Biostatistics

A Methodology for the Health Sciences

Second Edition

GERALD VAN BELLE
LLOYD D. FISHER
PATRICK J. HEAGERTY
THOMAS LUMLEY

Department of Biostatistics and
Department of Environmental and
Occupational Health Sciences
University of Washington
Seattle, Washington

WILEY-
INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication Data:

Biostatistics: a methodology for the health sciences / Gerald van Belle . . . [et al.]– 2nd ed.
p. cm. – (Wiley series in probability and statistics)
First ed. published in 1993, entered under Fisher, Lloyd.
Includes bibliographical references and index.
ISBN 0-471-03185-2 (cloth)
1. Biometry. I. Van Belle, Gerald. II. Fisher, Lloyd, 1939– Biostatistics. III. Series.

QH323.5.B562 2004
610′.1′5195–dc22

2004040491

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Ad majorem Dei gloriam

Contents

Preface to the First Edition

The purpose of this book is for readers to learn how to apply statistical methods to the biomedical sciences. The book is written so that those with no prior training in statistics and a mathematical knowledge through algebra can follow the text—although the more mathematical training one has, the easier the learning. The book is written for people in a wide variety of biomedical fields, including (alphabetically) biologists, biostatisticians, dentists, epidemiologists, health services researchers, health administrators, nurses, and physicians. The text appears to have a daunting amount of material. Indeed, there is a great deal of material, but most students will not cover it all. Also, over 30% of the text is devoted to notes, problems, and references, so that there is not as much material as there seems to be at first sight. In addition to not covering entire chapters, the following are optional materials: asterisks (*) preceding a section number or problem denote more advanced material that the instructor may want to skip; the notes at the end of each chapter contain material for extending and enriching the primary material of the chapter, but this may be skipped.

Although the order of authorship may appear alphabetical, in fact it is random (we tossed a fair coin to determine the sequence) and the book is an equal collaborative effort of the authors. We have many people to thank. Our families have been helpful and long-suffering during the writing of the book: for LF, Ginny, Brad, and Laura; for GvB, Johanna, Loeske, William John, Gerard, Christine, Louis, and Bud and Stacy. The many students who were taught with various versions of portions of this material were very helpful. We are also grateful to the many collaborating investigators, who taught us much about science as well as the joys of collaborative research. Among those deserving thanks are for LF: Ed Alderman, Christer Allgulander, Fred Applebaum, Michele Battie, Tom Bigger, Stan Bigos, Jeff Borer, Martial Bourassa, Raleigh Bowden, Bob Bruce, Bernie Chaitman, Reg Clift, Rollie Dickson, Kris Doney, Eric Foster, Bob Frye, Bernard Gersh, Karl Hammermeister, Dave Holmes, Mel Judkins, George Kaiser, Ward Kennedy, Tom Killip, Ray Lipicky, Paul Martin, George McDonald, Joel Meyers, Bill Myers, Michael Mock, Gene Passamani, Don Peterson, Bill Rogers, Tom Ryan, Jean Sanders, Lester Sauvage, Rainer Storb, Keith Sullivan, Bob Temple, Don Thomas, Don Weiner, Bob Witherspoon, and a large number of others. For GvB: Ralph Bradley, Richard Cornell, Polly Feigl, Pat Friel, Al Heyman, Myles Hollander, Jim Hughes, Dave Kalman, Jane Koenig, Tom Koepsell, Bud Kukull, Eric Larson, Will Longstreth, Dave Luthy, Lorene Nelson, Don Martin, Duane Meeter, Gil Omenn, Don Peterson, Gordon Pledger, Richard Savage, Kirk Shy, Nancy Temkin, and many others. In addition, GvB acknowledges the secretarial and moral support of Sue Goleeke. There were many excellent and able typists over the years; special thanks to Myrna Kramer, Pat Coley, and Jan Alcorn. We owe special thanks to Amy Plummer for superb work in tracking down authors and publishers for permission to cite their work. We thank Robert Fisher for help with numerous figures. Rob Christ did an excellent job of using LaTeX for the final version of the text. Finally, several people assisted with running particular examples and creating the tables; we thank Barry Storer, Margie Jones, and Gary Schoch.

Our initial contact with Wiley was the indefatigable Beatrice Shube. Her enthusiasm for our effort carried over to her successor, Kate Roach. The associate managing editor, Rose Ann Campise, was of great help during the final preparation of this manuscript.

With a work this size there are bound to be some errors, inaccuracies, and ambiguous statements. We would appreciate receiving your comments. We have set up a special electronic-mail account for your feedback:

http://www.biostat-text.info

Lloyd D. Fisher
Gerald van Belle

Preface to the Second Edition

Biostatistics did not spring fully formed from the brow of R. A. Fisher, but evolved over many years. This process is continuing, although it may not be obvious from the outside. It has been ten years since the first edition of this book appeared (and rather longer since it was begun). Over this time, new areas of biostatistics have been developed and emphases and interpretations have changed.

The original authors, faced with the daunting task of updating a 1000-page text, decided to invite two colleagues to take the lead in this task. These colleagues, experts in longitudinal data analysis, survival analysis, computing, and all things modern and statistical, have given a twenty-first-century thrust to the book.

The author sequence for the first edition was determined by the toss of a coin (see the Preface to the First Edition). For the second edition it was decided to switch the sequence of the first two authors and add the new authors in alphabetical sequence.

This second edition adds a chapter on randomized trials and another on longitudinal data analysis. Substantial changes have been made in discussing robust statistics, model building, survival analysis, and discrimination. Notes have been added, throughout, and many graphs redrawn. We have tried to eliminate errata found in the first edition, and while more have undoubtedly been added, we hope there has been a net improvement. When you find mistakes we would appreciate hearing about them at *http://www.vanbelle.org/biostatistics/*.

Another major change over the past decade or so has been technological. Statistical software and the computers to run it have become much more widely available—many of the graphs and new analyses in this book were produced on a laptop that weighs only slightly more than a copy of the first edition—and the Internet provides ready access to information that used to be available only in university libraries. In order to accommodate the new sections and to attempt to keep up with future changes, we have shifted some material to a set of Web appendices. These may be found at *http://www.biostat-text.info*. The Web appendices include notes, data sets and sample analyses, links to other online resources, all but a bare minimum of the statistical tables from the first edition, and other material for which ink on paper is a less suitable medium.

These advances in technology have not solved the problem of deadlines, and we would particularly like to thank Steve Quigley at Wiley for his equanimity in the face of schedule slippage.

GERALD VAN BELLE
LLOYD FISHER
PATRICK HEAGERTY
THOMAS LUMLEY

Seattle, June 15, 2003

CHAPTER 1

Introduction to Biostatistics

1.1 INTRODUCTION

We welcome the reader who wishes to learn biostatistics. In this chapter we introduce you to the subject. We define statistics and biostatistics. Then examples are given where biostatistical techniques are useful. These examples show that biostatistics is an important tool in advancing our biological knowledge; biostatistics helps evaluate many life-and-death issues in medicine.

We urge you to read the examples carefully. Ask yourself, "what can be inferred from the information presented?" How would you design a study or experiment to investigate the problem at hand? What would you do with the data after they are collected? We want you to realize that biostatistics is a tool that can be used to benefit you and society.

The chapter closes with a description of what you may accomplish through use of this book. To paraphrase Pythagoras, there is no royal road to biostatistics. You need to be involved. You need to work hard. You need to think. You need to analyze actual data. The end result will be a tool that has immediate practical uses. As you thoughtfully consider the material presented here, you will develop thought patterns that are useful in evaluating information in all areas of your life.

1.2 WHAT IS THE FIELD OF STATISTICS?

Much of the joy and grief in life arises in situations that involve considerable uncertainty. Here are a few such situations:

1. Parents of a child with a genetic defect consider whether or not they should have another child. They will base their decision on the chance that the next child will have the same defect.
2. To choose the best therapy, a physician must compare the *prognosis*, or future course, of a patient under several therapies. A therapy may be a success, a failure, or somewhere in between; the evaluation of the chance of each occurrence necessarily enters into the decision.
3. In an experiment to investigate whether a food additive is *carcinogenic* (i.e., causes or at least enhances the possibility of having cancer), the U.S. Food and Drug Administration has animals treated with and without the additive. Often, cancer will develop in both the treated and untreated groups of animals. In both groups there will be animals that do

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

not develop cancer. There is a need for some method of determining whether the group treated with the additive has “too much” cancer.

4. It is well known that “smoking causes cancer.” Smoking does not cause cancer in the same manner that striking a billiard ball with another causes the second billiard ball to move. Many people smoke heavily for long periods of time and do not develop cancer. The formation of cancer subsequent to smoking is not an invariable consequence but occurs only a fraction of the time. Data collected to examine the association between smoking and cancer must be analyzed with recognition of an uncertain and variable outcome.
5. In designing and planning medical care facilities, planners take into account differing needs for medical care. Needs change because there are new modes of therapy, as well as demographic shifts, that may increase or decrease the need for facilities. All of the uncertainty associated with the future health of a population and its future geographic and demographic patterns should be taken into account.

Inherent in all of these examples is the idea of uncertainty. Similar situations do not always result in the same outcome. Statistics deals with this variability. This somewhat vague formulation will become clearer in this book. Many definitions of statistics explicitly bring in the idea of variability. Some definitions of statistics are given in the Notes at the end of the chapter.

1.3 WHY BIOSTATISTICS?

Biostatistics is the study of statistics as applied to biological areas. Biological laboratory experiments, medical research (including clinical research), and health services research all use statistical methods. Many other biological disciplines rely on statistical methodology.

Why should one study biostatistics rather than statistics, since the methods have wide applicability? There are three reasons for focusing on biostatistics:

1. Some statistical methods are used more heavily in biostatistics than in other fields. For example, a general statistical textbook would not discuss the life-table method of analyzing survival data—of importance in many biostatistical applications. The topics in this book are tailored to the applications in mind.
2. Examples are drawn from the biological, medical, and health care areas; this helps you maintain motivation. It also helps you understand how to apply statistical methods.
3. A third reason for a biostatistical text is to teach the material to an audience of health professionals. In this case, the interaction between students and teacher, but especially among the students themselves, is of great value in learning and applying the subject matter.

1.4 GOALS OF THIS BOOK

Suppose that we wanted to learn something about drugs; we can think of four different levels of knowledge. At the first level, a person may merely know that drugs act chemically when introduced into the body and produce many different effects. A second, higher level of knowledge is to know that a specific drug is given in certain situations, but we have no idea why the particular drug works. We do not know whether a drug might be useful in a situation that we have not yet seen. At the next, third level, we have a good idea why things work and also know how to administer drugs. At this level we do not have complete knowledge of all the biochemical principles involved, but we do have considerable knowledge about the activity and workings of the drug.

Finally, at the fourth and highest level, we have detailed knowledge of all of the interactions of the drug; we know the current research. This level is appropriate for researchers: those seeking to develop new drugs and to understand further the mechanisms of existing drugs. Think of the field of biostatistics in analogy to the drug field discussed above. It is our goal that those who complete the material in this book should be on the third level. This book is written to enable you to do more than apply statistical techniques mindlessly.

The greatest danger is in statistical analysis untouched by the human mind. We have the following objectives:

1. You should understand specified statistical concepts and procedures.
2. You should be able to identify procedures appropriate (and inappropriate) to a given situation. You should also have the knowledge to recognize when you do not know of an appropriate technique.
3. You should be able to carry out appropriate specified statistical procedures.

These are high goals for you, the reader of the book. But experience has shown that professionals in a wide variety of biological and medical areas can and do attain this level of expertise. The material presented in the book is often difficult and challenging; time and effort will, however, result in the acquisition of a valuable and indispensable tool that is useful in our daily lives as well as in scientific work.

1.5 STATISTICAL PROBLEMS IN BIOMEDICAL RESEARCH

We conclude this chapter with several examples of situations in which biostatistical design and analysis have been or could have been of use. The examples are placed here to introduce you to the subject, to provide motivation for you if you have not thought about such matters before, and to encourage thought about the need for methods of approaching variability and uncertainty in data.

The examples below deal with clinical medicine, an area that has general interest. Other examples can be found in Tanur et al. [1989].

1.5.1 Example 1: Treatment of King Charles II

This first example deals with the treatment of King Charles II during his terminal illness. The following quote is taken from Haggard [1929]:

> Some idea of the nature and number of the drug substances used in the medicine of the past may be obtained from the records of the treatment given King Charles II at the time of his death. These records are extant in the writings of a Dr. Scarburgh, one of the twelve or fourteen physicians called in to treat the king. At eight o'clock on Monday morning of February 2, 1685, King Charles was being shaved in his bedroom. With a sudden cry he fell backward and had a violent convulsion. He became unconscious, rallied once or twice, and after a few days died. Seventeenth-century autopsy records are far from complete, but one could hazard a guess that the king suffered with an embolism—that is, a floating blood clot which has plugged up an artery and deprived some portion of his brain of blood—or else his kidneys were diseased. As the first step in treatment the king was bled to the extent of a pint from a vein in his right arm. Next his shoulder was cut into and the incised area "cupped" to suck out an additional eight ounces of blood. After this homicidal onslaught the drugging began. An emetic and purgative were administered, and soon after a second purgative. This was followed by an enema containing antimony, sacred bitters, rock salt, mallow leaves, violets, beet root, camomile flowers, fennel seeds, linseed, cinnamon, cardamom seed, saphron, cochineal, and aloes. The enema was repeated in two hours and a purgative given. The king's head was shaved and a blister raised on his scalp. A sneezing powder of hellebore root was administered, and also a powder of cowslip flowers "to strengthen his brain." The cathartics were repeated at frequent intervals and interspersed with a soothing drink composed of barley water, licorice and sweet almond. Likewise

> white wine, absinthe and anise were given, as also were extracts of thistle leaves, mint, rue, and angelica. For external treatment a plaster of Burgundy pitch and pigeon dung was applied to the king's feet. The bleeding and purging continued, and to the medicaments were added melon seeds, manna, slippery elm, black cherry water, an extract of flowers of lime, lily-of-the-valley, peony, lavender, and dissolved pearls. Later came gentian root, nutmeg, quinine, and cloves. The king's condition did not improve, indeed it grew worse, and in the emergency forty drops of extract of human skull were administered to allay convulsions. A rallying dose of Raleigh's antidote was forced down the king's throat; this antidote contained an enormous number of herbs and animal extracts. Finally bezoar stone was given. Then says Scarburgh: "Alas! after an ill-fated night his serene majesty's strength seemed exhausted to such a degree that the whole assembly of physicians lost all hope and became despondent: still so as not to appear to fail in doing their duty in any detail, they brought into play the most active cordial." As a sort of grand summary to this pharmaceutical debauch a mixture of Raleigh's antidote, pearl julep, and ammonia was forced down the throat of the dying king.

From this time and distance there are comical aspects about this observational study describing the "treatment" given to King Charles. It should be remembered that his physicians were doing their best according to the state of their knowledge. Our knowledge has advanced considerably, but it would be intellectual pride to assume that all modes of medical treatment in use today are necessarily beneficial. This example illustrates that there is a need for sound scientific development and verification in the biomedical sciences.

1.5.2 Example 2: Relationship between the Use of Oral Contraceptives and Thromboembolic Disease

In 1967 in Great Britain, there was concern about higher rates of thromboembolic disease (disease from blood clots) among women using oral contraceptives than among women not using oral contraceptives. To investigate the possibility of a relationship, Vessey and Doll [1969] studied existing cases with thromboembolic disease. Such a study is called a *retrospective study* because retrospectively, or after the fact, the cases were identified and data accumulated for analysis. The study began by identifying women aged 16 to 40 years who had been discharged from one of 19 hospitals with a diagnosis of deep vein thrombosis, pulmonary embolism, cerebral thrombosis, or coronary thrombosis.

The idea of the study was to interview the cases to see if more of them were using oral contraceptives than one would "expect." The investigators needed to know how much oral contraceptive us to expect assuming that such us does not predispose people to thromboembolic disease. This is done by identifying a group of women "comparable" to the cases. The amount of oral contraceptive use in this *control*, or *comparison*, *group* is used as a standard of comparison for the cases. In this study, two control women were selected for each case: The control women had suffered an acute surgical or medical condition, or had been admitted for elective surgery. The controls had the same age, date of hospital admission, and parity (number of live births) as the cases. The controls were selected to have the absence of any predisposing cause of thromboembolic disease.

If there is no relationship between oral contraception and thromboembolic disease, the cases with thromboembolic disease would be no more likely than the controls to use oral contraceptives. In this study, 42 of 84 cases, or 50%, used oral contraceptives. Twenty-three of the 168 controls, or 14%, of the controls used oral contraceptives. After deciding that such a difference is unlikely to occur by chance, the authors concluded that there is a relationship between oral contraceptive use and thromboembolic disease.

This study is an example of a case–control study. The aim of such a study is to examine potential risk factors (i.e., factors that may dispose a person to have the disease) for a disease. The study begins with the identification of cases with the disease specified. A control group is then selected. The control group is a group of subjects comparable to the cases except for the presence of the disease and the possible presence of the risk factor(s). The case and control groups are then examined to see if a risk factor occurs more often than would be expected by chance in the cases than in the controls.

1.5.3 Example 3: Use of Laboratory Tests and the Relation to Quality of Care

An important feature of medical care are laboratory tests. These tests affect both the quality and the cost of care. The frequency with which such tests are ordered varies with the physician. It is not clear how the frequency of such tests influences the quality of medical care. Laboratory tests are sometimes ordered as part of "defensive" medical practice. Some of the variation is due to training. Studies investigating the relationship between use of tests and quality of care need to be designed carefully to measure the quantities of interest reliably, without bias. Given the expense of laboratory tests and limited time and resources, there clearly is a need for evaluation of the relationship between the use of laboratory tests and the quality of care.

The study discussed here consisted of 21 physicians serving medical internships as reported by Schroeder et al. [1974]. The interns were ranked independently on overall clinical capability (i.e., quality of care) by five faculty internists who had interacted with them during their medical training. Only patients admitted with uncomplicated acute myocardial infarction or uncomplicated chest pain were considered for the study. "Medical records of all patients hospitalized on the coronary care unit between July 1, 1971 and June 20, 1972, were analyzed and all patients meeting the eligibility criteria were included in the study. . . . " The frequency of laboratory utilization ordered during the first three days of hospitalization was translated into cost. Since daily EKGs and enzyme determinations (SGOT, LDH, and CPK) were ordered on all patients, the costs of these tests were excluded. Mean costs of laboratory use were calculated for each intern's subset of patients, and the interns were ranked in order of increasing costs on a per-patient basis.

Ranking by the five faculty internists and by cost are given in Table 1.1. There is considerable variation in the evaluations of the five internists; for example, intern K is ranked seventeenth in clinical competence by internists I and III, but first by internist II. This table still does not clearly answer the question of whether there is a relationship between clinical competence and the frequency of use of laboratory tests and their cost. Figure 1.1 shows the relationship between cost and one measure of clinical competence; on the basis of this graph and some statistical calculations, the authors conclude that "at least in the setting measured, no overall correlation existed between cost of medical care and competence of medical care."

This study contains good examples of the types of (basically statistical) problems facing a researcher in the health administration area. First, what is the population of interest? In other words, what population do the 21 interns represent? Second, there are difficult measurement problems: Is level of clinical competence, as evaluated by an internist, equivalent to the level of quality of care? How reliable are the internists? The variation in their assessments has already been noted. Is cost of laboratory use synonymous with cost of medical care as the authors seem to imply in their conclusion?

1.5.4 Example 4: Internal Mammary Artery Ligation

One of the greatest health problems in the world, especially in industrialized nations, is coronary artery disease. The coronary arteries are the arteries around the outside of the heart. These arteries bring blood to the heart muscle (myocardium). Coronary artery disease brings a narrowing of the coronary arteries. Such narrowing often results in chest, neck, and arm pain (angina pectoris) precipitated by exertion. When arteries block off completely or *occlude*, a portion of the heart muscle is deprived of its blood supply, with life-giving oxygen and nutrients. A myocardial infarction, or heart attack, is the death of a portion of the heart muscle.

As the coronary arteries narrow, the body often compensates by building *collateral circulation*, circulation that involves branches from existing coronary arteries that develop to bring blood to an area of restricted blood flow. The internal mammary arteries are arteries that bring

Table 1.1 Independent Assessment of Clinical Competence of 21 Medical Interns by Five Faculty Internists and Ranking of Cost of Laboratory Procedures Ordered, George Washington University Hospital, 1971–1972

	Clinical Competence[a]							Rank of Costs of
Intern	I	II	III	IV	V	Total	Rank	Procedures Ordered[b]
A	1	2	1	2	1	7	1	10
B	2	6	2	1	2	13	2	5
C	5	4	11	5	3	28	3	7
D	4	5	3	12	7	31	4	8
E	3	9	8	9	8	37	5	16
F	13	11	7	3	5	39	7	9
G	7	12	5	4	11	39	7	13
H	11	3	9	10	6	39	7	18
I	9	15	6	8	4	42	9	12
J	16	8	4	7	14	49	10	1
K	17	1	17	11	9	55	11	20
L	6	7	21	16	10	60	12	19
M	8	20	14	6	17	65	13	21
N	18	10	13	13	13	67	14	14
O	12	14	12	18	15	71	15	17
P	19	13	10	17	16	75	16	11
Q	20	16	16	15	12	77	17	4
R	14	18	19	14	19	84	18	15
S	10	19	18	20	20	87	19	3
T	15	17	20	21	21	94	20.5	2
U	21	21	15	19	18	94	20.5	5

Source: Data from Schroeder et al. [1974]; by permission of Medical Care.

[a] 1 = most competent.

[b] 1 = least expensive.

blood to the chest. The tributaries of the internal mammary arteries develop collateral circulation to the coronary arteries. It was thus reasoned that by tying off, or *ligating*, the internal mammary arteries, a larger blood supply would be forced to the heart. An operation, internal mammary artery ligation, was developed to implement this procedure.

Early results of the operation were most promising. Battezzati et al. [1959] reported on 304 patients who underwent internal mammary artery ligation: 94.8% of the patients reported improvement; 4.9% reported no appreciable change. It would seem that the surgery gave great improvement [Ratcliff, 1957; Time, 1959]. Still, the possibility remained that the improvement resulted from a placebo effect. A *placebo effect* is a change, or perceived change, resulting from the psychological benefits of having undergone treatment. It is well known that inert tablets will cure a substantial portion of headaches and stomach aches and afford pain relief. The placebo effect of surgery might be even more substantial.

Two studies of internal mammary artery ligation were performed using a sham operation as a control. Both studies were *double blind*: Neither the patients nor physicians evaluating the effect of surgery knew whether the ligation had taken place. In each study, incisions were made in the patient's chest and the internal mammary arteries exposed. In the sham operation, nothing further was done. For the other patients, the arteries were ligated. Both studies selected patients having the ligation or sham operation by random assignment [Hitchcock et al., 1966; Ruffin et al., 1969].

Cobb et al. [1959] reported on the subjective patient estimates of "significant" improvement. Patients were asked to estimate the percent improvement after the surgery. Another indication

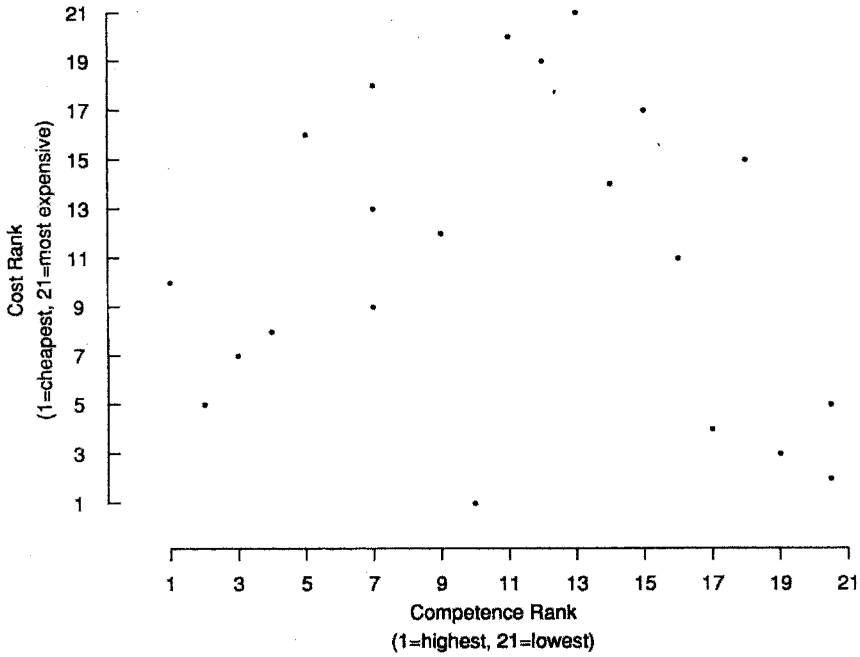


Figure 1.1 Rank order of clinical competence vs. rank order of cost of laboratory tests orders for 21 interns, George Washington University Hospital, 1971–1972. (Data from Schroeder et al. [1974].)

of the amount of pain experienced is the number of nitroglycerin tablets taken for anginal pain. Table 1.2 reports these data.

Dimond et al. [1960] reported a study of 18 patients, of whom five received the sham operation and 13 received surgery. Table 1.3 presents the patients' opinion of the percentage benefit of surgery.

Both papers conclude that it is unlikely that the internal mammary artery ligation has benefit, beyond the placebo effect, in the treatment of coronary artery disease. Note that 12 of the 14, or 86%, of those receiving the sham operation reported improvement in the two studies. These studies point to the need for appropriate comparison groups when making scientific inferences.

Table 1.2 Subjective Improvement as Measured by Patient Reporting and Number of Nitroglycerin Tablets

	Ligated	Nonligated
Number of patients	8	9
Average percent improvement reported	32	43
Subjects reporting 40% or more improvement	5	5
Subjects reporting no improvement	3	2
Nitroglycerin tablets taken		
Average before operation (no./week)	43	30
Average after operation (no./week)	25	17
Average percent decrease (no./week)	34	43

Source: Cobb et al. [1959].

Table 1.3 Patients' Opinions of Surgical Benefit

Patients' Opinions of the Benefit of Surgery	Patient Number[a]
Cured (90–100%)	4, 10, 11, 12*, 14*
Definite benefit (50–90%)	2, 3*, 6, 8, 9*, 13*, 15, 17, 18
Improved but disappointed (25–50%)	7
Improved for two weeks, now same or worse	1, 5, 16

Source: Dimond et al. [1960].

[a] The numbers 1–18 refer to the individual patients as they occurred in the series, grouped according to their own evaluation of their benefit, expressed as a percentage. Those numbers followed by an asterisk indicate a patient on whom a sham operation was performed.

The use of clinical trials has greatly enhanced medical progress. Examples are given throughout the book, but this is not the primary emphasis of the text. Good references for learning much about clinical trials are Meinert [1986], Friedman et al. [1981], Tanur et al. [1989], and Fleiss [1986].

NOTES

1.1 Some Definitions of Statistics

- "The science of statistics is essentially a branch of Applied Mathematics, and may be regarded as mathematics applied to observational data. ... Statistics may be regarded (i) as the study of populations, (ii) as the study of variation, (iii) as the study of methods of the reduction of data." Fisher [1950]
- "Statistics is the branch of the scientific method which deals with the data obtained by counting or measuring the properties of populations of natural phenomena." Kendall and Stuart [1963]
- "The science and art of dealing with variation in such a way as to obtain reliable results." Mainland [1963]
- "Statistics is concerned with the inferential process, in particular with the planning and analysis of experiments or surveys, with the nature of observational errors and sources of variability that obscure underlying patterns, and with the efficient summarizing of sets of data." Kruskal [1968]
- "Statistics = Uncertainty and Behavior." Savage [1968]
- "... the principal object of statistics [is] to make inference on the probability of events from their observed frequencies." von Mises [1957]
- "The technology of the scientific method." Mood [1950]
- "The statement, still frequently made, that statistics is a branch of mathematics is no more true than would be a similar claim in respect of engineering ... [G]ood statistical practice is equally demanding of appreciation of factors outside the formal mathematical structure, essential though that structure is." Finney [1975]

There is clearly no complete consensus in the definitions of statistics. But certain elements reappear in all the definitions: variation, uncertainty, inference, science. In previous sections we have illustrated how the concepts occur in some typical biomedical studies. The need for biostatistics has thus been shown.

REFERENCES

Battezzati, M., Tagliaferro, A., and Cattaneo, A. D. [1959]. Clinical evaluation of bilateral internal mammary artery ligation as treatment of coronary heart disease. *American Journal of Cardiology*, **4**: 180–183.

Cobb, L. A., Thomas, G. I., Dillard, D. H., Merendino, K. A., and Bruce, R. A. [1959]. An evaluation of internal-mammary-artery ligation by a double blind technique. *New England Journal of Medicine*, **260**: 1115–1118.

Dimond, E. G., Kittle, C. F., and Crockett, J. E. [1960]. Comparison of internal mammary artery ligation and sham operation for angina pectoris. *American Journal of Cardiology*, **5**: 483–486.

Finney, D. J. [1975]. Numbers and data. *Biometrics*, **31**: 375–386.

Fisher, R. A. [1950]. *Statistical Methods for Research Workers*, 11th ed. Hafner, New York.

Fleiss, J. L. [1986]. *The Design and Analysis of Clinical Experiments*. Wiley, New York.

Friedman, L. M., Furberg, C. D., and DeMets, D. L. [1981]. *Fundamentals of Clinical Trials*. John Wright, Boston.

Haggard, H. W. [1929]. *Devils, Drugs, and Doctors*. Blue Ribbon Books, New York.

Hitchcock, C. R., Ruiz, E., Sutherland, R. D., and Bitter, J. E. [1966]. Eighteen-month follow-up of gastric freezing in 173 patients with duodenal ulcer. *Journal of the American Medical Association*, **195**: 115–119.

Kendall, M. G., and Stuart, A. [1963]. *The Advanced Theory of Statistics*, Vol. 1, 2nd ed. Charles Griffin, London.

Kruskal, W. [1968]. In *International Encyclopedia of the Social Sciences*, D. L. Sills (ed). Macmillan, New York.

Mainland, D. [1963]. *Elementary Medical Statistics*, 2nd ed. Saunders, Philadelphia.

Meinert, C. L. [1986]. *Clinical Trials: Design, Conduct and Analysis*. Oxford University Press, New York.

Mood, A. M. [1950]. *Introduction to the Theory of Statistics*. McGraw-Hill, New York.

Ratcliff, J. D. [1957]. New surgery for ailing hearts. *Reader's Digest*, **71**: 70–73.

Ruffin, J. M., Grizzle, J. E., Hightower, N. C., McHarcy, G., Shull, H., and Kirsner, J. B. [1969]. A cooperative double-blind evaluation of gastric "freezing" in the treatment of duodenal ulcer. *New England Journal of Medicine*, **281**: 16–19.

Savage, I. R. [1968]. *Statistics: Uncertainty and Behavior*. Houghton Mifflin, Boston.

Schroeder, S. A., Schliftman, A., and Piemme, T. E. [1974]. Variation among physicians in use of laboratory tests: relation to quality of care. *Medical Care*, **12**: 709–713.

Tanur, J. M., Mosteller, F., Kruskal, W. H., Link, R. F., Pieters, R. S., and Rising, G. R. (eds.) [1989]. *Statistics: A Guide to the Unknown*, 3rd ed. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA.

Time [1962]. Frozen ulcers. *Time*, May 18: 45–47.

Vessey, M. P., and Doll, R. [1969]. Investigation of the relation between use of oral contraceptives and thromboembolic disease: a further report. *British Medical Journal*, **2**: 651–657.

von Mises, R. [1957]. *Probability, Statistics and Truth*, 2nd ed. Macmillan, New York.

CHAPTER 2

Biostatistical Design of Medical Studies

2.1 INTRODUCTION

In this chapter we introduce some of the principles of biostatistical design. Many of the ideas are expanded in later chapters. This chapter also serves as a reminder that statistics is not an end in itself but a tool to be used in investigating the world around us. The study of statistics should serve to develop critical, analytical thought and common sense as well as to introduce specific tools and methods of processing data.

2.2 PROBLEMS TO BE INVESTIGATED

Biomedical studies arise in many ways. A particular study may result from a sequence of experiments, each one leading naturally to the next. The study may be triggered by observation of an interesting case, or observation of a mold (e.g., penicillin in a petri dish). The study may be instigated by a governmental agency in response to a question of national importance. The basic ideas of the study may be defined by an advisory panel. Many of the critical studies and experiments in biomedical science have come from one person with an idea for a radical interpretation of past data.

Formulation of the problem to be studied lies outside the realm of statistics per se. Statistical considerations may suggest that an experiment is too expensive to conduct, or may suggest an approach that differs from that planned. The need to evaluate data from a study statistically forces an investigator to sharpen the focus of the study. It makes one translate intuitive ideas into an analytical model capable of generating data that may be evaluated statistically.

To answer a given scientific question, many different studies may be considered. Possible studies may range from small laboratory experiments, to large and expensive experiments involving humans, to observational studies. It is worth spending a considerable amount of time thinking about alternatives. In most cases your first idea for a study will not be your best—unless it is your only idea.

