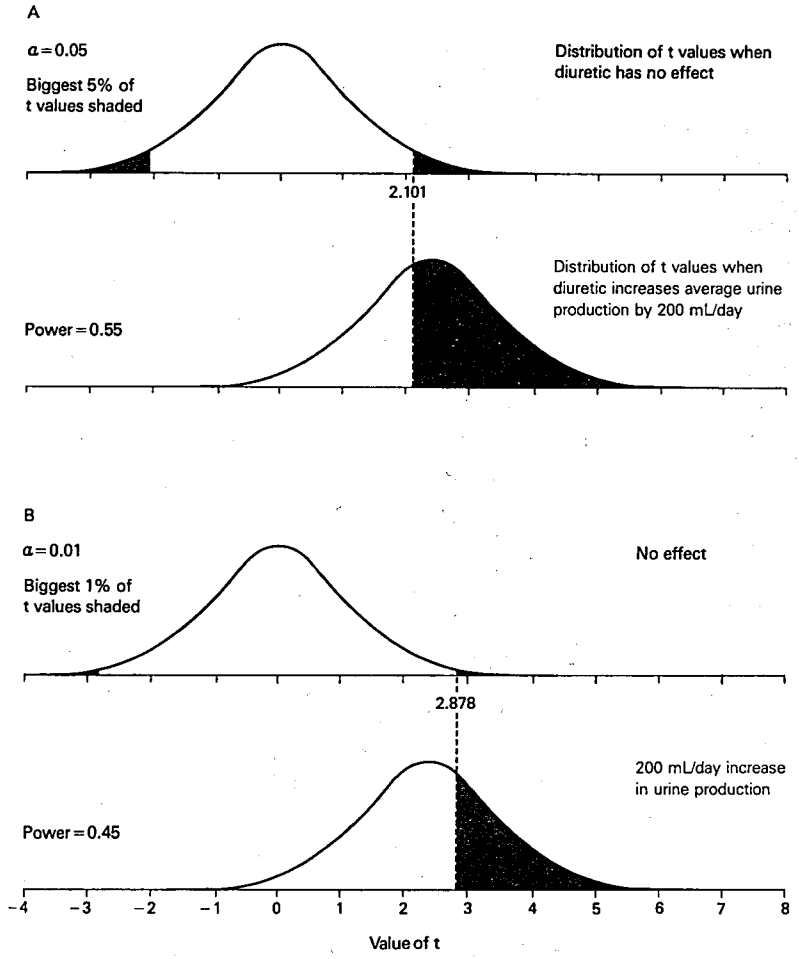


Figure 6-4 (*A*) $t = 2.101$ defines the 5 percent most extreme values of the t statistic we would expect to be associated with our experiment if the drug did not increase urine production. Of the possible values of the t statistic that will actually be associated with our experiment, 55 percent fall above this point because the drug increased urine production by 200 mL/day on the average. (*B*) If we increase the confidence with which we wish to reject the hypothesis that the drug had no effect (make α smaller), the proportion of possible t values that will actually arise from the experiment also goes down, so the power of the test to detect an effect of a given size decreases. For example, if we change α from .05 to .01, the power of the test drops from 55 to 45 percent because the cutoff for a "big" t increases from 2.101 to 2.878.

them fall above 2.878, so the power of the test has fallen to .45. In other words, there is less than an even chance that we will report that the drug is effective even though it actually is.

By requiring stronger evidence that there be a treatment effect before reporting it we have decreased the chances of erroneously reporting an effect (the Type I error), but we have increased the chances of failing to detect a difference when one actually exists (the type II error) because we decreased the power of the test. This trade-off always exists.

The Size of the Treatment Effect

We just demonstrated that the power of a test decreases as we reduce the acceptable risk of making a Type I error, α. The entire discussion was based on the fact that the drug increased average urine production by 200 mL/day, from 1200 to 1400 mL/day. Had this change been different, the actual distribution of t values connected with the experiment also would have been different. In other words, the power of a test depends on the size of the difference to be detected.

Let us consider three specific examples. Figure 6-5*A* shows the t distribution (the distribution of possible values of the t statistic) for a sample size of 10 if the diuretic had no effect and the two treatment groups could be considered two random samples drawn from the same population. The most extreme 5 percent of the values are shaded, just as in Fig. 6-4. Figure 6-5*B* shows the distribution of t values we would expect if the drug increased urine production an average of 200 mL/day over the placebo; 55 percent of the possible values are beyond -2.101 or $+2.101$, so the power of the test is .55. (So far we are just recapitulating the results in Fig. 6-4). Figure 6-5*C* shows the distribution of t values that would occur if the drug increased urine production only by 100 mL/day on the average. Now only 17 percent of the possible values (corresponding to 17 percent of the area under the curve) fall above 2.101, that is, the power of the test to detect a difference of only 100 mL/day is only 0.17. In other words, there is less than 1 chance in 5 that doing a study of two groups of 10 people would detect a change in urine production of 100 mL/day if we required that $P < .05$ before reporting an effect. Finally, Fig. 6-5*D* shows the distribution of t values that would occur if the drug increased urine production by an average of 400 mL/day. Now, 99 percent of all possible t values fall above 2.101; the power of the test

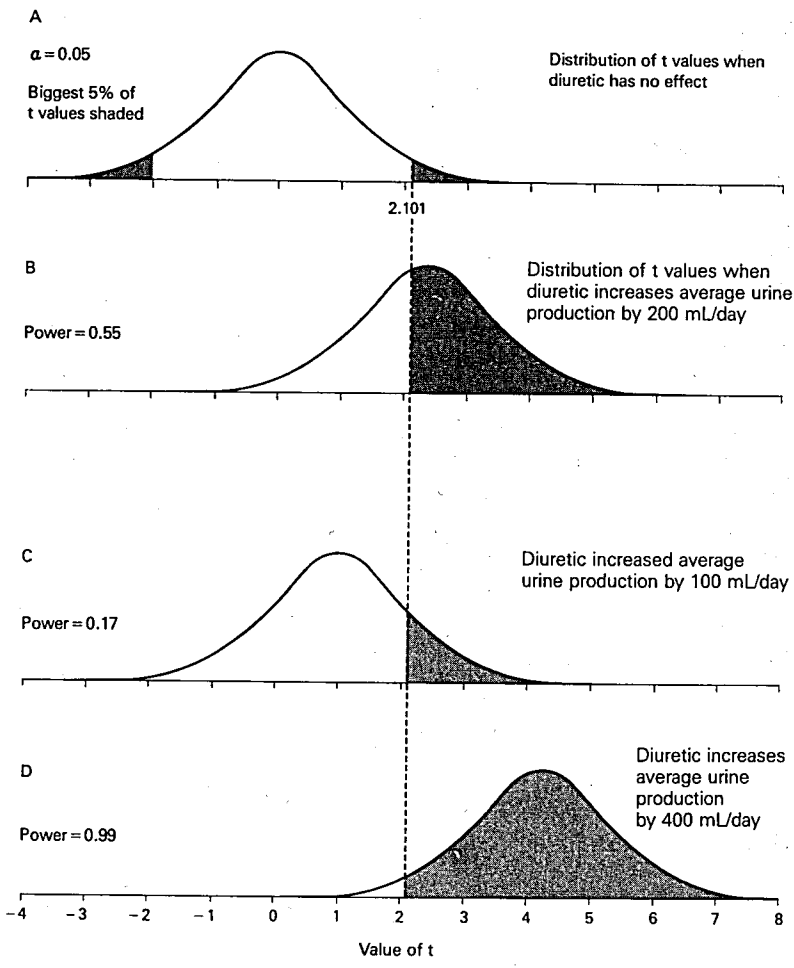


Figure 6-5 As the size of the treatment effect increases, the proportion of *t* values that will actually arise from the experiment increases, so the power of the test increases.

to detect a difference this large is .99. The chances are quite good that our experiment will produce accurate results. Figure 6-5 illustrates the general rule: *It is easier to detect big differences than small ones.*

We could repeat this process for all possible sizes of the treatment effect, from no effect at all up to very large effects, then plot the power

of the test as it varies with the change in urine production actually produced by the drug. Figure 6-6 shows a plot of the results, called a *power function,* of the test. It quantifies how much easier it is to detect a change (when we require a value of t corresponding to $P < .05$ and two samples of 10 people each) in urine production as the actual drug effect

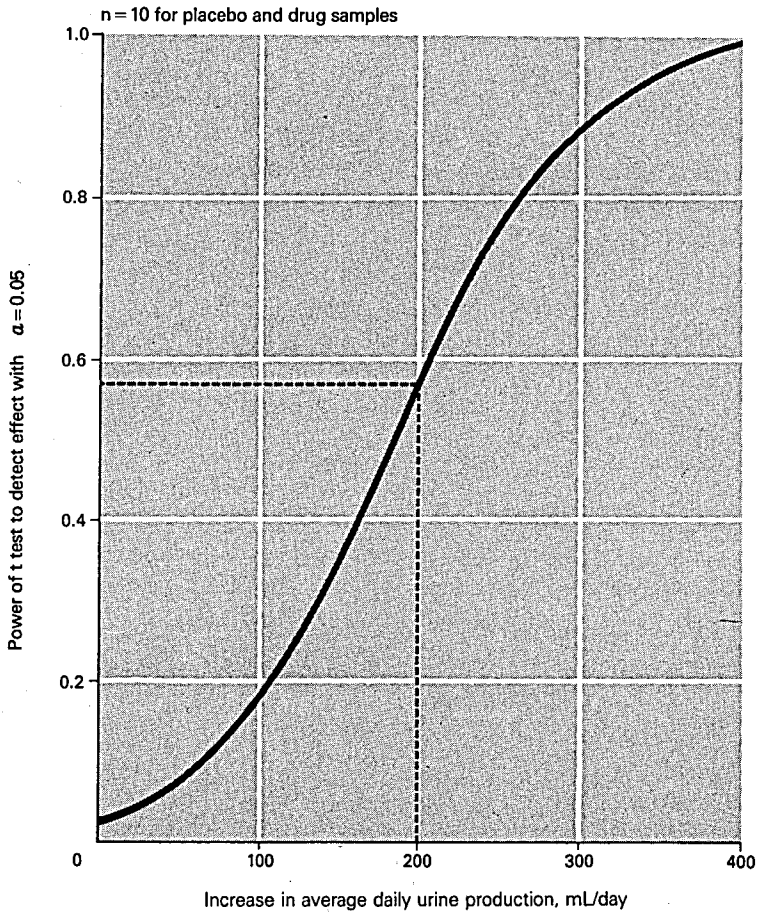


Figure 6-6 The power of a t test to detect a change in urine production based on experiments with two groups of people, each containing 10 individuals. The dashed line indicates how to read the graph. A t test has a power of .55 for detecting a 200 mL per day change in urine production.

gets larger and larger. This plot shows that if the drug increases urine production by 200 mL/day, there is a 55 percent chance that we will detect this change with the experiment designed as we have it; if urine production increases by 350 mL/day, the chance of our detecting this effect improves to 95 percent.

The Population Variability

The power of a test increases as the size of the treatment effect increases, but the variability in the population under study also affects the likelihood with which we can detect a treatment effect of a given size. In particular, recall from Chapter 4 that the *t*-test statistic, used to compare two means, is

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{(s^2/n_1) + (s^2/n_2)}}$$

in which $\overline{X}_1$ and $\overline{X}_2$ are the means, s is the pooled estimate of the population standard deviation σ, and n_1 and n_2 are the sizes of the two samples. $\overline{X}_1$ and $\overline{X}_2$ are estimates of μ_1 and μ_2, the two (different) population means. In the interest of simplicity, let us assume that the two samples are the same size; that is, $n_1 = n_2 = n$. Then t computed from our observations is an estimate of

$$t' = \frac{\mu_1 - \mu_2}{\sqrt{(\sigma^2/n) + (\sigma^2/n)}} = \frac{\mu_1 - \mu_2}{\sigma\sqrt{2/n}}$$

Denote the change in population mean value with the treatment by δ, Greek delta; then $\mu_1 - \mu_2 = \delta$, and

$$t' = \frac{\delta/\sigma}{\sqrt{2/n}} = \frac{\delta}{\sigma}\sqrt{\frac{n}{2}}$$

Therefore, t' depends on the change in the mean response normalized by the population standard deviation.

For example, the standard deviation in urine production in the population we are studying is 200 mL/day (from Fig. 6-1). In this context, an increase in urine production of 200 or 400 mL/day can be seen to be 1 or 2 standard deviations, a fairly substantial change. These same absolute changes in urine production would be even more striking if the popula-

tion standard deviation were only 50 mL/day, in which case a 200 mL/day absolute change would be 4 standard deviations. On the other hand, these changes in urine production would be hard to detect—indeed one wonders if you would want to detect them—if the population standard deviation were 500 mL/day. In this case, 200 mL/day would be only 0.4 standard deviation of the normal population.

As the variability in the population σ decreases, the power of the test for detecting a fixed absolute size of treatment effects δ increases and vice versa. In fact, we can combine the influence of these two factors by considering the dimensionless ratio $\phi = \delta/\sigma$, known as the *noncentrality parameter,* rather than each one separately.

Bigger Samples Mean More Powerful Tests

So far we have seen two things: (1) The power of a test to correctly reject the hypothesis that a treatment has no effect decreases as the confidence with which you wish to reject that hypothesis increases; (2) the power increases as the size of the treatment effect, measured with respect to the population standard deviation, increases. In most cases, investigators cannot control either of these factors and for a given sample size are stuck with whatever the power of the test is. However, the situation is not totally beyond their control. They can increase the power of the test without sacrificing the confidence with which they reject the hypothesis of no treatment effect (α) by *increasing the sample size.*

Increasing the sample size generally increases the power, for two reasons. First, as the sample size grows, the number of degrees of freedom increases, and the value of the test statistic that defines the "biggest" 100α percent of possible values under the assumption of no treatment effect generally decreases. Second, as the equation for t' above shows, the value of t (and many other test statistics) increases as sample size n increases. As a result, the distribution of t values that occur when the treatment has an effect of a given size δ/σ is located at higher t values as sample size increases.

For example, Fig. 6-7*A* shows the same information as Fig. 6-4*A*, with the sample size equal to 10 in each of the two groups. Figure 6-7*B* shows the distribution of possible t values if the hypothesis of no effect were true as well as the distribution of t values that would appear if the drug still increased urine production by 200 mL/day but now based on an experiment with 20 people in each group. Since there are 20 people in each group, the experiment has $\nu = 2(20 - 1) = 38$ degrees of

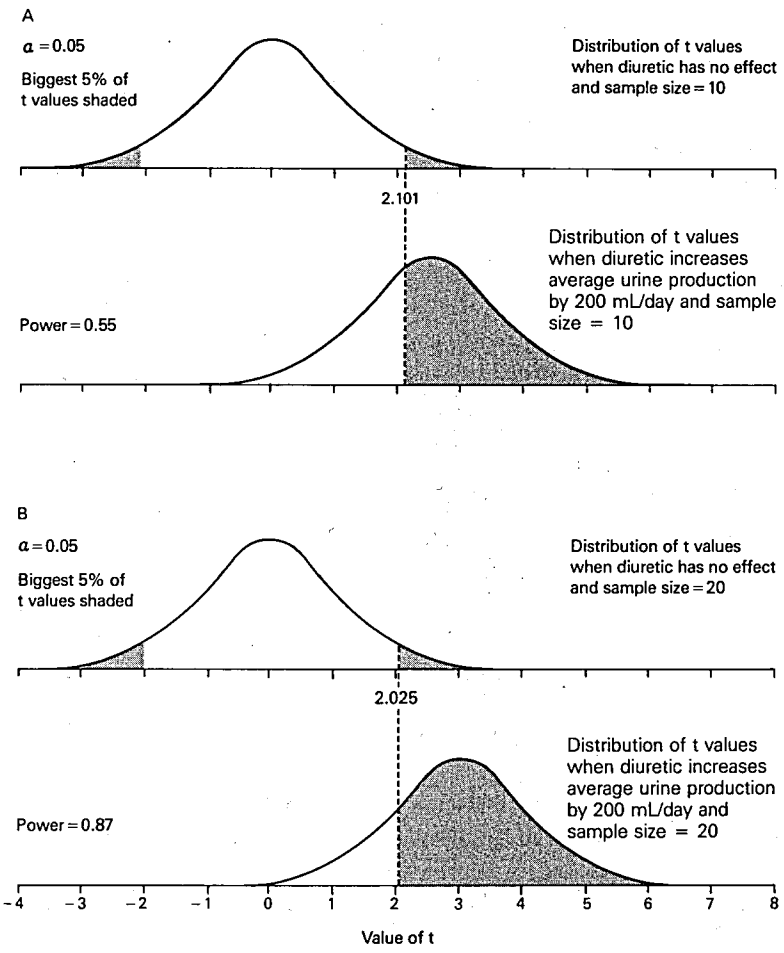


Figure 6-7 As the sample size increases, the power of the test increases for two reasons: (1) the critical value of t for a given confidence level in concluding that the treatment had an effect decreases, and (2) the values of the t statistic associated with the experiment increase.

freedom. From Table 4-1 the critical value of t defining the most extreme 5 percent of possible values is 2.025 (compared with 2.101 when there were 10 people in each group). The larger sample size also produces larger values of t, on the average, when the treatment increases urine production by 200 mL/day (with a population standard

deviation of 200 mL/day, as before) than it did with the smaller sample size. These two factors combine to make 87 percent of the possible values of t fall above 2.025. Therefore, there is an 87 percent chance of concluding that the drug has an effect; the power of the test under these circumstances is .87, up substantially from the value of .55 associated with the smaller sample size.

We could repeat this analysis over and over again to compute the power of this test to detect a 200 mL/day increase in urine production for a variety of sample sizes. Figure 6-8 shows the results of such computations. As the sample size increases, so does the test's power. In fact, estimating the sample size required to detect an effect large enough to be clinically significant is probably the major practical use to which power computations are put. Such computations are especially important in planning randomized clinical trials to estimate how many patients will have to be recruited and how many centers will have to be involved to accumulate enough patients to obtain a large enough sample to complete a meaningful analysis.

What Determines Power? A Summary

Figure 6-9 shows a general power curve for the t test, allowing for a variety of sample sizes and differences of interest. All these curves assume that we will reject the hypothesis of no treatment effect whenever we compute a value of t from the data that corresponds to $P < 0.5$ (so $\alpha = .05$). If we were more or less stringent in our requirement concerning the size of t necessary to report a difference, we would obtain a family of curves different from those in Fig. 6-9.

There is one curve for each value of the sample size n in Fig. 6-9. This value of n represents the size of *each* of the two sample groups being compared with the t test. Most power charts (and tables) present the results assuming that each of the experimental groups is of the same size, because, for a given total sample size, power is greatest when there are equal numbers of subjects in each treatment group. Thus, when using power analysis to estimate the sample size for an experiment, the result actually yields the size of each of the sample groups. Power analysis also can be used to estimate the power of a test that yielded a negative finding; in the case of unequal sample sizes, use the size of the smaller sample in the power analysis. This procedure will give you a conservative (low) estimate for the power of the test.

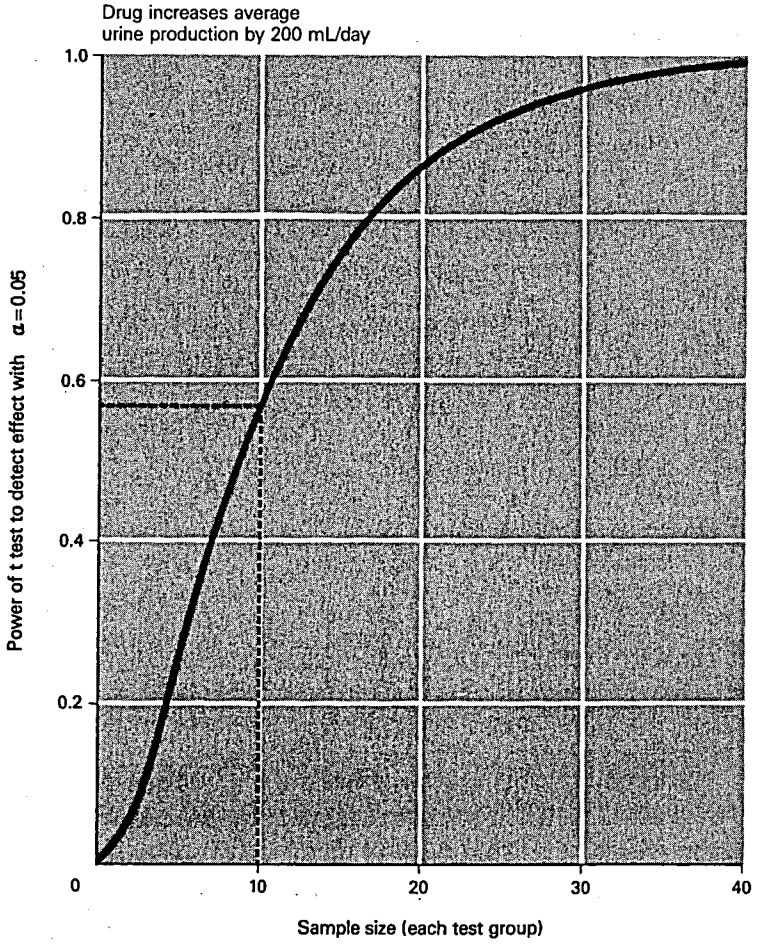


Figure 6-8 The effect of sample size on the power of a *t* test to detect a 200 mL per day increase in urine production with $\alpha = .05$ and a population standard deviation in urine production of 200 mL per day. The dashed line illustrates how to read the graph. A sample size of 10 yields a power of .55 for a *t* test to detect a 200 mL/day change in urine production.

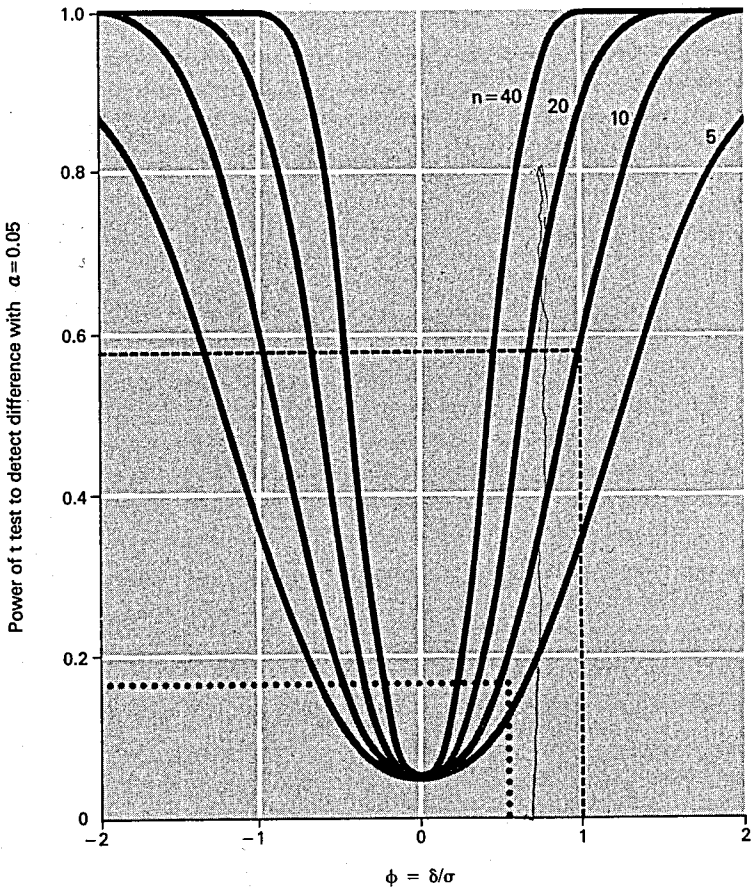


Figure 6-9 The power function for a test for comparing two experimental groups, each of size n, with $\alpha = .05$. δ is the size of the change we wish to detect, σ is the population standard deviation. If we had taken $\alpha = .01$ or any other value, we would have obtained a different set of curves. The dashed line indicates how to read the power of a test to detect a $\delta = 200$ mL/day change in urine production with a $\sigma = 200$ mL/day standard deviation in the underlying population with a sample size of $n = 10$ in each test group; the power of this test is .55. The dotted line indicates how to find the power of an experiment designed to study the effects of anesthesia on the cardiovascular system in which $\phi = \delta/\sigma = .55$ with a sample size of 9; the power of this test is only .16.

To illustrate the use of Fig. 6-9, again consider the effects of diuretic presented in Fig. 6-1. We wish to compute the power of a t test (with a 5 percent risk of a Type I error, $\alpha = .05$) to detect a mean change in urine production of 200 mL/day when the population has a standard deviation of 200 mL/day. Hence

$$\phi = \frac{\delta}{\sigma} = \frac{200 \text{ mL/day}}{200 \text{ mL/day}} = 1$$

Since the sample size is $n = 10$ (in both the placebo and drug groups), we use the "$n = 10$" line in Fig. 6-9 to find that this test will have a power of .55.

All the examples in this chapter so far deal with estimating the power of an experiment that is analyzed with a t test. It is also possible to compute the power for all the other statistical procedures described in this book. Although the details of the computations are different, the same variables are important and play the same general roles in the computation.

Another Look at Halothane versus Morphine for Open-Heart Surgery

Table 4-2 presented data on the effects of anesthesia on the cardiovascular system. When we analyzed these data with a t test, we did not conclude that halothane and morphine anesthesia produced significantly different values of cardiac index, which is defined as the rate at which the heart pumps blood (the cardiac output) divided by body surface area. This conclusion, however, was based on relatively small samples ($n = 9$ for the halothane group and $n = 16$ for the morphine group), and there was a 15 percent change in mean cardiac index (from 2.08 L/m^2 for halothane to 1.75 L/m^2 for morphine) between these two anesthetic regimes. While a 15 percent change in cardiac index may not be clinically important, a 25 percent change could be. The question then becomes: What is the power of this experiment to detect a 25 percent change in cardiac index?

We have already decided that a 25 percent change in cardiac index, 0.52 L/m^2 (25 percent of 2.08 L/m^2), is the size of the treatment effect worth detecting. From the data in Table 4-2, the pooled estimate of the variance in the underlying population is $s^2_{wit} = 0.88\ (L/m^2)^2$; take the square root of this number to obtain the estimate of the population

standard deviation of 0.94 L/m^2. Hence

$$\phi = \frac{\delta}{\sigma} = \frac{.52 \text{ L/m}^2}{.94 \text{ L/m}^2} = 0.553$$

Since the two sample groups have different sizes, we estimate the power of the test based on the size of the smaller group, 9. From Fig. 6-9, the power is only .16! Thus it is very unlikely that this experiment would be able to detect a 25 percent change in cardiac index.

We summarize our discussion of the power of hypothesis-testing procedures with these five statements.

- *The power of a test tells the likelihood that the hypothesis of no treatment effect will be rejected when the treatment has an effect.*
- *The more stringent our requirement for reporting that the treatment produced an effect (i.e., the smaller the chances of erroneously reporting that the treatment was effective), the lower the power of the test.*
- *The smaller the size of the treatment effect (with respect to the population standard deviation), the harder it is to detect.*
- *The larger the sample size, the greater the power of the test.*
- *The exact procedure to compute the power of a test depends on the test itself.*

POWER AND SAMPLE SIZE FOR ANALYSIS OF VARIANCE*

The issues underlying power and sample size calculations in analysis of variance are no different than for the *t* test. The only difference is the way in which the size of the minimum detectable treatment effect is quantified and the mathematical relationship relating this magnitude and the risk of erroneously concluding a treatment effect. The measure of the treatment effect to be detected is more complicated than in a *t* test because it must be expressed as more than a simple difference of two groups (because there are generally more than two groups in an analysis of variance). The size of the treatment effect is again quantified

*In an introductory course, this section can be skipped without interfering with the remaining material in the book.

by the *noncentrality parameter,* ϕ, although it is defined differently than for a *t* test. To estimate the power of an analysis of variance, you specify the number of treatment groups, sample size, risk of a false-positive (α) you are willing to accept, and size of the treatment effect you wish to detect (ϕ), then look the power up in charts for analysis of variance, just as we used Figure 6-9 for *t* tests.

The first step is to define the size of the treatment effect with the noncentrality parameter. We specify the minimum difference between any two treatment groups we wish to detect, δ, just as when computing the power of the *t* test. In this case, we define

$$\phi = \frac{\delta}{\sigma}\sqrt{\frac{n}{2k}}$$

where σ is the standard deviation within the underlying population, k is the number of treatment groups, and n is the sample size of each treatment group.* (Note the similarity with the definition of $\phi = \delta/\sigma$ for the *t* test.) Once ϕ is determined, obtain the power by looking in a power chart such as Figure 6-10 with the appropriate number of numerator degrees of freedom, $v_n = k - 1$ and denominator degrees of freedom $v_d = k(n - 1)$. (A more complete set of power charts for analysis of variance appears in Appendix B.)