In laboratory research, many different experiments may shed light on a given hypothesis or question. Sometimes, less-than-optimal execution of a well-conceived experiment sheds more light than arduous and excellent experimentation unimaginatively designed. One mark of a good scientist is that he or she attacks important problems in a clever manner.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

2.3 VARIOUS TYPES OF STUDIES

A problem may be investigated in a variety of ways. To decide on your method of approach, it is necessary to understand the types of studies that might be done. To facilitate the discussion of design, we introduce definitions of commonly used types of studies.

Definition 2.1. An *observational study* collects data from an existing situation. The data collection does not intentionally interfere with the running of the system.

There are subtleties associated with observational studies. The act of observation may introduce change into a system. For example, if physicians know that their behavior is being monitored and charted for study purposes, they may tend to adhere more strictly to procedures than would be the case otherwise. Pathologists performing autopsies guided by a study form may invariably look for a certain finding not routinely sought. The act of sending out questionnaires about health care may sensitize people to the need for health care; this might result in more demand. Asking constantly about a person's health can introduce hypochondria.

A side effect introduced by the act of observation is the *Hawthorne effect*, named after a famous experiment carried out at the Hawthorne works of the Western Electric Company. Employees were engaged in the production of electrical relays. The study was designed to investigate the effect of better working conditions, including increased pay, shorter hours, better lighting and ventilation, and pauses for rest and refreshment. All were introduced, with "resulting" increased output. As a control, working conditions were returned to original conditions. Production continued to rise! The investigators concluded that increased morale due to the attention and resulting *esprit de corps* among workers resulted in better production. Humans and animals are not machines or passive experimental units [Roethlisberger, 1941].

Definition 2.2. An *experiment* is a study in which an investigator deliberately sets one or more factors to a specific level.

Experiments lead to stronger scientific inferences than do observational studies. The "cleanest" experiments exist in the physical sciences; nevertheless, in the biological sciences, particularly with the use of randomization (a topic discussed below), strong scientific inferences can be obtained. Experiments are superior to observational studies in part because in an observational study one may not be observing one or more variables that are of crucial importance to interpreting the observations. Observational studies are always open to misinterpretation due to a lack of knowledge in a given field. In an experiment, by seeing the change that results when a factor is varied, the causal inference is much stronger.

Definition 2.3. A *laboratory experiment* is an experiment that takes place in an environment (called a *laboratory*) where experimental manipulation is facilitated.

Although this definition is loose, the connotation of the term *laboratory experiment* is that the experiment is run under conditions where most of the variables of interest can be controlled very closely (e.g., temperature, air quality). In laboratory experiments involving animals, the aim is that animals be treated in the same manner in all respects except with regard to the factors varied by the investigator.

Definition 2.4. A *comparative experiment* is an experiment that compares two or more techniques, treatments, or levels of a variable.

There are many examples of comparative experiments in biomedical areas. For example, it is common in nutrition to compare laboratory animals on different diets. There are many experiments comparing different drugs. Experiments may compare the effect of a given treatment with that of no treatment. (From a strictly logical point of view, "no treatment" is in itself a type of treatment.) There are also comparative observational studies. In a comparative *study* one might, for example, observe women using and women not using birth control pills and examine the incidence of complications such as thrombophlebitis. The women themselves would decide whether or not to use birth control pills. The user and nonuser groups would probably differ in a great many other ways. In a comparative *experiment*, one might have women selected by chance to receive birth control pills, with the control group using some other method.

Definition 2.5. An *experimental unit* or *study unit* is the smallest unit on which an experiment or study is performed.

In a clinical study, the experimental units are usually humans. (In other cases, it may be an eye; for example, one eye may receive treatment, the other being a control.) In animal experiments, the experimental unit is usually an animal. With a study on teaching, the experimental unit may be a class—as the teaching method will usually be given to an entire class. Study units are the object of consideration when one discusses sample size.

Definition 2.6. An experiment is a *crossover experiment* if the same experimental unit receives more than one treatment or is investigated under more than one condition of the experiment. The different treatments are given during nonoverlapping time periods.

An example of a crossover experiment is one in which laboratory animals are treated sequentially with more than one drug and blood levels of certain metabolites are measured for each drug. A major benefit of a crossover experiment is that each experimental unit serves as its own control (the term *control* is explained in more detail below), eliminating subject-to-subject variability in response to the treatment or experimental conditions being considered. Major disadvantages of a crossover experiment are that (1) there may be a carryover effect of the first treatment continuing into the next treatment period; (2) the experimental unit may change over time; (3) in animal or human experiments, the treatment introduces permanent physiological changes; (4) the experiment may take longer so that investigator and subject enthusiasm wanes; and (5) the chance of dropping out increases.

Definition 2.7. A *clinical study* is one that takes place in the setting of clinical medicine.

A study that takes place in an organizational unit dispensing health care—such as a hospital, psychiatric clinic, well-child clinic, or group practice clinic—is a clinical study.

We now turn to the concepts of prospective studies and retrospective studies, usually involving human populations.

Definition 2.8. A *cohort* of people is a group of people whose membership is clearly defined.

Examples of cohorts are all persons enrolling in the Graduate School at the University of Washington for the fall quarter of 2003; all females between the ages of 30 and 35 (as of a certain date) whose residence is within the New York City limits; all smokers in the United States as of January 1, 1953, where a person is defined to be a smoker if he or she smoked one or more cigarettes during the preceding calendar year. Often, cohorts are followed over some time interval.

Definition 2.9. An *endpoint* is a clearly defined outcome or event associated with an experimental or study unit.

An endpoint may be the presence of a particular disease or five-year survival after, say, a radical mastectomy. An important characteristic of an endpoint is that it can be clearly defined and observed.

Definition 2.10. A *prospective study* is one in which a cohort of people is followed for the occurrence or nonoccurrence of specified endpoints or events or measurements.

In the analysis of data from a prospective study, the occurrence of the endpoints is often related to characteristics of the cohort measured at the beginning of the study.

Definition 2.11. *Baseline characteristics* or *baseline variables* are values collected at the time of entry into the study.

The Salk polio vaccine trial is an example of a prospective study, in fact, a prospective experiment. On occasion, you may be able to conduct a prospective study from existing data; that is, some unit of government or other agency may have collected data for other purposes, which allows you to analyze the data as a prospective study. In other words, there is a well-defined cohort for which records have already been collected (for some other purpose) which can be used for your study. Such studies are sometimes called *historical prospective studies*.

One drawback associated with prospective studies is that the endpoint of interest may occur infrequently. In this case, extremely large numbers of people need to be followed in order that the study will have enough endpoints for statistical analysis. As discussed below, other designs, help get around this problem.

Definition 2.12. A *retrospective study* is one in which people having a particular outcome or endpoint are identified and studied.

These subjects are usually compared to others without the endpoint. The groups are compared to see whether the people with the given endpoint have a higher fraction with one or more of the factors that are conjectured to increase the risk of endpoints.

Subjects with particular characteristics of interest are often collected into registries. Such a registry usually covers a well-defined population. In Sweden, for example, there is a twin registry. In the United States there are cancer registries, often defined for a specified metropolitan area. Registries can be used for retrospective as well as prospective studies. A cancer registry can be used retrospectively to compare the presence or absence of possible causal factors of cancer after generating appropriate controls—either randomly from the same population or by some matching mechanism. Alternatively, a cancer registry can be used prospectively by comparing survival times of cancer patients having various therapies.

One way of avoiding the large sample sizes needed to collect enough cases prospectively is to use the case–control study, discussed in Chapter 1.

Definition 2.13. A *case–control study* selects all cases, usually of a disease, that meet fixed criteria. A group, called *controls*, that serve as a comparison for the cases is also selected. The cases and controls are compared with respect to various characteristics.

Controls are sometimes selected to match the individual case; in other situations, an entire group of controls is selected for comparison with an entire group of cases.

Definition 2.14. In a *matched case–control study*, controls are selected to match characteristics of individual cases. The cases and control(s) are associated with each other. There may be more than one control for each case.

Definition 2.15. In a *frequency-matched case–control study*, controls are selected to match characteristics of the entire case sample (e.g., age, gender, year of event). The cases and controls are not otherwise associated. There may be more than one control for each case.

Suppose that we want to study characteristics of cases of a disease. One way to do this would be to identify new cases appearing during some time interval. A second possibility would be to identify all known cases at some fixed time. The first approach is *longitudinal*; the second approach is *cross-sectional.*

Definition 2.16. A *longitudinal study* collects information on study units over a specified time period. A *cross-sectional study* collects data on study units at a fixed time.

Figure 2.1 illustrates the difference. The longitudinal study might collect information on the six new cases appearing over the interval specified. The cross-sectional study would identify the nine cases available at the fixed time point. The cross-sectional study will have proportionately more cases with a long duration. (Why?) For completeness, we repeat the definitions given informally in Chapter 1.

Definition 2.17. A *placebo treatment* is designed to appear exactly like a comparison treatment but to be devoid of the active part of the treatment.

Definition 2.18. The *placebo effect* results from the belief that one has been treated rather than having experienced actual changes due to physical, physiological, and chemical activities of a treatment.

Definition 2.19. A study is *single blind* if subjects being treated are unaware of which treatment (including any control) they are receiving. A study is *double blind* if it is single blind

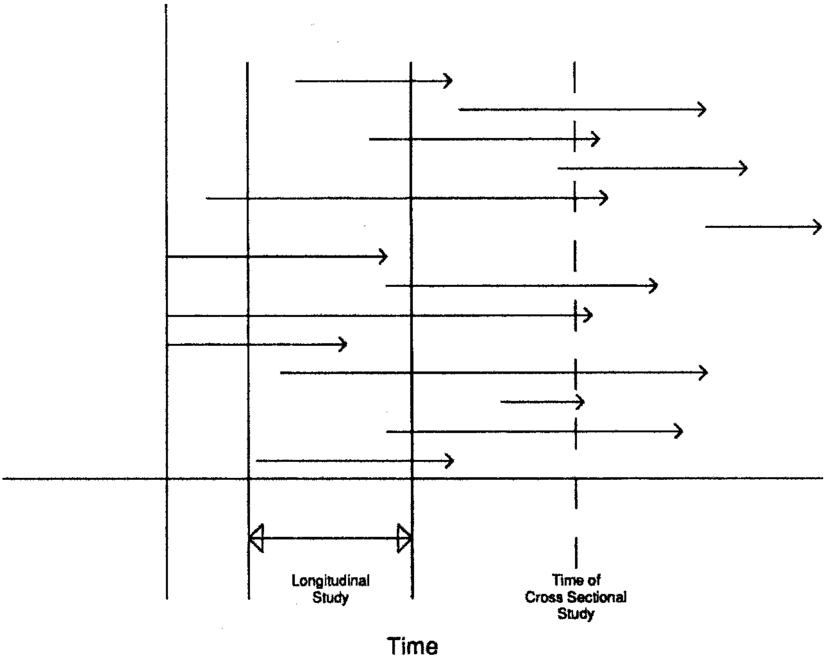


Figure 2.1 Longitudinal and cross-sectional study of cases of a disease.

and the people who are evaluating the outcome variables are also unaware of which treatment the subjects are receiving.

2.4 STEPS NECESSARY TO PERFORM A STUDY

In this section we outline briefly the steps involved in conducting a study. The steps are interrelated and are oversimplified here in order to isolate various elements of scientific research and to discuss the statistical issues involved:

1. A question or problem area of interest is considered. This does not involve biostatistics per se.
2. A study is to be designed to answer the question. The design of the study must consider at least the following elements:
   - **a.** Identify the data to be collected. This includes the variables to be measured as well as the number of experimental units, that is, the size of the study or experiment.
   - **b.** An appropriate analytical model needs to be developed for describing and processing data.
   - **c.** What inferences does one hope to make from the study? What conclusions might one draw from the study? To what population(s) is the conclusion applicable?
3. The study is carried out and the data are collected.
4. The data are analyzed and conclusions and inferences are drawn.
5. The results are used. This may involve changing operating procedures, publishing results, or planning a subsequent study.

2.5 ETHICS

Many studies and experiments in the biomedical field involve animal and/or human participants. Moral and legal issues are involved in both areas. Ethics must be of primary concern. In particular, we mention five points relevant to experimentation with humans:

1. It is our opinion that all investigators involved in a study are responsible for the conduct of an ethical study to the extent that they may be expected to know what is involved in the study. For example, we think that it is unethical to be involved in the analysis of data that have been collected in an unethical manner.
2. Investigators are close to a study and often excited about its potential benefits and advances. It is difficult for them to consider all ethical issues objectively. For this reason, in proposed studies involving humans (or animals), there should be review by people not concerned or connected with the study or the investigators. The reviewers should not profit directly in any way if the study is carried out. Implementation of the study should be contingent on such a review.
3. People participating in an experiment should understand and sign an informed consent form. The *principle of informed consent* says that a participant should know about the conduct of a study and about any possible harm and/or benefits that may result from participation in the study. For those unable to give informed consent, appropriate representatives may give the consent.
4. Subjects should be free to withdraw at any time, or to refuse initial participation, without being penalized or jeopardized with respect to current and future care and activities.
5. Both the Nuremberg Code and the Helsinki Accord recommend that, when possible, animal studies be done prior to human experimentation.

References relevant to ethical issues include the U.S. Department of Health, Education, and Welfare's (HEW's) statement on *Protection of Human Subjects* [1975], Papworth's book, *Human Guinea Pigs* [1967], and Spicker et al. [1988]; Papworth is extremely critical of the conduct of modern biological experimentation. There are also guidelines for studies involving animals. See, for example, *Guide for the Care and Use of Laboratory Animals* [HEW, 1985] and *Animal Welfare* [USDA, 1989]. Ethical issues in randomized trials are discussed further in Chapter 19.

2.6 DATA COLLECTION: DESIGN OF FORMS

2.6.1 What Data Are to Be Collected?

In studies involving only one or two investigators, there is often almost complete agreement as to what data are to be collected. In this case it is very important that good laboratory records be maintained. It is especially important that variations in the experimental procedure (e.g., loss of power during a time period, unexpected change in temperature in a room containing laboratory animals) be recorded. If there are peculiar patterns in the data, detailed notes may point to possible causes. The necessity for keeping detailed notes is even more crucial in large studies or experiments involving many investigators; it is difficult for one person to have complete knowledge of a study.

In a large collaborative study involving a human population, it is not always easy to decide what data to collect. For example, often there is interest in getting prognostic information. How many potentially prognostic variables should you record?

Suppose that you are measuring pain relief or quality of life; how many questions do you need to characterize these abstract ideas reasonably? In looking for complications of drugs, should you instruct investigators to enter all complications? This may be an unreliable procedure if you are dependent on a large, diverse group of observers. In studies with many investigators, each investigator will want to collect data relating to her or his special interests. You can arrive rapidly at large, complex forms. If too many data are collected, there are various "prices" to be paid. One obvious price is the expense of collecting and handling large and complex data sets. Another is reluctance (especially by volunteer subjects) to fill out long, complicated forms, leading to possible biases in subject recruitment. If a study lasts a long time, the investigators may become fatigued by the onerous task of data collection. Fatigue and lack of enthusiasm can affect the quality of data through a lack of care and effort in its collection.

On the other hand, there are many examples where too few data were collected. One of the most difficult tasks in designing forms is to remember to include all necessary items. The more complex the situation, the more difficult the task. It is easy to look at existing questions and to respond to them. If a question is missing, how is one alerted to the fact? One of the authors was involved in the design of a follow-up form where mortality could not be recorded. There was an explanation for this: The patients were to fill out the forms. Nevertheless, it was necessary to include forms that would allow those responsible for follow-up to record mortality, the primary endpoint of the study.

To assure that all necessary data are on the form, you are advised to follow four steps:

1. Perform a thorough review of all forms *with a written response* by all participating investigators.
2. Decide on the statistical analyses beforehand. Check that *specific* analyses involving *specific* variables can be run. Often, the analysis is changed during processing of the data or in the course of "interactive" data analysis. This preliminary step is still necessary to ensure that data are available to answer the primary questions.
3. Look at other studies and papers in the area being studied. It may be useful to mimic analyses in the most outstanding of these papers. If they contain variables not recorded

in the new study, find out why. The usual reason for excluding variables is that they are not needed to answer the problems addressed.

4. If the study includes a pilot phase, as suggested below, analyze the data of the pilot phase to see if you can answer the questions of interest when more data become available.

2.6.2 Clarity of Questions

The task of designing clear and unambiguous questions is much greater than is generally realized. The following points are of help in designing such questions:

1. Who is filling out the forms? Forms to be filled out by many people should, as much as possible, be self-explanatory. There should not be another source to which people are required to go for explanation—often, they would not take the trouble. This need not be done if trained technicians or interviewers are being used in certain phases of the study.
2. The degree of accuracy and the units required should be specified where possible. For example, data on heights should not be recorded in both inches and centimeters in the same place. It may be useful to allow both entries and to have a computer adjust to a common unit. In this case have two possible entries, one designated as centimeters and the other designated as inches.
3. A response should be required on all sections of a form. Then if a portion of the form has no response, this would indicate that the answer was missing. (If an answer is required only under certain circumstances, you cannot determine whether a question was missed or a correct "no answer" response was given; a blank would be a valid answer. For example, in pathology, traditionally the pathologist reports only "positive" findings. If a finding is absent in the data, was the particular finding not considered, and missed, or was a positive outcome not there?)
4. There are many alternatives when collecting data about humans: forms filled out by a subject, an in-person interview by a trained interviewer, a telephone interview, forms filled out by medical personnel after a general discussion with the subject, or forms filled out by direct observation. It is an eye-opening experience to collect the "same" data in several different ways. This leads to a healthy respect for the amount of variability in the data. It may also lead to clarification of the data being collected. In collecting subjective opinions, there is usually interaction between a subject and the method of data collection. This may greatly influence, albeit unconsciously, a subject's response.

The following points should also be noted. A high level of formal education of subjects and/or interviewer is not necessarily associated with greater accuracy or reproducibility of data collected. The personality of a subject and/or interviewer can be more important than the level of education. The effort and attention given to a particular part of a complex data set should be proportional to its importance. Prompt editing of data for mistakes produces higher-quality data than when there is considerable delay between collecting, editing, and correction of forms.

2.6.3 Pretesting of Forms and Pilot Studies

If it is extremely difficult, indeed almost impossible, to design a satisfactory form, how is one to proceed? It is necessary to have a pretest of the forms, except in the simplest of experiments and studies. In a *pretest*, forms are filled out by one or more people prior to beginning an actual study and data collection. In this case, several points should be considered. People filling out forms should be representative of the people who will be filling them out during the study. You can be misled by having health professionals fill out forms that are designed for the "average" patient. You should ask the people filling out the pretest forms if they have any questions or are not sure about parts of the forms. However, it is important not to interfere while the forms are being used but to let them be used in the same context as will pertain in the study; then ask the questions. Preliminary data should be analyzed; you should look for differences in responses from different clinics or individuals. Such analyses may indicate that a variable is being interpreted differently by different groups. The pretest forms should be edited by those responsible for the design. Comments written on the forms or answers that are not legitimate can be important in improving the forms. During this phase of the study, one should pursue vigorously the causes of missing data.

A more complete approach is to have a *pilot study*, which consists of going through the actual mechanics of a proposed study. Thus, a pilot study works out both the "bugs" from forms used in data collection and operational problems within the study. Where possible, data collected in a pilot study should be compared with examples of the "same" data collected in other studies. Suppose that there is recording of data that are not quantitative but categorical (e.g., the amount of impairment of an animal, whether an animal is losing its hair, whether a patient has improved morale). There is a danger that the investigator(s) may use a convention that would not readily be understood by others. To evaluate the extent to which the data collected are understood, it is good procedure to ask others to examine some of the same study units and to record their opinion without first discussing what is meant by the categories being recorded. If there is great variability, this should lead to a need for appropriate caution in the interpretation of the data. This problem may be most severe when only one person is involved in data collection.

2.6.4 Layout and Appearance

The physical appearance of forms is important if many people are to fill them out. People attach more importance to a printed page than to a mimeographed page, even though the layout is the same. If one is depending on voluntary reporting of data, it may be worthwhile to spend a bit more to have forms printed in several colors with an attractive logo and appearance.

2.7 DATA EDITING AND VERIFICATION

If a study involves many people filling out forms, it will be necessary to have a manual and/or computer review of the content of the forms before beginning analysis. In most studies there are inexplicably large numbers of mistakes and missing data. If missing and miscoded data can be attacked vigorously *from the beginning* of a study, the quality of data can be vastly improved. Among checks that go into data editing are the following:

1. *Validity checks*. Check that only allowable values or codes are given for answers to the questions. For example, a negative weight is not allowed. A simple extension of this idea is to require that most of the data fall within a given range; range checks are set so that a small fraction of the valid data will be outside the range and will be "flagged"; for example, the height of a professional basketball team center (who happens to be a subject in the study) may fall outside the allowed range even though the height is correct. By checking out-of-range values, many incorrectly recorded values can be detected.
2. *Consistency checks*. There should be internal consistency of the data. Following are some examples:
   - **a.** If more than one form is involved, the dates on these forms should be consistent with each other (e.g., a date of surgery should precede the date of discharge for that surgery).
   - **b.** Consistency checks can be built into the study by collecting crucial data in two different ways (e.g., ask for both date of birth and age).
   - **c.** If the data are collected sequentially, it is useful to examine unexpected changes between forms (e.g., changes in height, or drastic changes such as changes of weight by 70%). Occasionally, such changes are correct, but they should be investigated.

d. In some cases there are certain combinations of replies that are mutually inconsistent; checks for these should be incorporated into the editing and verification procedures.

3. *Missing forms.* In some case–control studies, a particular control may refuse to participate in a study. Some preliminary data on this control may already have been collected. Some mechanism should be set up so that it is clear that no further information will be obtained for that control. (It will be useful to keep the preliminary information so that possible selection bias can be detected.) If forms are entered sequentially, it will be useful to decide when missing forms will be labeled "overdue" or "missing."

2.8 DATA HANDLING

All except the smallest experiments involve data that are eventually processed or analyzed by computer. Forms should be designed with this fact in mind. It should be easy to enter the form by keyboard. Some forms are called *self-coding*: Columns are given next to each variable for data entry. Except in cases where the forms are to be entered by a variety of people at different sites, the added cluttering of the form by the self-coding system is not worth the potential ease in data entry. Experienced persons entering the same type of form over and over soon know which columns to use. Alternatively, it is possible to overlay plastic sheets that give the columns for data entry.

For very large studies, the logistics of collecting data, putting the data on a computer system, and linking records may hinder a study more than any other factor. Although it is not appropriate to discuss these issues in detail here, the reader should be aware of this problem. In any large study, people with expertise in data handling and computer management of data should be consulted during the design phase. Inappropriately constructed data files result in unnecessary expense and delay during the analytic phase. In projects extending over a long period of time and requiring periodic reports, it is important that the timing and management of data collection and management be specified. Experience has shown that even with the best plans there will be inevitable delays. It is useful to allow some slack time between required submission of forms and reports, between final submission and data analysis.

Computer files or tapes will occasionally be erased accidentally. In the event of such a disaster it is necessary to have backup computer tapes and documentation. If information on individual subject participants is required, there are confidentiality laws to be considered as well as the investigator's ethical responsibility to protect subject interests. During the design of any study, everyone will underestimate the amount of work involved in accomplishing the task. Experience shows that caution is necessary in estimating time schedules. During a long study, constant vigilance is required to maintain the quality of data collection and flow. In laboratory experimentation, technicians may tend to become bored and slack off unless monitored. Clinical study personnel will tire of collecting the data and may try to accomplish this too rapidly unless monitored.

Data collection and handling usually involves almost all participants of the study and should not be underestimated. It is a common experience for research studies to be planned without allowing sufficient time or money for data processing and analysis. It is difficult to give a rule of thumb, but in a wide variety of studies, 15% of the expense has been in data handling, processing, and analysis.

2.9 AMOUNT OF DATA COLLECTED: SAMPLE SIZE

It is part of scientific folklore that one of the tasks of a statistician is to determine an appropriate sample size for a study. Statistical considerations do have a large bearing on the selection of a sample size. However, there is other scientific input that must be considered in order to arrive at the number of experimental units needed. If the purpose of an experiment is to estimate some quantity, there is a need to know how precise an estimate is desired and how confident the investigator wishes to be that the estimate is within a specified degree of precision. If the purpose of an experiment is to compare several treatments, it is necessary to know what difference is considered important and how certain the investigator wishes to be of detecting such a difference. Statistical calculation of sample size requires that all these considerations be quantified. (This topic is discussed in subsequent chapters.) In a descriptive observational study, the size of the study is determined by specifying the needed accuracy of estimates of population characteristics.

2.10 INFERENCES FROM A STUDY

2.10.1 Bias

The statistical term *bias* refers to a situation in which the statistical method used does not estimate the quantity thought to be estimated or test the hypothesis thought to be tested. This definition will be made more precise later. In this section the term is used on a intuitive level. Consider some examples of biased statistical procedures:

1. A proposal is made to measure the average amount of health care in the United States by means of a personal health questionnaire that is to be passed out at an American Medical Association convention. In this case, the AMA respondents constitute a biased sample of the overall population.
2. A famous historical example involves a telephone poll made during the Dewey–Truman presidential contest. At that time—and to some extent today—a large section of the population could not afford a telephone. Consequently, the poll was conducted among more well-to-do citizens, who constituted a biased sample with respect to presidential preference.
3. In a laboratory experiment, animals receiving one treatment are kept on one side of the room and animals receiving a second treatment are kept on another side. If there is a large differential in lighting and heat between the two sides of the room, one could find "treatment effects" that were in fact ascribable to differences in light and/or heat. Work by Riley [1975] suggests that level of stress (e.g., bottom cage vs. top cage) affects the resistance of animals to carcinogens.

In the examples of Section 1.5, methods of minimizing bias were considered. Single- and double-blind experiments reduce bias.

2.10.2 Similarity in a Comparative Study

If physicists at Berkeley perform an experiment in electron physics, it is expected that the same experiment could be performed successfully (given the appropriate equipment) in Moscow or London. One expects the same results because the current physical model is that all electrons are precisely the same (i.e., they are identical) and the experiments are truly similar experiments. In a comparative experiment, we would like to try out experiments on similar units.

We now discuss similarity where it is assumed for the sake of discussion that the experimental units are humans. The ideas and results, however, can be extended to animals and other types of experimental units. The experimental situations being compared will be called *treatments*. To get a fair comparison, it is necessary that the treatments be given to similar units. For example, if cancer patients whose disease had not progressed much receive a new treatment and their survival is compared to the standard treatment administered to all types of patients, the comparison would not be justified; the treatments were not given to similar groups.

Of all human beings, *identical twins* are the most alike, by having identical genetic background. Often, they are raised together, so they share the same environment. Even in an observational twin study, a strong scientific inference can be made if enough appropriate pairs of identical twins can be found. For example, suppose that the two "treatments" are smoking and nonsmoking. If one had identical twins raised together where one of the pair smoked and the other did not, the incidence of lung cancer, the general health, and the survival experience could provide quite strong scientific inferences as to the health effect of smoking. (In Sweden there is a twin registry to aid public health and medical studies.) It is difficult to conduct twin studies because sufficient numbers of identical twins need to be located, such that one member of the pair has one treatment and the other twin, another treatment. It is expensive to identify and find them. Since they have the same environment, in a smoking study it is most likely, that either both would smoke or both would not smoke. Such studies are logistically not possible in most circumstances.

A second approach is that of matching or pairing individuals. The rationale behind *matched* or *matched pair studies* is to find two persons who are identical with regard to all "pertinent" variables under consideration except the treatment. This may be thought of as an attempt to find a surrogate identical twin. In many studies, people are matched with regard to age, gender, race, and some indicator of socioeconomic status. In a prospective study, the two matched individuals receive differing treatments. In a retrospective study, the person with the endpoint is identified first (the person usually has some disease); as we have seen, such studies are called case–control studies. One weakness of such studies is that there may not be a sufficient number of subjects to make "good" matches. Matching on too many variables makes it virtually impossible to find a sufficient number of control subjects. No matter how well the matching is done, there is the possibility that the groups receiving the two treatments (the case and control groups) are not sufficiently similar because of unrecognized variables.

A third approach is not to match on specific variables but to try to select the subjects on an intuitive basis. For example, such procedures often select the next person entering the clinic, or have the patient select a friend of the same gender. The rationale here is that a friend will tend to belong to the same socioeconomic environment and have the same ethnic characteristics.

Still another approach, even farther removed from the "identical twins" approach, is to select a group receiving a given treatment and then to select in its entirety a second group as a control. The hope is that by careful consideration of the problem and good intuition, the control group will, in some sense, mirror the first treatment group with regard to "all pertinent characteristics" except the treatment and endpoint. In a retrospective study, the first group usually consists of cases and a control group selected from the remaining population.

The final approach is to select the two groups in some manner realizing that they will not be similar, and to measure pertinent variables, such as the variables that one had considered matching upon, as well as the appropriate endpoint variables. The idea is to make statistical adjustments to find out what would have happened had the two groups been comparable. Such adjustments are done in a variety of ways. The techniques are discussed in following chapters.

None of the foregoing methods of obtaining "valid" comparisons are totally satisfactory. In the 1920s, Sir Ronald A. Fisher and others made one of the great advances in scientific methodology—they assigned treatments to patients by chance; that is, they assigned treatments *randomly*. The technique is called *randomization*. The statistical or chance rule of assignment will satisfy certain properties that are best expressed by the concepts of probability theory. These concepts are described in Chapter 4. For assignment to two therapies, a coin toss could be used. A head would mean assignment to therapy 1; a tail would result in assignment to therapy 2. Each patient would have an equal chance of getting each therapy. Assignments to past patients would not have any effect on the therapy assigned to the next patient. By the laws of probability, on the average, treatment groups will be similar. *The groups will even be similar with respect to variables not measured or even thought about!* The mathematics of probability allow us to estimate whether differences in the outcome might be due to the chance assignment to the two groups or whether the differences should be ascribed to true differences between treatments. These points are discussed in more detail later.

2.10.3 Inference to a Larger Population

Usually, it is desired to apply the results of a study to a population beyond the experimental units. In an experiment with guinea pigs, the assumption is that if other guinea pigs had been used, the "same" results would have been found. In reporting good results with a new surgical procedure, it is implicit that this new procedure is probably good for a wide variety of patients in a wide variety of clinical settings. To extend results to a larger population, experimental units should be *representative* of the larger population. The best way to assure this is to select the experimental units *at random*, or by chance, from the larger population. The mechanics and interpretation of such random sampling are discussed in Chapter 4. Random sampling assures, on the average, a representative sample. In other instances, if one is willing to make assumptions, the extension may be valid. There is an implicit assumption in much clinical research that a treatment is good for almost everyone or almost no one. Many techniques are used initially on the subjects available at a given clinic. It is assumed that a result is true for all clinics if it works in one setting.

Sometimes, the results of a technique are compared with "historical" controls; that is, a new treatment is compared with the results of previous patients using an older technique. The use of historical controls can be hazardous; patient populations change with time, often in ways that have much more importance than is generally realized. Another approach with weaker inference is the use of an animal model. The term *animal model* indicates that the particular animal is susceptible to, or suffers from, a disease similar to that experienced by humans. If a treatment works on the animal, it may be useful for humans. There would then be an investigation in the human population to see whether the assumption is valid.

The results of an observational study carried out in one country may be extended to other countries. This is not always appropriate. Much of the "bread and butter" of epidemiology consists of noting that the same risk factor seems to produce different results in different populations, or in noting that the particular endpoint of a disease occurs with differing rates in different countries. There has been considerable advance in medical science by noting different responses among different populations. This is a broadening of the topic of this section: extending inferences in one population to another population.

2.10.4 Precision and Validity of Measurements

Statistical theory leads to the examination of variation in a method of measurement. The variation may be estimated by making repeated measurements on the same experimental unit. If instrumentation is involved, multiple measurements may be taken using more than one of the instruments to note the variation between instruments. If different observers, interviewers, or technicians take measurements, a quantification of the variability between observers may be made. It is necessary to have information on the precision of a method of measurement in calculating the sample size for a study. This information is also used in considering whether or not variables deserve repeated measurements to gain increased precision about the true response of an experimental unit.

Statistics helps in thinking about alternative methods of measuring a quantity. When introducing a new apparatus or new technique to measure a quantity of interest, validation against the old method is useful. In considering subjective ratings by different people (even when the subjective rating is given as a numerical scale), it often turns out that a quantity is not measured in the same fashion if the measurement method is changed. A new laboratory apparatus may measure consistently higher than an old one. In two methods of evaluating pain relief, one way of phrasing a question may tend to give a higher percentage of improvement. Methodologic statistical studies are helpful in placing interpretations and inferences in the proper context.

2.10.5 Quantification and Reduction of Uncertainty

Because of variability, there is uncertainty associated with the interpretation of study results. Statistical theory allows quantification of the uncertainty. If a quantity is being estimated, the amount of uncertainty in the estimate must be assessed. In considering a hypothesis, one may give numerical assessment of the chance of occurrence of the results observed when the hypothesis is true.