These same charts can be used to estimate the sample size necessary to detect a given effect with a specified power. The situation is a little more complicated than it was in the *t* test because the sample size,

*We present the analysis for equal sample sizes in all treatment groups and the case where all the means but one are equal and the other differs by δ. This arrangement produces the maximum power for a given total sample size. An alternative definition of ϕ involves specifying the means for the different treatment groups that you expect to detect, μ_i for each of the k groups. In this case

$$\phi = \sqrt{\frac{n\Sigma(\mu_i - \mu)^2}{k\sigma^2}}$$

where

$$\mu = \frac{\Sigma\mu_i}{k}$$

is the grand population mean. The definition of ϕ in terms of the minimum detectable difference is generally easier to use because it requires fewer assumptions.

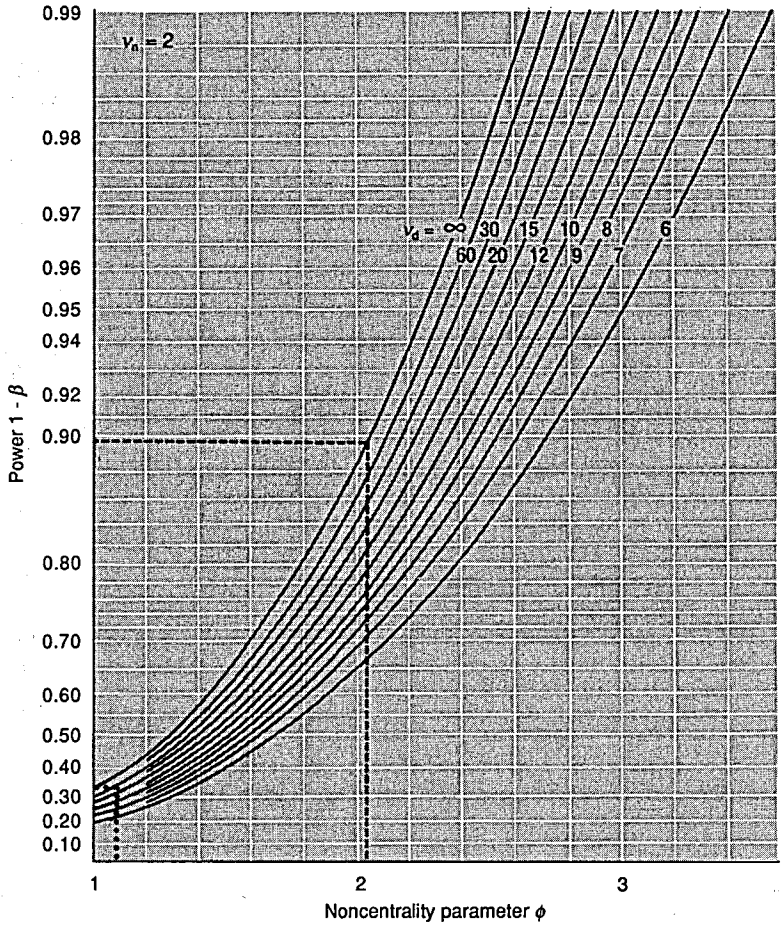


Figure 6-10 The power function for analysis of variance for $v_n = 2$ and $\alpha = .05$. Appendix B contains a complete set of power charts for a variety of values of v_n and $\alpha = .05$ and .01. *(Adapted from E. S. Pearson and H. O. Hartley, "Charts for the Power Function for Analysis of Variance Tests, Derived from the Non-Central F Distribution,"* Biometrika, *38:112–130, 1951.)*

n, appears in the noncentrality parameter, ϕ, and the denominator degrees of freedom, v_d. As a result, you must apply successive guesses to find n. You first guess n, compute the power, then adjust the guess until the computer power is close to the desired value. The example below illustrates this process.

Power, Menstruation, and Running

To illustrate power and sample size computations for analysis of variance, let us return to the study of the effect of running on menstruation we discussed in conjunction with Figure 3-9. The essential question is whether women who jog or are serious runners have menstrual patterns different from sedentary women. Suppose we wish to detect a change of $\delta = 1$ menses per year when there is an underlying variation of $\sigma = 2$ menses per year among $k = 3$ groups of women (controls, joggers, and long-distance runners), and $n = 26$ women in each group with 95 percent confidence ($\alpha = .05$). (This is about the magnitude of the effect observed in the example in Chapter 3.) To find the power of this test, we first compute the noncentrality parameter

$$\phi = \frac{1}{2}\sqrt{\frac{26}{2 \cdot 3}} = 1.04$$

There are $v_n = k - 1 = 3 - 1 = 2$ numerator and $v_d = k(n - 1) = 3(26 - 1) = 75$ denominator degrees of freedom. From Fig. 6-10, the power is only about .32!

As with most power computations, this result is sobering. Suppose that we wanted to increase the power of the test to .80; how big will the samples have to be? We already know that 26 women per group is too small. Examining Fig. 6-10 suggests that we need to get ϕ up to about 2. Since the sample size, n, appears under a square root in the definition of ϕ, let us increase n by about a factor of 4 to 100 women per group. Now,

$$\phi = \frac{1}{2}\sqrt{\frac{100}{2 \cdot 3}} = 2.04$$

and $v = k(n - 1) = 3(100 - 1) = 297$. From Fig. 6-10, the power is .90. Given the uncertainties in the estimates of σ before actually conducting the experiment, this is probably close enough to stop working. The problem is that getting large samples is often difficult (and expensive). To get closer to the desired power of .80, let us try a smaller sample size, say 75. Now,

$$\phi = \frac{1}{2}\sqrt{\frac{75}{2 \cdot 3}} = 1.77$$

and $v_d = 3(75 - 1) = 222$. From Fig. 6-10, the power is .80. Therefore, to have an 80 percent chance of detecting a change of 1 menses per year among three groups of women when the standard deviation of the underlying population is 2 menses per year with 95 percent confidence, we need 75 women in each group.

POWER AND SAMPLE SIZE FOR COMPARING TWO PROPORTIONS*

The development of formulas for power and sample size when comparing two proportions is similar to the procedure that we used for the t test, except that we will be basing the computations on the normal distribution. We wish to find the power of a z test to detect a difference between two proportions, p_1 and p_2 with sample sizes n_1 and n_2. Recall, from Chapter 5, that the z test statistic used to compare two observed proportions, is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{s_{p_1 - p_2}}$$

This test statistic is distributed according to a normal distribution with mean $(p_1 - p_2)$ and standard deviation

$$s_{p_1 - p_2} = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$

(Note that we do not base our estimate of the standard deviation on sample estimates, since p_1 and p_2 are specified as part of the problem.) Figure 6-11 *A* shows this expected distribution of observed differences, $\hat{p}_1 - \hat{p}_2$. To achieve a power of $100(1 - \beta)$ percent, $100(1 - \beta)$ percent of this distribution must be above $(p_1 - p_2) + z_{\beta(1)} s_{p_1 - p_2}$, where $z_{\beta(1)}$ is the lower-tail (one tail) value on the standard normal distribution that defines the lowest 100β percent of the distribution (Table 6-2). For example, to obtain 80 percent power, $z_\beta = z_{.20} = -.842$.

Under the null hypothesis $p_1 = p_2$, they would be simply two different estimates of the proportion of the underlying population that has the characteristic of interest, p. We can estimate p with

*In introductory courses, this material can be skipped without loss of continuity.

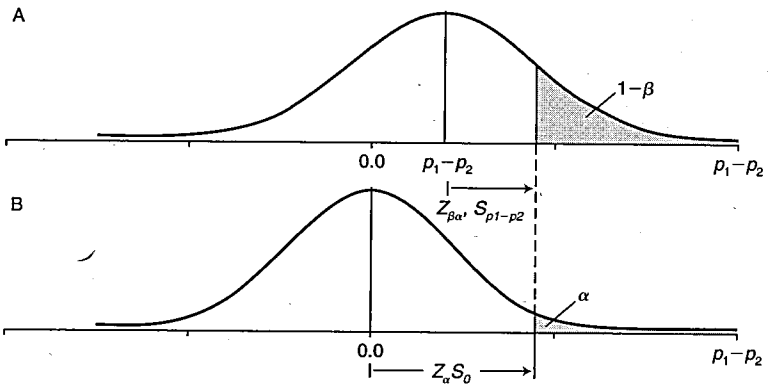


Figure 6-11 (*A*) The distribution of all possible values of the observed differences between the two proportions, $\hat{p}_1 - \hat{p}_2$, follows a normal distribution centered on the true mean difference of the two populations, $p_1 - p_2$, and with standard deviation $s_{p_1-p_2}$. (*B*) If the null hypothesis of no difference between the two populations is true, then the distribution of all observed differences of the two proportions will be centered on 0 (because $p_1 - p_2 = 0$) with standard deviation s_0. If you are willing to accept a false positive (Type I) risk of α, then α percent of the possible observed values of $\hat{p}_1 - \hat{p}_2$ will be at or above $z_\alpha s_0$. At the same time, 100(1−β) percent of the possible observed values of $\hat{p}_1 - \hat{p}_2$ will be above this value when the actual difference between the two population proportions is $p_1 - p_2 \neq 0$, which is the power of the test to detect this difference.

$$\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

so the resulting sampling distribution of the z test statistic will be distributed normally with mean $p_1 - p_2 = 0$ and standard deviation

$$s_0 = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}$$

Figure 6-11*B* shows this distribution. We reject the null hypothesis of no difference at the 100α percent level of confidence when the observed value of $\hat{p}_1 - \hat{p}_2$ is large enough to be above $z_\alpha s_0$.

We will achieve 100(1 − β) percent power for hypotheses rejected with 100α percent level of confidence when the two points in panels *A* and *B* of Fig. 6-11 line up. (Compare with Fig. 6-4.) In mathematical terms, this occurs when

Table 6-2 Percentile Points of the Standard Normal Distribution (One Tail)

Fraction of distribution *below z* (β)	Fraction of distribution *above z* ($1-\beta$) Power	z
.001	.999	−3.0902
.005	.995	−2.5758
.010	.990	−2.3264
.020	.980	−2.0538
.050	.950	−1.6449
.100	.900	−1.2816
.200	.800	−0.8416
.300	.700	−0.5244
.400	.600	−0.2534
.500	.500	0.0000
.600	.400	0.2534
.700	.300	0.5244
.800	.200	0.8416
.900	.100	1.2816
.950	.050	1.6449
.980	.020	2.0538
.990	.010	2.3264
.995	.005	2.5758
.999	.001	3.0902

$$(p_1 - p_2) + z_{\beta(1)} s_{p_1 - p_2} = z_\alpha s_0$$

Finally, the power of the test is the probability (from the standard normal distribution) that z exceeds*

$$z_{\beta(1)} = \frac{z_\alpha s_0 - (p_1 - p_2)}{s_{p_1 - p_2}}$$

*Technically, we should also include the part of the distribution in Fig. 6-11*A* that falls below the lower z_α tail of the distribution in Fig. 6-11*B*, but this tail of the distribution rarely contributes anything of consequence. Note that these calculations do not include the Yates correction. It is possible to include the Yates correction by replacing $(p_1 - p_2)$ with $|p_1 - p_2| - (1/n_1 + 1/n_2)$. Doing so makes the arithmetic more difficult, but does not represent a theoretical change. Including the Yates correction lowers the power or increases the sample size.

Mortality Associated with Anesthesia for Open-Heart Surgery

When we studied the mortality associated with halothane (13.1 percent of 61 patients) and morphine (14.9 percent of 67 patients) anesthesia in open heart surgery in Chapter 5, we did not find a significant difference. What is the power of this study to detect a 30 percent difference in mortality, from 14 to 10 percent with 95 percent confidence?

In this case, $p_1 = .14$ and $p_2 = .10$; $n_1 = 61$ and $n_2 = 67$, so

$$s_{p_1 - p_2} = \sqrt{\frac{.14\,(1 - .14)}{61} + \frac{.10\,(1 - .10)}{67}} = .0576$$

To compute s_0 we first compute

$$\hat{p} = \frac{.14 \cdot 61 + .10 \cdot 67}{61 + 67} = .119$$

in which case

$$s_0 = \sqrt{\frac{.119\,(1 - .119)}{61} + \frac{.119\,(1 - .119)}{67}} = .0573$$

The two-tail 95 percent critical value of the normal distribution, $z_{.05}$ is 1.96, so the power of the test is the fraction of the normal distribution above

$$z_{\beta(1)} = \frac{1.96 \cdot .0573 - (.14 - .10)}{.0576} = 1.255$$

From Table 6-2, the power of the test is only 11 percent, so you should be careful about negative conclusions in such a study!

Sample Size for Comparing Two Proportions

To obtain the sample size to compare two proportions, simply take $z_{\beta(1)}$ as given and solve the resulting equations for n, the size of each group.

Assuming that the two groups are the same size, this process yields

$$n = \frac{A\left[1 + \sqrt{1 + \frac{4\delta}{A}}\right]^2}{4\delta^2}$$

where

$$A = {}_2[z_\alpha \sqrt{2\,\hat{p}\,(1 - \hat{p})} + z_{\beta(1)} \sqrt{P_1\,(1 - p_1) + p_2\,(1 - p_2)}]^2$$

$$\hat{p} = \frac{p_1 + p_2}{2}$$

$$\delta = |p_1 - p_2|$$

POWER AND SAMPLE SIZE FOR RELATIVE RISK AND ODDS RATIO

The formulas developed above can be used to estimate power and sample sizes for relative risks and odds ratios. Instead of specifying both proportions, you simply specify one proportion, the desired relative risk or odds ratio, and compute the other proportion. Let p_1 be the probability of disease in the unexposed members of the population and p_2 be the probability of disease in the exposed members of the population.

The relative risk is the ratio of the probability of disease in those exposed to the toxin of interest over those not exposed,

$$RR = \frac{p_{\text{exposed}}}{p_{\text{unexposed}}} = \frac{p_2}{p_1}$$

so use the formulas above with

$$p_2 = RR \cdot p_1$$

Likewise, the odds ratio is

$$OR = \frac{p_{\text{exposed}}/(1 - p_{\text{exposed}})}{p_{\text{unexposed}}/(1 - p_{\text{unexposed}})} = \frac{p_2/(1 - p_2)}{p_1/(1 - p_1)}$$

so

$$p_2 = \frac{OR \cdot p_1}{1 + p_1 (OR - 1)}$$

POWER AND SAMPLE SIZE FOR CONTINGENCY TABLES*

Figure 6-10 (and the corresponding charts in Appendix B) can also be used to compute the power and sample size for contingency tables. As with other power computations, the first step is to define the pattern you wish to be able to detect. This effect is specified by selecting the proportions of row and column observations that appear in each cell of the contingency table. Table 6-3 shows the notation for the computation for a 3 × 2 contingency table, p_{11} is the proportion of all observations expected in the upper left cell of the table, p_{12} the proportion in the upper right corner, and so on. All the proportions must add up to 1. The r row and c column sums are denoted with R's and C's with subscripts corresponding to the rows and columns. The noncentrality parameter for such a contingency table is defined as

$$\phi = \sqrt{\frac{N}{(r-1)(c-1)+1} \sum \frac{(p_{ij} - R_i C_j)^2}{R_i C_j}}$$

where r is the number of rows, c is the number of columns, and N is the total number of observations. This value of ϕ is used with Fig. 6-10 with $v_n = (r-1)(c-1)$ and $v_d = \infty$ degrees of freedom.

To compute the sample size necessary to achieve a given power, simply reverse this process. Determine the necessary value of ϕ to

Table 6-3 Notation for Computing Power for Contingency Tables

p_{11}	p_{12}	R_1
p_{21}	p_{22}	R_2
p_{31}	p_{32}	R_3
C_1	C_2	1.00

*In introductory courses, this section can be skipped without loss of continuity.

achieve the desired power with $v_n = (r - 1)(c - 1)$ and $v_d = \infty$ from Fig. 6-10 (or the power charts in Appendix B). We obtain the sample size by solving the equation above for N, to obtain

$$N = \frac{\phi^2[(r-1)(c-1)+1]}{\sum \frac{(p_{ij} - R_i C_j)^2}{R_i C_j}}$$

Physicians, Perspiration, and Power

In addition to studying the effect of running on menstruation, Dale and colleagues also studied how likely women were to consult a physician depending on their running status. (This example was discussed in conjunction with Table 5-5.) Let us examine the power of a contingency table to detect the pattern of proportions shown in Table 6-4 with 95 percent confidence (α) from a sample of $N = 165$ women. Substituting into the equation above

$$\phi = \left[\frac{165}{(3-1)(2-1)+1}\left[\frac{(.025 - .250 \cdot .350)^2}{.250 \cdot .350} + \frac{(.225 - .250 \cdot .650)^2}{.250 \cdot .650} + \frac{(.100 - .300 \cdot .350)^2}{.300 \cdot .350} + \frac{(.200 - .300 \cdot .650)^2}{.300 \cdot .650} + \frac{(.225 - .450 \cdot .350)^2}{.450 \cdot .350} + \frac{(.225 - .450 \cdot .650)^2}{.450 \cdot .650}\right]\right]^{1/2}$$

$$\phi = 2.50$$

Table 6-4 Pattern of Physician Consultation for Menstrual Problems to be Detected

Group	Yes	No	Total
Controls	.025	.225	.250
Joggers	.100	.200	.300
Runners	.225	.225	.450
Total	.350	.650	1.00

Consult Fig. 6-10 with $\phi = 2.50$, $v_n = (r - 1)(c - 1) = (3 - 1)(2 - 1) = 2$ and $v_d = \infty$ degrees of freedom to obtain a power of .98 to detect this pattern with 95 percent confidence.

PRACTICAL PROBLEMS IN USING POWER

If you know the size of the treatment effect, population standard deviation, α, and sample size, you can use graphs like Fig. 6-9 to estimate the power of a t test after the fact. Unfortunately, in practice, one does not know how large an effect a given treatment will have (finding that out is usually the reason for the study in the first place); so you must specify how large a change is *worth detecting* to compute the power of the test.

This requirement to go on record about how small a change is worth detecting may be one reason that very few people report the power of the tests they use. While such information is not especially important when investigators report that they detected a difference, it can be quite important when they report that they failed to detect one. If the power of the test to detect a clinically significant effect is small, say 25 percent, this report will mean something quite different than if the test was powerful enough to detect a clinically significant difference 85 percent of the time.

These difficulties are even more acute when using power computations to decide on the sample size for a study in advance. Completing this computation requires that investigators estimate not only the size of the effect they think is worth detecting and the confidence with which they hope to accept (β) or reject (α) the hypothesis that the treatment is effective but also the standard deviation of the population being studied. Sometimes existing information can be used to estimate these numbers; sometimes investigators do a pilot study to estimate them; sometimes they simply guess.

WHAT DIFFERENCE DOES IT MAKE?

In Chapter 4 we discussed the most common error in the use of statistical methods in the medical literature, inappropriate use of the t test. Repeated use of t tests increases the chances of reporting a "statistically significant" difference above the nominal levels one obtains from the t distribution. In the language of this chapter, it increases the Type I error. In practical terms, this increases the chances that an investigator

will report some procedure or therapy capable of producing an effect beyond what one would expect from chance variation when the evidence does not actually support this conclusion.

This chapter examined the other side of the coin, the fact that perfectly correctly designed studies employing statistical methods correctly may fail to detect real, perhaps clinically important, differences simply because the sample sizes are too small to give the procedure enough power to detect the effect. This chapter shows how you can estimate the power of a given test after the results are reported in the literature and also how investigators can estimate the number of subjects they need to study to detect a specified difference with a given level of confidence (say, 95 percent; that is, $\alpha = .05$). Such computations are often quite distressing because they often reveal the need for a large number of experimental subjects, especially compared with the relatively few patients who typically form the basis for clinical studies.* Sometimes the investigators increase the size of the difference they say they wish to detect, decrease the power they find acceptable, or ignore the whole problem in an effort to reduce the necessary sample size. Most medical investigators never confront these problems because they have never heard of power.

In 1979, Jennie Freiman and colleagues† examined 71 randomized clinical trials published between 1960 and 1977 in journals, such as *The Lancet,* the *New England Journal of Medicine,* and the *Journal of the American Medical Association,* reporting that the treatment studied did not produce a "statistically significant" ($P < .05$) improvement in clinical outcome. Only 20 percent of these studies included enough subjects to detect a 25 percent improvement in clinical outcome with a power of .50 or better. In other words, if the treatment produced a 25 percent reduction in mortality rate or other clinically important endpoint, there was less than a 50:50 chance that the clinical trial would be able to detect it with $P < .05$. Moreover, Freiman and colleagues found that *only one* of the 71 papers stated that α and β were considered at the start of the study; 18

*R. A. Fletcher and S. W. Fletcher ("Clinical Research in General Medical Journals: A 30-Year Perspective," *N. Engl. J. Med.,* **301**:180–183, 1979) report the median number of subjects included in clinical studies published in the *Journal of the American Medical Association, The Lancet,* and the *New England Journal of Medicine* in 1946 to 1976 ranged from 16 to 36 people.

†J. A. Freiman, T. C. Chalmers, H. Smith, Jr., and R. R. Kuebler, "The Importance of Beta, the Type II Error and Sample Size in the Design and Interpretation of the Randomized Controlled Trial," *N. Engl. J. Med.,* **299**:690–694, 1978.

recognized a trend in the results, whereas 14 commented on the need for a larger sample size.

Fifteen years later, in 1994, Mohler and colleagues* revisited this question by examining randomized controlled trials in these same journals published in 1975, 1980, 1985, and 1990. While the number of randomized controlled trials published in 1990 was more than twice the number published in 1975, the proportion reporting negative results remained reasonably constant, at about 27 percent of all the trials. Only 16 percent and 36 percent of the negative studies had an adequate power (.80) to detect a 25 percent or 50 percent change in outcome, respectively. Only one third of the studies with negative results reported information regarding how the sample sizes were computed. While disappointing, this result is better than Freiman and colleagues observed in 1979, when no one reported on sample size computations.

The fact remains, however, that publication of "negative" studies without adequate attention to having a large enough sample size to draw definitive conclusions remains a problem. Thus, in this area, like the rest of statistical applications in the medical literature, it is up to responsible readers to interpret what they read rather than take it at face value.

Other than throwing your hands up when a study with low power fails to detect a statistically significant effect, is there anything an investigator or clinician reading the literature can learn from the results? Yes. Instead of focusing on the accept–reject logic of statistical hypothesis testing,† one can try to estimate how strongly the observations *suggest* an effect by estimating the size of the hypothesized effect together with the uncertainty of this

*D. Mohler, C. S. Dulberg, G. A. Wells, "Statistical Power, Sample Size, and Their Reporting in Randomized Clinical Trials," *JAMA* **272:**122–124, 1994.

†There is another approach that can be used in some clinical trials to avoid this accept–reject problem. In a *sequential trial* the data are analyzed after each new individual is added to the study and the decision made to (1) accept the hypothesis of no treatment effect, (2) reject the hypothesis, or (3) study another individual. Sequential tests generally allow one to achieve the same levels of α and β for a given size treatment effect with a smaller sample size than the methods discussed in this book. This smaller sample size is purchased at the cost of increased complexity of the statistical procedures. Sequential analyses are often performed by repeated use of the statistical procedures presented in this book, such as the t test. This procedure is incorrect because it produces overoptimistic P values, just as the repeated use of t tests (without the Bonferroni correction) produces erroneous results when one should do an analysis of variance. See W.J. Dixon and F.J. Massey, *Introduction to Statistical Analysis* (4 ed.), McGraw-Hill, New York, 1983, chapter 18, "Sequential Analysis," for an introduction to sequential analysis.

estimate.* We laid the groundwork for this procedure in Chapters 2, 4, and 5 when we discussed the standard error and the t distribution. The next chapter builds on this base to develop the idea of confidence limits.

PROBLEMS

6-1 Use the data in Table 4-2 to find the power of a t test to detect a 50 percent difference in cardiac index between halothane and morphine anesthesia.

6-2 How large a sample size would be necessary to have an 80 percent chance of detecting a 25 percent difference in cardiac index between halothane and morphine anesthesia?

6-3 Use the data in Table 4-2 to find the power of the experiments reported there to detect a 25 percent change in mean arterial blood pressure and total peripheral resistance.

6-4 In Prob. 3-5 (and again in Prob. 4-4), we decided that there was insufficient evidence to conclude that men and women who have had at least one vertebral fracture differ in vertebral bone density. What is the power of this test to detect average (with $\alpha = .05$) bone density in men 20 percent lower than the average bone density for women?

6-5 How large a sample would be necessary to be 90 percent confident that men have vertebral bone densities of at least 30 percent of the values for women when you wish to be 95 percent confident in any conclusion that vertebral bone densities differ between men and women?

6-6 Use the data in Prob. 3-2 to find the power of detecting a change in mean forced midexpiratory flow of 0.25 L/s with 95 percent confidence.

6-7 Use the data in Prob. 3-3 to find the power of detecting an increase in HDL of 5 mg/dL and 10 mg/dL with 95 percent confidence.

6-8 How large must each sample group be to have an 80 percent power to detect a change of 5 mg/dL with 95 percent confidence?

6-9 What is the power of the experiment in Prob. 5-4 to detect a situation in which ampicillin and trimethoprim-sulfamethoxazole both prevent recurrence of urinary tract infections one-third of the time and cephalexin prevents recurrence two-thirds of the time. Assume that the same number of people take each drug as in Prob. 5-4. Use $\alpha = .05$.

6-10 How large would the sample size need to be in Prob. 6-9 to reach 80 percent power?

*One quick way to use a computerized statistical package to estimate if getting more cases would resolve a power problem is to simply copy the data twice and rerun the analysis on the doubled data set. If the results become less ambiguous, it suggests that obtaining more cases (on the assumption that the data will be similar to that which you have already obtained) will yield less ambiguous results. This procedure is, of course, not a substitute for a formal power analysis and would certainly not be reportable in a scientific paper, but it is an easy way to get an idea of whether gathering more data would be worthwhile.

Chapter 7

Confidence Intervals

All the statistical procedures developed so far were designed to help decide whether or not a set of observations is compatible with some hypothesis. These procedures yielded *P* values to estimate the chance of reporting that a treatment has an effect when it really does not and the power to estimate the chance that the test would detect a treatment effect of some specified size. This decision-making paradigm does not characterize the size of the difference or illuminate results that may not be statistically significant (i.e., not associated with a value of *P* below .05) but does nevertheless suggest an effect. In addition, since *P* depends not only on the magnitude of the treatment effect but also the sample size, it is common for experiments to yield very small values of *P* (what investigators often call "highly significant" results) when the magnitude of the treatment effect is so small that it is clinically or scientifically unimportant. As Chapter 6 noted, it can be more informative to think not only in terms of the accept–reject approach of statistical hypothesis testing but also to estimate the size of the

treatment effect together with some measure of the uncertainty in that estimate.

This approach is not new; we used it in Chapter 2 when we defined the standard error of the mean to quantify the certainty with which we could estimate the population mean from a sample. We observed that since the population of all sample means at least approximately follows a normal distribution, the true (and unobserved) population mean will lie within about 2 standard errors of the mean of the sample mean 95 percent of the time. We now develop the tools to make this statement more precise and generalize it to apply to other estimation problems, such as the size of the effect a treatment produces. The resulting estimates, called *confidence intervals*, can also be used to test hypotheses.* This approach yields exactly the same conclusions as the procedures we discussed earlier because it simply represents a different perspective on how to use concepts like the standard error, t, and normal distributions. Confidence intervals are also used to estimate the range of values that include a specified proportion of all members of a population, such as the "normal range" of values for a laboratory test.

THE SIZE OF THE TREATMENT EFFECT MEASURED AS THE DIFFERENCE OF TWO MEANS

In Chapter 4, we defined the t statistic to be

$$t = \frac{\text{difference of sample means}}{\text{standard error of difference of sample means}}$$

then computed its value for the data observed in an experiment. Next, we compared the result with the value t_α that defined the most extreme 100α percent of the possible values to t that would occur (in both tails) if the two samples were drawn from a single population. If the observed value of t exceeded t_α, we reported a "statistically significant" difference, with $P < \alpha$. As Fig. 4-5 showed, the distribution of possible

*Some statisticians believe that confidence intervals provide a better way to think about the results of experiments than traditional hypothesis testing. For a brief exposition from this perspective, see K. J. Rothman, "A Show of Confidence," *N. Engl. J. Med.*, **299**:1362–1363, 1978.

values of t has a mean of zero and is symmetric about zero when the two samples are drawn from the *same* population.

On the other hand, if the two samples are drawn from populations with *different* means, the distribution of values of t associated with all possible experiments involving two samples of a given size is *not* centered on zero; it does not follow the t distribution. As Figs. 6-3 and 6-5 showed, the actual distribution of possible values of t has a nonzero mean that depends on the size of the treatment effect. It is possible to review the definition of t so that it will be distributed according to the t distribution in Fig. 4-5 *regardless of whether or not the treatment actually has an effect.* This modified definition of t is

$$t = \frac{\text{difference of sample means} - \text{true difference in population means}}{\text{standard error of difference of sample means}}$$

Notice that if the hypothesis of no treatment effect is correct, the difference in population means is zero and this definition of t reduces to the one we used before.