Appreciation of statistical methodology often leads to the design of a study with increased precision and consequently, a smaller sample size. An example of an efficient technique is the statistical idea of blocking. Blocks are subsets of relatively homogeneous experimental units. The strategy is to apply all treatments randomly to the units within a particular block. Such a design is called a *randomized block design*. The advantage of the technique is that comparisons of treatments are intrablock comparisons (i.e., comparisons within blocks) and are more precise because of the homogeneity of the experimental units within the blocks, so that it is easier to detect treatment differences. As discussed earlier, simple randomization does ensure similar groups, but the variability within the treatment groups will be greater if no blocking of experimental units has been done. For example, if age is important prognostically in the outcome of a comparative trial of two therapies, there are two approaches that one may take. If one ignores age and randomizes the two therapies, the therapies will be tested on similar groups, but the variability in outcome due to age will tend to mask the effects of the two treatments. Suppose that you place people whose ages are close into blocks and assign each treatment by a chance mechanism within each block. If you then compare the treatments within the blocks, the effect of age on the outcome of the two therapies will be largely eliminated. A more precise comparison of the therapeutic effects can be gained. This increased precision due to statistical design leads to a study that requires a smaller sample size than does a completely randomized design. However, see Meier et al. [1968] for some cautions.

A good statistical design allows the investigation of several factors at one time with little added cost (Sir R. A. Fisher as quoted by Yates [1964]):

> No aphorism is more frequently repeated with field trials than we must ask Nature a few questions, or ideally, one question at a time. The writer is convinced that this view is wholly mistaken. Nature, he suggests, will best respond to a logical and carefully thought out questionnaire; indeed if we ask her a single question, she will often refuse to answer until some other topic has been discussed.

PROBLEMS

2.1 Consider the following terms defined in Chapters 1 and 2: single blind, double blind, placebo, observational study, experiment, laboratory experiment, comparative experiment, crossover experiment, clinical study, cohort, prospective study, retrospective study, case–control study, and matched case–control study. In the examples of section 1.5, which terms apply to which parts of these examples?

2.2 List possible advantages and disadvantages of a double-blind study. Give some examples where a double-blind study clearly cannot be carried out; suggest how virtues of "blinding" can still be retained.

2.3 Discuss the ethical aspects of a randomized placebo-controlled experiment. Can you think of situations where it would be extremely difficult to carry out such an experiment?

2.4 Discuss the advantages of randomization in a randomized placebo-controlled experiment. Can you think of alternative, possibly better, designs? Consider (at least) the aspects of bias and efficiency.

2.5 This problem involves the design of two questions on "stress" to be used on a data collection form for the population of a group practice health maintenance organization. After a few years of follow-up, it is desired to assess the effect of physical and psychological stress.

- **(a)** Design a question that classifies jobs by the amount of physical work involved. Use eight or fewer categories. Assume that the answer to the question is to be based on job title. That is, someone will code the answer given a job title.
- **(b)** Same as part (a), but now the classification should pertain to the amount of psychological stress.
- **(c)** Have yourself and (independently) a friend answer your two questions for the following occupational categories: student, college professor, plumber, waitress, homemaker, salesperson, unemployed, retired, unable to work (due to illness), physician, hospital administrator, grocery clerk, prisoner.
- **(d)** What other types of questions would you need to design to capture the total amount of stress in the person's life?

2.6 In designing a form, careful distinction must be made between the following categories of nonresponse to a question: (1) not applicable, (2) not noted, (3) don't know, (4) none, and (5) normal. If nothing is filled in, someone has to determine which of the five categories applies—and often this cannot be done after the interview or the records have been destroyed. This is particularly troublesome when medical records are abstracted. Suppose that you are checking medical records to record the number of pregnancies (*gravidity*) of a patient. Unless the gravidity is specifically given, you have a problem. If no number is given, any one of the four categories above could apply. Give two other examples of questions with ambiguous interpretation of "blank" responses. Devise a scheme for interview data that is unambiguous and does not require further editing.

REFERENCES

Meier, P., Free, S. M., Jr., and Jackson, G. L. [1968]. Reconsideration of methodology in studies of pain relief. *Biometrics*, **14**: 330–342.

Papworth, M. H. [1967]. *Human Guinea Pigs*. Beacon Press, Boston.

Riley, V. [1975]. Mouse mammary tumors: alteration of incidence as apparent function of stress. *Science*, **189**: 465–467.

Roethlisberger, F. S. [1941]. *Management and Morals*. Harvard University Press, Cambridge, MA.

Spicker, S. F., et al. (eds.) [1988]. *The Use of Human Beings in Research, with Special Reference to Clinical Trials*. Kluwer Academic, Boston.

U.S. Department of Agriculture [1989]. Animal welfare: proposed rules, part III. *Federal Register*, Mar. 15, 1989.

U.S. Department of Health, Education, and Welfare [1975]. Protection of human subjects, part III. *Federal Register*, Aug. 8, 1975, **40**: 11854.

U.S. Department of Health, Education, and Welfare [1985]. *Guide for the Care and Use of Laboratory Animals*. DHEW Publication (NIH) 86–23. U.S. Government Printing Office, Washington, DC.

Yates, F. [1964]. Sir Ronald Fisher and the design of experiments. *Biometrics*, **20**: 307–321. Used with permission from the Biometric Society.

CHAPTER 3

Descriptive Statistics

3.1 INTRODUCTION

The beginning of an introductory statistics textbook usually contains a few paragraphs placing the subject matter in encyclopedic order, discussing the limitations or wide ramifications of the topic, and tends to the more philosophical rather than the substantive–scientific. Briefly, we consider science to be a study of the world emphasizing qualities of permanence, order, and structure. Such a study involves a drastic reduction of the real world, and often, numerical aspects only are considered. If there is no obvious numerical aspect or ordering, an attempt is made to impose it. For example, quality of medical care is not an immediately numerically scaled phenomenon but a scale is often induced or imposed. Statistics is concerned with the estimation, summarization, and obtaining of reliable numerical characteristics of the world. It will be seen that this is in line with some of the definitions given in the Notes in Chapter 1.

It may be objected that a characteristic such as the gender of a newborn baby is not numerical, but it can be coded (arbitrarily) in a numerical way; for example, 0 = male and 1 = female. Many such characteristics can be *labeled* numerically, and as long as the code, or the dictionary, is known, it is possible to go back and forth.

Consider a set of measurements of head circumferences of term infants born in a particular hospital. We have a quantity of interest—head circumference—which varies from baby to baby, and a collection of actual values of head circumferences.

Definition 3.1. A *variable* is a quantity that may vary from object to object.

Definition 3.2. A *sample* (or data set) is a collection of values of one or more variables. A member of the sample is called an *element*.

We distinguish between a variable and the value of a variable in the same way that the label "title of a book in the library" is distinguished from the title *Gray's Anatomy*. A variable will usually be represented by a capital letter, say, Y, and a value of the variable by a lowercase letter, say, y.

In this chapter we discuss briefly the types of variables typically dealt with in statistics. We then go on to discuss ways of *describing* samples of values of variables, both numerically and graphically. A key concept is that of a *frequency distribution*. Such presentations can be considered part of *descriptive statistics*. Finally, we discuss one of the earliest challenges to statistics, how to *reduce* samples to a few summarizing numbers. This will be considered under the heading of descriptive statistics.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

3.2 TYPES OF VARIABLES

3.2.1 Qualitative (Categorical) Variables

Some examples of qualitative (or categorical) variables and their values are:

1. Color of a person's hair (black, gray, red, . . . , brown)
2. Gender of child (male, female)
3. Province of residence of a Canadian citizen (Newfoundland, Nova Scotia, . . . , British Columbia)
4. Cause of death of newborn (congenital malformation, asphyxia, . . .)

Definition 3.3. A *qualitative variable* has values that are intrinsically nonnumerical (categorical).

As suggested earlier, the values of a qualitative variable can always be put into numerical form. The simplest numerical form is consecutive labeling of the values of the variable. The values of a qualitative variable are also referred to as *outcomes* or *states*.

Note that examples 3 and 4 above are ambiguous. In example 3, what shall we do with Canadian citizens living outside Canada? We could arbitrarily add another "province" with the label "Outside Canada." Example 4 is ambiguous because there may be more than one cause of death. Both of these examples show that it is not always easy to anticipate all the values of a variable. Either the list of values must be changed or the variable must be redefined.

The arithmetic operation associated with the values of qualitative variables is usually that of counting. Counting is perhaps the most elementary—but not necessarily simple—operation that organizes or abstracts characteristics. A *count* is an answer to the question: How many? (Counting assumes that whatever is counted shares some characteristics with the other "objects." Hence it disregards what is unique and reduces the objects under consideration to a common category or class.) Counting leads to statements such as "the number of births in Ontario in 1979 was 121,655."

Qualitative variables can often be ordered or ranked. *Ranking* or *ordering* places a set of objects in a sequence according to a specified scale. In Chapter 2, clinicians ranked interns according to the quality of medical care delivered. The "objects" were the interns and the scale was "quality of medical care delivered." The interns could also be ranked according to their height, from shortest to tallest—the "objects" are again the interns and the scale is "height." The provinces of Canada could be ordered by their population sizes from lowest to highest. Another possible ordering is by the latitudes of, say, the capitals of each province. Even hair color could be ordered by the wavelength of the dominant color. Two points should be noted in connection with ordering or qualitative variables. First, as indicated by the example of the provinces, there is more than one ordering that can be imposed on the outcomes of a variable (i.e., there is no natural ordering); the type of ordering imposed will depend on the nature of the variable and the purpose for which it is studied—if we wanted to study the impact of crowding or pollution in Canadian provinces, we might want to rank them by population size. If we wanted to study rates of melanoma as related to amount of ultraviolet radiation, we might want to rank them by the latitude of the provinces as summarized, say by the latitudes of the capitals or most populous areas. Second, the ordering need not be complete; that is, we may not be able to rank each outcome above or below another. For example, two of the Canadian provinces may have virtually identical populations, so that it is not possible to order them. Such orderings are called *partial*.

3.2.2 Quantitative Variables

Some examples of quantitative variables (with scale of measurement; values) are the following:

1. Height of father ($\frac{1}{2}$ inch units; 0.0, 0.5, 1.0, 1.5, . . . , 99.0, 99.5, 100.0)

2. Number of particles emitted by a radioactive source (counts per minute; 0, 1, 2, 3, . . .)
3. Total body calcium of a patient with osteoporosis (nearest gram; 0, 1, 2, . . . , 9999, 10,000)
4. Survival time of a patient diagnosed with lung cancer (nearest day; 0, 1, 2, . . . , 19,999, 20,000)
5. Apgar score of infant 60 seconds after birth (counts; 0, 1, 2, . . . , 8, 9, 10)
6. Number of children in a family (counts; 0, 1, 2, 3, . . .)

Definition 3.4. A *quantitative variable* has values that are intrinsically numerical.

As illustrated by the examples above, we must specify two aspects of a variable: the scale of measurement and the values the variable can take on. Some quantitative variables have numerical values that are integers, or discrete. Such variables are referred to as *discrete variables*. The variable "number of particles emitted by a radioactive source" is such an example; there are "gaps" between the successive values of this variable. It is not possible to observe 3.5 particles. (It is sometimes a source of amusement when discrete numbers are manipulated to produce values that cannot occur—for example, "the average American family" has 2.125 children). Other quantitative variables have values that are potentially associated with real numbers—such variables are called *continuous variables*. For example, the survival time of a patient diagnosed with lung cancer may be expressed to the nearest day, but this phrase implies that there has been rounding. We could refine the measurement to, say, hours, or even more precisely, to minutes or seconds. The exactness of the values of such a variable is determined by the precision of the measuring instrument as well as the usefulness of extending the value. Usually, a reasonable unit is assumed and it is considered *pedantic* to have a unit that is too refined, or *rough* to have a unit that does not permit distinction between the objects on which the variable is measured. Examples 1, 3, and 4 above deal with continuous variables; those in the other examples are discrete. Note that with quantitative variables there is a natural ordering (e.g., from lowest to highest value) (see Note 3.7 for another taxonomy of data).

In each illustration of qualitative and quantitative variables, we listed all the possible values of a variable. (Sometimes the values could not be listed, usually indicated by inserting three dots ". . . " into the sequence.) This leads to:

Definition 3.5. The *sample space* or *population* is the set of all possible values of a variable.

The definition or listing of the sample space is not a trivial task. In the examples of qualitative variables, we already discussed some ambiguities associated with the definitions of a variable and the sample space associated with the variable. Your definition must be reasonably precise without being "picky." Consider again the variable "province of residence of a Canadian citizen" and the sample space (Newfoundland, Nova Scotia, . . . , British Columbia). Some questions that can be raised include:

1. What about citizens living in the Northwest Territories? (Reasonable question)
2. Are landed immigrants who are not yet citizens to be excluded? (Reasonable question)
3. What time point is intended? Today? January 1, 2000? (Reasonable question)
4. If January 1, 2000 is used, what about citizens who died on that day? Are they to be included? (Becoming somewhat "picky")

3.3 *DESCRIPTIVE* STATISTICS

3.3.1 Tabulations and Frequency Distributions

One of the simplest ways to summarize data is by tabulation. John Graunt, in 1662, published his observations on bills of mortality, excerpts of which can be found in Newman [1956].

Table 3.1 Diseases and Casualties in the City of London 1632

Disease	Casualties
Abortive and stillborn	445
Affrighted	1
Aged	628
Ague	43
⋮	
Crisomes and infants	2268
⋮	
Tissick	34
Vomiting	1
Worms	27
In all	9535

Source: A selection from Graunt's tables; from Newman [1956].

Table 3.1 is a condensation of Graunt's list of 63 diseases and casualties. Several things should be noted about the table. To make up the table, three ingredients are needed: (1) a *collection* of objects (in this case, humans), (2) a *variable* of interest (the cause of death), and (3) the *frequency* of occurrence of each category. These are defined more precisely later. Second, we note that the disease categories are arranged alphabetically (ordering number 1). This may not be too helpful if we want to look at the most common causes of death. Let us rearrange Graunt's table by listing disease categories by greatest frequencies (ordering number 2).

Table 3.2 lists the 10 most common disease categories in Graunt's table and summarizes $8274/9535 = 87\%$ of the data in Table 3.1. From Table 3.2 we see at once that "crisomes" is the most frequent cause of death. (A *crisome* is an infant dying within one month of birth. Gaunt lists the number of "christenings" [births] as 9584, so a crude estimate of neonatal mortality is $2268/9584 \doteq 24\%$. The symbol "$\doteq$" means "approximately equal to.") Finally, we note that data for 1633 almost certainly would not have been identical to that of 1632. However, the number in the category "crisomes" probably would have remained the largest. An example of a statistical question is whether this predominance of "crisomes and infants" has a quality of permanence from one year to the next.

A second example of a tabulation involves keypunching errors made by a data-entry operator. To be entered were 156 lines of data, each line containing data on the number of crib deaths for a particular month in King County, Washington, for the years 1965–1977. Other data on

Table 3.2 Rearrangement of Graunt's Data (Table 3.1) by the 10 Most Common Causes of Death

Disease	Casualties	Disease	Casualties
Crisomes and infants	2268	Bloody flux, scowring, and flux	348
Consumption	1797	Dropsy and swelling	267
Fever	1108	Convulsion	241
Aged	628	Childbed	171
Flocks and smallpox	531		
Teeth	470	Total	8274
Abortive and stillborn	445		

Table 3.3 Number of Keypunching Errors per Line for 156 Consecutive Lines of Data Entered[a]

0	0	1	0	2	0	0	0	1	0	0	0
0	0	0	0	1	0	0	1	2	0	0	1
1	0	0	2	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	1	1	1	1	0	0	0	0	0	0	1
0	1	0	0	1	0	0	0	0	2	0	0
1	0	0	0	2	0	0	0	0	0	0	0
1	0	0	0	1	0	1	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0	0
0	1	0	1	1	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0

[a] Each digit represents the number of errors in a line.

a line consisted of meteorological data as well as the total number of births for that month in King County. Each line required the punching of 47 characters, excluding the spaces. The numbers of errors per line starting with January 1965 and ending with December 1977 are listed in Table 3.3.

One of the problems with this table is its bulk. It is difficult to grasp its significance. You would not transmit this table over the phone to explain to someone the number of errors made. One way to summarize this table is to specify how many times a particular combination of errors occurred. One possibility is the following:

Number of Errors per Line	Number of Lines
0	124
1	27
2	5
3 or more	0

This list is again based on three ingredients: a *collection* of lines of data, a *variable* (the number of errors per line), and the *frequency* with which values of the variable occur. Have we lost something in going to this summary? Yes, we have lost the order in which the observations occurred. That could be important if we wanted to find out whether errors came "in bunches" or whether there was a learning process, so that fewer errors occurred as practice was gained. The original data are already a condensation. The "number of errors per line" does not give information about the location of the errors in the line or the type of error. (For educational purposes, the latter might be very important.)

A difference between the variables of Tables 3.2 and 3.3 is that the variable in the second example was *numerically valued* (i.e., took on numerical values), in contrast with the *categorically valued* variable of the first example. Statisticians typically mean the former when *variable* is used by itself, and we will specify *categorical variable* when appropriate. [As discussed before, a categorical variable can always be made numerical by (as in Table 3.1) arranging the values alphabetically and numbering the observed categories 1, 2, 3, ... This is not biologically meaningful because the ordering is a function of the language used.]

The data of the two examples above were discrete. A different type of variable is represented by the age at death of crib death, or SIDS (sudden infant death syndrome), cases. Table 3.4

Table 3.4 Age at Death (in Days) of 78 Cases of SIDS Occurring in King County, Washington, 1976–1977

225	174	274	164	130	96	102	80	81	148	130	48
68	64	234	24	187	117	42	38	28	53	120	66
176	120	77	79	108	117	96	80	87	85	61	65
68	139	307	185	150	88	108	60	108	95	25	80
143	57	53	90	76	99	29	110	113	67	22	118
47	34	206	104	90	157	80	171	23	92	115	87
42	77	65	45	32	44						

Table 3.5 Frequency Distribution of Age at Death of 78 SIDS Cases Occurring in King County, Washington, 1976–1977

Age Interval (days)	Number of Deaths	Age Interval (days)	Number of Deaths
1–30	6	211–240	1
31–60	13	241–270	0
61–90	23	271–300	1
91–120	18	301–330	1
121–150	7		
151–180	5	Total	78
181–210	3		

displays ages at death in days of 78 cases of SIDS in King County, Washington, during the years 1976–1977. The variable, age at death, is continuous. However, there is rounding to the nearest whole day. Thus, “68 days” could represent 68.438... or 67.8873..., where the three dots indicate an unending decimal sequence.

Again, the table staggers us by its bulk. Unlike the preceding example, it will not be too helpful to list the number of times that a particular value occurs: There are just too many different ages. One way to reduce the bulk is to define intervals of days and count the number of observations that fall in each interval. Table 3.5 displays the data grouped into 30-day intervals (months). Now the data make more sense. We note, for example, that many deaths occur between the ages of 61 and 90 days (two to three months) and that very few deaths occur after 180 days (six months). Somewhat surprisingly, there are relatively few deaths in the first month of life. This age distribution pattern is unique to SIDS.

We again note the three characteristics on which Table 3.5 is based: (1) a *collection* of 78 objects—SIDS cases, (2) a *variable* of interest—age at death, and (3) the *frequency* of occurrence of values falling in specified intervals. We are now ready to define these three characteristics more explicitly.

Definition 3.6. An *empirical frequency distribution* (EFD) of a variable is a listing of the values or ranges of values of the variable together with the frequencies with which these values or ranges of values occur.

The adjective *empirical* emphasizes that an *observed* set of values of a variable is being discussed; if this is obvious, we may use just “frequency distribution” (as in the heading of Table 3.5).

The choice of interval width and interval endpoint is somewhat arbitrary. They are usually chosen for convenience. In Table 3.5, a “natural” width is 30 days (one month) and convenient endpoints are 1 day, 31 days, 61 days, and so on. A good rule is to try to produce between seven and 10 intervals. To do this, divide the range of the values (*largest to smallest*) by 7, and then adjust to make a simple interval. For example, suppose that the variable is "weight of adult male" (expressed to the nearest kilogram) and the values vary from 54 to 115 kg. The range is $115 - 54 = 61$ kg, suggesting intervals of width $61/7 \doteq 8.7$ kg. This is clearly not a very good width; the closest "natural" width is 10 kg (producing a slightly coarser grid). A reasonable starting point is 50 kg, so that the intervals have endpoints 50 kg, 60 kg, 70 kg, and so on.

To compare several EFDs it is useful to make them comparable with respect to the total number of subjects. To make them comparable, we need:

Definition 3.7. The *size* of a sample is the number of elements in the sample.

Definition 3.8. An *empirical relative frequency distribution* (ERFD) is an empirical frequency distribution where the frequencies have been divided by the sample size.

Equivalently, the relative frequency of the value of a variable is the proportion of times that the value of the variable occurs. (The context often makes it clear that an *empirical* frequency distribution is involved. Similarly, many authors omit the adjective *relative* so that "frequency distribution" is shorthand for "empirical relative frequency distribution.")

To illustrate ERFDs, consider the data in Table 3.6, consisting of systolic blood pressures of three groups of Japanese men: native Japanese, first-generation immigrants to the United States (Issei), and second-generation Japanese in the United States (Nisei). The sample sizes are 2232, 263, and 1561, respectively.

It is difficult to compare these distributions because the sample sizes differ. The *relative* frequencies (proportions) are obtained by dividing each frequency by the corresponding sample size. The ERFD is presented in Table 3.7. For example, the (empirical) relative frequency of native Japanese with systolic blood pressure less than 106 mmHg is $218/2232 = 0.098$.

It is still difficult to make comparisons. One of the purposes of the study was to determine how much variables such as blood pressure were affected by environmental conditions. To see if there is a *shift* in the blood pressures, we could consider the proportion of men with blood pressures less than a specified value and compare the groups that way. Consider, for example, the proportion of men with systolic blood pressures less than or equal to 134 mmHg. For the native Japanese this is (Table 3.7) $0.098 + 0.122 + 0.151 + 0.162 = 0.533$, or 53.3%. For the Issei and Nisei these figures are 0.413 and 0.508, respectively. The latter two figures are somewhat lower than the first, suggesting that there has been a shift to higher systolic

Table 3.6 Empirical Frequency Distribution of Systolic Blood Pressure of Native Japanese and First- and Second-Generation Immigrants to the United States, Males Aged 45–69 Years

Blood Pressure (mmHg)	Native Japanese	Issei	California Nisei
<106	218	4	23
106–114	272	23	132
116–124	337	49	290
126–134	362	33	347
136–144	302	41	346
146–154	261	38	202
156–164	166	23	109
>166	314	52	112
Total	2232	263	1561

Source: Data from Winkelstein et al. [1975].

Table 3.7 Empirical Relative Frequency Distribution of Systolic Blood Pressure of Native Japanese and First- and Second-Generation Immigrants to the United States, Males Aged 45–69 Years

Blood Pressure (mmHg)	Native Japanese	Issei	California Nisei
<106	0.098	0.015	0.015
106–114	0.122	0.087	0.085
116–124	0.151	0.186	0.186
126–134	0.162	0.125	0.222
136–144	0.135	0.156	0.222
146–154	0.117	0.144	0.129
156–164	0.074	0.087	0.070
>166	0.141	0.198	0.072
Total	1.000	0.998	1.001
Sample size	(2232)	(263)	(1561)

Source: Data from Winkelstein et al. [1975].

blood pressure among the immigrants. Whether this shift represents sampling variability or a genuine shift in these groups can be determined by methods developed in the next three chapters.

The concept discussed above is formalized in the empirical cumulative distribution.

Definition 3.9. The *empirical cumulative distribution* (ECD) of a variable is a listing of values of the variable together with the *proportion* of observations less than or equal to that value (cumulative proportion).

Before we construct the ECD for a sample, we need to clear up one problem associated with rounding of values of continuous variables. Consider the age of death of the SIDS cases of Table 3.4. The first age listed is 225 days. Any value between 224.5+ and 225.5− is rounded off to 225 (224.5+ indicates a value greater than 224.5 by some arbitrarily small amount, and similarly, 225.5− indicates a value less than 225.5). Thus, the upper endpoint of the interval 1–30 days in Table 3.5 is $30.4\overline{9}$, or 30.5.

The ECD associated with the data of Table 3.5 is presented in Table 3.8, which contains (1) the age intervals, (2) endpoints of the intervals, (3) EFD, (4) ERFD, and (5) ECD.

Two comments are in order: (1) there is a slight rounding error in the last column because the relative frequencies are rounded to three decimal places—if we had calculated from the frequencies rather than the relative frequencies, this problem would not have occurred; and (2) given the cumulative proportions, the original proportions can be recovered. For example, consider the following endpoints and their cumulative frequencies:

150.5	0.860
180.5	0.924

Subtracting, $0.924 - 0.860 = 0.064$ produces the proportion in the interval 151–180. Mathematically, the ERFD and the ECD are equivalent.

Table 3.8 Frequency Distribution of Age at Death of 78 SIDS Cases Occurring in King County, Washington, 1976–1977

Age Interval (days)	Endpoint of Interval (days)	Number of Deaths	Relative Frequency (Proportion)	Cumulative Proportion
1–30	30.5	6	0.077	0.077
31–60	60.5	13	0.167	0.244
61–90	90.5	23	0.295	0.539
91–120	120.5	18	0.231	0.770
121–150	150.5	7	0.090	0.860
151–180	180.5	5	0.064	0.924
181–210	210.5	3	0.038	0.962
211–240	240.5	1	0.013	0.975
241–270	270.5	0	0.000	0.975
271–300	300.5	1	0.013	0.988
301–330	330.5	1	0.013	1.001
Total		78	1.001	

3.3.2 Graphs

Graphical displays frequently provide very effective descriptions of samples. In this section we discuss some very common ways of doing this and close with some examples that are innovative. Graphs can also be used to enhance certain features of data as well as to distort them. A good discussion can be found in Huff [1993].

One of the most common ways of describing a sample pictorially is to plot on one axis values of the variable and on another axis the frequency of occurrence of a value or a measure related to it. In constructing a *histogram* a number of cut points are chosen and the data are tabulated. The relative frequency of observations in each category is divided by the width of the category to obtain the *probability density*, and a bar is drawn with this height. The area of a bar is proportional to the frequency of occurrence of values in the interval.

The most important choice in drawing a histogram is the number of categories, as quite different visual impressions can be conveyed by different choices. Figure 3.1 shows measurements of albumin, a blood protein, in 418 patients with the liver disease *primary biliary cirrhosis*, using

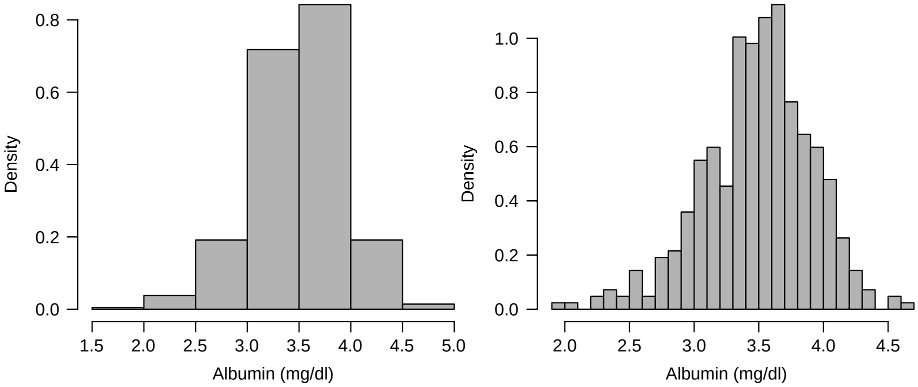


Figure 3.1 Histograms of serum albumin concentration in 418 PBC patients, using two different sets of categories.

data made available on the Web by T. M. Therneau of the Mayo Clinic. With five categories the distribution appears fairly symmetric, with a single peak. With 30 categories there is a definite suggestion of a second, lower peak. Statistical software will usually choose a sensible default number of categories, but it may be worth examining other choices.

The values of a variable are usually plotted on the abscissa (x-axis), the frequencies on the ordinate (y-axis). The ordinate on the left-hand side of Figure 3.1 contains the probability densities for each category. Note that the use of probability density means that the two histograms have similar vertical scales despite having different category widths: As the categories become narrower, the numerator and denominator of the probability density decrease together.

Histograms are sometimes defined so that the y-axis measures absolute or relative frequency rather than the apparently more complicated probability density. Two advantages arise from the use of a probability density rather than a simple count. The first is that the categories need not have the same width: It is possible to use wider categories in parts of the distribution where the data are relatively sparse. The second advantage is that the height of the bars does not depend systematically on the sample size: It is possible to compare on the same graph histograms from two samples of different sizes. It is also possible to compare the histogram to a hypothesized mathematical distribution by drawing the mathematical density function on the same graph (an example is shown in Figure 4.7.

Figure 3.2 displays the empirical cumulative distribution (ECD). This is a *step function* with jumps at the endpoints of the interval. The height of the jump is equal to the relative frequency of the observations in the interval. The ECD is nondecreasing and is bounded above by 1. Figure 3.2 emphasizes the discreteness of data. A *frequency polygon* and *cumulative frequency polygon* are often used with continuous variables to emphasize the continuity of the data. A frequency polygon is obtained by joining the heights of the bars of the histogram at their midpoints. The frequency polygon for the data of Table 3.8 is displayed in Figure 3.3. A question arises: Where is the midpoint of the interval? To calculate the midpoint for the interval 31–60 days, we note

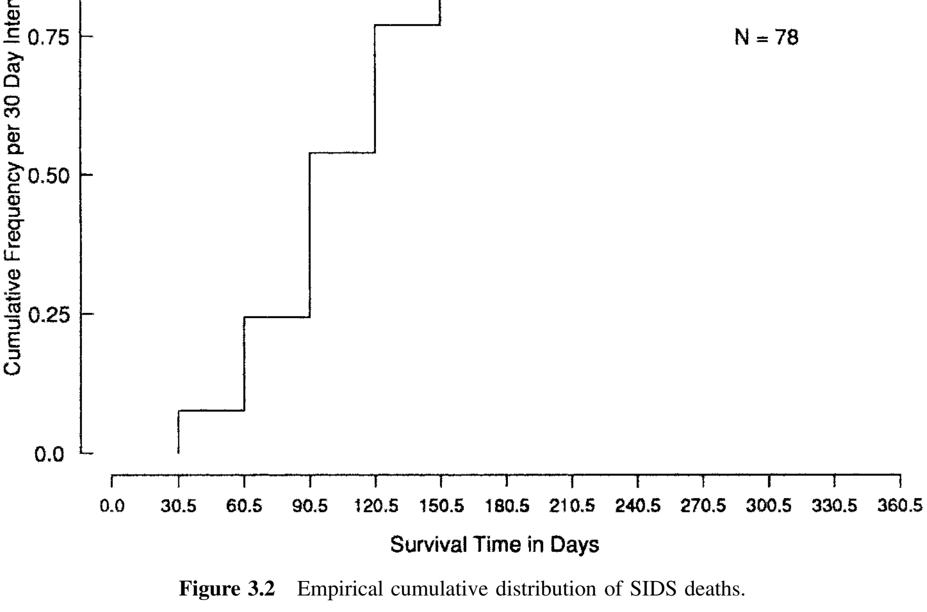


Figure 3.2 Empirical cumulative distribution of SIDS deaths.

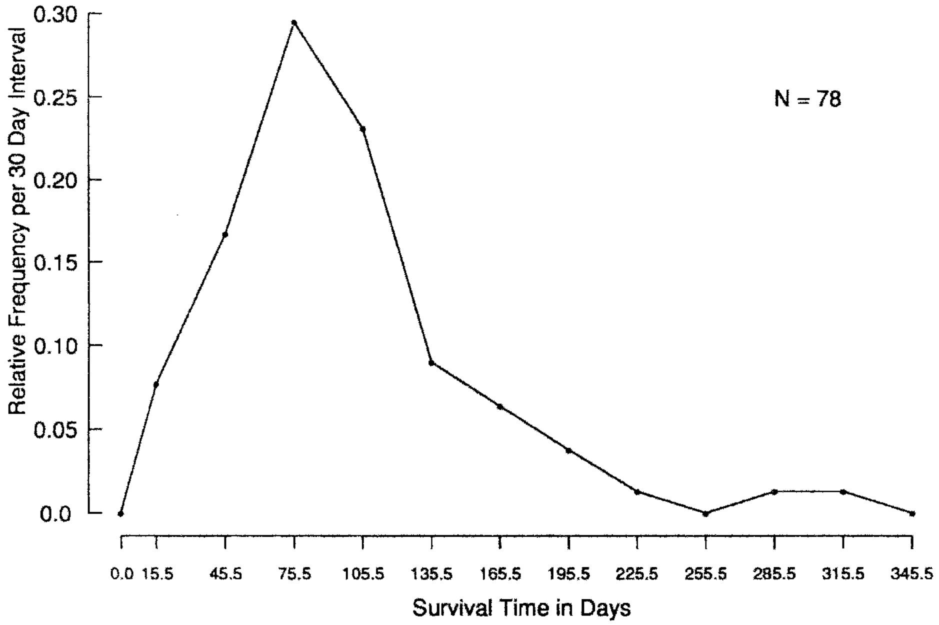


Figure 3.3 Frequency polygon of SIDS deaths.

that the limits of this interval are 30.5–60.5. The midpoint is halfway between these endpoints; hence, *midpoint* $= (30.5 + 60.5)/2 = 45.5$ days.