The equivalent mathematical statement is

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}}$$

In Chapter 4 we computed t from the observations, then compared it with the critical value for a "big" value of t with $\nu = n_1 + n_2 - 2$ degrees of freedom to obtain a P value. Now, however, we cannot follow this approach since we do not know all the terms of the right side of the equation. Specifically, *we do not know the true difference in mean values of the two populations* from which the samples were drawn, $\mu_1 - \mu_2$. We can, however, use this equation to estimate the size of the treatment effect, $\mu_1 - \mu_2$.

Instead of using the equation to determine t, we will select an appropriate value of t and use the equation to estimate $\mu_1 - \mu_2$. The only problem is that of selecting an appropriate value for t.

By definition, 100α percent of all possible values of t are more negative than $-t_\alpha$ or more positive than $+t_\alpha$. For example, only 5 percent of all possible t values will fall outside the interval between $-t_{.05}$

and $+t_{.05}$, where $t_{.05}$ is the critical value of t that defines the most extreme 5 percent of the t distribution (tabulated in Table 4-1). Therefore, $100(1-\alpha)$ percent of all possible values of t fall between $-t_\alpha$ and $+t_\alpha$. For example, 95 percent of all possible values of t will fall between $-t_{.05}$ and $+t_{.05}$.

Every different pair of random samples we draw in our experiment will be associated with different values of $\overline{X}_1, \overline{X}_2$, and $s_{\overline{X}_1-\overline{X}_2}$; and $100(1-\alpha)$ percent of all possible experiments involving samples of a given size will yield values of t that fall between $-t_\alpha$ and $+t_\alpha$. Therefore, for $100(1-\alpha)$ percent of all possible experiments

$$-t_\alpha < \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_{\overline{X}_1-\overline{X}_2}} < +t_\alpha$$

Solve this equation for the true difference in sample means

$$(\overline{X}_1 - \overline{X}_2) - t_\alpha s_{\overline{X}_1-\overline{X}_2} < \mu_1 - \mu_2 < (\overline{X}_1 - \overline{X}_2) + t_\alpha s_{\overline{X}_1-\overline{X}_2}$$

In other words, the actual difference of the means of the two populations from which the samples were drawn will fall within t_α standard errors of the difference of the sample means of the observed difference in the sample means (t_α has $\nu = n_1 + n - 2$ degrees of freedom, just as when we used the t distribution in hypothesis testing.) This range is called the $100(1-\alpha)$ percent *confidence interval for the difference of the means*. For example, the 95 percent confidence interval for the true difference of the sample means is

$$(\overline{X}_1 - \overline{X}_2) - t_{.05} s_{\overline{X}_1-\overline{X}_2} < \mu_1 - \mu_2 < (\overline{X}_1 - \overline{X}_2) + t_{.05} s_{\overline{X}_1-\overline{X}_2}$$

This equation defines the range that will include the true difference in the means for 95 percent of all possible experiments that involve drawing samples from the two populations under study.

Since this procedure to compute the confidence interval for the difference of two means uses the t distribution, it is subject to the same limitations as the t test. In particular, the samples must be drawn from populations that follow a normal distribution at least approximately.*

*It is also possible to define confidence intervals for differences in means when there are multiple comparisons, using q and q' in place of t. For a detailed discussion of these computations, see J. H. Zar, *Biostatistical Analysis*, 2d ed., Prentice-Hall, Englewood Cliffs, N.J., 1984, p. 191–192, 195.

THE EFFECTIVE DIURETIC

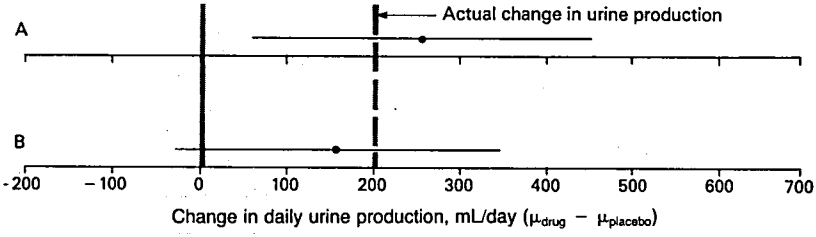
Figure 6-1 showed the distributions of daily urine production for a population of 200 individuals when they are taking a placebo or a drug that is an effective diuretic. The mean urine production of the entire population when all members are taking the placebo is $\mu_{pla} = 1200$ mL/day. The mean urine production for the population when all members are taking the drug is $\mu_{dr} = 1400$ mL/day. Therefore, the drug increases urine production by an average of $\mu_{dr} - \mu_{pla} = 1400 - 1200 = 200$ mL/day. An investigator, however, cannot observe every member of the population and must estimate the size of this effect from samples of people observed when they are taking the placebo or the drug. Figure 6-1 shows one pair of such samples, each of 10 individuals. The people who received the placebo had a mean urine output of 1150 mL/day, and the people receiving the drug had a mean urine output of 1400 mL/day. Thus, these two samples suggest that the drug increased urine production by $\bar{X}_{dr} - \bar{X}_{pla} = 1400 - 1150 = 250$ mL/day. The random variation associated with the sampling procedure led to a different estimate of the size of the treatment effect from that really present. Simply presenting this single estimate of 250 mL/day increase in urine output ignores the fact that there is some uncertainty in the estimates of the true mean urine output in the two populations, so there will be some uncertainty in the estimate of the true difference in urine output. We now use the confidence interval to present an alternative description of how large a change in urine output accompanies the drug. This interval describes the average change seen in the people included in the experiment and also reflects the uncertainty introduced by the random sampling process.

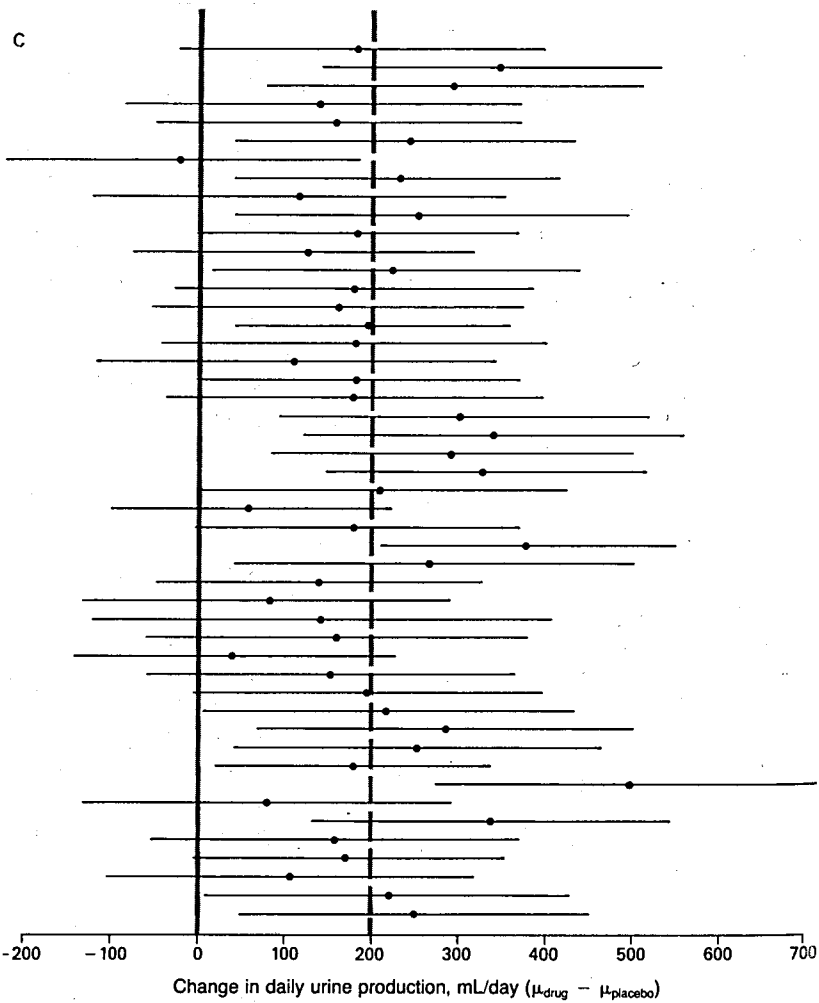
To estimate the standard error of the difference of the means $s_{\bar{X}_{dr}-\bar{X}_{pla}}$ we first compute a pooled estimate of the population variance. The standard deviations of observed urine production were 245 and 144 mL/day for people taking the drug and the placebo, respectively. Both samples included 10 people; therefore,

$$s^2 = \tfrac{1}{2}(s_{dr}^2 + s_{pla}^2) = \tfrac{1}{2}(245^2 + 144^2) = 201^2$$

and

$$s_{\bar{X}_{dr}-\bar{X}_{pla}} = \sqrt{\frac{s^2}{n_{dr}} + \frac{s^2}{n_{pla}}} = \sqrt{\frac{201^2}{10} + \frac{201^2}{10}} = 89.9 \text{ mL/day}$$


A
Actual change in urine production
B
-200
-100
0
100
200
300
400
500
600
700
Change in daily urine production, mL/day ($\mu_{drug} - \mu_{placebo}$)


C
-200
-100
0
100
200
300
400
500
600
700
Change in daily urine production, mL/day ($\mu_{drug} - \mu_{placebo}$)

To compute the 95 percent confidence interval, we need the value of $t_{.05}$ from Table 4-1. Since each sample contains $n = 10$ individuals, we use the value of $t_{.05}$ corresponding to $\nu = 2(n - 1) = 2(10 - 1) = 18$ degrees of freedom. From Table 4-1, $t_{.05} = 2.101$.

Now we are ready to compute the 95 percent confidence interval for the mean change in urine production that accompanies use of the drug

$$(\overline{X}_{dr} - \overline{X}_{pla}) - t_{.05}s_{\overline{X}_{dr}-\overline{X}_{pla}} < \mu_{dr} - \mu_{pla} < (\overline{X}_{dr} - \overline{X}_{pla}) + t_{.05}s_{\overline{X}_{dr}-\overline{X}_{pla}}$$

$$250 - 2.101(89.9) < \mu_{dr} - \mu_{pla} < 250 + 2.101(89.9)$$

$$61 \text{ mL/day} < \mu_{dr} - \mu_{pla} < 439 \text{ mL/day}$$

Thus, on the basis of this particular experiment, we can be 95 percent confident that the drug increases average urine production somewhere between 61 and 439 mL/day. The *range* of values from 61 to 439 *is* the 95 percent *confidence interval* corresponding to this experiment. As Fig. 7-1*A* shows, this interval includes the actual change in mean urine production, $\mu_{dr} - \mu_{pla}$, 200 mL/day.

More Experiments

Of course, there is nothing special about the two samples of 10 people each selected in the study we just analyzed. Just as the values of the sample mean and standard deviation vary with the specific random sample of people we happen to draw, so will the confidence interval we

---

Figure 7-1 (*A*) The 95 percent confidence interval for the change in urine production produced by the drug using the random samples shown in Fig. 6-1. The interval contains the true change in urine production, 200 mL/day (indicated by the dashed line). Since the interval does not include zero (indicated by the solid line), we can conclude the that drug increases urine output ($P < .05$). (*B*) The 95 percent confidence interval for change in urine production computed for the random samples shown in Fig. 6-2. The interval includes the actual change in urine production (200 mL/day), but it also includes zero, so that it is not possible to reject the hypothesis of no drug effect (at the 5 percent level). (*C*) The 95 percent confidence intervals for 48 more sets of random samples, e.g., experiments, drawn from the two populations in Fig. 6-1*A*. All but 3 of the 50 intervals shown in this figure include the actual change in urine production; 5 percent of *all* possible 95 percent confidence intervals will not include the 200 mL/day. Of the 50 confidence intervals, 22 include zero, meaning that the data do not permit rejecting the hypothesis of no difference at the 5 percent level. In these cases, we would make a Type II error. Since 45 percent of *all* possible 95 percent confidence intervals include zero, the probability of detecting a change in urine production is $1 - \beta = .55$.

compute from the resulting observations. (This should not be surprising, since the confidence interval is computed from the sample means and standard deviations.) The confidence interval we just computed corresponds to the specific random sample of individuals shown in Fig. 6-1. Had we selected a *different random sample* of people, say those in Fig. 6-2, we would have obtained a *different 95 percent confidence interval* for the size of the treatment effect.

The individuals selected at random for the experiment in Fig. 6-2 show a mean urine production of 1216 mL/day for the people taking the placebo and 1368 mL/day for the people taking the drug. The standard deviations of the two samples are 97 and 263 mL/day, respectively. In these two samples, the drug increased average urine production by $\overline{X}_{dr} - \overline{X}_{pla} = 1368 - 1216 = 152$ mL/day. The pooled estimate of the population variance is

$$s^2 = \tfrac{1}{2}(97^2 + 263^2) = 198^2$$

in which case,

$$s_{\overline{X}_{dr}-\overline{X}_{pla}} = \sqrt{\frac{198^2}{10} + \frac{198^2}{10}} = 89.9 \text{ mL/day}$$

So the 95 percent confidence interval for the mean change in urine production associated with the sample shown in Fig. 6-2 is

$$152 - 2.101(89.9) < \mu_{dr} - \mu_{pla} < 152 + 2.101(89.9)$$

$$-35 \text{ mL/day} < \mu_{dr} - \mu_{pla} < 339 \text{ mL/day}$$

This interval, while different from the first one we computed, also includes the actual mean increase in urine production, 200 mL/day (Fig. 7-1*B*). Had we drawn this sample rather than the one in Fig. 6-1, we would have been 95 percent confident that the drug increased average urine production somewhere between −35 and 339 mL/day. (Note that this interval includes negative values, indicating that the data do not permit us to exclude the possibility that the drug decreased as well as increased average urine production. This observation is the basis for using confidence intervals to test hypotheses later in this chapter.) In sum, *the specific 95 percent confidence interval we obtain depends on the specific random sample we happen to select for observation.*

So far, we have seen two such intervals that could arise from random sampling of the populations in Fig. 6-1; there are more than 10^{27} possible samples of 10 people each, so there are more than 10^{27} possible 95 percent confidence intervals. Figure 7-1*C* shows 48 more of them, computed by selecting two samples of 10 people each from the populations of placebo and drug takers. Of the 50 intervals shown in Fig. 7-1, all but 3 (about 5 percent) include the value of 200 mL/day, the actual change in average urine production associated with the drug.

WHAT DOES "CONFIDENCE" MEAN?

We are now ready to attach a precise meaning to the term *95 percent confident.* The specific 95 percent confidence interval associated with a given set of data will or will not actually include the true size of the treatment effect, but in the long run 95 percent of *all possible 95 percent confidence intervals* will include the true difference of mean values associated with the treatment. As such, it describes not only the size of the effect but quantifies the certainty with which one can estimate the size of the treatment effect.

The size of the interval depends on the level of confidence you want to have that it will actually include the true treatment effect. Since t_α increases as α decreases, requiring a greater and greater fraction of all possible confidence intervals to cover the true effect will make the intervals larger. To see this, let us compute the 90 percent, 95 percent, and 99 percent confidence intervals associated with the data in Fig. 6-1. To do so, we need only substitute the values of $t_{.10}$ and $t_{.01}$ corresponding to $\nu = 18$ from Table 4-1 for t_α in the formula derived above. (We have already solved the problem for $t_{.05}$.)

For the 90 percent confidence interval, $t_{.10} = 1.734$, so the interval associated with the samples in Fig. 6-1 is

$$250 - 1.734(89.9) < \mu_{dr} - \mu_{pla} < 250 + 1.734(89.9)$$

$$90 \text{ mL/day} < \mu_{dr} - \mu_{pla} < 410 \text{ mL/day}$$

which, as Fig. 7-2 shows, is narrower than the 95 percent interval. Does this mean the data now magically yield a more precise estimate of the

treatment effect? No. If you are willing to accept the risk that 10 percent of all possible confidence intervals will not include the true change in mean values, you can get by with a narrower interval.

On the other hand, if you want to specify an interval selected from a population of confidence intervals, 99 percent of which include the true change in population means, you compute the confidence interval with $t_{.01} = 2.552$. The 99 percent confidence interval associated with the samples in Fig. 6-1 is

$$250 - 2.552(89.9) < \mu_{dr} - \mu_{pla} < 250 + 2.552(89.9)$$

$$21 \text{ mL/day} < \mu_{dr} - \mu_{pla} < 479 \text{ mL/day}$$

This interval is wider than the other two in Fig. 7-2.

In sum, the confidence interval gives a range that is computed in the hope that it will include the parameter of interest (in this case, the

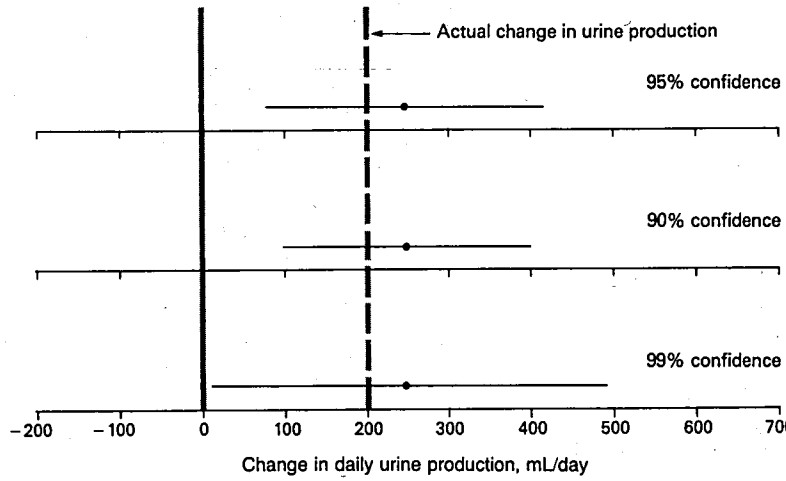


Figure 7-2 Increasing the level of confidence you wish to have that a confidence interval includes the true treatment effect makes the interval wider. All the confidence intervals in this figure were computed from the two random samples shown in Fig. 6-1. The 90 percent confidence interval is narrower than the 95 percent confidence interval, and the 99 percent confidence interval is wider. The actual change in urine production, 200 mL/day, is indicated with the dashed line.

difference of two population means). The confidence level associated with the interval (say 95, 90, or 99 percent) gives the percentage of all such possible intervals that will actually include the true value of the parameter. A *particular* interval will or will not include the true value of the parameter. Unfortunately, you can never know whether or not that interval does. All you can say is that the chances of selecting an interval that does not include the true value is small (say 5, 10, or 1 percent). The more confidence you wish to have that the interval will cover the true value, the wider the interval.

CONFIDENCE INTERVALS CAN BE USED TO TEST HYPOTHESES

As already noted, confidence intervals can provide another route to testing statistical hypotheses. This fact should not be surprising because we use all the same ingredients, the difference of the sample means, the standard error of the difference of sample means, and the value of t that corresponds to the biggest α fraction of the possible values defined by the t distribution with ν degrees of freedom.

Given a confidence interval, one cannot say where within the interval the true difference in population means lies. If the confidence interval contains zero, the evidence represented by the experimental observations is not sufficient to rule out the possibility that $\mu_1 - \mu_2 = 0$, that is, that $\mu_1 = \mu_2$, the hypothesis that the t test tests. Hence, we can state the following rule as follows.

If the 100(1 − α) percent confidence interval associated with a set of data includes zero, there is not sufficient evidence to reject the hypothesis of no effect with $P < \alpha$. If the confidence interval does not include zero, there is sufficient evidence to reject the hypothesis of no effect with $P < \alpha$.

Apply this rule to the two examples just discussed. The 95 percent confidence interval in Fig. 7-1*A* does not include zero, so we can report that the drug produced a statistically significant change in urine production ($P < .05$), just as we did using the t test. The 95 percent confidence interval in Fig. 7-1*B* includes zero, so the random sample (shown in Fig. 6-2) used to compute it does not provide sufficient evidence to reject the hypothesis that the drug has no effect. This, too, is the same conclusion we reached before.

Of the fifty 95 percent confidence intervals shown in Fig. 7-1, twenty-two include zero. Hence $^{22}/_{50} = 44$ percent of these random samples do not permit reporting a difference with 95 percent confidence, i.e., with $P < .05$. If we looked at all possible 95 percent confidence intervals computed for these two populations with two samples of 10 people each, we would find that 45 percent of them include zero, meaning that we would fail to report a true difference, that is, would make a Type II error, 45 percent of the time. Hence, $\beta = .45$, and the power of the test is .55, just as before (compare Fig. 6-4).

The confidence-interval approach to hypothesis testing offers two potential advantages. In addition to permitting you to reject the hypothesis of no effect when the interval does not include zero, it also gives information about the size of the effect. Thus, if a result reaches statistical significance more because of a large sample size than because of a large treatment effect, the confidence interval will show it. In other words, it will make it easier to recognize effects that can be detected with confidence but are too small to be of clinical or scientific significance.

For example, suppose we wish to study the potential value of a proposed antihypertensive drug. We select two samples of 100 people each and administer a placebo to one group and the drug to the other. The treated group has a mean diastolic pressure of 81 mmHg and a standard deviation of 11 mmHg; the control (placebo) group has a mean blood pressure of 85 mmHg and a standard deviation of 9 mmHg. Are these data consistent with the hypothesis that the diastolic blood pressure among people taking the drug and placebo were actually no different? To answer this question, we use the data to complete a t test. The pooled-variance estimate is

$$s^2 = {}^1/_2(11^2 + 9^2) = 10^2$$

so

$$t = \frac{\overline{X}_{dr} - \overline{X}_{pla}}{s_{\overline{X}_{dr} - \overline{X}_{pla}}} = \frac{81 - 85}{\sqrt{(10^2/100) + (10^2/100)}} = -2.83$$

This value is more negative than -2.61, the critical value of t that defines the 1 percent most extreme of the t distribution with $\nu = 2(n - 1) = 198$ degrees of freedom (from Table 4-1). Thus, we can assert that the drug lowers diastolic blood pressure ($P < .01$).

But is this result clinically significant? To gain a feeling for this, compute the 95 percent confidence interval for the mean difference in diastolic blood pressure for people taking placebo versus the drug. Since $t_{.05}$ for 198 degrees of freedom is (from Table 4-1) 1.973, the confidence interval is

$$-4 - 1.973(1.42) < \mu_{dr} - \mu_{pla} < -4 + 1.973(1.42)$$
$$-6.88 \text{ mmHg} < \mu_{dr} - \mu_{pla} < -1.2 \text{ mmHg}$$

In other words, we can be 95 percent confident that the drug lowers blood pressure between 1.2 and 6.8 mmHg. This is not a very large effect, especially when compared with standard deviations of the blood pressures observed within each of the samples, which are around 10 mmHg. Thus, while the drug does seem to lower blood pressure on the average, examining the confidence interval permitted us to see that the size of the effect is not very impressive. The small value of P was more a reflection of the sample size than the size of the effect on blood pressure.

This example also points up the importance of examining not only the P values reported in a study but also the *size* of the treatment effect compared with the variability within each of the treatment groups. Usually this requires converting the standard errors of the mean reported in the paper to standard deviations by multiplying them by the square root of the sample size. This simple step often shows clinical studies to be of potential interest in illuminating physiological mechanisms but of little value in diagnosing or managing a specific patient because of person-to-person variability.

CONFIDENCE INTERVAL FOR THE POPULATION MEAN

The procedure we developed above can be used to compute a confidence interval for the mean of population from which a sample was drawn. The resulting confidence interval is the origin of the rule, stated in Chapter 2, that the true (and unobserved) mean of the original population will lie within 2 standard errors of the mean of the sample mean for about 95 percent of all possible samples.

The confidence intervals we computed up to this point are based on the fact that

$$t = \frac{\text{difference of sample means} - \text{difference in population means}}{\text{standard error of difference of sample means}}$$

follows the t distribution. It is also possible to show that

$$t = \frac{\text{sample mean} - \text{population mean}}{\text{standard error of mean}}$$

follows the t distribution. The equivalent mathematical statement is

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}}$$

We can compute the $100(1 - \alpha)$ percent confidence interval for the population mean by obtaining the value of t_α corresponding to $\nu = n - 1$ degrees of freedom, in which n is the sample size. Substitute this value for t in the equation and solve for μ (just as we did for $\mu_1 - \mu_2$ earlier).

$$\overline{X} - t_\alpha s_{\overline{X}} < \mu < \overline{X} + t_\alpha s_{\overline{X}}$$

The interpretation of the confidence interval for the mean is analogous to the interpretation of the confidence interval for the difference of two means: every possible random sample of a given size can be used to compute a, say, 95 percent confidence interval for the population mean, and this same percentage (95 percent) of all such intervals will include the true population mean.

It is common to approximate the 95 percent confidence interval with the sample mean plus or minus twice the standard error of the mean because the values of $t_{.05}$ are approximately 2 for sample sizes above about 20 (see Table 4-1). This approximate rule of thumb does underestimate the size of the confidence interval for the mean, however, especially for the small sample sizes common in biomedical research.

THE SIZE OF THE TREATMENT EFFECT MEASURED AS THE DIFFERENCE OF TWO RATES OR PROPORTIONS

It is easy to generalize the procedures we just developed to permit us to compute confidence intervals for rates and proportions. In Chapter 5 we

used the statistic

$$z = \frac{\text{difference of sample proportions}}{\text{standard error of difference of proportions}}$$

to test the hypothesis that the observed proportions of events in two samples were consistent with the hypothesis that the event occurred at the same rate in the two populations. It is possible to show that even when the two populations have different proportions of members with the attribute, the ratio

$$z = \frac{\text{difference of sample proportions} - \text{difference of population proportions}}{\text{standard error of difference of sample proportions}}$$

is distributed approximately according to the normal distribution so long as the sample sizes are large enough.

If p_1 and p_2 are the actual proportions of members of each of the two populations with the attribute, and if the corresponding estimates computed form the samples are $\hat{p}_1$ and $\hat{p}_2$, respectively,

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{s_{\hat{p}_1 - \hat{p}_2}}$$

We can use this equation to define the $100(1 - \alpha)$ percent confidence interval for the difference in proportions by substituting z_α for z in this equation and solving just as we did before. z_α is the value that defines the most extreme α proportion of the values in the normal distribution;* $z_\alpha = z_{.05} = 1.960$ is commonly used, since it is used to define the 95 percent confidence interval. Thus,

$$(\hat{p}_1 - \hat{p}_2) - z_\alpha s_{\hat{p}_1 - \hat{p}_2} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_\alpha s_{\hat{p}_1 - \hat{p}_2}$$

for $100(1 - \alpha)$ percent of all possible samples.

Difference in Mortality Associated with Anesthesia for Open-Heart Surgery

In Chapter 5 we tested the hypothesis that the mortality rates associated with halothane and morphine anesthesia were no different. What is the

*This value can also be obtained from a t table, e.g., Table 4-1, by taking the value of t corresponding to an infinite number of degrees of freedom.

95 percent confidence interval for the difference in mortality rate for these two agents?

The mortality rates observed with these two anesthetic agents were 13.1 percent (8 of 61 people) and 14.9 percent (10 of 67 people). Therefore, the difference in observed mortality rates is $\hat{p}_{\text{hlo}} - \hat{p}_{\text{mor}} = 0.131 - 0.15 = -0.020$ and the standard error of the difference, based on a pooled estimate of the proportion of all patients who died is,

$$\hat{p} = \frac{8 + 10}{61 + 67} = .14$$

$$s_{\hat{p}_{\text{hlo}} - \hat{p}_{\text{mor}}} = \sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_{\text{hlo}}} + \frac{1}{n_{\text{mor}}}\right)}$$

$$= \sqrt{.14(1 - .14)\left(\frac{1}{61} + \frac{1}{67}\right)} = .062 = 6.2\%$$

Therefore, the 95 percent confidence interval for the difference in mortality rates is

$$(\hat{p}_{\text{hlo}} - \hat{p}_{\text{mor}}) - z_{.05}s_{\hat{p}_{\text{hlo}} + \hat{p}_{\text{mor}}} < s_{\hat{p}_{\text{hlo}} - \hat{p}_{\text{mor}}} < (\hat{p}_{\text{hlo}} - \hat{p}_{\text{mor}}) + z_{.05}s_{\hat{p}_{\text{hlo}} + \hat{p}_{\text{mor}}}$$

$$(\hat{p}_{\text{hlo}} - \hat{p}_{\text{mor}}) - z_{.05}s_{\hat{p}_{\text{hlo}} + \hat{p}_{\text{mor}}} < s_{\hat{p}_{\text{hlo}} - \hat{p}_{\text{mor}}} < (\hat{p}_{\text{hlo}} - \hat{p}_{\text{mor}})$$

$$-.020 - 1.960(.062) < s_{\hat{p}_{\text{hlo}} - \hat{p}_{\text{mor}}} < -.020 + 1.960(.062)$$

$$-.142 < s_{\hat{p}_{\text{hlo}} - \hat{p}_{\text{mor}}} < .102$$

We can be 95 percent confident that the true difference in mortality rate lies between a 14.2 percent better rate for morphine and a 10.2 percent better rate for halothane.* Since the confidence interval contains zero, there is not sufficient evidence to reject the hypothesis that the two anesthetic agents are associated with the same mortality rate. Furthermore, the confidence interval ranges about equally on both sides of zero, so there is not even a suggestion that one agent is superior to the other.