All midpoints are spaced in intervals of 30 days, so that the midpoints are 15.5, 45.5, 75.5, and so on. To close the polygon, the midpoints of two additional intervals are needed: one to the left of the first interval (1–30) and one to the right of the last interval observed (301–330), both of these with zero observed frequencies.

A cumulative frequency polygon is constructed by joining the cumulative relative frequencies observed at the endpoints of their respective intervals. Figure 3.4 displays the cumulative relative frequency of the SIDS data of Table 3.8. The curve has the value 0.0 below 0.5 and the value 1.0 to the right of 330.5. Both the histograms and the cumulative frequency graphs implicitly assume that the observations in our interval are evenly distributed over that interval.

One advantage of a cumulative frequency polygon is that the proportion (or percentage) of observations less than a specified value can be read off easily from the graph. For example, from Figure 3.4 it can be seen that 50% of the observations have a value of less than 88 days (this is the median of the sample). See Section 3.4.1 for further discussion.

EFDs can often be graphed in an innovative way to illustrate a point. Consider the data in Figure 3.5, which contains the frequency of births per day as related to phases of the moon. Data were collected by Schwab [1975] on the number of births for two years, grouped by each day of the 29-day lunar cycle, presented here as a circular distribution where the lengths of the sectors are proportional to the frequencies. (There is clearly no evidence supporting the hypothesis that the cycle of the moon influences birth rate.)

Sometimes more than one variable is associated with each of the objects under study. Data arising from such situations are called *multivariate data.* A moment's reflection will convince you that most biomedical data are multivariate in nature. For example, the variable "blood pressure of a patient" is usually expressed by two numbers, systolic and diastolic blood pressure. We often specify age and gender of patients to characterize blood pressure more accurately.

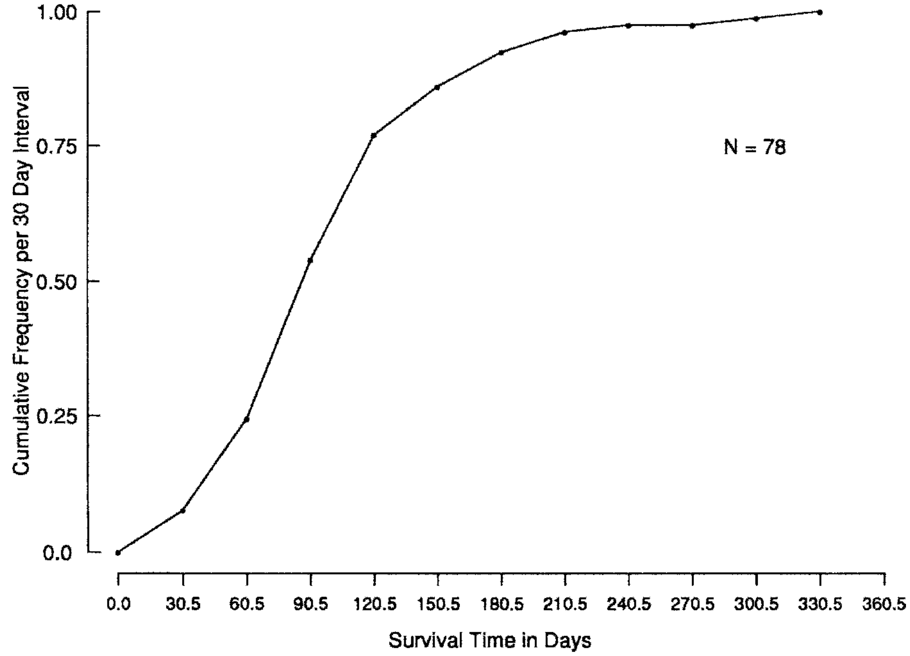


Figure 3.4 Cumulative frequency polygon of SIDS deaths.

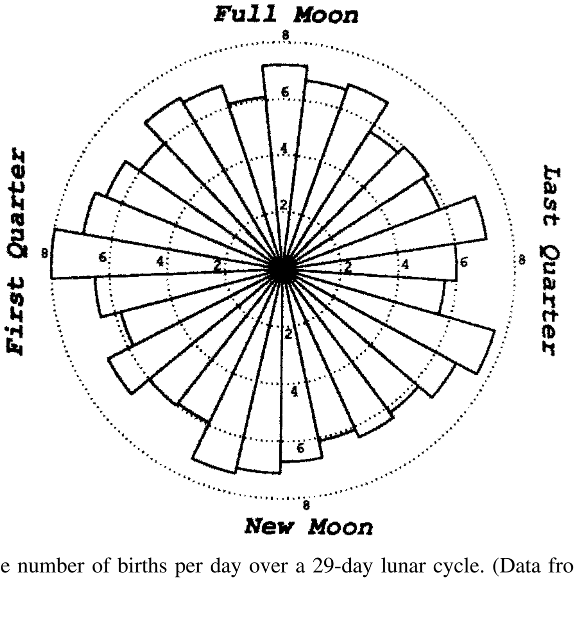


Figure 3.5 Average number of births per day over a 29-day lunar cycle. (Data from Schwab [1975].)

In the multivariate situation, in addition to describing the frequency with which each value of each variable occurs, we may want to study the relationships among the variables. For example, Table 1.2 and Figure 1.1 attempt to assess the relationship between the variables "clinical competence" and "cost of laboratory procedures ordered" of interns. Graphs of multivariate data will be found throughout the book.

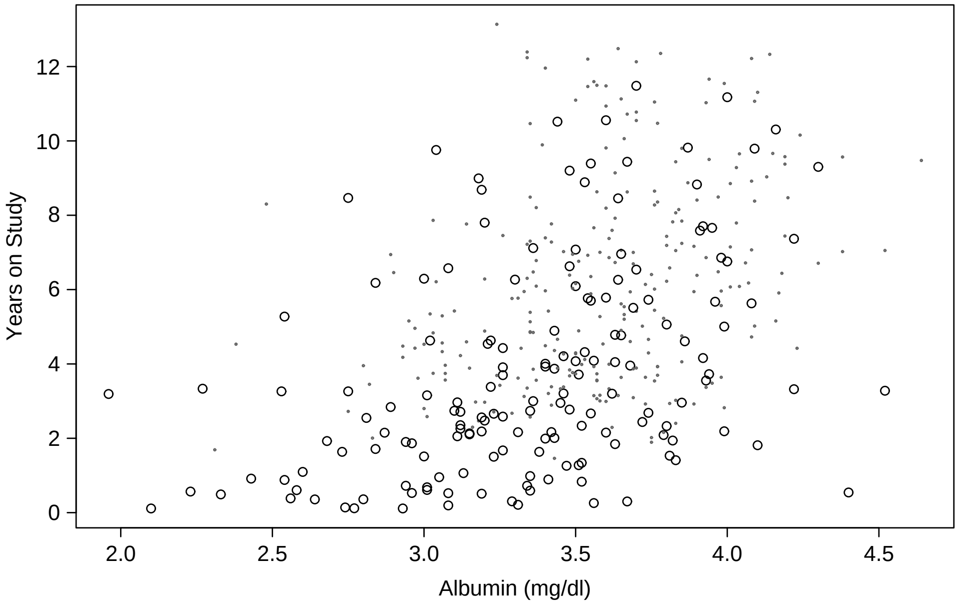


Figure 3.6 Survival time in primary biliary cirrhosis by serum albumin concentrations. Large circles are deaths, small circles are patients alive at last contact. (Data from Fleming and Harrington [1991].)

Here we present a few examples of visually displaying values of several variables at the same time. A simple one relates the serum albumin values from Figure 3.1 to survival time in the 418 patients. We do not know the survival times for everyone, as some were still alive at the end of the study. The statistical analysis of such data occupies an entire chapter of this book, but a simple descriptive graph is possible. Figure 3.6 shows large circles at survival time for patients who died. For those still alive it shows small circles at the last time known alive. For exploratory analysis and presentation these could be indicated by different colors, something that is unfortunately still not feasible for this book.

Another simple multivariate example can be found in our discussion of factor analysis. Figure 14.7 shows a matrix of correlations between variables using shaded circles whose size shows the strength of the relationship and whose shading indicates whether the relationship is positive or negative. Figure 14.7 is particularly interesting, as the graphical display helped us find an error that we missed in the first edition.

A more sophisticated example of multivariate data graphics is the *conditioning plot* [Cleveland, 1993]. This helps you examine how the relationship between two variables depends on a third. Figure 3.7 shows daily data on ozone concentration and sunlight in New York, during the summer of 1973. These should be related monotonically; ozone is produced from other pollutants by chemical reactions driven by sunlight. The four panels show four plots of ozone concentration vs. solar radiation for various ranges of temperature. The shaded bar in the title of each plot indicates the range of temperatures. These ranges overlap, which allows more panels to be shown without the data becoming too sparse. Not every statistical package will produce these coplots with a single function, but it is straightforward to draw them by taking appropriate subsets of your data.

The relationship clearly varies with temperature. At low temperatures there is little relationship, and as the temperature increases the relationship becomes stronger. Ignoring the effect of temperature and simply graphing ozone and solar radiation results in a more confusing relationship (examined in Figure 3.9). In Problem 10 we ask you to explore these data further.

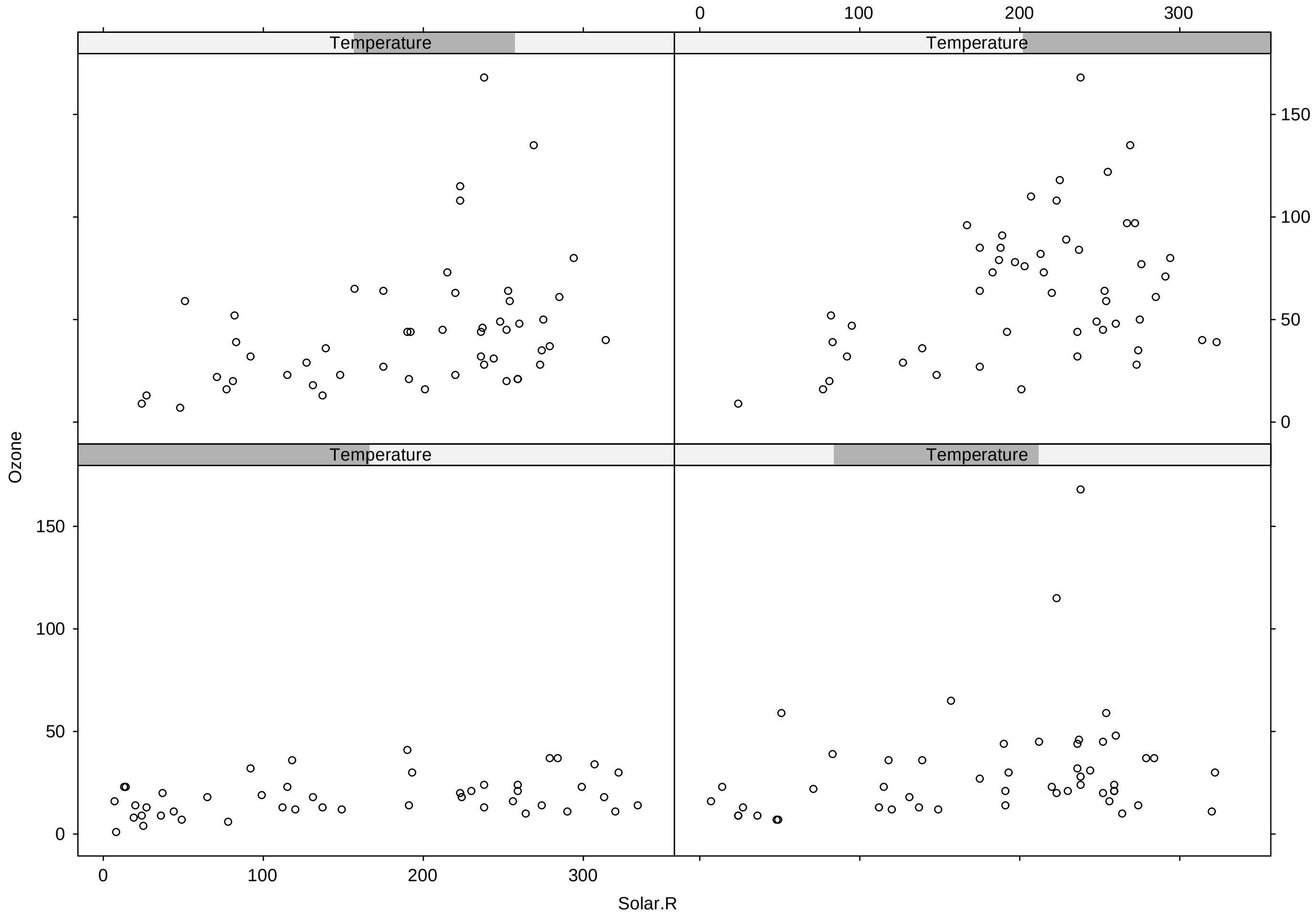


Figure 3.7 Ozone concentration by solar radiation intensity in New York, May–September 1973, conditioned on temperature. (From R Foundation [2002].)

For beautiful books on the visual display of data, see Tufte [1990, 1997, 2001]. A very readable compendium of graphical methods is contained in Moses [1987], and more recent methods are described by Cleveland [1994]. Wilkinson [1999] discusses the structure and taxonomy of graphs.

3.4 DESCRIPTIVE *STATISTICS*

In Section 3.3 our emphasis was on tabular and visual display of data. It is clear that these techniques can be used to great advantage when summarizing and highlighting data. However, even a table or a graph takes up quite a bit of space, cannot be summarized in the mind too easily, and particularly for a graph, represents data with some imprecision. For these and other reasons, numerical characteristics of data are calculated routinely.

Definition 3.10. A *statistic* is a numerical characteristic of a sample.

One of the functions of statistics as a field of study is to describe samples by as few numerical characteristics as possible. Most numerical characteristics can be classified broadly into statistics derived from percentiles of a frequency distribution and statistics derived from moments of a frequency distribution (both approaches are explained below). Roughly speaking, the former approach tends to be associated with a statistical methodology usually termed *nonparametric*, the latter with *parametric* methods. The two classes are used, contrasted, and evaluated throughout the book.

3.4.1 Statistics Derived from Percentiles

A *percentile* has an intuitively simple meaning—for example, the 25th percentile is that value of a variable such that 25% of the observations are less than that value and 75% of the observations are greater. You can supply a similar definition for, say, the 75th percentile. However, when we apply these definitions to a particular sample, we may run into three problems: (1) small sample size, (2) tied values, or (3) nonuniqueness of a percentile. Consider the following sample of four observations:

$$22, 22, 24, 27$$

How can we define the 25th percentile for this sample? There is no value of the variable with this property. But for the 75th percentile, there is an infinite number of values—for example, 24.5, 25, and 26.9378 all satisfy the definition of the 75th percentile. For large samples, these problems disappear and we will define percentiles for small samples in a way that is consistent with the intuitive definition. To find a particular percentile in practice, we would rank the observations from smallest to largest and count until the proportion specified had been reached. For example, to find the 50th percentile of the four numbers above, we want to be somewhere between the second- and third-largest observation (between the values for ranks 2 and 3). Usually, this value is taken to be halfway between the two values. This could be thought of as the value with rank 2.5—call this a *half rank*. Note that

$$2.5 = \left(\frac{50}{100}\right)(1 + \text{sample size})$$

You can verify that the following definition is consistent with your intuitive understanding of percentiles:

Definition 3.11. The Pth *percentile* of a sample of n observations is that value of the variable with rank $(P/100)(1 + n)$. If this rank is not an integer, it is rounded to the nearest half rank.

The following data deal with the aflatoxin levels of raw peanut kernels as described by Quesenberry et al. [1976]. Approximately 560 g of ground meal was divided among 16 centrifuge bottles and analyzed. One sample was lost, so that only 15 readings are available (measurement units are not given). The values were

$$30, 26, 26, 36, 48, 50, 16, 31, 22, 27, 23, 35, 52, 28, 37$$

The 50th percentile is that value with rank $(50/100)(1 + 15) = 8$. The eighth largest (or smallest) observation is 30. The 25th percentile is the observation with rank $(25/100)(1+15) = 4$, and this is 26. Similarly, the 75th percentile is 37. The 10th percentile (or decile) is that value with rank $(10/100)(1 + 15) = 1.6$, so we take the value halfway between the smallest and second-smallest observation, which is $(1/2)(16 + 22) = 19$. The 90th percentile is the value with rank $(90/100)(1 + 15) = 14.4$; this is rounded to the nearest half rank of 14.5. The value with this half rank is $(1/2)(50 + 52) = 51$.

Certain percentile or functions of percentiles have specific names:

Percentile	**Name**
50	Median
25	Lower quartile
75	Upper quartile

All these statistics tell something about the location of the data. If we want to describe how spread out the values of a sample are, we can use the range of values (largest minus smallest), but a problem is that this statistic is very dependent on the sample size. A better statistic is given by:

Definition 3.12. The *interquartile range* (IQR) is the difference between the 75th and 25th percentiles.

For the aflatoxin example, the interquartile range is $37 - 26 = 11$. Recall the *range* of a set of numbers is the largest value minus the smallest value. The data can be summarized as follows:

Median	30	Measures of location
Minimum	16	
Maximum	52	
Interquartile range	11	Measures of spread
Range	36	

The first three measures describe the location of the data; the last two give a description of their spread. If we were to add 100 to each of the observations, the median, minimum, and maximum would be shifted by 100, but the interquartile range and range would be unaffected.

These data can be summarized graphically by means of a *box plot* (also called a *box-and-whisker plot*). A rectangle with upper and lower edges at the 25th and 75th percentiles is drawn with a line in the rectangle at the median (50th percentile). Lines (whiskers) are drawn from the rectangle (box) to the highest and lowest values that are within 1.5 × IQR of the median; any points more extreme than this are plotted individually. This is Tukey's [1977] definition of the box plot; an alternative definition draws the whiskers from the quartiles to the maximum and minimum.

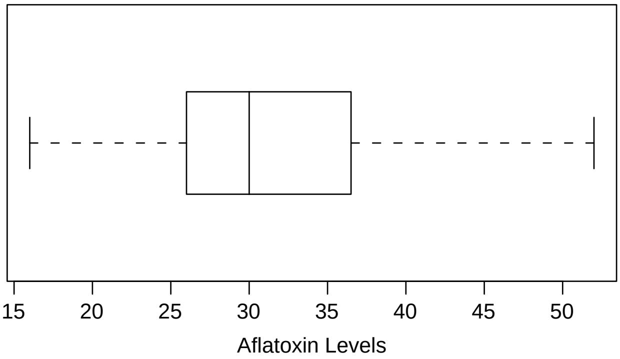


Figure 3.8 Box plot.

The box plot for these data (Figure 3.8) indicates that 50% of the data between the lower and upper quartiles is distributed over a much narrower range than the remaining 50% of the data. There are no extreme values outside the "fences" at median ± 1.5 × IQR.

3.4.2 Statistics Derived from Moments

The statistics discussed in Section 3.4.1 dealt primarily with describing the location and the variation of a sample of values of a variable. In this section we introduce another class of statistics, which have a similar purpose. In this class are the ordinary average, or arithmetic mean, and standard deviation. The reason these statistics are said to be derived from *moments* is that they are based on powers or moments of the observations.

Definition 3.13. The *arithmetic mean* of a sample of values of a variable is the average of all the observations.

Consider the aflatoxin data mentioned in Section 3.4.1. The arithmetic mean of the data is

$$\frac{30 + 26 + 26 + \cdots + 28 + 37}{15} = \frac{487}{15} = 32.4\overline{6} \doteq 32.5$$

A reasonable rule is to express the mean with one more significant digit than the observations, hence we round $32.4\overline{6}$—a nonterminating decimal—to 32.5. (See also Note 3.2 on significant digits and rounding.)

Notation. The specification of some of the statistics to be calculated can be simplified by the use of notation. We use a capital letter for the name of a variable and the corresponding lowercase letter for a value. For example, $Y =$ *aflatoxin level* (the name of the variable); $y = 30$ (the value of aflatoxin level for a particular specimen). We use the Greek symbol $\sum$ to mean "sum all the observations." Thus, for the aflatoxin example, $\sum y$ is shorthand for the statement "sum all the aflatoxin levels." Finally, we use the symbol $\overline{y}$ to denote the arithmetic mean of the sample. The arithmetic mean of a sample of n values of a variable can now be written as

$$\overline{y} = \frac{\sum y}{n}$$

For example, $\sum y = 487, n = 15$, and $\overline{y} = 487/15 \doteq 32.5$. Consider now the variable of Table 3.3: the number of keypunching errors per line. Suppose that we want the average

Table 3.9 Calculation of Arithmetic Average from Empirical Frequency and Empirical Relative Frequency Distribution[a]

Number of Errors per Line, y	Number of Lines, f	Proportion of Lines, p	$p \times y$
0	124	0.79487	0.00000
1	27	0.17308	0.17308
2	5	0.03205	0.06410
3	0	0.00000	0.00000
Total	156	1.00000	0.23718

[a] Data from Table 3.3.

number of errors per line. By definition, this is $(0+0+1+0+2+\cdots+0+0+0+0)/156 = 37/156 \doteq 0.2$ error per line. But this is a tedious way to calculate the average. A simpler way utilizes the frequency distribution or relative frequency distribution.

The total number of errors is $(124 \times 0) + (27 \times 1) + (5 \times 2) + (0 \times 3) = 37$; that is, there are 124 lines without errors; 27 lines each of which contains one error, for a total of 27 errors for these types of lines; and 5 lines with two errors, for a total of 10 errors for these types of lines; and finally, no lines with 3 errors (or more). So the arithmetic mean is

$$\overline{y} = \frac{\sum fy}{\sum f} = \frac{\sum fy}{n}$$

since the frequencies, f, add up to n, the sample size. Here, the sum $\sum fy$ is over observed values of y, each value appearing once.

The arithmetic mean can also be calculated from the empirical relative frequencies. We use the following algebraic property:

$$\overline{y} = \frac{\sum fy}{n} = \sum \frac{fy}{n} = \sum \frac{f}{n} y = \sum py$$

The *f/n* are precisely the empirical relative frequencies or proportions, p. The calculations using proportions are given in Table 3.9. The value obtained for the sample mean is the same as before. The formula $\overline{y} = \sum py$ will be used extensively in Chapter 4 when we come to probability distributions. If the values y represent the midpoints of intervals in an empirical frequency distribution, the mean of the grouped data can be calculated in the same way.

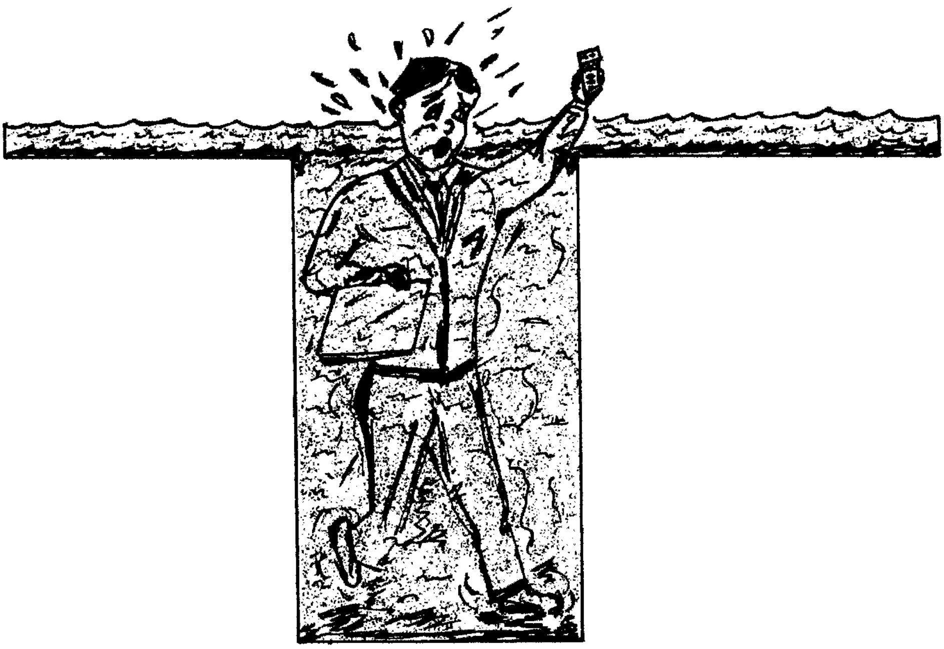
Analogous to the interquartile range there is a measure of spread based on sample moments.

Definition 3.14. The *standard deviation* of a sample of n values of a variable Y is

$$s = \sqrt{\frac{\sum(y - \overline{y})^2}{n-1}}$$

Roughly, the standard deviation is the square root of the average of the square of the deviations from the sample mean. The reason for dividing by $n - 1$ is explained in Note 3.5. Before giving an example, we note the following properties of the standard deviation:

1. The standard deviation has the same units of measurement as the variable. If the observations are expressed in centimeters, the standard deviation is expressed in centimeters.



Cartoon 3.1 Variation is important: statistician drowning in a river of average depth 10.634 inches.

2. If a constant value is added to each of the observations, the value of the standard deviation is unchanged.
3. If the observations are multiplied by a positive constant value, the standard deviation is multiplied by the same constant value.
4. The following two formulas are sometimes computationally more convenient in calculating the standard deviation by hand:

$$s = \sqrt{\frac{\sum y^2 - n\overline{y}^2}{n-1}} = \sqrt{\frac{\sum y^2 - (\sum y)^2/n}{n-1}}$$

   Rounding errors accumulate more rapidly using these formulas; care should be taken to carry enough significant digits in the computation.
5. The square of the standard deviation is called the *variance*.
6. In many situations the standard deviation can be approximated by

$$s \doteq \frac{\text{interquartile range}}{1.35}$$

7. In many cases it is true that approximately 68% of the observations fall within one standard deviation of the mean; approximately 95% within two standard deviations.

3.4.3 Graphs Based on Estimated Moments

One purpose for drawing a graph of two variables X and Y is to decide how Y changes as X changes. Just as statistics such as the mean help summarize the location of one or two samples, they can be used to summarize how the location of Y changes with X. A simple way to do this is to divide the data into *bins* and compute the mean or median for each bin.

Example 3.1. Consider the New York air quality data in Figure 3.7. When we plot ozone concentrations against solar radiation without conditioning variables, there is an apparent triangular relationship. We might want a summary of this relationship rather than trying to assess it purely by eye. One simple summary is to compute the mean ozone concentration for various ranges of solar radiation. We compute the mean ozone for days with solar radiation 0–50 langleys, 50–150, 100–200, 150–250, and so on. Plotting these means at the midpoint of the interval and joining the dots gives the dotted line shown in Figure 3.9.

Modern statistical software provides a variety of different *scatter plot smoothers* that perform more sophisticated versions of this calculation. The technical details of these are complicated, but they are conceptually very similar to the local means that we used above. The solid line in Figure 3.9 is a popular scatter plot smoother called *lowess* [Cleveland, 1981].

3.4.4 Other Measures of Location and Spread

There are many other measures of location and spread. In the former category we mention the mode and the geometric mean.

Definition 3.15. The *mode* of a sample of values of a variable Y is that value that occurs most frequently.

The mode is usually calculated for large sets of discrete data. Consider the data in Table 3.10, the distribution of the number of boys per family of eight children. The most frequently occurring value of the variable Y, the number of boys per family of eight children, is 4. There are more families with that number of boys than any other specified number of boys. For data arranged in histograms, the mode is usually associated with the midpoint of the interval having the highest frequency. For example, the mode of the systolic blood pressure of the native Japanese men listed in Table 3.6 is 130 mmHg; the modal value for Issei is 120 mmHg.

Definition 3.16. The *geometric mean* of a sample of nonnegative values of a variable Y is the nth root of the product of the n values, where n is the sample size.

Equivalently, it is the antilogarithm of the arithmetic mean of the logarithms of the values. (See Note 3.1 for a brief discussion of logarithms.)

Consider the following four observations of systolic blood pressure in mmHg:

$$118, 120, 122, 160$$

The arithmetic mean is 130 mmHg, which is larger than the first three values because the 160 mmHg value “pulls” the mean to the right. The geometric mean is $(118 \times 120 \times 122 \times 160)^{1/4} \doteq 128.9$ mmHg. The geometric mean is less affected by the extreme value of 160 mmHg. The median is 121 mmHg. If the value of 160 mmHg is changed to a more extreme value, the mean will be affected the most, the geometric mean somewhat less, and the median not at all.

Two other measures of spread are the average deviation and median absolute deviation (MAD). These are related to the standard deviation in that they are based on a location measure applied to deviations. Where the standard deviation squares the deviations to make them all positive, the average deviation takes the absolute value of the deviations (just drops any minus signs).

Definition 3.17. The *average deviation* of a sample of values of a variable is the arithmetic average of the absolute values of the deviations about the sample mean.

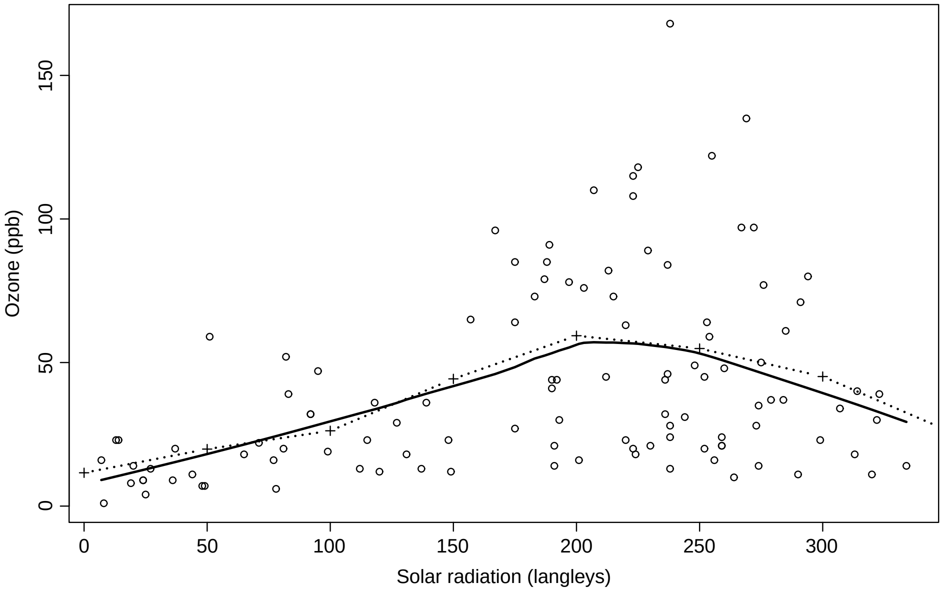


Figure 3.9 Ozone and solar radiation in New York during the summer of 1973, with scatter plot smoothers.

Table 3.10 Number of Boys in Families of Eight Children

Number of Boys per Family of Eight Children	Empirical Frequency (Number of Families)	Empirical Relative Frequency of Families
0	215	0.0040
1	1,485	0.0277
2	5,331	0.0993
3	10,649	0.1984
4	14,959	0.2787
5	11,929	0.2222
6	6,678	0.1244
7	2,092	0.0390
8	342	0.0064
Total	53,680	1.0000

Source: Geissler's data reprinted in Fisher [1958].

Using symbols, the average deviation can be written as

$$\text{average deviation} = \frac{\sum |y - \overline{y}|}{n}$$

The median absolute deviation takes the deviations from the median rather than the mean, and takes the median of the absolute values of these deviations.

Definition 3.18. The *median absolute deviation* of a sample of values of a variable is the median of the absolute values of the deviations about the sample median.

Using symbols, the median absolute deviation can be written as

$$\text{MAD} = \text{median}\{|y - \text{median}\{y\}|\}$$

The average deviation and the MAD are substantially less affected by extreme values than is the standard deviation.

3.4.5 Which Statistics?

Table 3.11 lists the statistics that have been defined so far, categorized by their use. The question arises: Which statistic should be used for a particular situation? There is no simple answer because the choice depends on the data and the needs of the investigator. Statistics derived from percentiles and those derived from moments can be compared with respect to:

1. *Scientific relevance*. In some cases the scientific question dictates or at least restricts the choice of statistic. Consider a study conducted by the Medicare program being on the effects of exercise on the amount of money expended on medical care. Their interest is in whether exercise affects total costs, or equivalently, whether it affects the arithmetic mean. A researcher studying serum cholesterol levels and the risk of heart disease might be more interested in the proportions of subjects whose cholesterol levels fell in the various categories defined by the National Cholesterol Education Program. In a completely different field, Gould [1996] discusses the absence of batting averages over 0.400 in baseball in recent years and shows that considering a measure of spread rather than a measure of location provides a much clearer explanation

2. *Robustness*. The robustness of a statistic is related to its resistance to being affected by extreme values. In Section 3.4.4 it was shown that the mean—as compared to the median and geometric mean—is most affected by extreme values. The median is said to be more robust. Robustness may be beneficial or harmful, depending on the application: In sampling pollution levels at an industrial site one would be interested in a statistic that was very much affected by extreme values. In comparing cholesterol levels between people on different diets, one might care more about the typical value and not want the results affected by an occasional extreme.