*To include the Yates correction, widen the upper and lower bounds of the confidence interval by $\frac{1}{2}(1/n_{\text{hlo}} + 1/n_{\text{mor}})$.

Difference in Thrombosis with Aspirin in People Receiving Hemodialysis

Chapter 5 also discussed the evidence that administering low-dose aspirin to people receiving regular kidney dialysis reduces the proportion of people who develop thrombosis. Of the people taking the placebo, 72 percent developed thrombosis, and 32 percent of the people taking aspirin did. Given only this information, we would report that aspirin reduced the proportion of patients who developed thrombosis by 40 percent. What is the 95 percent confidence interval for the improvement?

The standard error of the difference in proportion of patients who developed thrombosis is .15 (from Chapter 5). So the 95 percent confidence interval for the true difference in proportion of patients who developed thrombosis is

$$.40 - 1.96(.15) < p_{pla} - p_{asp} < .40 + 1.96(.15)$$
$$.11 < p_{pla} - p_{asp} < .69$$

We can be 95 percent confident that aspirin reduces the rate of thrombosis somewhere between 11 and 69 percent compared with placebo.

How Negative Is a "Negative" Clinical Trial?

Chapter 6 discussed the study of 71 randomized clinical trials that did not demonstrate a statistically significant improvement in clinical outcome (mortality, complications, or the number of patients who showed no improvement, depending on the study). Most of these trials involved too few patients to have sufficient power to be confident that the failure to detect a treatment effect was not due to an inadequate sample size. To get a feeling for how compatible the data are with the hypothesis of no treatment effect, let us examine the 90 percent confidence intervals for the proportion of "successful" cases (the definition of success varied with the study) for all 71 trials. Figure 7-3 shows these confidence intervals.

All the confidence intervals include zero, so we cannot rule out the possibility that the treatments had no effect. Note, however, that some

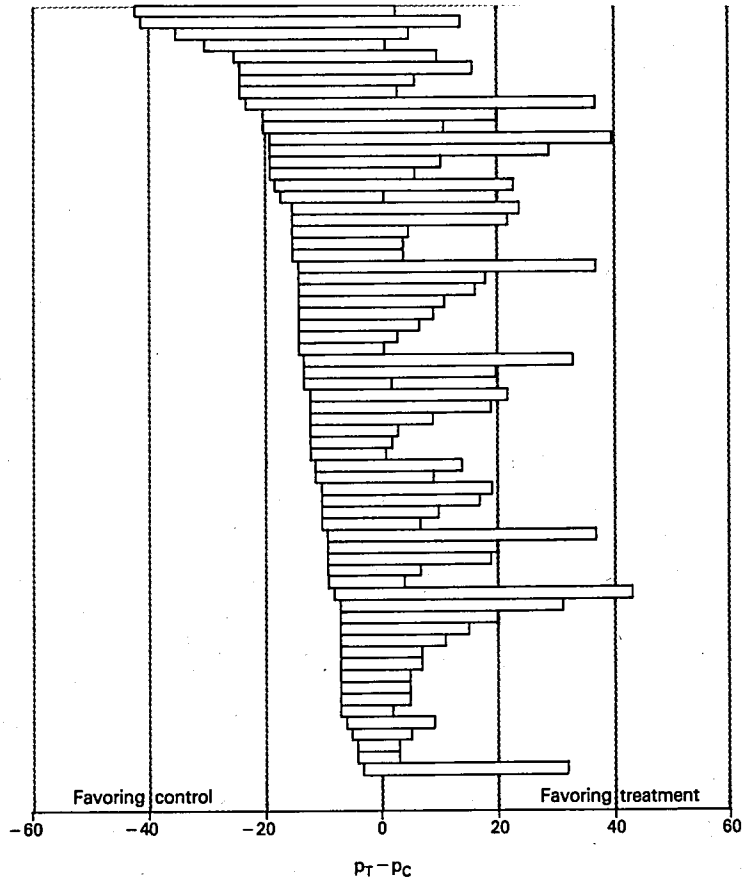


Figure 7-3 The 90 percent confidence intervals for 71 negative clinical trials. Since all the intervals contain zero, there is not sufficient evidence that the success rate is different for the treatment and control groups. Nevertheless, the data are also compatible with the treatment producing a substantial improvement in success rate in many of the trials. *(Data from Fig. 2 of J. A. Freiman, T. C. Chalmers, H. Smith, Jr., and R. R. Keubler, "The Importance of Beta, the Type II Error and Sample Size in the Design and Interpretation of the Randomized Control Trial: Survey of 71 'Negative' Trials,"* N. Engl. J. Med., *299:690–694, 1978.)*

of the trials are also compatible with the possibility that the treatments produced sizable improvements in the success rate. Remember that while we can be 90 percent confident that the true change in proportion of successes lies in the interval, it could be anywhere. Does this prove that some of these treatments improved clinical outcome? No. The important point is that the confidence with which we can assert that there was no treatment effect is often the same as the confidence with which we can assert that the treatment produced a sizable improvement. While the size and location of the confidence interval cannot be used as part of a formal statistical argument to prove that the treatment had an effect, it certainly can help you look for trends in the data.

CONFIDENCE INTERVAL FOR RATES AND PROPORTIONS

It is possible to use the normal distribution to compute approximate confidence intervals for proportions from observations, so long as the sample size is large enough to make the approximation reasonably accurate.* When it is not possible to use this approximation, we will compute the exact confidence intervals based on the binomial distribution. While we will not go into the computational details of this procedure, we will present the necessary results in graphical form because papers often present results based on small numbers of subjects. Examining the confidence intervals as opposed to only the observed proportion of patients with a given attribute is especially useful in thinking about such studies, because a change of *a single patient* from one group to the other often makes a large difference in the observed proportion of patients with the attribute of interest.

Just as there was an analogous way to use the t distribution to relate the difference of means and the confidence interval for a single sample mean, it is possible to show that if the sample size is large enough

$$z = \frac{\text{observed proportion} - \text{true proportion}}{\text{standard error of proportion}}$$

*As discussed in Chapter 5, $n\hat{p}$ and $n(1-\hat{p})$ must both exceed about 5, where $\hat{p}$ is the proportion of the observed sample having the attribute of interest.

In other words

$$z = \frac{\hat{p} - p}{s_{\hat{p}}}$$

approximately follows the normal distribution (in Table 6-4). Hence, we can use this equation to define the $100(1 - \alpha)$ percent confidence interval for the true proportion p with

$$\hat{p} - z_\alpha s_{\hat{p}} < p < \hat{p} + z_\alpha s_{\hat{p}}$$

The Fraction of Articles with Statistical Errors

From the 1950s through the 1970s, the fraction of articles published in medical journals that include errors in the use of statistical procedures has remained around 50 percent. For example, Sheila Gore and colleagues* found that 32 of the 77 original papers published in the *British Medical Journal* between January and March 1976 contained at least one error in their use of statistical procedures. What is the 95 percent confidence interval for the proportion of all articles published in journals of comparable quality at that time?

The proportion of articles with errors is $\hat{p} = {}^{32}/_{77} = .42$, and the standard error of the proportion is $s_{\hat{p}} = \sqrt{.42(1 - .42)/77} = .056$. Therefore, the 95 percent confidence interval is

$$.42 - 1.96(.056) < p < .42 + 1.96(.056)$$
$$.31 < p < .53$$

How potentially serious are these errors? Gore and colleagues reported that 5 of the 62 analytical reports that made some use of statistical procedures made some claim in the summary that was

*S. M. Gore, I. G. Jones, and E. C. Rytter, "Misuse of Statistical Methods: Critical Assessment of Articles in BMJ from January to March 1976," *Br. Med. J.*, 1(6053):85–87, 1977.

not supported by the data presented. Thus $\hat{p} = {}^{5}/_{62} = .081$ and $s_{\hat{p}} = \sqrt{.081(1 - .081)/62} = .035$, so the 95 percent confidence interval for the proportion of papers with conclusions not supported by the data is

$$.081 - 1.960(.035) < p < .081 + 1.960(.035)$$

or from 1 to 15 percent.

Exact Confidence Intervals for Rates and Proportions

When the sample size or observed proportion is too small for the approximate confidence interval based on the normal distribution to be reliable, you have to compute the confidence interval based on the exact theoretical distribution of a proportion, the *binomial distribution.** Since results based on small sample sizes with low observed rates of events turn up frequently in the medical literature, we present the results of computation of confidence intervals using the binomial distri-bution.

To illustrate how the procedure we followed above can fall apart when $n\hat{p}$ is below about 5, we consider an example. Suppose a surgeon says that he has done 30 operations without a single complication. His observed complication rate $\hat{p}$ is ${}^{0}/_{30} = 0$ percent for the 30 specific patients he operated on. Impressive as this is, it is unlikely that the surgeon will continue operating forever without a complication, so the fact that $\hat{p} = 0$ probably reflects good luck in the randomly selected patients who happened to be operated on during the period in question. To obtain a better estimate of p, the surgeon's true complication rate, we will compute the 95 percent confidence interval for p.

*The reason we could use the normal distribution here and in Chapter 5 is that for large enough sample sizes there is little difference between the binomial and normal distributions. This result is a consequence of the central-limit theorem, discussed in Chapter 2. For the actual derivation of these results, see W. J. Dixon and F. J. Massey, *Introduction to Statistical Analysis* (4th ed.), McGraw-Hill, New York, 1983, sec. 13–5, "Binomial Distribution: Proportion," or B. W. Brown, Jr., and M. Hollander, *Statistics: A Biomedical Introduction*, Wiley, New York, 1977, chapter 7, "Statistical Inference for Dichotomous Variables."

Let us try to apply our existing procedure. Since $\hat{p} = 0$,

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = \sqrt{\frac{0(1 - 0)}{30}} = 0$$

and the 95 percent confidence interval is from zero to zero. This result does not make sense. There is no way that a surgeon can *never* have a complication. Obviously, the approximation breaks down.

Figure 7-4 gives a graphical presentation of the 95 percent confidence intervals for proportions. The upper and lower limits are read off the vertical axis using the pair of curves corresponding to the size of the sample n used to estimate $\hat{p}$ at the point on the horizontal axis corresponding to the observed $\hat{p}$. For our surgeon, $\hat{p} = 0$ and $n = 30$, so the 95 percent confidence interval for his true complication rate is from 0 to .10. In other words, we can be 95 percent confident that his true complication rate, based on the 30 cases we happened to observe, is somewhere between 0 and 10 percent.

Now, suppose the surgeon had a single complication. Then $\hat{p} = 1/30 = 0.033$ and

$$s_{\hat{p}} = \sqrt{.033(1 - .033)/30} = .033$$

so the 95 percent confidence interval for the true complication rate, computed using the approximate method, is

$$.33 - 1.96(.033) < p < .33 + 1.96(.033)$$

$$-.032 < p < .098$$

Think about this result for a moment. There is no way a surgeon can have a *negative* complication rate.

Figure 7-4 gives the exact confidence interval, from 0 to .13, or 0 to 13 percent.* This confidence interval is not too different from that

*When there are no "failures" observed, the approximate upper end of the 95% confidence interval for the true failure rate is approximately $3/n$, where n is the sample size. For a more extensive discussion of interpreting results when there are no "failures," see J. A. Hanley and A. Lippman-Hand, "If Nothing Goes Wrong, Is Everything All Right? Interpreting Zero Numerators," *JAMA* **249:**1743–1745, 1983.

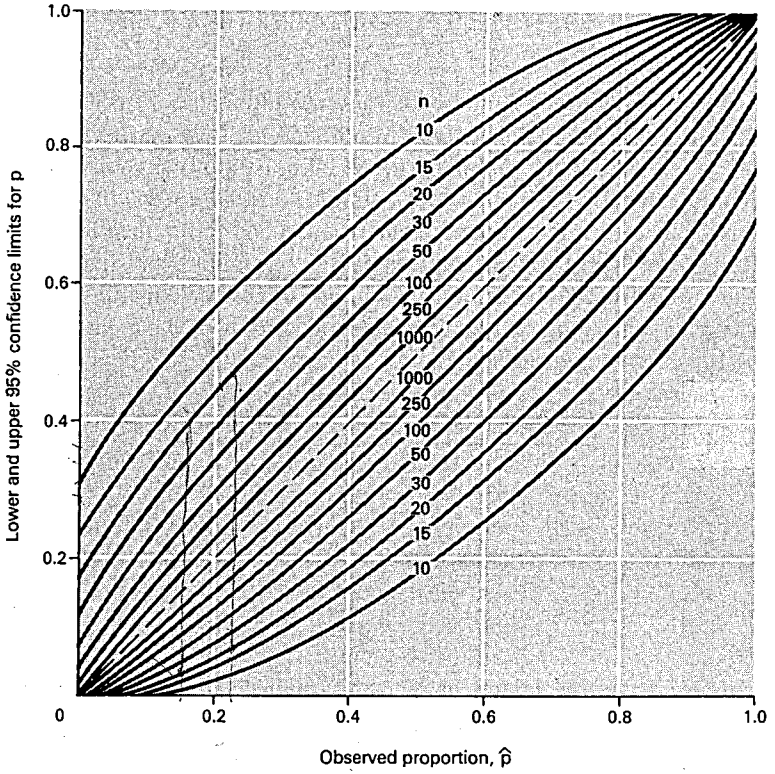


Figure 7-4 Graphical presentation of the exact 95 percent confidence intervals (based on the binomial distribution) for the population proportion. You read this plot by reading the two limits of the lines defined by the sample size at the point on the horizontal axis at the proportion of the sample with the attribute of interest $\hat{p}$. *(Adapted from C. J. Clopper and E. S. Pearson, "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial,"* Biometrika, *26:404, 1934.)*

computed when there were no complications, as it should be, since there is little real difference between not having any complications and having only one complication in such a small sample.

Notice how important sample size is, especially for small sample sizes. Had the surgeon been bragging that he had a zero complication rate on the basis of only 10 cases, the 95 percent confidence interval for his true complication rate would have extended from zero all the way to 33 percent!