3. *Mathematical simplicity*. The arithmetic mean is more appropriate if the data can be described by a particular mathematical model: the normal or Gaussian frequency distribution, which is the basis for a large part of the theory of statistics. This is described in Chapter 4.

4. *Computational Ease*. Historically, means were easier to compute by hand for moderately large data sets. Concerns such as this vanished with the widespread availability of computers but may reappear with the very large data sets produced by remote sensing or high-throughput genomics. Unfortunately, it is not possible to give general guidelines as to which statistics

Table 3.11 Statistics Defined in This Chapter

Location	Spread
Median	Interquartile range
Percentile	Range
Arithmetic mean	Standard deviation
Geometric mean	Average deviation
Mode	Median absolute deviation

will impose less computational burden. You may need to experiment with your hardware and software if speed or memory limitations become important.

5. *Similarity*. In many samples, the mean and median are not too different. If the empirical frequency distribution of the data is almost symmetrical, the mean and the median tend to be close to each other.

In the absence of specific reasons to chose another statistic, it is suggested that the median and mean be calculated as measures of location and the interquartile range and standard deviation as measures of spread. The other statistics have limited or specialized use. We discuss robustness further in Chapter 8.

NOTES

3.1 Logarithms

A *logarithm* is an exponent on a base. The base is usually 10 or e (2.71828183...). Logarithms with base 10 are called *common logarithms*; logarithms with base e are called *natural logarithms*. To illustrate these concepts, consider

$$100 = 10^2 = (2.71828183\ldots)^{4.605170\ldots} = e^{4.605170\ldots}$$

That is, the logarithm to the base 10 of 100 is 2, usually written

$$\log_{10}(100) = 2$$

and the logarithm of 100 to the base e is

$$\log_e(100) = 4.605170\ldots$$

The three dots indicate that the number is an unending decimal expansion. Unless otherwise stated, logarithms herein will always be natural logarithms. Other bases are sometimes useful—in particular, the base 2. In determining hemagglutination levels, a series of dilutions of serum are set, each dilution being half of the preceding one. The dilution series may be 1 : 1, 1 : 2, 1 : 4, 1 : 8, 1 : 16, 1 : 32, and so on. The logarithm of the dilution factor using the base 2 is then simply

$$\begin{aligned} \log_2(1) &= 0 \\ \log_2(2) &= 1 \\ \log_2(4) &= 2 \\ \log_2(8) &= 3 \\ \log_2(16) &= 4 \qquad \text{etc.} \end{aligned}$$

The following properties of logarithms are the only ones needed in this book. For simplicity, we use the base e, but the operations are valid for any base.

1. Multiplication of numbers is equivalent to adding logarithms ($e^a \times e^b = e^{a+b}$).
2. The logarithm of the reciprocal of a number is the negative of the logarithm of the number ($1/e^a = e^{-a}$).
3. Rule 2 is a special case of this rule: Division of numbers is equivalent to subtracting logarithms ($e^a/e^b = e^{a-b}$).

Most pocket calculators permit rapid calculations of logarithms and antilogarithms. Tables are also available. You should verify that you can still use logarithms by working a few problems both ways.

3.2 *Stem-and-Leaf Diagrams*

An elegant way of describing data by hand consists of *stem-and-leaf diagrams* (a phrase coined by J. W. Tukey [1977]; see his book for some additional innovative methods of describing data). Consider the aflatoxin data from Section 3.4.1. We can tabulate these data according to their first digit (the "stem") as follows:

Stem (tens)	**Leaf (units)**	**Stem (tens)**	**Leaf (units)**
1	6	4	8
2	6 6 2 7 3 8	5	0 2
3	0 6 1 5 7		

For example, the row 3|06157 is a description of the observations 30, 36, 31, 35, and 37. The most frequently occurring category is the 20s. The smallest value is 16, the largest value, 52.

A nice feature of the stem-and-leaf diagram is that all the values can be recovered (but not in the sequence in which the observations were made). Another useful feature is that a quick ordering of the observations can be obtained by use of a stem-and-leaf diagram. Many statistical packages produce stem-and-leaf plots, but there appears to be little point to this, as the advantages over histograms or empirical frequency distributions apply only to hand computation.

3.3 *Color and Graphics*

With the wide availability of digital projectors and inexpensive color inkjet printers, there are many more opportunities for statisticians to use color to annotate and extend graphs. Differences in color are processed "preattentively" by the brain—they "pop out" visually without a conscious search. It is still important to choose colors wisely, and many of the reference books we list discuss this issue. Colored points and lines can be bright, intense colors, but large areas should use paler, less intense shades. Choosing colors to represent a quantitative variable is quite difficult, and it is advisable to make use of color schemes chosen by experts, such as those at *http://colorbrewer.org*.

Particular attention should be paid to limitations on the available color range. Color graphs may be photocopied in black and white, and might need to remain legible. LCD projectors may have disappointing color saturation. Ideas and emotions associated with a particular color might vary in different societies. Finally, it is important to remember that about 7% of men (and almost no women) cannot distinguish red and green. The Web appendix contains a number of links on color choice for graphics.

3.4 *Significant Digits: Rounding and Approximation*

In working with numbers that are used to estimate some quantity, we are soon faced with the question of the number of significant digits to carry or to report. A typical rule is to report the mean of a set of observations to one more place and the standard deviation to two more places than the original observation. But this is merely a guideline—which may be wrong. Following DeLury [1958], we can think of two ways in which approximation to the value of a quantity can arise: (1) through arithmetical operations only, or (2) through measurement. If we express the mean of the three numbers 140, 150, and 152 as 147.3, we have approximated the exact mean, $147\frac{1}{3}$, so that there is *rounding error*. This error arises purely as the result of the arithmetical operation of division. The rounding error can be calculated exactly: $147.\overline{3} - 147.3 = 0.0\overline{3}$.

But this is not the complete story. If the above three observations are the weights of three teenage boys measured to the nearest pound, the true average weight can vary all the way from $146.8\overline{3}$ to $147.8\overline{3}$ pounds; that is, the recorded weights (140, 150, 152) could vary from the three lowest values (139.5, 149.5, 151.5) to the three highest values (140.5, 150.5, 152.5), producing the two averages above. This type of rounding can be called *measurement rounding*. Knowledge of the measurement operation is required to assess the extent of the measurement rounding error: If the three numbers above represent systolic blood pressure readings in mmHg expressed to the nearest *even* number, you can verify that the actual arithmetic mean of these three observations can vary from 146.33 to 148.33, so that even the third "significant" digit could be in error.

Unfortunately, we are not quite done yet with assessing the extent of an approximation. If the weights of the three boys are a sample from populations of boys and the population mean is to be estimated, we will also have to deal with *sampling variability* (a second aspect of the measurement process), and the effect of sampling variability is likely to be much larger than the effect of rounding error and measurement roundings. Assessing the extent of sampling variability is discussed in Chapter 4.

For the present time, we give you the following guidelines: When calculating by hand, minimize the number of rounding errors in intermediate arithmetical calculations. So, for example, instead of calculating

$$\sum (y - \overline{y})^2$$

in the process of calculating the standard deviation, use the equivalent relationship

$$\sum y^2 - \frac{(\sum y)^2}{n}$$

You should also note that we are more likely to use approximations with the arithmetical operations of division and the taking of square roots, less likely with addition, multiplication, and subtraction. So if you can sequence the calculations with division and square root being last, rounding errors due to arithmetical calculations will have been minimized. Note that the guidelines for a computer would be quite different. Computers will keep a large number of digits for all intermediate results, and guidelines for minimizing errors depend on keeping the size of the rounding errors small rather than the number of occasions of rounding.

The rule stated above is reasonable. In Chapter 4 you will learn a better way of assessing the extent of approximation in measuring a quantity of interest.

3.5 Degrees of Freedom

The concept of degrees of freedom appears again and again in this book. To make the concept clear, we need the idea of a linear constraint on a set of numbers; this is illustrated by several examples. Consider the numbers of girls, X, and the number of boys, Y, in a family. (Note that X and Y are variables.) The numbers X and Y are free to vary and we say that there are two degrees of freedom associated with these variables. However, suppose that the total number of children in a family, as in the example, is specified to be precisely 8. Then, given that the number of girls is 3, the number of boys is fixed—namely, $8 - 3 = 5$. Given the constraint on the total number of children, the two variables X and Y are no longer both free to vary, but fixing one determines the other. That is, now there is only one degree of freedom. The constraint can be expressed as

$$X + Y = 8 \qquad \text{so that} \quad Y = 8 - X$$

Constraints of this type are called *linear constraints*.

Table 3.12 Frequency Distribution of Form and Color of 556 Garden Peas

	Variable 1: Form		
Variable 2: Color	Round	Wrinkled	Total
Yellow	315	101	416
Green	108	32	140
Total	423	133	556

Source: Data from Mendel [1911].

A second example is based on Mendel's work in plant propagation. Mendel [1911] reported the results of many genetic experiments. One data set related two variables: form and color. Table 3.12 summarizes these characteristics for 556 garden peas. Let A, B, C, and D be the numbers of peas as follows:

	Form	
Color	**Round**	**Wrinkled**
Yellow	A	B
Green	C	D

For example, A is the number of peas that are round and yellow. Without restrictions, the numbers A, B, C and D can be any nonnegative integers: There are four degrees of freedom. Suppose now that the total number of peas is fixed at 556 (as in Table 3.12). That is, $A + B + C + D = 556$. Now only three of the numbers are free to vary. Suppose, in addition, that the number of yellows peas is fixed at 416. Now only two numbers can vary; for example, fixing A determines B, and fixing C determines D. Finally, if the numbers of round peas is also fixed, only one number in the table can be chosen. If, instead of the last constraint on the number of round peas, the number of green peas had been fixed, two degrees would have remained since the constraints "number of yellow peas fixed" and "number of green peas fixed" are not independent, given that the total number of peas is fixed.

These results can be summarized in the following rule: Given a set of N quantities and $M(\leq N)$ linear, independent constraints, the number of degrees of freedom associated with the N quantities is $N - M$. It is often, but not always, the case that degrees of freedom can be defined in the same way for nonlinear constraints.

Calculations of averages will almost always involve the number of degrees of freedom associated with a statistic rather than its number of components. For example, the quantity $\sum(y-\overline{y})^2$ used in calculating the standard deviation of a sample of, say, n values of a variable Y has $n-1$ degrees of freedom associated with it because $\sum(y-\overline{y}) = 0$. That is, the sum of the deviations about the mean is zero.

3.6 Moments

Given a sample of observations $y_1, y_2, \ldots, y_n$ of a variable Y, the rth sample moment about zero, m_r^*, is defined to be

$$m_r^* = \frac{\sum y^r}{n} \qquad \text{for } r = 1, 2, 3, \ldots$$

For example, $m_1^* = \sum y^1/n = \sum y/n = \overline{y}$ is just the arithmetic mean.

The rth sample moment about the mean, m_r, is defined to be

$$m_r = \frac{\sum(y - \overline{y})^r}{n} \qquad \text{for } r = 1, 2, 3, \ldots$$

The value of m_1 is zero (see Problem 3.15). It is clear that m_2 and s^2 (the sample variance) are closely connected. For a large number of observations, m_2 will be approximately equal to s^2. One of the earliest statistical procedures (about 1900) was the *method of moments* of Karl Pearson. The method specified that all estimates derived from a sample should be based on sample moments. Some properties of moments are:

- $m_1 = 0$.
- Odd-numbered moments about the mean of symmetric frequency distributions are equal to zero.
- A unimodal frequency distribution is skewed to the right if the mean is greater than the mode; it is skewed to the left if the mean is less than the mode. For distributions skewed to the right, $m_3 > 0$; for distributions skewed to the left, $m_3 < 0$.

The latter property is used to characterize the *skewness of a distribution*, defined by

$$a_3 = \frac{\sum(y - \overline{y})^3}{[\sum(y - \overline{y})^2]^{3/2}} = \frac{m_3}{(m_2)^{3/2}}$$

The division by $(m_2)^{3/2}$ is to standardize the statistic, which now is unitless. Thus, a set of observations expressed in degrees Fahrenheit will have the same value of a_3 when expressed in degrees Celsius. Values of $a_3 > 0$ indicate positive skewness, skewness to the right, whereas values of $a_3 < 0$ indicate negative skewness. Some typical curves and corresponding values for the skewness statistics are illustrated in Figure 3.10. Note that all but the last two frequency distributions are symmetric; the last figure, with skewness $a_3 = -2.71$, is a mirror image of the penultimate figure, with skewness $a_3 = 2.71$.

The fourth moment about the mean is involved in the characterization of the flatness or peakedness of a distribution, labeled *kurtosis* (degree of archedness); a measure of kurtosis is defined by

$$a_4 = \frac{\sum(y - \overline{y})^4}{[\sum(y - \overline{y})^2]^2} = \frac{m_4}{(m_2)^2}$$

Again, as in the case of a_3, the statistic is unitless. The following terms are used to characterize values of a_4.

$a_4 = 3$	*mesokurtic*: the value for a bell-shaped distribution (Gaussian or normal distribution)
$a_4 < 3$	*leptokurtic*: thin or peaked shape (or “light tails”)
$a_4 > 3$	*platykurtic*: flat shape (or “heavy tails”)

Values of this statistic associated with particular frequency distribution configurations are illustrated in Figure 3.10. The first figure is similar to a bell-shaped curve and has a value $a_4 = 3.03$, very close to 3. Other frequency distributions have values as indicated. It is meaningful to speak of kurtosis only for symmetric distributions.

3.7 Taxonomy of Data

Social scientists have thought hard about types of data. Table 3.13 summarizes a fairly standard taxonomy of data based on the four scales nominal, ordinal, interval, and ratio. This table is to

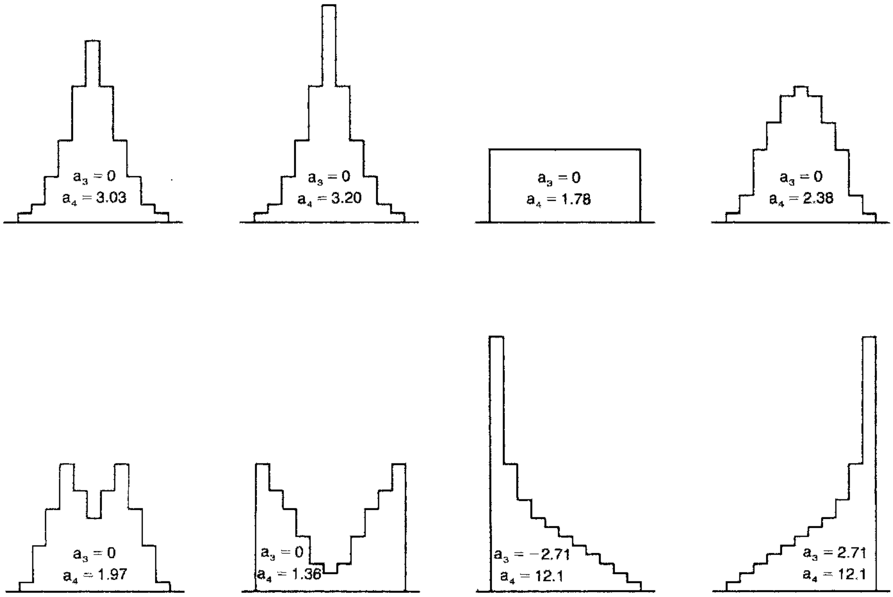


Figure 3.10 Values of skewness (a_3) and kurtosis (a_4) for selected data configurations.

Table 3.13 Standard Taxonomy of Data

Scale	Characteristic Question	Statistic	Statistic to Be Used
Nominal	Do *A* and *B* differ?	List of diseases; marital status	Mode
Ordinal	Is *A* bigger (better) than *B*?	Quality of teaching (unacceptable/acceptable)	Median
Interval	How much do *A* and *B* differ?	Temperatures; dates of birth	Mean
Ratio	How many times is *A* bigger than *B*?	Distances; ages; heights	Mean

be used as a guide only. You can be too rigid in applying this scheme (as unfortunately, some academic journals are). Frequently, ordinal data are coded in increasing numerical order and averages are taken. Or, interval and ratio measurements are ranked (i.e., reduced to ordinal status) and averages taken at that point. Even with nominal data, we sometimes calculate averages. For example: coding male = 0, female = 1 in a class of 100 students, the average is the proportion of females in the class. Most statistical procedures for ordinal data implicitly use a numerical coding scheme, even if this is not made clear to the user. For further discussion, see Luce and Narens [1987], van Belle [2002], and Velleman and Wilkinson [1993].

PROBLEMS

3.1 Characterize the following variables and classify them as qualitative or quantitative. If qualitative, can the variable be ordered? If quantitative, is the variable discrete or continuous? In each case define the values of the variable: (1) race, (2) date of birth, (3) systolic blood pressure, (4) intelligence quotient, (5) Apgar score, (6) white blood count, (7) weight, and (8) quality of medical care.

3.2 For each variable listed in Problem 3.1, define a suitable sample space. For two of the sample spaces so defined, explain how you would draw a sample. What statistics could be used to summarize such a sample?

3.3 Many variables of medical interest are derived from (functions of) several other variables. For example, as a measure of obesity there is the body mass index (BMI), which is given by weight/height^2. Another example is the dose of an anticonvulsant to be administered, usually calculated on the basis of milligram of medicine per kilogram of body weight. What are some assumptions when these types of variables are used? Give two additional examples.

3.4 Every row of 12 observations in Table 3.3 can be summed to form the number of keypunching errors per year of data. Calculate the 13 values for this variable. Make a stem-and-leaf diagram. Calculate the (sample) mean and standard deviation. How do this mean and standard deviation compare with the mean and standard deviation for the number of keypunching errors per line of data?

3.5 The precise specification of the value of a variable is not always easy. Consider the data dealing with keypunching errors in Table 3.3. How is an error defined? A fairly frequent occurrence was the transposition of two digits—for example, a value of "63" might have been entered as "36." Does this represent one or two errors? Sometimes a zero was omitted, changing, for example, 0.0317 to 0.317. Does this represent four errors or one? Consider the list of qualitative variables at the beginning of Section 3.2, and name some problems that you might encounter in defining the values of some of the variables.

3.6 Give three examples of frequency distributions from areas of your own research interest. Be sure to specify (1) what constitutes the sample, (2) the variable of interest, and (3) the frequencies of values or ranges of values of the variables.

3.7 A constant is added to each observation in a set of data (relocation). Describe the effect on the median, lower quartile, range, interquartile range, minimum, mean, variance, and standard deviation. What is the effect on these statistics if each observation is multiplied by a constant (rescaling)? Relocation and rescaling, called linear *transformations*, are frequently used: for example, converting from °C to °F, defined by $°\text{F} = 1.8 \times °\text{C} + 32$. What is the rescaling constant? Give two more examples of rescaling and relocation. An example of nonlinear transformation is going from the radius of a circle to its area: $A = \pi r^2$. Give two more examples of nonlinear transformations.

3.8 Show that the geometric mean is always smaller than the arithmetic mean (unless all the observations are identical). This implies that the mean of the logarithms is not the same as the logarithm of the mean. Is the median of the logarithms equal to the logarithm of the median? What about the interquartile range? How do these results generalize to other nonlinear transformations?

3.9 The data in Table 3.14 deal with the treatment of essential hypertension (*essential* is a technical term meaning that the cause is unknown; a synonym is *idiopathic*). Seventeen patients received treatments C, A, and B, where C = control period, A = propranolol + phenoxybenzamine, and B = propranolol + phenoxybenzamine + hydrochlorothiazide. Each patient received C first, then either A or B, and finally, B or A. The data consist of the systolic blood pressure in the recumbent position. (Note that in this example blood pressures are not always even-numbered.)

Table 3.14 Treatment Data for Hypertension

	C	*A*	*B*		*C*	*A*	*B*
1	185	148	132	10	180	132	136
2	160	128	120	11	176	140	135
3	190	144	118	12	200	165	144
4	192	158	115	13	188	140	115
5	218	152	148	14	200	140	126
6	200	135	134	15	178	135	140
7	210	150	128	16	180	130	130
8	225	165	140	17	150	122	132
9	190	155	138				

Source: Vlachakis and Mendlowitz [1976].

(a) Construct stem-and-leaf diagrams for each of the three treatments. Can you think of some innovative way of displaying the three diagrams together to highlight the data?

(b) Graph as a single graph the ECDFs for each of treatments *C, A*, and *B*.

(c) Construct box plots for each of treatments *C, A*, and *B*. State your conclusions with respect to the systolic blood pressures associated with the three treatments.

(d) Consider the difference between treatments *A* and *B* for each patient. Construct a box plot for the difference. Compare this result with that of part (b).

(e) Calculate the mean and standard deviation for each of the treatments *C, A*, and *B*.

(f) Consider, again, the difference between treatments *A* and *B* for each patient. Calculate the mean and standard deviation for the difference. Relate the mean to the means obtained in part (d). How many standard deviations is the mean away from zero?

3.10 The New York air quality data used in Figure 3.7 are given in the Web appendix to this chapter. Using these data, draw a simple plot of ozone vs. Solar radiation and compare it to conditioning plots where the subsets are defined by temperature, by wind speed, and by both variables together (i.e., one panel would be high temperature and high wind speed). How does the visual impression depend on the number of panels and the conditioning variables?

3.11 Table 3.15 is a frequency distribution of fasting serum insulin (μU/mL) of males and females in a rural population of Jamaican adults. (Serum insulin levels are expressed as whole numbers, so that "7-" represents the values 7 and 8.) The last frequencies are associated with levels greater than 45. Assume that these represent the levels 45 and 46.

(a) Plot both frequency distributions as histograms.

(b) Plot the relative frequency distributions.

(c) Calculate the ECDF.

(d) Construct box plots for males and females. State your conclusions.

(e) Assume that all the observations are concentrated at the midpoints of the intervals. Calculate the mean and standard deviation for males and females.

(f) The distribution is obviously skewed. Transform the levels for males to logarithms and calculate the mean and standard deviation. The transformation can be carried in at least two ways: (1) consider the observations to be centered at the midpoints,

Table 3.15 Frequency Distribution of Fasting Serum Insulin

Fasting Serum Insulin ($\mu U/mL$)	Males	Females	Fasting Serum Insulin ($\mu U/mL$)	Males	Females
7−	1	3	29−	8	14
9−	9	3	31−	8	11
11−	20	9	33−	4	10
13−	32	21	35−	4	8
15−	32	23	37−	3	7
17−	22	39	39−	1	2
19−	23	39	41−	1	3
21−	19	23	43−	1	1
23−	20	27	≥ 45	6	11
25−	13	23			
27−	8	19	Total	235	296

Source: Data from Florey et al. [1977].

transform the midpoints to logarithms, and group into six to eight intervals; and (2) set up six to eight intervals on the logarithmic scale, transform to the original scale, and estimate by interpolation the number of observations in the interval. What type of mean is the antilogarithm of the logarithmic mean? Compare it with the median and arithmetic mean.

3.12 There has been a long-held belief that births occur more frequently in the "small hours of the morning" than at any other time of day. Sutton [1945] collected the time of birth at the King George V Memorial Hospital, Sydney, for 2654 consecutive births. (*Note:* The total number of observations listed is 2650, not 2654 as stated by Sutton.) The frequency of births by hour in a 24-hour day is listed in Table 3.16.

(a) Sutton states that the data "confirmed the belief ... that more births occur in the small hours of the morning than at any other time in the 24 hours." Develop a graphical display that illustrates this point.

(b) Is there evidence of Sutton's statement: "An interesting point emerging was the relatively small number of births during the meal hours of the staff; this suggested either hastening or holding back of the second stage during meal hours"?

Table 3.16 Frequency of Birth by Hour of Birth

Time	Births	Time	Births	Time	Births
6–7 pm	92	2 am	151	10 am	101
7 pm	102	3 am	110	11 am	107
8 pm	100	4 am	144	12 pm	97
9 pm	101	5–6 am	136	1 pm	93
10 pm	127	6–7 am	117	2 pm	100
11 pm	118	7 am	80	3 pm	93
12 am	97	8 am	125	4 pm	131
1 am	136	9 am	87	5–6 pm	105

(c) The data points in fact represent frequencies of values of a variable that has been divided into intervals. What is the variable?

3.13 At the International Health Exhibition in Britain, in 1884, Francis Galton, a scientist with strong statistical interests, obtained data on the strength of pull. His data for 519 males aged 23 to 26 are listed in Table 3.17. Assume that the smallest and largest categories are spread uniformly over a 10-pound interval.

Table 3.17 Strength of Pull

Pull Strength (lb)	Cases Observed	Pull Strength (lb)	Cases Observed
Under 50	10	Under 90	113
Under 60	42	Under 100	22
Under 70	140	Above 100	24
Under 80	168		
		Total	519

(a) The description of the data is exactly as in Galton [1889]. What are the intervals, assuming that strength of pull is measured to the nearest pound?

(b) Calculate the median and 25th and 75th percentiles.

(c) Graph the ECDF.

(d) Calculate the mean and standard deviation assuming that the observations are centered at the midpoints of the intervals.

(e) Calculate the proportion of observations within one standard deviation of the mean.

3.14 The aflatoxin data cited at the beginning of Section 3.2 were taken from a larger set in the paper by Quesenberry et al. [1976]. The authors state:

> Aflatoxin is a toxic material that can be produced in peanuts by the fungus *Aspergillus flavus*. As a precautionary measure all commercial lots of peanuts in the United States (approximately 20,000 each crop year) are tested for aflatoxin.... Because aflatoxin is often highly concentrated in a small percentage of the kernels, variation among aflatoxin determinations is large.... Estimation of the distribution (of levels) is important. ... About 6200g of raw peanut kernels contaminated with aflatoxin were comminuted (ground up). The ground meal was then divided into 11 subsamples (lots) weighing approximately 560g each. Each subsample was blended with 2800ml methanol-water-hexane solution for two minutes, and the homogenate divided equally among 16 centrifuge bottles. One observation was lost from each of three subsamples leaving eight subsamples with 16 determinations and three subsamples with 15 determinations.

The original data were given to two decimal places; they are shown in Table 3.18 rounded off to the nearest whole number. The data are listed by lot number, with asterisks indicating lost observations.

(a) Make stem-and-leaf diagrams of the data of lots 1, 2, and 10. Make box plots and histograms for these three lots, and discuss differences among these lots with respect to location and spread.

(b) The data are analyzed by means of a MINITAB computer program. The data are entered by columns and the command DESCRIBE is used to give standard

Table 3.18 Aflatoxin Data by Lot Number

1	2	3	4	5	6	7	8	9	10	11
121	95	20	22	30	11	29	34	17	8	53
72	56	20	33	26	19	33	28	18	6	113
118	72	25	23	26	13	37	35	11	7	70
91	59	22	68	36	13	25	33	12	5	100
105	115	25	28	48	12	25	32	25	7	87
151	42	21	27	50	17	36	29	20	7	83
125	99	19	29	16	13	49	32	17	12	83
84	54	24	29	31	18	38	33	9	8	65
138	90	24	52	22	18	29	31	15	9	74
83	92	20	29	27	17	29	32	21	14	112
117	67	12	22	23	16	32	29	17	13	98
91	92	24	29	35	14	40	26	19	11	85
101	100	15	37	52	11	36	37	23	5	82
75	77	15	41	28	15	31	28	17	7	95
137	92	23	24	37	16	32	31	15	4	60
146	66	22	36	*	12	*	32	17	12	*

Table 3.19 MINITAB Analysis of Aflatoxin Data[a]

```
MTB > desc c1–c11
       N     N*     MEAN     MEDIAN     STDEV     MIN     MAX      Q1      Q3
C1     16    0     109.69    111.00     25.62     72      151     85.75   134.00
C2     16    0      79.25     83.50     20.51     42      115     60.75    94.25
C3     16    0      20.687    21.500     3.860    12       25     19.25    24.00
C4     16    0      33.06     29.00     12.17     22       68     24.75    36.75
C5     15    1      32.47     30.00     10.63     16       52     26.00    37.00
C6     16    0      14.688    14.500     2.651    11       19     12.25    17.00
C7     15    1      33.40     32.00      6.23     25       49     29.00    37.00
C8     16    0      31.375    32.000     2.849    26       37     29.00    33.00
C9     16    0      17.06     17.00      4.19      9       25     15.00    19.75
C10    16    0       8.438     7.500     3.076     4       14      6.25    11.75
C11    15    1      84.00     83.00     17.74     53      113     70.00    98.00
```

[a] N*, number of missing observations; Q1 and Q3, 25th and 75th percentiles, respectively.

descriptive statistics for each lot. The output from the program (slightly modified) is given in Table 3.19.

(c) Verify that the statistics for lot 1 are correct in the printout.

(d) There is an interesting pattern between the means and their standard deviations. Make a plot of the means vs. standard deviation. Describe the pattern.

(e) One way of describing the pattern between means and standard deviations is to calculate the ratio of the standard deviation to the mean. This ratio is called the *coefficient of variation*. It is usually multiplied by 100 and expressed as the percent coefficient of variation. Calculate the coefficients of variation in percentages for each of the 11 lots, and make a plot of their value with the associated means. Do you see any pattern now? Verify that the average of the coefficients of variation is about 24%. A reasonable number to keep in mind for many biological measurements is that the variability as measured by the standard deviation is about 30% of the mean.

Table 3.20 Plasma Prostaglandin E Levels

Patient Number	Mean Plasma iPGE (pg/mL)	Mean Serum Calcium (ml/dL)
	Patients with Hypercalcemia	
1	500	13.3
2	500	11.2
3	301	13.4
4	272	11.5
5	226	11.4
6	183	11.6
7	183	11.7
8	177	12.1
9	136	12.5
10	118	12.2
11	60	18.0
	Patients without Hypercalcemia	
12	254	10.1
13	172	9.4
14	168	9.3
15	150	8.6
16	148	10.5
17	144	10.3
18	130	10.5
19	121	10.2
20	100	9.7
21	88	9.2

3.15 A paper by Robertson et al. [1976] discusses the level of plasma prostaglandin E (iPGE) in patients with cancer with and without hypercalcemia. The data are given in Table 3.20. Note that the variables are the mean plasma iPGE and mean serum Ca levels—presumably, more than one assay was carried out for each patient's level. The number of such tests for each patient is not indicated, nor is the criterion for the number.

(a) Calculate the mean and standard deviation of plasma iPGE level for patients with hypercalcemia; do the same for patients without hypercalcemia.

(b) Make box plots for plasma iPGE levels for each group. Can you draw any conclusions from these plots? Do they suggest that the two groups differ in plasma iPGE levels?

(c) The article states that normal limits for serum calcium levels are 8.5 to 10.5 mg/dL. It is clear that patients were classified as hypercalcemic if their serum calcium levels exceeded 10.5 mg/dL. Without classifying patients it may be postulated that high plasma iPGE levels tend to be associated with high serum calcium levels. Make a plot of the plasma iPGE and serum calcium levels to determine if there is a suggestion of a pattern relating these two variables.

3.16 Prove or verify the following for the observations $y_1, y_2, \ldots, y_n$.

(a) $\sum 2y = 2\sum y$.

(b) $\sum(y - \overline{y}) = 0$.

(c) By means of an example, show that $\sum y^2 \neq (\sum y)^2$.

(d) If a is a constant, $\sum ay = a \sum y$.

(e) If a is a constant, $\sum(a + y) = na + \sum y$.

(f) $\sum(y/n) = (1/n) \sum y$.

(g) $\sum(a + y)^2 = na^2 + 2a \sum y + \sum y^2$.

(h) $\sum(y - \overline{y})^2 = \sum y^2 - (\sum y)^2/n$.

(i) $\sum(y - \overline{y})^2 = \sum y^2 - n\overline{y}^2$.

3.17 A variable Y is grouped into intervals of width h and represented by the midpoint of the interval. What is the maximum error possible in calculating the mean of all the observations?

3.18 Prove that the two definitions of the geometric mean are equivalent.

3.19 Calculate the average number of boys per family of eight children for the data given in Table 3.10.

3.20 The formula $\overline{Y} = \sum py$ is also valid for observations not arranged in a frequency distribution as follows: If we let $1/N = p$, we get back to the formula $\overline{Y} = \sum py$. Show that this is so for the following four observations: 3, 9, 1, 7.

3.21 Calculate the average systolic blood pressure of native Japanese men using the frequency data of Table 3.6. Verify that the same value is obtained using the relative frequency data of Table 3.7.

3.22 Using the taxonomy of data described in Note 3.6, classify each of the variables in Problem 3.1 according to the scheme described in the note.

REFERENCES

Cleveland, W. S. [1981]. LOWESS: a program for smoothing scatterplots by robust locally weighted regression. *American Statistician*, **35**: 54.

Cleveland, W. S. [1993]. *Visualizing Data*. Hobart Press, Summit, NJ.

Cleveland, W. S. [1994]. *The Elements of Graphing Data*. Hobart Press, Summit, NJ.

DeLury, D. B. [1958]. Computations with approximate numbers. *Mathematics Teacher*, **51**: 521–530. Reprinted in Ku, H. H. (ed.) [1969]. *Precision Measurement and Calibration*. NBS Special Publication 300. U.S. Government Printing Office, Washington, DC.

Fisher, R. A. [1958]. *Statistical Methods for Research Workers*, 13th ed. Oliver & Boyd, London.

Fleming, T. R. and Harrington, D. P. [1991]. *Counting Processes and Survival Analysis*. John Wiley & Sons, New York.

Florey, C. du V., Milner, R. D. G., and Miall, W. E. [1977]. Serum insulin and blood sugar levels in a rural population of Jamaican adults. *Journal of Chronic Diseases*, **30**: 49–60. Used with permission from Pergamon Press, Inc.

Galton, F. [1889]. *Natural Inheritance*. Macmillan, London.

Gould, S. J. [1996]. *Full House: The Spread of Excellence from Plato to Darwin*. Harmony Books, New York.

Graunt, J. [1662]. Natural and political observations mentioned in a following index and made upon the Bills of Mortality. In Newman, J. R. (ed.) [1956]. *The World of Mathematics*, Vol. 3. Simon & Schuster, New York, pp. 1421–1435.

Huff, D. [1993]. *How to Lie with Statistics*. W. W. Norton, New York.

Luce, R. D. and Narens, L. [1987]. Measurement scales on the continuum. *Science*, **236**: 1527–1532.

Mendel, G. [1911]. *Versuche über Pflanzenhybriden*. Wilhelm Engelmann, Leipzig, p. 18.

Moses, L. E. [1987]. Graphical methods in statistical analysis. *Annual Reviews of Public Health*, **8**: 309–353.

Newman, J. R. (ed.) [1956]. *The World of Mathematics*, Vol. 3. Simon & Schuster, New York, pp. 1421–1435.

Quesenberry, P. D., Whitaker, T. B., and Dickens, J. W. [1976]. On testing normality using several samples: an analysis of peanut aflatoxin data. *Biometrics*, **32**: 753–759. With permission of the Biometric Society.

R Foundation for Statistical Computing [2002]. *R, Version 1.7.0*, Air quality data set. *http://cran.r-project.org*.

Robertson, R. P., Baylink, D. J., Metz, S. A., and Cummings, K. B. [1976]. Plasma prostaglandin E in patients with cancer with and without hypercalcemia. *Journal of Clinical Endocrinology and Metabolism*, **43**: 1330–1335.

Schwab, B. [1975]. Delivery of babies and full moon (letter to the editor). *Canadian Medical Association Journal*, **113**: 489, 493.

Sutton, D. H. [1945]. Gestation period. *Medical Journal of Australia*, Vol. I, **32**: 611–613. Used with permission.

Tufte, E. R. [1990]. *Envisioning Information*. Graphics Press, Cheshire, CT.

Tufte, E. R. [1997]. *Visual Explanations*. Graphics Press, Cheshire, CT.

Tufte, E. R. [2001]. *The Visual Display of Quantitative Information*. 2nd ed. Graphics Press, Cheshire, CT.

Tukey, J. W. [1977]. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

van Belle, G. [2002]. *Statistical Rules of Thumb*. Wiley, New York.

Velleman, P. F. and Wilkinson, L. [1993]. Nominal, ordinal, interval, and ratio typologies are misleading. *American Statistician* **46**: 193–197.

Vlachakis, N. D. and Mendlowitz, M. [1976]. Alpha- and beta-adrenergic receptor blocking agents combined with a diuretic in the treatment of essential hypertension. *Journal of Clinical Pharmacology*, **16**: 352–360.

Wilkinson, L. [1999]. *The Grammar of Graphics*. Springer, New York.

Winkelstein, W., Jr., Kagan, A., Kato, H., and Sacks, S. T. [1975]. Epidemiological studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii and California: blood pressure distributions. *American Journal of Epidemiology*, **102**: 502–513.

CHAPTER 4

Statistical Inference: Populations and Samples

4.1 INTRODUCTION

Statistical inference has been defined as “the attempt to reach a conclusion concerning all members of a class from observations of only some of them” [Runes, 1959]. In statistics, “all members of a class” form the *population* or *sample space*, and the subset observed forms a *sample*; we discussed this in Sections 3.1 and 3.2. We now discuss the *process* of obtaining a valid sample from a population; specifically, when is it valid to make a statement about a population on the basis of a sample? One of the assumptions in any scientific investigation is that valid inferences can be made—that the results of a study can apply to a larger population. For example, we can assume that a new therapy developed at the Memorial Sloan–Kettering Cancer Center in New York is applicable to cancer patients in Great Britain. You can easily supply additional examples.

In the next section we note which characteristics of a population are of interest and illustrate this with two examples. In Section 4.3 we introduce probability theory as a way by which we can define valid sampling procedures. In Section 4.4 we apply the theory to a well-known statistical model for a population, the normal frequency distribution, which has practical as well as theoretical interest. One reason for the importance of the normal distribution is given in Section 4.5, which discusses the concept of sampling distribution. In the next three sections we discuss inferences about population means and variances on the basis of a single sample.

4.2 POPULATION AND SAMPLE

4.2.1 Definition and Examples

You should review Chapter 3 for the concepts of *variable*, *sample space* or *population*, and *statistic*.

Definition 4.1. A *parameter* is a numerical characteristic of a population.

Analogous to numerical characteristics of a sample (statistics), we will be interested in numerical characteristics of populations (parameters). The population characteristics are usually unknown because the entire population cannot be enumerated or studied. The problem of

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

statistical inference can then be stated as follows: On the basis of a sample from a population, what can be said about the population from which the sample came? In this section we illustrate the four concepts of population and its corresponding parameters, and sample and its corresponding statistics.

Example 4.1. We illustrate those four concepts with an example from Chapter 3, systolic blood pressure for Japanese men, aged 45–69, living in Japan. The "population" can be considered to be the collection of blood pressures of all Japanese men. The blood pressures are assumed to have been taken under standardized conditions. Clearly, Winkelstein et al. [1975] could not possibly measure all Japanese men, but a subset of 2232 eligible men were chosen. This is the sample. A numerical quantity of interest could be the average systolic blood pressure. This average for the population is a *parameter*; the average for the sample is the *statistic*. Since the total population cannot be measured, the parameter value is unknown. The statistic, the average for the sample, can be calculated. You are probably assuming now that the sample average is a good estimate of the population average. You may be correct. Later in this chapter we specify under what conditions this is true, but note for now that all the elements of inference are present.

Example 4.2. Consider this experimental situation. We want to assess the effectiveness of a new special diet for children with phenylketonuria (PKU). One effect of this condition is that untreated children become mentally retarded. The diet is used with a set of PKU children and their IQs are measured when they reach 4 years of age. What is the population? It is hypothetical in this case: all PKU children who could potentially be treated with the new diet. The variable of interest is the IQ associated with each child. The sample is the set of children actually treated. A parameter could be the median IQ of the hypothetical population; a statistic might be the median IQ of the children in the sample. The question to be answered is whether the median IQ of this treated hypothetical population is the same or comparable to that of non-PKU children.

A sampling situation has the following components: A population of measurement is specified, a sample is taken from the population, and measurements are made. A statistic is calculated which—in some way—makes a statement about the corresponding population parameter. Some practical questions that come up are:

1. Is the population defined unambiguously?
2. Is the variable clearly observable?
3. Is the sample "valid"?
4. Is the sample "big enough"?

The first two questions have been discussed in previous chapters. In this chapter we begin to answer the last two.

Conventionally, parameters are indicated by Greek letters and the estimate of the parameter by the corresponding Roman letter. For example, μ is the population mean, and m is the sample mean. Similarly, the population standard deviation will be indicated by σ and the corresponding sample estimate by s.

4.2.2 Estimation and Hypothesis Testing

Two approaches are commonly used in making statements about population parameters: estimation and hypothesis testing. *Estimation*, as the name suggests, attempts to estimate values of parameters. As discussed before, the sample mean is thought to estimate, in some way, the mean of the population from which the sample was drawn. In Example 4.1 the mean of the blood pressures is considered an estimate of the corresponding population value. *Hypothesis testing* makes inferences about (population) parameters by supposing that they have certain values, and then testing whether the data observed are consistent with the hypothesis. Example 4.2 illustrates this framework: Is the mean IQ of the population of PKU children treated with the special diet the same as that of the population of non-PKU children? We could hypothesize that it is and determine, in some way, whether the data are inconsistent with this hypothesis.

You could argue that in the second example we are also dealing with estimation. If one could estimate the mean IQ of the treated population, the hypothesis could be dealt with. This is quite true. In Section 4.7 we will see that in many instances hypothesis testing and estimation are but two sides of the same coin.

One additional comment about estimation: A distinction is usually made between point estimate and interval estimate. A sample mean is a *point estimate*. An *interval estimate* is a range of values that is reasonably certain to straddle the value of the parameter of interest.

4.3 VALID INFERENCE THROUGH PROBABILITY THEORY

4.3.1 Precise Specification of Our Ignorance

Everyone "knows" that the probability of heads coming up in the toss of a coin is 1/2 and that the probability of a 3 in the toss of a die is 1/6. More subtly, the probability that a randomly selected patient has systolic blood pressure less than the population median is 1/2, although some may claim, after the measurement is made, that it is either 0 or 1—that is, the systolic blood pressure of the patient is either below the median or greater than or equal to the median.

What do we mean by the phrase "the probability of"? Consider one more situation. We toss a thumbtack on a hard, smooth surface such as a table, if the outcome is ⊥, we call it "up"; if the outcome is ⊤, we call it "down." What is the probability of "up"? It is clear that in this example we do not know, a priori, the probability of "up"—it depends on the physical characteristics of the thumbtack. How would you *estimate* the probability of "up"? Intuitively, you would toss the thumbtack a large number of times and observe the proportion of times the thumbtack landed "up"—and that is the way we define probability. Mathematically, we define the probability of "up" as the relative frequency of the occurrence of "up" as the number of tosses become indefinitely large. This is an illustration of the *relative frequency* concept of probability. Some of its ingredients are: (1) a trial or experiment has a set of specified outcomes; (2) the outcome of one trial does not influence the outcome of another trial; (3) the trials are identical; and (4) the probability of a specified outcome is the limit of its relative frequency of occurrence as the number of trials becomes indefinitely large.

Probabilities provide a link between a population and samples. A *probability* can be thought of as a numerical statement about what we know and do not know: a precise specification of our ignorance [Fisher, 1956]. In the thumbtack-tossing experiment, we know that the relative frequency of occurrences of "up" will approach some number: the probability of "up." What we do not know is what the outcome will be on the next toss. A probability, then, is a characteristic of a population of outcomes. When we say that the probability of a head in a coin toss is 1/2, we are making a statement about a population of tosses. For alternative interpretations of probability, see Note 4.1. On the basis of the relative frequency interpretation of probability, we deduce that probabilities are numbers between zero and 1 (including zero and 1).

The outcome of a trial such as a coin toss will be denoted by a capital letter; for example, $H =$ "coin toss results in head" and $T =$ "coin toss results in tail." Frequently, the letter can be chosen as a mnemonic for the outcome. The probability of an outcome, O, in a trial will be denoted by $P[O]$. Thus, in the coin-tossing experiment, we have $P[H]$ and $P[T]$ for the probabilities of "head" and "tail," respectively.

4.3.2 Working with Probabilities

Outcomes of trials can be categorized by two criteria: statistical independence and mutual exclusiveness.

Definition 4.2. Two outcomes are *statistically independent* if the probability of their joint occurrence is the product of the probabilities of occurrence of each outcome.

Using notation, let C be one outcome and D be another outcome; $P[C]$ is the probability of occurrence of C, and $P[D]$ is the probability of occurrence of D. Then C and D are statistically independent if

$$P[CD] = P[C]P[D]$$

where $[CD]$ means that both C and D occur.

Statistically independent events are the model for events that "have nothing to do with each other." In other words, the occurrence of one event does not change the probability of the other occurring. Later this is explained in more detail.

Models of independent outcomes are the outcomes of successive tosses of a coin, die, or the spinning of a roulette wheel. For example, suppose that the outcomes of two tosses of a coin are statistically independent. Then the probability of two heads, $P[HH]$, by statistical independence is

$$P[HH] = P[H]P[H] = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

Similarly,

$$P[HT] = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

$$P[TH] = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

and

$$P[TT] = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

Note that the outcome HT means "head on toss 1 and tail on toss 2."

You may wonder why we refer to coin tossing and dice throws so much. One reason has been given already: These activities form patterns of probabilistic situations. Second, they can be models for many experimental situations. Suppose that we consider the Winkelstein et al. [1975] study dealing with blood pressures of Japanese men. What is the probability that each of two men has a blood pressure less than the median of the population? We can use the coin-toss model: By definition, half of the population has blood pressure less than the median. The populations can then be thought of as a very large collection of trials each of which has two outcomes: less than the median, and greater than or equal to the median. If the selection of two men can be modeled by the coin-tossing experiment, the probability that both men have blood pressures less than the median is $1/2 \times 1/2 = 1/4$. We now formalize this:

Definition 4.3. Outcomes of a series of repetitions of a trial are a *random sample* of outcomes if the probability of their joint occurrence is the product of the probabilities of each occurring separately. If every possible sample of k outcomes has the same probability of occurrence, the sample is called a *simple random sample*. This is the most common type of random sample.

Suppose that we are dealing with the outcomes of trials. We label the outcomes O_k, where the subscript is used to denote the order in the sequence; O_1 is the outcome specified for the first trial, O_2 is the outcome for the second trial, and so on. Then the outcomes form a random sample if

$$P[O_1O_2O_3 \cdots O_k] = P[O_1]P[O_2]P[O_3] \cdots P[O_k].$$

The phrase "a random sample" is therefore not so much a statement about the sample as a statement about the method that produced the sample. The randomness of the sample allows us to make valid statements about the population from which it came. It also allows us to quantify what we know and do not know. (See Note 4.6 for another type of random sampling.)

How can we draw a random sample? For the coin tosses and dice throws, this is fairly obvious. But how do we draw a random sample of Japanese men? Theoretically, we could have their names on slips of paper in a very large barrel. The contents are stirred and slips of paper drawn out—a random sample. Clearly, this is not done in practice. In fact, often, a sample is claimed to be random by default: "There is no reason to believe that it is not random." Thus, college students taking part in a experiment are implicitly assumed to be a "random sample of people." Sometimes this is reasonable; as mentioned earlier, cancer patients treated in New York are considered very similar with respect to cancer to cancer patients in California. There is a gradation in the seriousness of nonrandomness of samples: "Red blood cells from healthy adult volunteers" are apt to be similar in many respects the world over (and dissimilar in others); "diets of teenagers," on the other hand, will vary from region to region.

Obtaining a truly random sample is a difficult task that is rarely carried out successfully. A standard criticism of any study is that the sample of data is not a random sample, so that the inference is not valid. Some problems in sampling were discussed in Chapter 2; here we list a few additional problems:

1. The population or sample space is not defined.
2. Part of the population of interest is not available for study.
3. The population is not identifiable or it changes with time.
4. The sampling procedure is faulty, due to limitations in time, money, and effort.
5. Random allocation of members of a group to two or more treatments does not imply that the group itself is necessarily a random sample.

Most of these problems are present in any study, sometimes in an unexpected way. For example, in an experiment involving rats, the animals were "haphazardly" drawn from a cage for assignment to one treatment, and the remaining rats were given another treatment. "Differences" between the treatments were due to the fact that the more agile and larger animals evaded "haphazard" selection and wound up in the second treatment. For some practical ways of drawing random samples, see Note 4.9.

Now we consider probabilities of mutually exclusive events:

Definition 4.4. Two outcomes are *mutually exclusive* if at most one of them can occur at a time; that is, the outcomes do not overlap.

Using notation, let C be one outcome and D another; then it can be shown (using the relative frequency definition) that $P[C \text{ or } D] = P[C] + P[D]$ if the outcomes are mutually exclusive. Here, the connective "or" is used in its inclusive sense, "either/or, or both."

Some examples of mutually exclusive outcomes are H and T on a coin toss; the race of a person for purposes of a study can be defined as "black," "white," or "other," and each subject can belong to only one category; the method of delivery can be either "vaginal" or by means of a "cesarean section."

Example 4.3. We now illustrate outcomes that are not mutually exclusive. Suppose that the Japanese men in the Winkelstein data are categorized by weight: "reasonable weight" or "overweight," and their blood pressures by "normal" or "high." Suppose that we have the following table:

	Blood Pressure		
Weight	**Normal (N)**	**High (H)**	
Reasonable (R)	0.6	0.1	0.7
Overweight (O)	0.2	0.1	0.3
Total	0.8	0.2	1.0

The entries in the table are the probabilities of outcomes for a person selected randomly from the population, so that, for example, 20% of Japanese men are considered overweight and have normal blood pressure. Consider the outcomes "overweight" and "high blood pressure." What is the probability of the outcome $[O$ or $H]$ (overweight, high blood pressure, or both)? This corresponds to the following data in boldface type:

	N	***H***	
R	0.6	**0.1**	0.7
O	**0.2**	**0.1**	0.3
Total	0.8	0.2	1.0

$$P[O \text{ or } H] = 0.2 + 0.1 + 0.1 = 0.4$$

But $P[O] + P[H] = 0.2 + 0.3 = 0.5$. Hence, O and H are not mutually exclusive. In terms of calculation, we see that we have added in the outcome $P[OH]$ twice:

	N	***H***	
R		0.1	
O	0.2	0.1	0.3
Total		0.2	

The correct value is obtained if we subtract $P[OH]$ as follows:

$$\begin{aligned} P[O \text{ or } H] &= P[O] + P[H] - P[OH] \\ &= 0.3 + 0.2 - 0.1 \\ &= 0.4 \end{aligned}$$

This example is an illustration of the addition rule of probabilities.

Definition 4.5. By the *addition rule*, for any two outcomes, the probability of occurrence of either outcome or both is the sum of the probabilities of each occurring minus the probability of their joint occurrence.

Using notation, for any two outcomes C and D,

$$P[C \text{ or } D] = P[C] + [D] - P[CD]$$

Two outcomes, C and D, are mutually exclusive if they cannot occur together. In this case, $P[CD] = 0$ and $P[C \text{ or } D] = P[C] + P[D]$, as stated previously.

We conclude this section by briefly discussing dependent outcomes. The outcomes O and H in Example 4.3 were not mutually exclusive. Were they independent? By Definition 4.2, O and H are statistically independent if $P[OH] = P[O]P[H]$.

From the table, we get $P[OH] = 0.1$, $P[O] = 0.3$, and $P[H] = 0.2$, so that

$$0.1 \neq (0.3)(0.2)$$

Of subjects with reasonable weight, only 1 in 7 has high blood pressure, but among overweight persons, 1 in 3 has high blood pressure. Thus, the probability of high blood pressure in overweight subjects is greater than the probability of high blood pressure in subjects of normal weight. The reverse statement can also be made: 2 of 8 persons with normal blood pressure are overweight; 1 of 2 persons with high blood pressure is overweight.

The statement "of subjects with reasonable weight, only 1 in 7 has high blood pressure" can be stated as a probability: "The probability that a person with reasonable weight has high blood pressure is 1/7." Formally, this is written as

$$P[H|R] = \frac{1}{7}$$

or P[high blood pressure *given* a reasonable weight] $= 1/7$. The probability $P[H|R]$ is called a *conditional* probability. You can verify that $P[H|R] = P[HR]/P[R]$.

Definition 4.6. For any two outcomes C and D, the *conditional probability* of the occurrence of C *given* the occurrence of D, $P[C|D]$, is given by

$$P[C|D] = \frac{P[CD]}{P[D]}$$

For completeness we now state the multiplication rule of probability (which is discussed in more detail in Chapter 6).

Definition 4.7. By the *multiplication rule*, for any two outcomes C and D, the probability of the joint occurrence of C and D, $P[CD]$, is given by

$$P[CD] = P[C]P[D|C]$$

or equivalently,

$$P[CD] = P[D]P[C|D]$$

Example 4.3. [continued] What is the probability that a randomly selected person is overweight and has high blood pressure? In our notation we want $P[OH]$. By the multiplication rule, this probability is

$$P[OH] = P[O]P[H|O]$$

Using Definition 4.6 gives us

$$P[H|O] = \frac{P[OH]}{P[O]} = \frac{0.1}{0.3} = \frac{1}{3}$$

so that

$$P[OH] = 0.3\left(\frac{1}{3}\right) = 0.1$$

Alternatively, we could have calculated $P[OH]$ by

$$P[OH] = P[H]P[O|H]$$

which becomes

$$P[OH] = 0.2\left(\frac{0.1}{0.2}\right) = 0.1$$

We can also state the criterion for statistical independence in terms of conditional probabilities. From Definition 4.2, two outcomes C and D are statistically independent if $P[CD] = P[C]P[D]$ (i.e., the probability of the joint occurrence of C and D is the product of the probability of C and the probability of D). The multiplication rule states that for *any* two outcomes C and D,

$$P[CD] = P[C]P[D|C]$$

Under independence,

$$P[CD] = P[C]P[D]$$

Combining the two, we see that C and D are independent if (and only if) $P[D|C] = P[D]$. In other words, the probability of occurrence of D is not altered by the occurrence of C. This has intuitive appeal.

When do we use the addition rule; when the multiplication rule? Use the addition rule to calculate the probability that either one or both events occur. Use the multiplication rule to calculate the probability of the joint occurrence of two events.

4.3.3 Random Variables and Distributions

Basic to the field of statistics is the concept of a random variable:

Definition 4.8. A *random variable* is a variable associated with a random sample.

The only difference between a *variable* defined in Chapter 3 and a *random variable* is the process that generates the value of the variable. If this process is random, we speak of a random variable. All the examples of variables in Chapter 3 can be interpreted in terms of random variables if the samples are random samples. The empirical relative frequency of occurrence of a value of the variable becomes an estimate of the probability of occurrence of that value. For example, the relative frequencies of the values of the variable "number of boys in families with eight children" in Table 3.12 become estimates of the probabilities of occurrence of these values.

The distinction between discrete and continuous variables carries over to random variables. Also, as with variables, we denote the label of a random variable by capital letters (say $X, Y, V, \ldots$) and a value of the random variable by the corresponding lowercase letter ($x, y, v, \ldots$).

We are interested in describing the probabilities with which values of a random variable occur. For discrete random variables, this is straightforward. For example, let Y be the outcome of the toss of a die. Then Y can take on the values 1, 2, 3, 4, 5, 6, and we write

$$P[Y = 1] = \frac{1}{6}, \quad P[Y = 2] = \frac{1}{6}, \ldots, \quad P[Y = 6] = \frac{1}{6}$$

This leads to the following definition:

Definition 4.9. A *probability function* is a function that for each possible value of a discrete random variable takes on the probability of that value occurring. The function is usually presented as a listing of the values with the probabilities of occurrence of the values.

Consider again the data of Table 3.12, the number of boys in families with eight children. The observed empirical relative frequencies can be considered estimates of probabilities if the 53,680 families are a random sample. The probability distribution is then estimated as shown in Table 4.1. The estimated probability of observing precisely two boys in a family of eight children is 0.0993 or, approximately, 1 in 10. Since the sample is very large, we will treat—in this discussion—the estimated probabilities as if they were the actual probabilities. If Y represents the number of boys in a family with eight children, we write

$$P[Y = 2] = 0.0993$$

What is the probability of two boys or fewer? This can be expressed as

$$P[Y \leq 2] = P[Y = 2 \text{ or } Y = 1 \text{ or } Y = 0]$$

Since these are mutually exclusive outcomes,

$$\begin{aligned} P[Y \leq 2] &= P[Y = 2] + P[Y = 1] + P[Y = 0] \\ &= 0.0993 + 0.0277 + 0.0040 \\ &= 0.1310 \end{aligned}$$

Approximately 13% of families with eight children will have two or fewer boys. A probability function can be represented graphically by a plot of the values of the variable against the probability of the value. The probability function for the Geissler data is presented in Figure 4.1.

How can we describe probabilities associated with continuous random variables? Somewhat paradoxically, the probability of a specified value for a continuous random variable is zero! For example, the probability of finding anyone with height 63.141592654 inches—and not 63.141592653 inches—is virtually zero. If we were to continue the decimal expansion, the probability becomes smaller yet. But we do find people with height, say, 63 inches. When we write 63 inches, however, we do not mean 63.000... inches (and we are almost certain not to find anybody with that height), but we have in mind *an interval* of values of height, anyone with height between 62.500... and 63.500... inches. We could then divide the values of the continuous random variable into intervals, treat the midpoints of the intervals as the values of a discrete variable, and list the probabilities associated with these values. Table 3.7 illustrates this approach with the division of the systolic blood pressure of Japanese men into discrete intervals.

We start with the histogram and the relative frequencies associated with the intervals of values in the histogram. The area under the "curve" is equal to 1 if the width of each interval

Table 4.1 Number of Boys in Eight-Child Families

Number of Boys	Probability	Number of Boys	Probability
0	0.0040	6	0.1244
1	0.0277	7	0.0390
2	0.0993	8	0.0064
3	0.1984		
4	0.2787		
5	0.2222	Total	1.0000

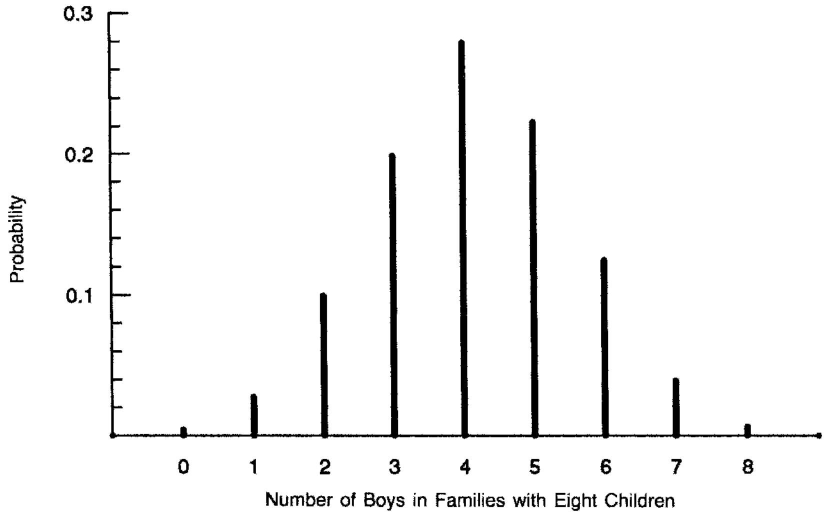


Figure 4.1 Probability function of the random variable "number of boys in families with eight children." (Geissler's date; reprinted in Fisher [1958]; see Table 3.10.)

is 1; or if we normalize (i.e., multiply by a constant so that the area is equal to 1). Suppose now that the interval widths are made smaller and smaller, and simultaneously, the number of cases increased. Normalize so that the area under the curve remains equal to 1; then the curve is assumed to take on a smooth shape. Such shapes are called *probability density functions* or, more briefly, *densities*:

Definition 4.10. *A probability density function* is a curve that specifies, by means of the area under the curve over an interval, the probability that a continuous random variable falls within the interval. The total area under the curve is 1.

Some simple densities are illustrated in Figure 4.2. Figure 4.2(*a*) and (*b*) represent uniform densities on the intervals $(-1, 1)$ and (0, 1), respectively. Figure 4.2(*c*) illustrates a triangular

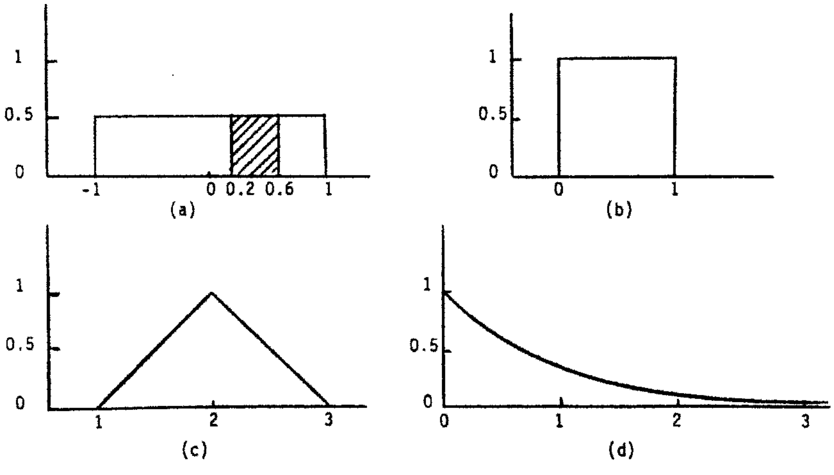


Figure 4.2 Examples of probability density functions. In each case, the area under the curve is equal to 1.

density, and Figure 4.2(*d*) an exponential density. The latter curve is defined over the entire positive axis. (It requires calculus to show that the area under this curve is 1.) The probability that a continuous random variable takes on a value in a specified interval is equal to the area over the interval. For example, the probability that the random variable in Figure 4.2(*a*) falls in the interval 0.2–0.6 is equal to the area over the interval. This is, $(0.6 - 0.2)(0.5) = 0.20$, so that we expect 20% of values of this random variable to fall in this interval. One of the most important probability density function is the normal distribution; it is discussed in detail in Section 4.4.

How can we talk about a random sample of observations of a continuous variable? The simplest way is to consider the drawing of an observation as a trial and the probability of observing an arbitrary (but specified) value or smaller of the random variable. Definition 4.3 can then be applied.

Before turning to the normal distribution, we introduce the concept of averages of random variables. In Section 3.4.2, we discussed the average of a discrete variable based on the empirical relative frequency distribution. The average of a discrete variable Y with values $y_1, y_2, \ldots, y_k$ occurring with relative frequencies $p_1, p_2, \ldots, p_k$, respectively, was shown to be

$$\overline{y} = \sum py$$

(We omit the subscripts since it is clear that we are summing over all the values.) Now, if Y is a *random* variable and $p_1, p_2, \ldots, p_k$ are the *probabilities* of occurrence of the values $y_1, y_2, \ldots, y_k$, we give the quantity $\sum py$ a special name:

Definition 4.11. The *expected value of a discrete random variable* Y, denoted by $E(Y)$, is

$$E(Y) = \sum py$$

where $p_1, \ldots, p_k$ are the probabilities of occurrence of the k possible values $y_1, \ldots, y_k$ of Y. The quantity $E(\,Y)$ is usually denoted by μ.

To calculate the expected value for the data of Table 3.12, the number of boys in families with eight children, we proceed as follows. Let $p_1, p_2, \ldots, p_k$ represent the probabilities $P[Y = 0]$, $P[Y = 1], \ldots, P[Y = 8]$. Then the expected value is

$$\begin{aligned} E(Y) &= p_0 \times 0 + p_1 \times 1 + \cdots + p_8 \times 8 \\ &= (0.0040)(0) + (0.0277)(1) + (0.0993)(2) + \cdots + (0.0064)(8) \\ &= 4.1179 \\ &= 4.12 \text{ boys} \end{aligned}$$

This leads to the statement: "A family with eight children will have an average of 4.12 boys."

Corresponding to the sample variance, s^2, is the variance associated with a discrete random variable:

Definition 4.12. The *variance of a discrete random variable* Y is

$$E(Y - \mu)^2 = \sum p(y - \mu)^2$$

where $p_1, \ldots, p_k$ are the probabilities of occurrence of the k possible values $y_1, \ldots, y_k$ of Y.

The quantity $E(Y-\mu)^2$ is usually denoted by σ^2, where σ is the Greek lowercase letter *sigma*. For the example above, we calculate

$$\begin{aligned}\sigma^2 &= (0.0040)(0-4.1179)^2 + (0.0277)(1-4.1179)^2 + \cdots + (0.0064)(1-4.1179)^2\\ &= 2.0666\end{aligned}$$

Several comments about $E(Y-\mu)^2$ can be made:

1. Computationally, it is equivalent to calculating the sample variance using a divisor of n rather than $n-1$, and probabilities rather than relative frequencies.
2. The square root of $\sigma^2(\sigma)$ is called the (population) *standard deviation* of the random variable.
3. It can be shown that $\sum p(y-\mu)^2 = \sum py^2 - \mu^2$. The quantity $\sum py^2$ is called the *second moment about the origin* and can be defined as the average value of the squares of Y or the expected value of Y^2. This can then be written as $E(Y^2)$, so that $E(Y-\mu)^2 = E(Y^2) - E^2(Y) = E(Y^2) - \mu^2$. See Note 4.9 for further development of the algebra of expectations.