CONFIDENCE INTERVALS FOR RELATIVE RISK AND ODDS RATIO*

Because the relative risk and odds ratio are ratios, the distributions of the values of these statistics are not normally distributed. It turns out, however, that the logarithm of these ratios is normally distributed. Therefore, we can use approaches similar to those used with proportions to the logarithms of the relative risk and odds ratio, then invert the results to return to the original scale. By convention, statisticians and epidemiologists use the natural logarithm for these calculations.† Using the notation in Table 5-14, the natural logarithm of the relative risk, ln RR, is normally distributed with standard error

$$S_{\ln RR} = \sqrt{\frac{1 - a/(a+b)}{a} + \frac{1 - c/(c+d)}{c}}$$

Therefore, the $100(1 - \alpha)$ percent confidence interval for the natural logarithm of the true population $\ln RR_{true}$ is

$$\ln RR - z_\alpha s_{s_{\ln RR}} < \ln RR_{true} < \ln RR + z_\alpha s_{s_{\ln RR}}$$

We convert these estimates back to the original units by applying the exponential function to the terms in this equation to obtain

$$e^{\ln RR - z_\alpha s_{s_{\ln RR}}} < RR_{true} < e^{\ln RR + z_\alpha s_{s_{\ln RR}}}$$

Thus, you could test the null hypothesis that the true RR = 1, that the treatment (or risk factor) had no effect, by computing this confidence interval and seeing if it included 1.0.

Likewise, the natural logarithm of the odds ratio, OR, is normally distributed. Using the notation in Table 5-15, the standard error is

$$S_{\ln OR} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

*In an introductory course, this section can be skipped without any loss of continuity.

†The natural logarithm has the base $e = 2.71828\ldots$ rather than 10, which is the base of the common logarithm. Because e is the base, the natural logarithm and exponential functions are *inverses*, i.e., $e^{\ln x} = x$ and $\ln e^x = x$.

and the 100(1-α) percent confidence interval for the true odds ratio is

$$e^{\ln \text{OR} - z_\alpha s_{\ln \text{OR}}} < \text{OR}_{\text{true}} < e^{\ln \text{OR} + z_\alpha s_{\ln \text{OR}}}$$

This confidence interval can also be used to test the null hypothesis that the true OR = 1, that exposure to the risk factor is not associated with an increase in the odds of having the disease.

Difference in Thrombosis with Aspirin in People Receiving Hemodialysis

Earlier in this chapter we saw how to use a confidence interval to test the null hypothesis that there was no difference in the probability of thrombosis in people receiving aspirin or a placebo. We can also test this hypothesis by examining the relative risk of thrombosis in this clinical trial. In Chapter 5, we estimated that the relative risk of a thrombosis was .44 in people receiving aspirin compared to people receiving a placebo. Using the data in Table 5-1, which shows that $a = 6$, $b = 18$, $c = 13$, and $d = 7$, we estimate the standard error of ln RR as

$$S_{\ln \text{RR}} = \sqrt{\frac{1 - 6/(6 + 18)}{6} + \frac{1 - 13/(13 + 7)}{13}} = .390$$

To estimate the 95 percent confidence interval, we note that $z_{.05} = 1.96$ and compute

$$e^{\ln.44 - 1.96 \cdot .390} < \text{RR}_{\text{true}} < e^{\ln.44 - 1.96 \cdot .390}$$

$$e^{-1.586} < \text{RR}_{\text{true}} < e^{-.057}$$

$$.20 < \text{RR}_{\text{true}} < .94$$

Hence, we can be 95 percent confident that the true relative risk of thrombosis for people taking aspirin compared with placebo is somewhere between .20 and .94. Because this range does not include 1, we can conclude that aspirin significantly changes the risk of thrombosis, and prevents thrombosis.

Passive Smoking and Breast Cancer

We can compute the confidence interval for the odds ratio of a premenopausal woman who is exposed to secondhand smoke developing

breast cancer using the data in Table 5-16. To compute the 95 percent confidence interval for this odds ratio, we note that the observed odds ratio is 2.91 and, from Table 5-16, $a = 50$, $b = 14$, $c = 43$, and $d = 35$. Therefore,

$$S_{\ln OR} = \sqrt{\frac{1}{50} + \frac{1}{14} + \frac{1}{43} + \frac{1}{35}} = .378$$

and, so,

$$e^{\ln 2.91 - 1.96 \cdot .378} < RR_{true} < e^{\ln 2.91 - 1.96 \cdot .378}$$

$$e^{.327} < RR_{true} < e^{1.809}$$

$$1.39 < RR_{true} < 6.10$$

Thus, we can be 95 percent confident that the true odds ratio is somewhere between 1.39 and 6.10. Because the 95 percent confidence interval for the true relative risk excludes 1, we conclude that passive smoking significantly increases the odds of breast cancer in premenopausal women.

CONFIDENCE INTERVAL FOR THE ENTIRE POPULATION*

So far computed intervals that we can have a high degree of confidence in will include a *population parameter*, such as μ or p. It is often desirable to determine a confidence interval for the *population itself*, most commonly when defining the normal range of some variable. The most common approach is to take the range defined by 2 standard deviations about the sample mean on the grounds that this interval contains 95 percent of the members of a population that follows the normal distribution (Fig. 2-5). In fact, in carefully worded language Chapter 2 suggested this rule. When the sample used to compute the mean and standard deviation is large (more than 100 to 200 members), this common rule of thumb is reasonably accurate. Unfortunately, most clinical

*Confidence intervals for the population are also called *tolerance limits*. The procedures derived in this section are appropriate for analyzing data obtained from a population that is normally distributed. If the population follows other distributions, there are alternate procedures for computing confidence intervals for the population.

studies are based on much smaller samples (of the order of 5 to 20 individuals). With such small samples, use of this two standard deviations rule of thumb seriously underestimates the range of values likely to be included in the population from which the samples were drawn.

For example, Fig. 2-6 showed the population of the heights of all 200 Martians, together with the results of three random samples of 10 Martians each. Figure 2-6*A* showed that 95 percent of all Martians have heights between 31 and 49 cm. The mean and standard deviation of the heights of population of all 200 Martians are 40 and 5 cm, respectively. The three samples illustrated in Fig. 2-6 yield estimates of the mean of 41.5, 36, and 40 cm, and of the standard deviation of 3.8, 5, and 5 cm, respectively. Suppose we simply compute the range defined by two *sample* standard deviations above and below the *sample* mean with the expectation that this range will include 95 percent of the population. Figure 7-5*A* shows the results of this computation for each

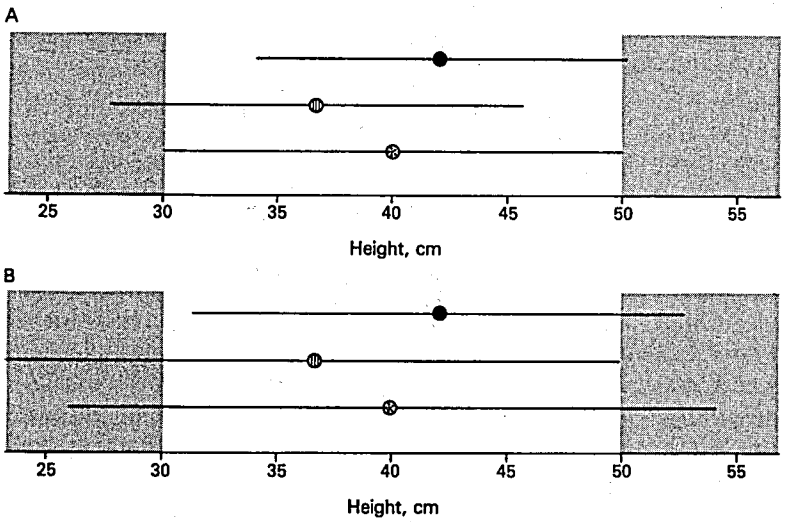


Figure 7-5 (*A*) The range defined by the sample mean ±2 standard deviations for the three samples of 10 Martians each shown in Fig. 2-6. Two of the three resulting ranges do *not* cover the entire range that includes 95 percent of the population members (indicated by the vertical lines). (*B*) The 95 percent confidence intervals for the population, computed as the sample mean $\pm K_{.05}$ times the sample standard deviation covers the actual range that includes 95 percent of the actual population; 95 percent of all such intervals will cover 95 percent of the actual population range.

of the three samples in Fig. 2-6. The light area defines the range of actual heights that covers 95 percent of the Martians' heights. Two of the three samples yield intervals that do not include 95 percent of the population.

This problem arises because both the sample mean and standard deviation are only *estimates* of the population mean and standard deviation and so cannot be used interchangeably with the population mean and standard deviation when computing the range of population values. To see why, consider the sample in Fig. 2-6*B* that yielded estimates of the mean and standard deviation of 36 and 5 cm, respectively. By good fortune, the estimate of the standard deviation computed from the sample equaled the population standard deviation. The estimate of the population mean, however, was low. As a result, the interval 2 standard deviations above and below the sample mean did not reach high enough to cover 95 percent of the entire population values. Because of the potential errors in the estimates of the population mean and standard deviation, we must be conservative and use a range greater than 2 standard deviations around the sample mean to be sure of including, say, 95 percent of the entire population. However, as the size of the sample used to estimate the mean and standard deviation increases, the certainty with which we can use these estimates to compute the range spanned by the entire population increases, so we do not have to be as conservative (i.e., take fewer multiples of the sample standard deviation) when computing an interval that contains a specified proportion of the population members.

Specifying the confidence interval for the entire population is more involved than specifying the confidence intervals we have discussed so far because you must specify both the *fraction of the population* f you wish the interval to cover and the *confidence you wish to have that any given interval will cover it.* The size of the interval depends on these two things and the size of the sample used to estimate the mean and standard deviation. The $100(1 - \alpha)$ percent confidence interval for $100f$ percent of the population is

$$\overline{X} - K_\alpha s < X < \overline{X} + K_\alpha s$$

in which $\overline{X}$ and s are the sample mean and standard deviation and K_α is the number of sample standard deviations about the sample mean needed to cover the desired part of the population. Figure 7-6 shows

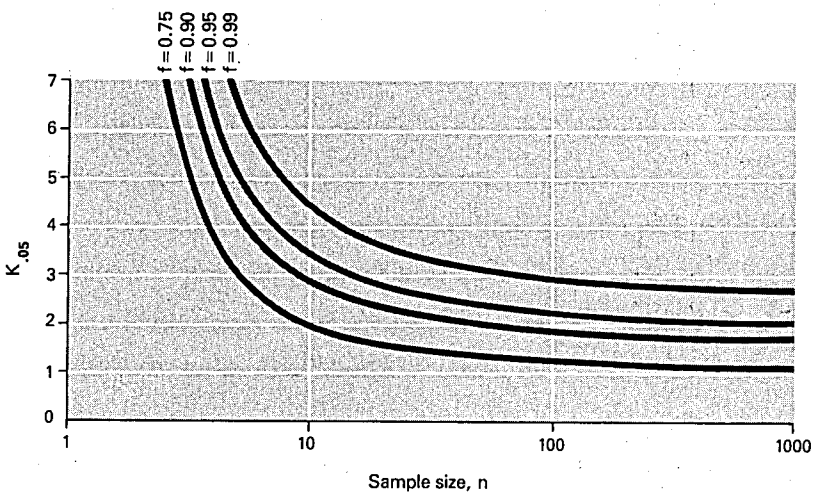


Figure 7-6 K_{05} depends on the size of the sample n used to estimate the mean and standard deviation and the fraction f of the population you want the interval to include.

K_{05} as a function of sample size for various values of f. It plays a role similar to t_α or z_α.

K_α is larger than t_α (which is larger than z_α) because it accounts for uncertainty in the estimates of both the mean *and* standard deviation, rather than the mean alone.*

Notice that K_α can be much larger than 2 for sample sizes in the range of 5 to 25, which are common in biomedical research. Thus, simply taking 2 standard deviations about the mean may substantially underestimate the range of the population from which the samples were drawn. Figure 7-5*B* shows the 95 percent confidence interval for 95 percent of the population of Martians' heights based on the three samples of 10 Martians each shown in Fig. 2-6. All three of the intervals include 95 percent of the population.

As Chapter 2 discussed, many people confuse the standard error of the mean with the standard deviation and consider the range defined by

*For a derivation of K_α that clearly shows how it is related to the confidence limits for the mean and standard deviation, see A. E. Lewis, *Biostatistics*, Reinhold, New York, 1966, chapter 12, "Tolerance Limits and Indices of Discrimination."

"sample mean $\pm$ 2 standard errors of the mean" to encompass about 95 percent of the population. This error leads them to seriously underestimate the possible range of values in the population from which the sample was drawn. We have seen that, for the relatively small sample sizes common in biomedical research, applying the 2 standard deviations rule may underestimate the range of values in the underlying population as well.

PROBLEMS

7-1 Find the 90 and 95 percent confidence intervals for the mean number of authors of articles published in the medical literature in 1946, 1956, 1966, and 1976 using the data from Prob. 2-6.

7-2 Problem 3-1 described an experiment in which women were treated with a gel containing prostaglandin E_2 and a placebo gel to see if the active gel would facilitate softening and dilation of the cervix during an induced labor. One reason for trying to facilitate cervical softening and dilation is to avoid having to do a cesarean section. C. O'Herlihy and H. MacDonald ("Influence of Preinduction Prostaglandin E_2 Vaginal Gel on Cervical Ripening and Labor," *Obstet. Gynecol.*, **54**:708–710, 1979) observed that 15 percent of the 21 women in the treatment group required cesarean sections and 23.9 percent of the 21 women in the control group required cesarean sections. Find the 95 percent confidence intervals for the percentage of all women having cesarean sections after receiving each treatment and the difference in cesarean section rate for all women in the two different groups. Can you be 95 percent confident that the prostaglandin E_2 gel reduces the chances that a woman whose labor is being induced will need to be delivered by a cesarean section?

7-3 Find the 95 percent confidence interval for the difference in the mean duration of labor in women treated with prostaglandin E_2 gel compared with women treated with placebo gel using the data in Prob. 3-1. Based on this confidence interval, is the difference statistically significant with $P < .05$?

7-4 Find the 95 percent confidence intervals for the proportion of both groups in Prob. 5-1 for which high-frequency neural modulation was an effective dental analgesic. Compare this result with the hypothesis tests computed in Prob. 5-1.

7-5 Find the 95 percent confidence intervals for the mean forced midexpiratory flows for the different test groups in Prob. 3-2. Use this information to identify people with different or similar lung function (as we did with Bonferroni t tests in Chapter 4).

7-6 Find the 95 percent confidence intervals for the percentage of articles that reported the results of research based on data collected before deciding on the question to be investigated. Use the data in Prob. 5-6.

7-7 Use the data in Prob. 2-3 to find the 95 percent confidence interval for 90 and 95 percent of the population of PCB concentrations in Japanese adults. Plot these intervals together with the observations.

7-8 Rework Prob. 5-11 using confidence intervals.

7-9 Rework Prob. 5-12 using confidence intervals.

7-10 Rework Prob. 5-13 using confidence intervals.

7-11 Rework Prob. 5-14 using confidence intervals.