What about the mean and variance of a continuous random variable? As before, we could divide the range of the continuous random variable into a number of intervals, calculate the associated probabilities of the variable, assume that the values are concentrated at the midpoints of the intervals, and proceed with Definitions 4.8 and 4.9. This is precisely what is done with one additional step: The intervals are made narrower and narrower. The mean is then the limit of a sequence of means calculated in this way, and similarly the variance. In these few sentences we have crudely summarized the mathematical process known as *integration*. We will only state the results of such processes but will not actually derive or demonstrate them. For the densities presented in Figure 4.2, the following results can be stated:

Figure	Name	μ	σ^2
4.2(a)	Uniform on $(-1, 1)$	0	1/3
4.2(b)	Uniform on (0, 1)	1/2	1/12
4.2(c)	Triangular on (1, 3)	2	1/6
4.2(d)	Exponential	1	1

The first three densities in Figure 4.2 are examples of *symmetric* densities. A symmetric density always has equality of mean and median. The exponential density is not symmetric; it is "skewed to the right." Such a density has a mean that is larger than the median; for Figure 4.2(d), the median is about 0.69.

It is useful at times to state the functional form for the density. If Y is the random variable, then for a value $Y = y$, the height of the density is given by $f(y)$. The densities in Figure 4.2 have the functional forms shown in Table 4.2. The letter e in $f(y) = e^{-y}$ is the base of the natural logarithms. The symbol ∞ stands for positive infinity.

4.4 NORMAL DISTRIBUTIONS

Statistically, a *population* is the set of all possible values of a variable; random selection of objects of the population makes the variable a random variable and the population is described completely (*modeled*) if the probability function or the probability density function is specified.

Table 4.2 Densities in Figure 4.2

Figure	Name of Density	Function	Range of Y
4.2(a)	Uniform on $(-1, 1)$	$f(y) = 0.5$	$(-1, 1)$
		$f(y) = 0$	elsewhere
4.2(b)	Uniform on (0, 1)	$f(y) = 1$	(0, 1)
		$f(y) = 0$	elsewhere
4.2(c)	Triangular on (1,3)	$f(y) = y - 1$	(1, 2)
		$f(y) = 3 - y$	(2, 3)
		$f(y) = 0$	elsewhere
4.2(d)	Exponential	$f(y) = e^{-y}$	$(0, \infty)$
		$f(y) = 0$	elsewhere

A statistical challenge is to find models of populations that use a few parameters (say, two or three), yet have wide applicability to real data. The *normal* or *Gaussian distribution* is one such statistical model.

The term *Gaussian* refers to Carl Friedrich Gauss, who developed and applied this model. The term *normal* appears to have been coined by Francis Galton. It is important to remember that there is nothing normal or abnormal about the normal distribution! A given data set may or may not be modeled adequately by the normal distribution. However, the normal distribution often proves to be a satisfactory model for data sets. The first and most important reason is that it "works," as will be indicated below. Second, there is a mathematical reason suggesting that a Gaussian distribution may adequately represent many data sets—the famous central limit theorem discussed in Section 4.5. Finally, there is a matter of practicality. The statistical theory and methods associated with the normal distribution work in a nice fashion and have many desirable mathematical properties. But no matter how convenient the theory, the assumptions that a data set is modeled adequately by a normal curve should be verified when looking at a particular data set. One such method is presented in Section 4.4.3.

4.4.1 Examples of Data That Might Be Modeled by a Normal Distribution

The first example is taken from a paper by Golubjatnikov et al. [1972]. Figure 4.3 shows serum cholesterol levels of Mexican and Wisconsin children in two different age groups. In each case

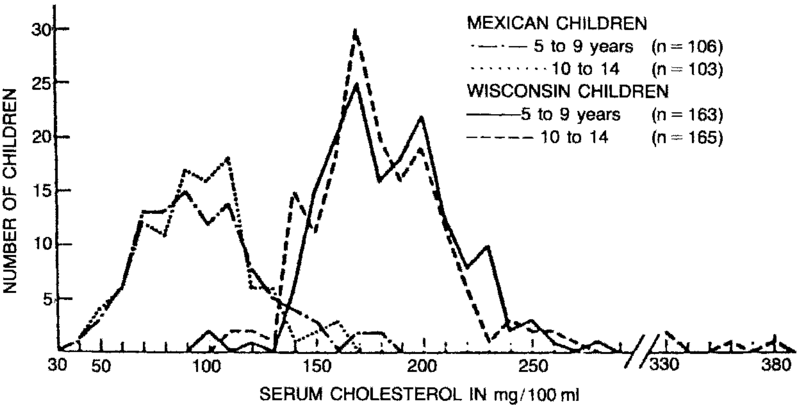


Figure 4.3 Distribution of serum cholesterol levels in Mexican and Wisconsin school children. (Data from Golubjatnikov et al. [1972].)

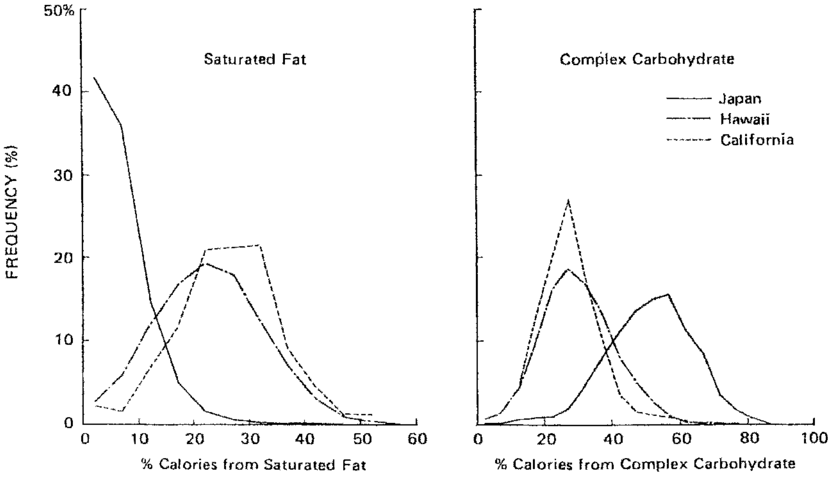


Figure 4.4 Frequency distribution of dietary saturated fat and dietary complex carbohydrate intake. (Data from Kato et al. [1973].)

there is considerable fluctuation in the graphs, probably due to the small numbers of people considered. However, it might be possible to model such data with a normal curve. Note that there seem to be possibly too many values in the right tail to model the data by a normal curve since normal curves are symmetric about their center point.

Figure 4.4 deals with epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii, and California. The curves present the frequency distribution of the percentage of calories from saturated fat and from complex carbohydrate in the three groups of men. Such percentages necessarily lie on the interval from 0 to 100. For the Hawaiian and Californian men with regard to saturated fat, the bell-shaped curve might be a reasonable model. Note, however, that for Japanese men, with a very low percentage of the diet from saturated fat, a bell-shaped curve would obviously be inappropriate.

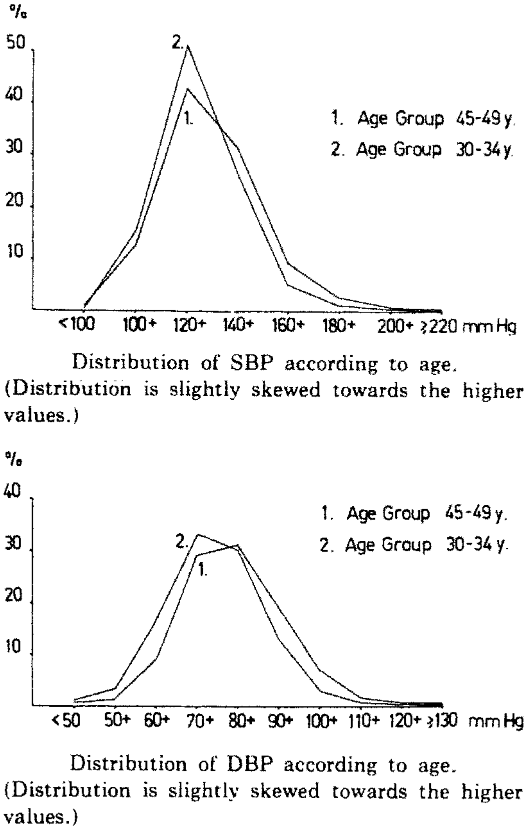
A third example from Kesteloot and van Houte [1973] examines blood pressure measurements on 42,000 members of the Belgian army and territorial police. Figure 4.5 gives two different age groups. Again, particularly in the graphs of the diastolic pressures, it appears that a bell-shaped curve might not be a bad model.

Another example of data that do not appear to be modeled very well by a symmetric bell-shaped curve is from a paper by Hagerup et al. [1972] dealing with serum cholesterol, serum triglyceride, and ABO blood groups in a population of 50-year-old Danish men and women. Figure 4.6 shows the distribution of serum triglycerides. There is a notable asymmetry to the distribution, there being too many values to the right of the peak of the distribution as opposed to the left.

A final example of data that are not normally distributed are the 2-hour plasma glucose levels (mg per 100 mL) in Pima Indians. The data in Figure 4.7 are the plasma glucose levels for male Pima Indians for each decade of age. The data become clearly bimodal (two modes) with increasing decade of age. Note also that the overall curve is shifting to the right with increasing decade: The first mode shifts from approximately 100 mg per 100 mL in the 5- to 14-year decade to about 170 mg per 100 mL in the 65- to 74-year decade.

4.4.2 Calculating Areas under the Normal Curve

A normal distribution is specified completely by its mean, μ, and standard deviation, σ. Figure 4.8 illustrates some normal distributions with specific means and standard deviations. Note that two



Distribution of SBP according to age. (Distribution is slightly skewed towards the higher values.)

Distribution of DBP according to age. (Distribution is slightly skewed towards the higher values.)

Figure 4.5 Distributions of systolic and diastolic blood pressures according to age. (Data from Kesteloot and van Houte [1973].)

normal distributions with the same standard deviation but different means have the same shape and are merely shifted; similarly, two normal distributions with the same means but different standard deviations are centered in the same place but have different shapes. Consequently, μ is called a *location parameter* and σ a *shape parameter*.

The *standard deviation* is the distance from the mean to the point of inflection of the curve. This is the point where a tangent to the curve switches from being over the curve to under the curve.

As with any density, the probability that a normally distributed random variable takes on a value in a specified interval is equal to the area over the interval. So we need to be able to calculate these areas in order to know the desired probabilities. Unfortunately, there is no simple algebraic formula that gives these areas, so tables must be used (see Note 4.15). Fortunately, we need only one table. For any normal distribution, we can calculate areas under its curve using a table for a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$ by expressing the variable in the number of standard deviations from the mean. Using algebraic notation, we get the following:

Definition 4.13. For a random variable Y with mean μ and standard deviation σ, the associated *standard score, Z,* is

$$Z = \frac{Y - \mu}{\sigma}$$

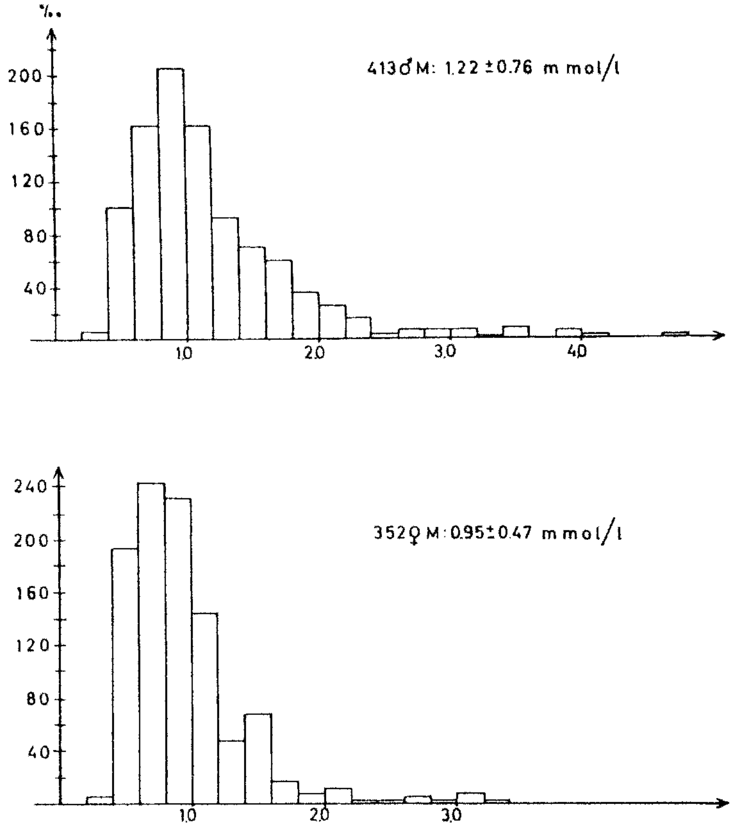


Figure 4.6 Serum triglycerides: 50-year survey in Glostrup. Fasting blood samples were drawn for determination of serum triglyceride by the method of Laurell. (Data from Hagerup et al. [1972].)

Given values for μ and σ, we can go from the "Y scale" to the "Z scale," and vice versa. Algebraically, we can solve for Y and get $Y = \mu + \sigma Z$. This is also the procedure that is used to get from degrees Celsius (°C) to degrees Fahrenheit (°F). The relationship is

$$^\circ\mathrm{C} = \frac{^\circ\mathrm{F} - 32}{1.8}$$

Similarly,

$$^\circ\mathrm{F} = 32 + 1.8 \times {}^\circ\mathrm{C}$$

Definition 4.14. A *standard normal distribution* is a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

Table A.1 in the Appendix gives standard normal probabilities. The table lists the area to the left of the stated value of the standard normal deviate under the columns headed "cum. dist." For example, the area to the left of $Z = 0.10$ is 0.5398, as shown in Figure 4.9.

In words, 53.98% of normally distributed observations have values less than 0.10 standard deviation above the mean. We use the notation $P[Z \leq 0.10] = 0.5398$, or in general, $P[Z \leq z]$. To indicate a value of Z associated with a specified area, p, to its left, we will use a subscript on the value Z_p. For example, $P[Z \leq z_{0.1}] = 0.10$; that is, we want that value of Z such that

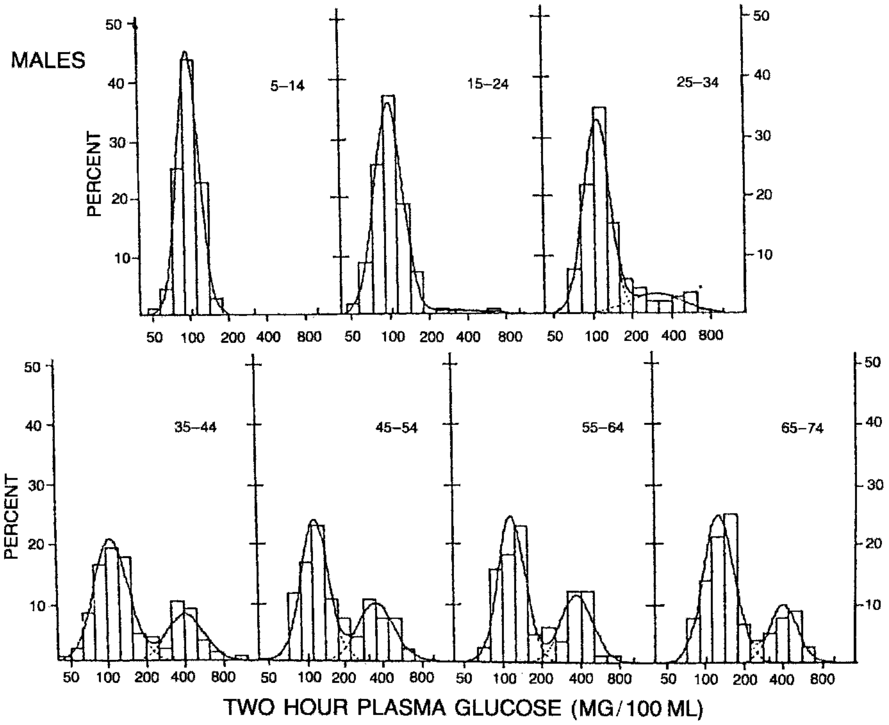


Figure 4.7 Distribution of 2-hour plasma glucose levels (mg/100 mL) in male Pima Indians by decade. (Data from Rushforth et al. [1971].)

0.1 of the area is to its left (call it $z_{0.1}$), or equivalently, such that a proportion 0.1 of Z values are less than or equal to $z_{0.1}$. By symmetry, we note that $z_{1-p} = -z_p$.

Since the total area under the curve is 1, we can get areas in the right-hand tail by subtraction. Formally,

$$P[Z > z] = 1 - P[Z \leq z]$$

In terms of the example above, $P[Z > 0.10] = 1 - 0.5398 = 0.4602$. By symmetry, areas to the left of $Z = 0$ can also be obtained. For example, $P[Z \leq -0.10] = P[Z > 0.10] = 0.4602$. These values are indicated in Figure 4.10.

We now illustrate use of the standard normal table with two word problems. When calculating areas under the normal curve, you will find it helpful to draw a rough normal curve and shade in the required area.

Example 4.4. Suppose that IQ is normally distributed with mean $\mu = 100$ and standard deviation $\sigma = 15$. A person with IQ > 115 has a *high IQ*. What proportion of the population has high IQs? The area required is shown in Figure 4.11. It is clear that IQ = 115 is one standard deviation above the mean, so the statement $P[IQ > 115]$ is equivalent to $P[Z > 1]$. This can be obtained from Table A.1 using the relationship $P[Z > 1] = 1 - P[Z \leq 1] = 1 - 0.8413 = 0.1587$. Thus, 15.87% of the population has a high IQ. By the same token, if an IQ below 85 is labeled *low IQ*, 15.87% of the population has a low IQ.

Example 4.5. Consider the serum cholesterol levels of Wisconsin children as pictured in Figure 4.3. Suppose that the population mean is 175 mg per 100 mL and the population standard

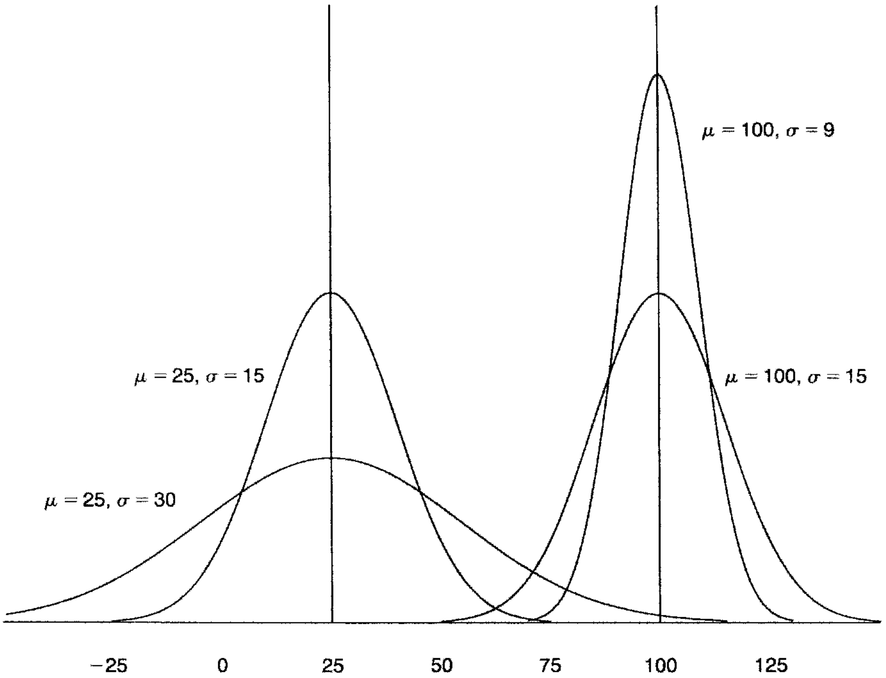


Figure 4.8 Examples of normal distributions.

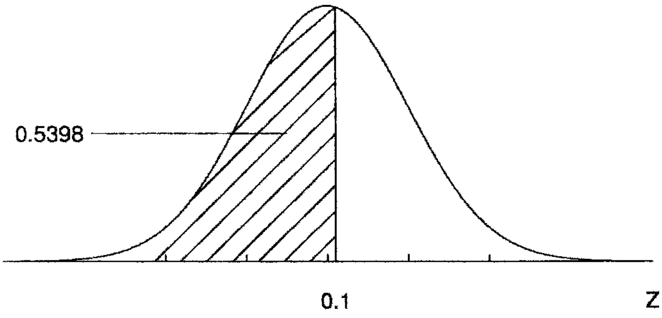


Figure 4.9 Area to the left of $Z = 0.10$ is 0.5398.

deviation is 30 mg per 100 mL. Suppose that a "normal cholesterol value" is taken to be a value within two standard deviations of the mean. What are the *normal limits*, and what proportion of Wisconsin children will be within normal limits?

We want the area within ±2 standard deviations of the mean (Figure 4.12). This can be expressed as $P[-2 \leq Z \leq +2]$. By symmetry and the property that the area under the normal curve is 1.0, we can express this as

$$P[-2 \leq Z \leq 2] = 1 - 2P[Z > 2]$$

(You should sketch this situation, to convince yourself.) From Table A.1, $P[Z \leq 2] = 0.9772$, so that $P[Z > 2] = 1 - 0.9772 = 0.0228$. (Note that this value is computed for you in the

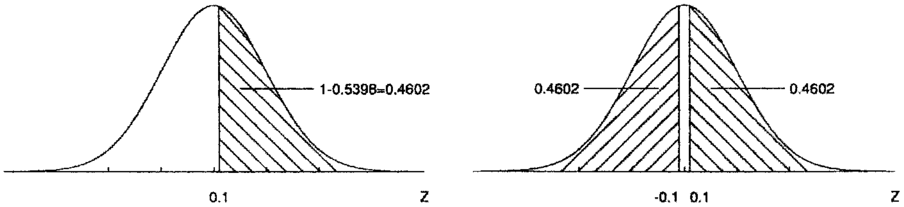


Figure 4.10 $P[Z \leq -0.10] = P[Z > 0.10] = 0.4602$.

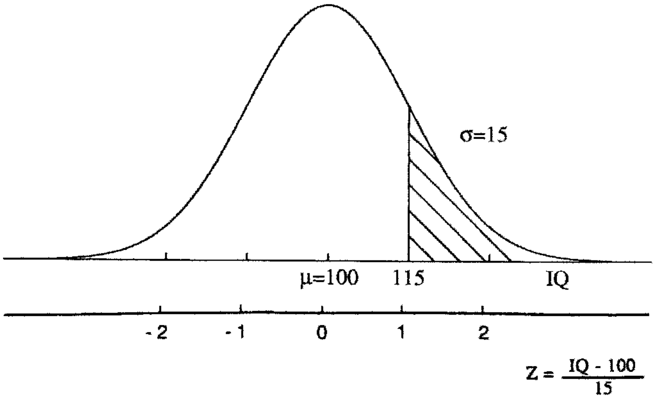


Figure 4.11 Proportion of the population with high IQs.

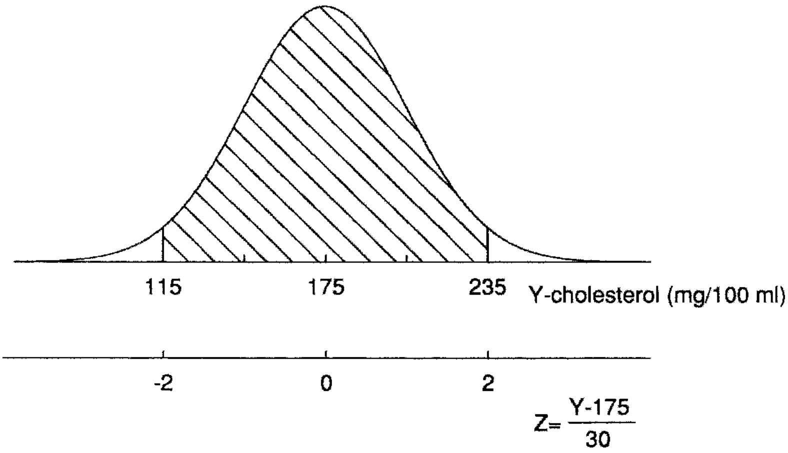


Figure 4.12 Area with ±2 standard deviations of the mean.

column labeled "one-sided.") The desired probability is

$$
\begin{aligned}
P[-2 \leq Z \leq 2] &= 1 - 2(0.0228) \\
&= 0.9544
\end{aligned}
$$

In words, 95.44% of the population of Wisconsin schoolchildren have cholesterol values within normal limits.

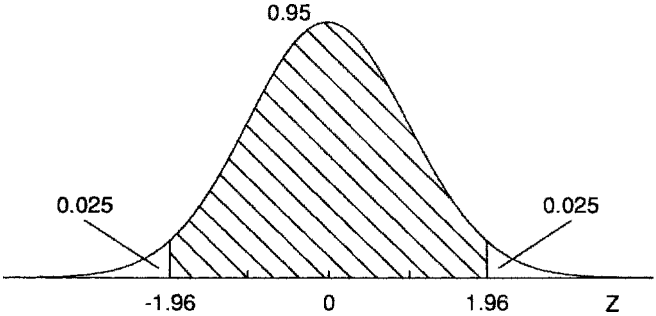


Figure 4.13 Ninety-five percent of normally distributed observations are within ±1.96 standard deviations of the mean.

Suppose that we change the question: Instead of defining normal limits and calculating the proportion within these limits, we define the limits such that, say, 95% of the population has cholesterol values within the stated limits. Before, we went from cholesterol level to Z-value to area; now we want to go from area to Z-value to cholesterol values. In this case, Table A.2 will be useful. Again, we begin with an illustration, Figure 4.13. From Table A.2 we get $P[Z > 1.96] = 0.025$, so that $P[-1.96 \leq Z \leq 1.96] = 0.95$; in words, 95% of normally distributed observations are within ±1.96 standard deviations of the mean. Or, translated to cholesterol values by the formula, $Y = 175 + 30Z$. For $Z = 1.96, Y = 175 + (30)(1.96) = 233.8 \doteq 234$, and for $Z = -1.96, Y = 175 + (30)(-1.96) = 116.2 \doteq 116$. On the basis of the model, 95% of cholesterol values of Wisconsin children are between 116 and 234 mg per 100 mL. If the mean and standard deviation of cholesterol values of Wisconsin children are 175 and 30 mg per 100 mL, respectively, the 95% limits (116, 234) are called 95% *tolerance limits*.

Often, it is useful to know the range of normal values of a substance (variable) in a normal population. A laboratory test can then be carried out to determine whether a subject's values are high, low, or within normal limits.

Example 4.6. An article by Zervas et al. [1970] provides a list of normal values for more than 150 substances ranging from ammonia to vitamin B_{12}. These values have been reprinted in *The Merck Manual of Diagnosis and Therapy* [Berkow, 1999]. The term *normal values* does not imply that variables are normally distributed (i.e., follow a Gaussian or bell-shaped curve). A paper by Elveback et al. [1970] already indicated that of seven common substances (calcium, phosphate, total protein, albumin, urea, magnesium, and alkaline phosphatase), only albumin values can be summarized adequately by a normal distribution. All the other substances had distributions of values that were skewed. The authors (correctly) conclude that "the distributions of values in healthy persons *cannot* be assumed to be normal." Admittedly, this leaves an unsatisfactory situation: What, then, do we mean by *normal limits*? What proportion of normal values will fall outside the normal limits as the result of random variation? None of these—and other—critical questions can now be answered, because a statistical model is not available. But that appears to be the best we can do at this point; as the authors point out, "good limits are hard to get, and bad limits hard to change."

4.4.3 Quantile–Quantile Plots

How can we know whether the normal distribution model fits a particular set of data? There are many tests for normality, some graphical, some numerical. In this section we discuss a simple graphical test, the *quantile–quantile* (QQ) *plot*. In this approach we plot the quantiles of the data distribution observed against the expected quantiles for the normal distribution. The resulting graph is a version of the cumulative frequency distribution but with distorted axes chosen so that a normal distribution would give a straight line. In precomputer days, quantile–quantile plots for the normal distribution were obtained by drawing the empirical cumulative frequency distribution on special *normal probability paper*, but it is now possible to obtain quantile–quantile plots for many different distributions from the computer.

A famous book by Galton [1889] contains data on the stature of parents and their adult children. Table 4.3 gives the frequency distributions of heights of 928 adult children. The

Table 4.3 Frequency Distribution of Stature of 928 Adult Children

Endpoint (in.)	Frequency	Cumulative Frequency	Cumulative Percentage
61.7[a]	5	5	0.5
62.2	7	12	1.3
63.2	32	44	4.7
64.2	59	103	11.1
65.2	48	151	16.3
66.2	117	268	28.9
67.2	138	406	43.8
68.2	120	526	56.7
69.2	167	693	74.7
70.2	99	792	85.3
71.2	64	856	92.2
72.2	41	897	96.7
73.2	17	914	98.5
73.7[a]	14	928	100

Source: Galton [1889].

[a] Assumed endpoint.

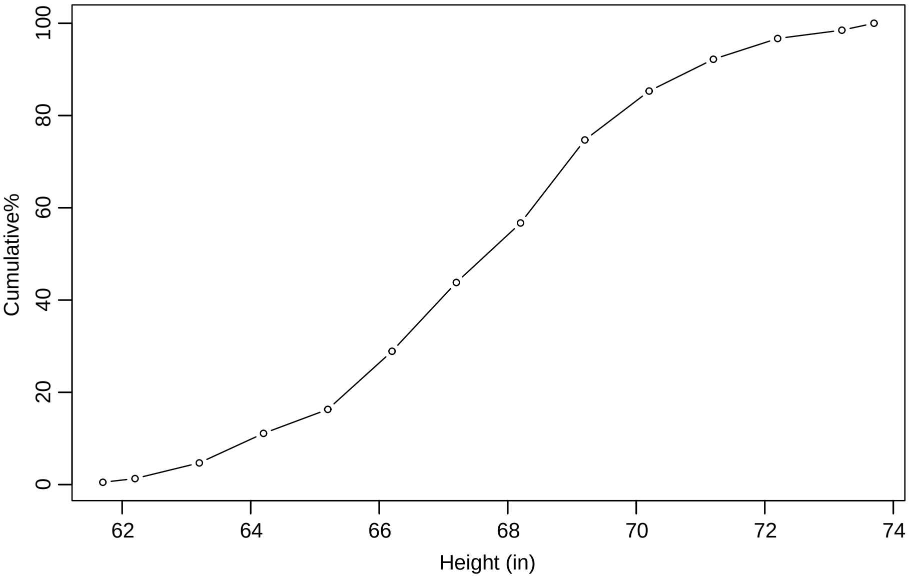


Figure 4.14 Empirical cumulative frequency polygon of heights of 928 adult children. (Data from Galton [1889].)

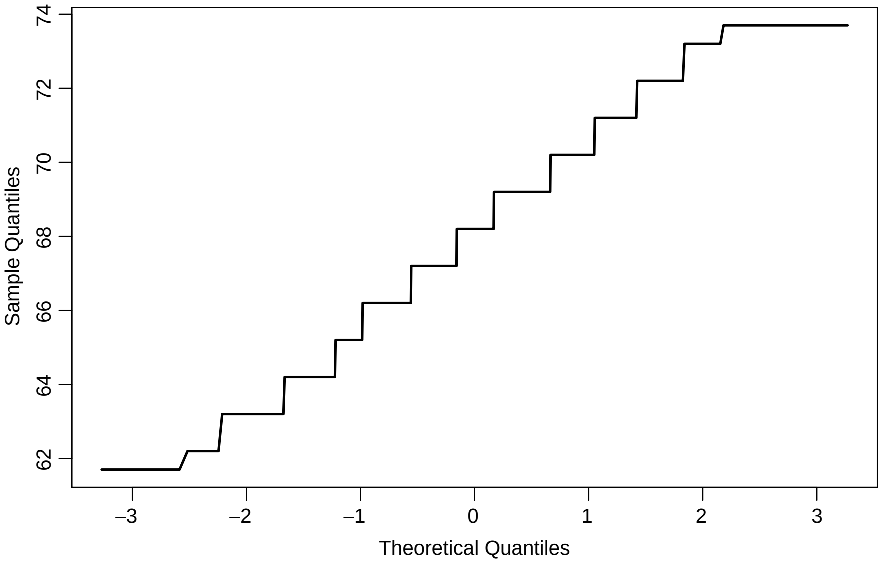


Figure 4.15 Quantile–quantile plot of heights of 928 adult children. (Data from Galton [1889].)

cumulative percentages plotted against the endpoints of the intervals in Figure 4.14 produce the usual sigmoid-shaped curve.

These data are now plotted on normal probability paper in Figure 4.15. The vertical scale has been stretched near 0% and 100% in such a way that data from a normal distribution should fall on a straight line. Clearly, the data are consistent with a normal distribution model.

4.5 SAMPLING DISTRIBUTIONS

4.5.1 Statistics Are Random Variables

Consider a large multicenter collaborative study of the effectiveness of a new cancer therapy. A great deal of care is taken to standardize the treatment from center to center, but it is obvious that the average survival time on the new therapy (or increased survival time if compared to a standard treatment) will vary from center to center. This is an illustration of a basic statistical fact: Sample statistics vary from sample to sample. The key idea is that a statistic associated with a random sample is a random variable. What we want to do in this section is to relate the variability of a statistic based on a random sample to the variability of the random variable on which the sample is based.

Definition 4.15. The probability (density) function of a statistic is called the *sampling distribution of the statistic*.

What are some of the characteristics of the sampling distribution? In this section we state some results about the sample mean. In Section 4.8 some properties of the sampling distribution of the sample variance are discussed.

4.5.2 Properties of Sampling Distribution

Result 4.1. If a random variable Y has population mean μ and population variance σ^2, the sampling distribution of sample means (of samples of size n) has population mean μ and population variance σ^2/n. Note that this result does not assume normality of the "parent" population.

Definition 4.16. The standard deviation of the sampling distribution is called the *standard error*.

Example 4.7. Suppose that IQ is a random variable with mean $\mu = 100$ and standard deviation $\sigma = 15$. Now consider the average IQ of classes of 25 students. What are the population mean and variance of these class averages? By Result 4.1, the class averages have population mean $\mu = 100$ and population variance $\sigma^2/n = 15^2/25 = 9$. Or, the standard error is $\sqrt{\sigma^2/n} = \sqrt{15^2/25} = \sqrt{9} = 3$.

To summarize:

	Population		
	Mean	**Variance**	**$\sqrt{\text{Variance}}$**
Single observation, Y	100	$15^2 = 225$	$15 = \sigma$
Mean of 25 observations, $\overline{Y}$	100	$15^2/25 = 9$	$3 = \sigma/\sqrt{n}$

The standard error of the sampling distribution of the sample mean $\overline{Y}$ is indicated by $\sigma_{\overline{Y}}$ to distinguish it from the standard deviation, σ, associated with the random variable Y. It is instructive to contemplate the formula for the standard error, $\sigma/\sqrt{n}$. This formula makes clear that a reduction in variability by, say, a factor of 2 requires a fourfold increase in sample size. Consider Example 4.7. How large must a class be to reduce the standard error from 3 to 1.5? We want $\sigma/\sqrt{n} = 1.5$. Given that $\sigma = 15$ and solving for n, we get $n = 100$. This is a fourfold increase in class size, from 25 to 100. In general, if we want to reduce the standard error by a factor of k, we must increase the sample size by a factor of k^2. This suggests that if a study consists of, say, 100 observations and with a great deal of additional effort (out of proportion to the effort of getting the 100 observations) another 10 observations can be obtained, the additional 10 may not be worth the effort.

The standard error based on 100 observations is $\sigma/\sqrt{100}$. The ratio of these standard errors is

$$\frac{\sigma/\sqrt{100}}{\sigma/\sqrt{110}} = \frac{\sqrt{100}}{\sqrt{110}} = 0.95$$

Hence a 10% increase in sample size produces only a 5% increase in precision. Of course, precision is not the only criterion we are interested in; if the 110 observations are randomly selected persons to be interviewed, it may be that the last 10 are very hard to locate or difficult to persuade to take part in the study, and not including them may introduce a serious *bias*. But with respect to *precision* there is not much difference between means based on 100 observations and means based on 110 observations (see Note 4.11).

4.5.3 Central Limit Theorem

Although Result 4.1 gives some characteristics of the sampling distribution, it does not permit us to calculate probabilities, because we do not know the form of the sampling distribution. To be able to do this, we need the following:

Result 4.2. If Y is normally distributed with mean μ and variance σ^2, then $\overline{Y}$, based on a random sample of n observations, is *normally distributed* with mean μ and variance σ^2/n.

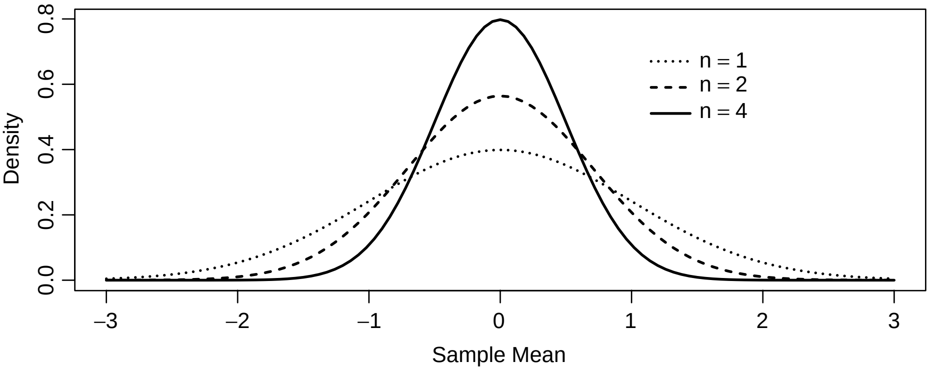


Figure 4.16 Three sampling distributions for means of random samples of size 1, 2, and 4 from a $N(0, 1)$ population.

Result 4.2 basically states that if Y is normally distributed, then $\overline{Y}$, the mean of a random sample, is normally distributed. Result 4.1 then specifies the mean and variance of the sampling distribution. Result 4.2 implies that as the sample size increases, the (normal) distribution of the sample mean becomes more and more "pinched." Figure 4.16 shows three sampling distributions for means of random samples of size 1, 2, and 4.

What is the probability that the average IQ of a class of 25 students exceeds 106? By Result 4.2, $\overline{Y}$, the average of 25 IQs, is normally distributed with mean $\mu = 100$ and standard error $\sigma/\sqrt{n} = 15/\sqrt{25} = 3$. Hence the probability that $\overline{Y} > 106$ can be calculated as

$$\begin{aligned} P[\overline{Y} \geq 106] &= P\left[Z \geq \frac{106 - 100}{3}\right] \\ &= P[Z \geq 2] \\ &= 1 - 0.9772 \\ &= 0.0228 \end{aligned}$$

So approximately 2% of average IQs of classes of 25 students will exceed 106. This can be compared with the probability that a single person's IQ exceeds 106:

$$P[Y > 106] = P\left[Z > \frac{6}{15}\right] = P[Z > 0.4] = 0.3446$$

The final result we want to state is known as the *central limit theorem.*

Result 4.3. If a random variable Y has population mean μ and population variance σ^2, the sample mean $\overline{Y}$, based on n observations, is approximately normally distributed with mean μ and variance σ^2/n, for sufficiently large n.

This is a remarkable result and the most important reason for the central role of the normal distribution in statistics. What this states basically is that means of random samples from *any* distribution (with mean and variance) will tend to be normally distributed as the sample size becomes sufficiently large. How large is "large"? Consider the distributions of Figure 4.2. Samples of six or more from the first three distributions will have means that are virtually normally

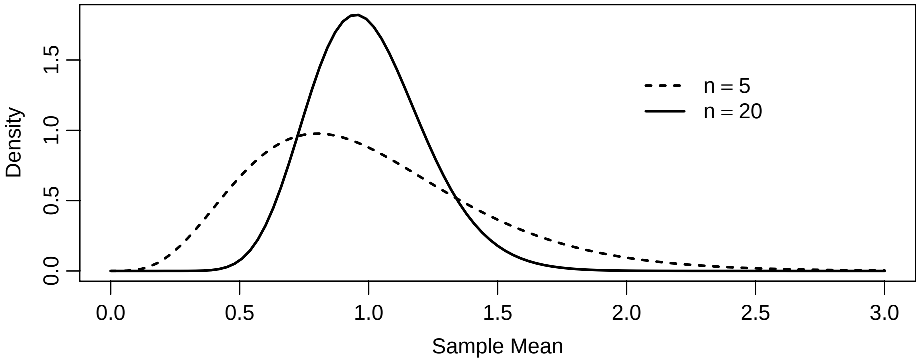


Figure 4.17 Sampling distributions of means of 5 and 20 observations when the parent distribution is exponential.

distributed. The fourth distribution will take somewhat larger samples before approximate normality is obtained; n must be around 25 or 30. Figure 4.17 is a more skewed figure that shows the sampling distributions of means of samples of various sizes drawn from Figure 4.2(d).

The central limit theorem provides some reassurance when we are not certain whether observations are normally distributed. The means of reasonably sized samples will have a distribution that is approximately normal. So inference procedures based on the sample means can often use the normal distribution. But you must be careful not to impute normality to the original observations.

4.6 INFERENCE ABOUT THE MEAN OF A POPULATION

4.6.1 Point and Interval Estimates

In this section we discuss inference about the mean of a population when the population variance is known. The assumption may seem artificial, but sometimes this situation will occur. For example, it may be that a new treatment alters the level of a response variable but not its variability, so that the variability can be assumed to be known from previous experiments. (In Section 4.8 we discuss a method for comparing the variability of an experiment with previous established variability; in Chapter 5 the problem of inference when both population mean and variance are unknown is considered.)

To put the problem more formally, we have a random variable Y with unknown population mean μ. A random sample of size n is taken and inferences about μ are to be made on the basis of the sample. We assume that the population variance is known; denote it by σ^2. Normality will also be assumed; even when the population is not normal, we may be able to appeal to the central limit theorem.

A "natural" estimate of the population mean μ is the sample mean $\overline{Y}$. It is a natural estimate of μ because we know that $\overline{Y}$ is normally distributed with the same mean, μ, and variance σ^2/n. Even if Y is not normal, $\overline{Y}$ is approximately normal on the basis of the central limit theorem. The statistic $\overline{Y}$ is called a *point estimate* since we estimate the parameter μ by a single value or point.

Now the question arises: How precise is the estimate? How can we distinguish between two samples of, say, 25 and 100 observations? Both may give the same—or approximately the same—sample mean, but we know that the mean based on the 100 observations is more accurate, that is, has a smaller standard error. One possible way of summarizing this information is to give the sample mean and its standard error. This would be useful for *comparing* two samples. But this does not seem to be a useful approach in considering one sample and its information about the parameter. To use the information in the sample, we set up an *interval* estimate as follows: Consider the quantity $\mu \pm (1.96)\sigma/\sqrt{n}$. It describes the spread of sample means; in particular, 95% of means of samples of size n will fall in the interval $[\mu - 1.96\sigma/\sqrt{n}, \mu + 1.96\sigma/\sqrt{n}]$. The interval has the property that as n increases, the width decreases (refer to Section 4.5 for further discussion). Suppose that we now replace μ by its point estimate, $\overline{Y}$. How can we interpret the resulting interval? Since the sample mean, $\overline{Y}$, varies from sample to sample, it cannot mean that 95% of the sample means will fall in the interval for a specific sample mean. The interpretation is that the probability is 0.95 that the interval *straddles* the population mean. Such an interval is referred to as a *95% confidence interval* for the population mean, μ. We now formalize this definition.

Definition 4.17. A $100(1-\alpha)\%$ *confidence interval* for the mean μ of a normal population (with variance known) based on a random sample of size n is

$$\overline{Y} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $z_{1-\alpha/2}$ is the value of the standard normal deviate such that $100(1-\alpha)\%$ of the area falls within $\pm z_{1-\alpha/2}$.

Strictly speaking, we should write

$$\left(\overline{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

but by symmetry, $z_{\alpha/2} = -z_{1-\alpha/2}$, so that it is quicker to use the expression above.

Example 4.8. In Section 3.3.1 we discussed the age at death of 78 cases of crib death (SIDS) occurring in King County, Washington, in 1976–1977. Birth certificates were obtained for these cases and birthweights were tabulated. Let Y = birthweight in grams. Then, for these 78 cases, $\overline{Y} = 2993.6 = 2994$ g. From a listing of all the birthweights, it is known that the standard deviation of birthweight is about 800 g (i.e., $\sigma = 800$ g). A 95% confidence interval for the mean birthweight of SIDS cases is calculated to be

$$2994 \pm (1.96)\left(\frac{800}{\sqrt{78}}\right) \quad \text{or} \quad 2994 \pm (1.96)(90.6) \quad \text{or} \quad 2994 \pm 178$$

producing a lower limit of 2816 g and an upper limit of 3172 g. Thus, on the basis of these data, we are 95% confident that we have straddled the population mean, μ, of birthweight of SIDS infants by the interval (2816, 3172).

Suppose that we had wanted to be more confident: say, a level of 99%. The value of Z now becomes 2.58 (from Table A.2), and the corresponding limits are $2994 \pm (2.58)(800/\sqrt{78})$, or (2760, 3228). The width of the 99% confidence interval is greater than that of the 95% confidence interval (468 g vs. 356 g), the price we paid for being more sure that we have straddled the population mean.

Several comments should be made about confidence intervals:

1. Since the population mean μ is fixed, it is not correct to say that the probability is $1-\alpha$ that μ is in the confidence interval *once it is computed*; that probability is zero or 1. Either the mean is in the interval and the probability is equal to 1, or the mean is not in the interval and the probability is zero.

2. We can increase our confidence that the interval straddles the population mean by decreasing α, hence increasing $Z_{1-\alpha/2}$. We can take values from Table A.2 to construct the following confidence levels:

Confidence Level	Z-Value
90%	1.64
95%	1.96
99%	2.58
99.9%	3.29

The effect of increasing the confidence level will be to increase the width of the confidence interval.

3. To decrease the width of the confidence interval, we can either decrease the confidence level or increase the sample size. The width of the interval is $2z_{1-\alpha/2}\sigma/\sqrt{n}$. For a fixed confidence level the width is essentially a function of $\sigma/\sqrt{n}$, the standard error of the mean. To decrease the width by a factor of, say, 2, the sample size must be increased by a factor of 4, analogous to the discussion in Section 4.5.2.
4. Confidence levels are usually taken to be 95% or 99%. These levels are a matter of convention; there are no theoretical reasons for choosing these values. A rough rule to keep in mind is that a 95% confidence interval is defined by the sample mean ±2 standard errors (*not* standard deviations).

4.6.2 Hypothesis Testing

In estimation, we start with a sample statistic and make a statement about the population parameter: A confidence interval makes a probabilistic statement about straddling the population parameter. In hypothesis testing, we start by assuming a value for a parameter, and a probability statement is made about the value of the corresponding statistic. In this section, as in Section 4.6.1, we assume that the population variance is known and that we want to make inferences about the mean of a normal population on the basis of a sample mean. The basic strategy in hypothesis testing is to measure how far an observed statistic is from a hypothesized value of the parameter. If the distance is "great" (Figure 4.18) we would argue that the hypothesized parameter value is inconsistent with the data and we would be inclined to reject the hypothesis (we could be wrong, of course; rare events do happen).

To interpret the distance, we must take into account the basic variability (σ^2) of the observations and the size of the sample (n) on which the statistic is based. As a rough rule of thumb that is explained below, if the observed value of the statistic is more than two standard errors from the hypothesized parameter value, we question the truth of the hypothesis.

To continue Example 4.8, the mean birthweight of the 78 SIDS cases was 2994 g. The standard deviation σ_0 was assumed to be 800 g, and the standard error $\sigma/\sqrt{n} = 800/\sqrt{78} = 90.6$ g. One question that comes up in the study of SIDS is whether SIDS cases tend to have a different birthweight than the general population. For the general population, the average birthweight is about 3300 g. Is the *sample* mean value of 2994 g consistent with this value? Figure 4.19 shows that the distance between the two values is 306 g. The standard error is 90.6,

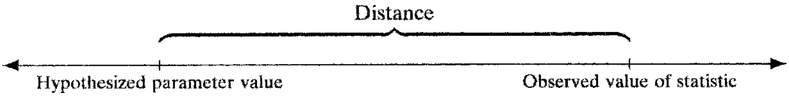


Figure 4.18 Great distance from a hypothesized value of a parameter.

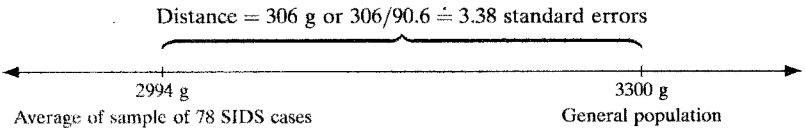

Distance = 306 g or 306/90.6 ≐ 3.38 standard errors
2994 g
Average of sample of 78 SIDS cases
3300 g
General population

Figure 4.19 Distance between the two values is 306 g.

so the observed value is 306/90.6 = 3.38 standard errors from the hypothesized population mean. By the rule we stated, the distance is so great that we would conclude that the mean of the *sample* of SIDS births is inconsistent with the mean value in the general population. Hence, we would conclude that the SIDS births come from a population with mean birthweight somewhat less than that of the general population. (This raises more questions, of course: Are the gestational ages comparable? What about the racial composition? and so on.) The best estimate we have of the mean birthweight of the population of SIDS cases is the sample mean: in this case, 2994 g, about 300 g lower than that for the normal population.

Before introducing some standard hypothesis testing terminology, two additional points should be made:

1. We have expressed "distance" in terms of number of standard errors from the hypothesized parameter value. Equivalently, we can associate a tail probability with the observed value of the statistic. For the sampling situation described above, we know that the sample mean $\overline{Y}$ is normally distributed with standard error $\sigma/\sqrt{n}$. As Figure 4.20 indicates, the farther away the observed value of the statistic is from the hypothesized parameter value, the smaller the area (probability) in the tail. This tail probability is usually called the *p-value*. For example (using Table A.2), the area to the right of 1.96 standard errors is 0.025; the area to the right of 2.58 standard errors is 0.005. Conversely, if we specify the area, the number of standard errors will be determined.
2. Suppose that we planned before doing the statistical test that we would not question the hypothesized parameter value if the observed value of the statistic fell within, say, two standard errors of the parameter value. We could divide the sample space for the statistic (i.e., the real line) into three regions as shown in Figure 4.21. These regions could have been set up before the value of the statistic was observed. All that needs to be determined then is in which region the observed value of the statistic falls to determine if it is consistent with the hypothesized value.

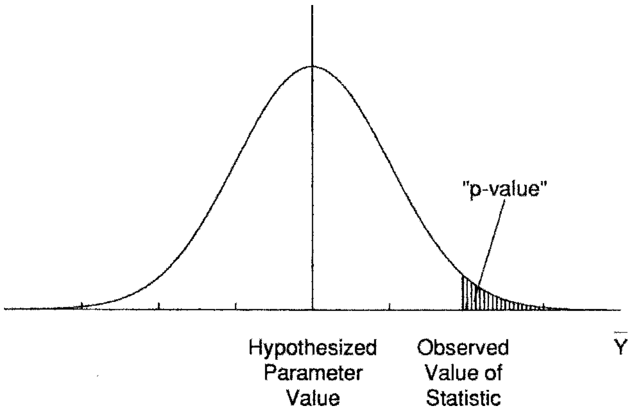


Figure 4.20 The farther away the observed value of a statistic from the hypothesized value of a parameter, the smaller the area in the tail.

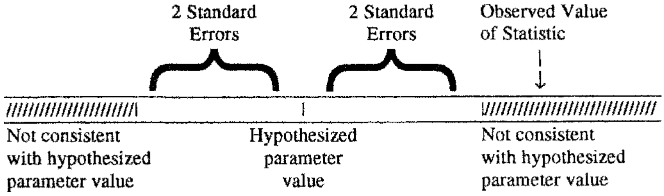

2 Standard Errors
2 Standard Errors
Observed Value of Statistic
Not consistent with hypothesized parameter value
Hypothesized parameter value
Not consistent with hypothesized parameter value

Figure 4.21 Sample space for the statistic.

We now formalize some of these concepts:

Definition 4.18. A *null hypothesis* specifies a hypothesized real value, or values, for a parameter (see Note 4.15 for further discussion).

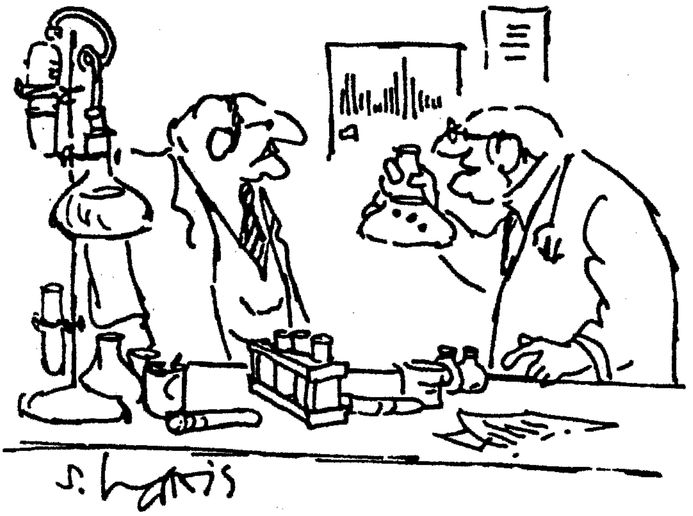
Definition 4.19. The *rejection region* consists of the set of values of a statistic for which the null hypothesis is rejected. The values of the boundaries of the region are called the *critical values*.

Definition 4.20. A *Type I error* occurs when the null hypothesis is rejected when, in fact, it is true. The *significance level* is the probability of a Type I error when the null hypothesis is true.

Definition 4.21. An *alternative hypothesis* specifies a real value or range of values for a parameter that will be considered when the null hypothesis is rejected.

Definition 4.22. A *Type II error* occurs when the null hypothesis is not rejected when it is false.

Definition 4.23. The *power of a test* is the probability of rejecting the null hypothesis when it is false.



"It may very well bring about immortality, but it will take forever to test it."

© 1976 by Sidney Harris — *American Scientist* Magazine

Cartoon 4.1 Testing some hypotheses can be tricky. (From *American Scientist*, March–April 1976.)

Definition 4.24. The *p-value* in a hypothesis testing situation is that value of p, $0 \leq p \leq 1$, such that for $\alpha > p$ the test rejects the null hypothesis at significance level α, and for $\alpha < p$ the test does not reject the null hypothesis. Intuitively, the p-value is the probability under the null hypothesis of observing a value as unlikely or more unlikely than the value of the test statistic. The p-value is a measure of the distance from the observed statistic to the value of the parameter specified by the null hypothesis.

Notation

1. The null hypothesis is denoted by H_0 the alternative hypothesis by H_A.
2. The probability of a Type I error is denoted by α, the probability of a Type II error by β. The power is then

$$\begin{aligned} \text{power} &= 1 - \text{probability of Type II error} \\ &= 1 - \beta \end{aligned}$$

Continuing Example 4.8, we can think of our assessment of the birthweight of SIDS babies as a type of decision problem illustrated in the following layout:

	State of Nature SIDS Birthweights	
Decision SIDS Birthweights	**Same as Normal**	**Not the Same**
Same as normal	Correct $(1 - \alpha)$	Type II error (β)
Not the same	Type I error (α)	Correct $(1 - \beta)$

This illustrates the two types of errors that can be made depending on our decision and the *state of nature*. The null hypothesis for this example can be written as

$$H_0 : \mu = 3300 \text{ g}$$

and the alternative hypothesis written as

$$H_A : \mu \neq 3300 \text{ g}$$

Suppose that we want to reject the null hypothesis when the sample mean $\overline{Y}$ is more than two standard errors from the H_0 value of 3300 g. The standard error is 90.6 g. The rejection region is then determined by $3300 \pm (2)(90.6)$ or 3300 ± 181.

We can then set up the hypothesis-testing framework as indicated in Figure 4.22. The rejection region consists of values to the left of 3119 g (i.e., $\mu - 2\sigma/\sqrt{n}$) and to the right of 3481 g (i.e., $\mu + 2\sigma/\sqrt{n}$). The observed value of the statistic, $\overline{Y} = 2994$ g, falls in the rejection region, and we therefore reject the null hypothesis that SIDS cases have the same mean birthweight as normal children. On the basis of the sample value observed, we conclude that SIDS babies tend to weigh less than normal babies.

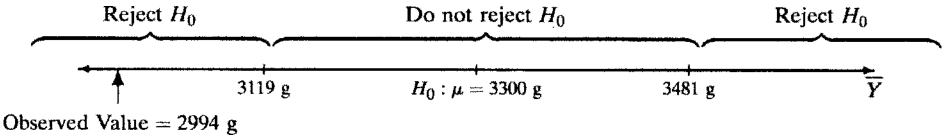


Figure 4.22 Hypothesis-testing framework for birthweight assessment.

The probability of a Type I error is the probability that the mean of a sample of 78 observations from a population with mean 3300 g is less than 3119 g or greater than 3481 g:

$$P[3119 \le \overline{Y} \le 3481] = P\left[\frac{3119-3300}{90.6} \le Z \le \frac{3481-3300}{90.6}\right]$$
$$= P[-2 \le Z \le +2]$$

where Z is a standard normal deviate.

From Table A.1,

$$P[Z \le 2] = 0.9772$$

so that

$$1 - P[-2 \le Z \le 2] = (2)(0.0228) = 0.0456$$

the probability of a Type I error. The probability is 0.0455 from the two-sided p-value of Table A.1. The difference relates to rounding.

The probability of a Type II error can be computed when a value for the parameter under the alternative hypothesis is specified. Suppose that for these data the alternative hypothesis is

$$H_A : \mu = 3000 \text{ g}$$

this value being suggested from previous studies. To calculate the probability of a Type II error—and the power—we assume that $\overline{Y}$, the mean of the 78 observations, comes from a normal distribution with mean 3000 g and standard error as before, 90.6 g. As Figure 4.23 indicates, the probability of a Type II error is the area over the interval (3119, 3481). This can be calculated as

$$\begin{aligned}
P[\text{Type II error}] &= P[3119 \le \overline{Y} \le 3481] \\
&= P\left[\frac{3119-3000}{90.6} \le Z \le \frac{3481-3000}{90.6}\right] \\
&\doteq P[1.31 \le Z \le 5.31] \\
&\doteq 1 - 0.905 \\
&\doteq 0.095
\end{aligned}$$

So $\beta = 0.095$ and the power is $1 - \beta = 0.905$. Again, these calculations can be made before any data are collected, and they say that if the SIDS population mean birthweight were 3000 g and the normal population birthweight 3300 g, the probability is 0.905 that a mean from a sample of 78 observations will be declared significantly different from 3300 g.

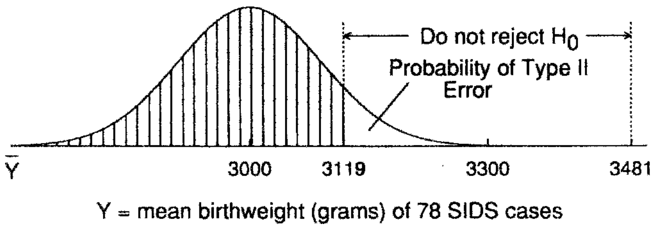


Figure 4.23 Probability of a Type II error.

Let us summarize the analysis of this example:

$$\begin{array}{c} \text{Hypothesis-testing setup} \\ \text{(no data taken)} \end{array} \begin{cases} H_0 : \mu = 3300 \text{ g} \\ H_A : \mu = 3000 \text{ g} \\ \sigma = 800 \text{ g (known)} \\ n = 78 \\ \text{rejection region: } \pm\, 2 \text{ standard errors from 3000 g} \\ \alpha = 0.0456 \\ \beta = 0.095 \\ 1 - \beta = 0.905 \end{cases}$$

Observe: $\overline{Y} = 2994$

Conclusion: Reject H_0

The value of α is usually specified beforehand: The most common value is 0.05, somewhat less common values are 0.01 or 0.001. Corresponding to the confidence level in interval estimation, we have the *significance level* in hypothesis testing. The significance level is often expressed as a percentage and defined to be $100\alpha\%$. Thus, for $\alpha = 0.05$, the hypothesis test is carried out at the 5%, or 0.05, significance level.

The use of a single symbol β for the probability of a Type II error is standard but a bit misleading. We expect β to stand for one number in the same way that α stands for one number. In fact, β is a function whose argument is the assumed true value of the parameter being tested. For example, in the context of $H_A : \mu = 3000$ g, β is a function of μ and could be written $\beta(\mu)$. It follows that the power is also a function of the true parameter: power $= 1 - \beta(\mu)$. Thus one must specify a value of μ to compute the power.

We finish this introduction to hypothesis testing with a discussion of the one- and two-tailed test. These are related to the choice of the rejection region. Even if α is specified, there is an infinity of rejection regions such that the area over the region is equal to α. Usually, only two types of regions are considered as shown in Figure 4.24. A *two-tailed test* is associated with a

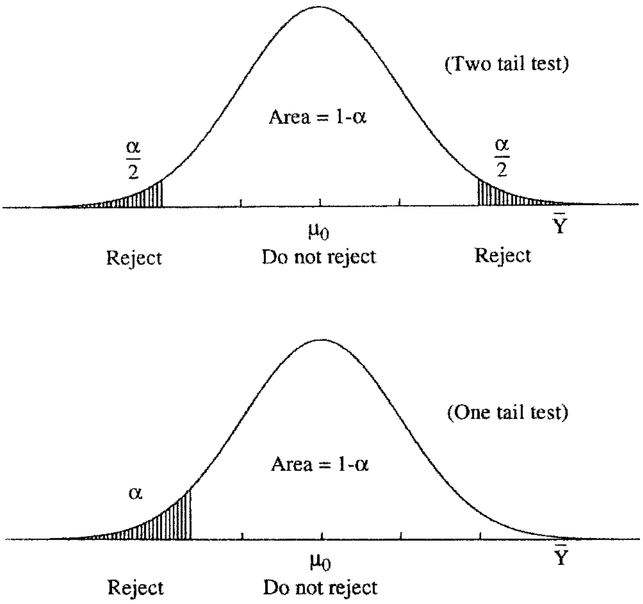


Figure 4.24 Two types of regions considered in hypothesis testing.

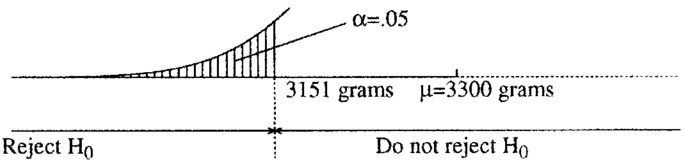


Figure 4.25 Start of the rejection region in a one-tailed test.

rejection region that extends both to the left and to the right of the hypothesized parameter value. A *one-tailed test* is associated with a region to one side of the parameter value. The alternative hypothesis determines the type of test to be carried out. Consider again the birthweight of SIDS cases. Suppose we know that if the mean birthweight of these cases is not the same as that of normal infants (3300 g), it must be less; it is not possible for it to be more. In that case, if the null hypothesis is false, we would expect the sample mean to be below 3300 g, and we would reject the null hypothesis for values of $\overline{Y}$ below 3300 g. We could then write the null hypothesis and alternative hypothesis as follows:

$$H_0 : \mu = 3300 \text{ g}$$
$$H_A : \mu < 3300 \text{ g}$$

We would want to carry out a one-tailed test in this case by setting up a rejection region to the left of the parameter value. Suppose that we want to test at the 0.05 level, and we only want to reject for values of $\overline{Y}$ below 3300 g. From Table A.2 we see that we must locate the start of the rejection region 1.64 standard errors to the left of $\mu = 3300$ g, as shown in Figure 4.25. The value is $3300 - (1.64)(800/\sqrt{78})$ or $3300 - (1.64)(90.6) = 3151$ g.

Suppose that we want a two-tailed test at the 0.05 level. The Z-value (Table A.2) is now 1.96, which distributes 0.025 in the left tail and 0.025 in the right tail. The corresponding values for the critical region are $3300 \pm (1.96)(90.6)$ or (3122, 3478), producing a region very similar to the region calculated earlier.

The question is: When should you do a one-tailed test and when a two-tailed test? As was stated, the alternative hypothesis determines this. An alternative hypothesis of the form $H_A : \mu \neq \mu_0$ is called *two-sided* and will require a two-tailed test. Similarly, the alternative $H_A : \mu < \mu_0$ is called one-sided and will lead to a one-tailed test. So should the alternative hypothesis be one- or two-sided? The experimental situation will determine this. For example, if nothing is known about the effect of a proposed therapy, the alternative hypothesis should be made two-sided. However, if it is suspected that a new therapy will do nothing or increase a response level, and if there is no reason to distinguish between no effect and a decrease in the response level, the test should be one-tailed. The general rule is: The more specific you can make the experiment, the greater the power of the test (see Fleiss et al. [2003, Sec. 2.4]). (See Problem 4.33 to convince yourself that the power of a one-tailed test is greater *if* the alternative hypothesis specifies the situation correctly.)

4.7 CONFIDENCE INTERVALS VS. TESTS OF HYPOTHESES

You may have noticed that there is a very close connection between the confidence intervals and the tests of hypotheses that we have constructed. In both approaches we have used the standard normal distribution and the quantity α.

In *confidence intervals* we:

1. Specify the confidence level $(1 - \alpha)$.

2. Read $z_{1-\alpha/2}$ from a standard normal table.
3. Calculate $\overline{Y} \pm z_{1-\alpha/2}\sigma/\sqrt{n}$.

In *hypothesis testing* we:

1. Specify the null hypothesis ($H_0 : \mu = \mu_0$).
2. Specify α, the probability of a Type I error.
3. Read $z_{1-\alpha/2}$ from a standard normal table.
4. Calculate $\mu_0 \pm z_{1-\alpha/2}\sigma/\sqrt{n}$.
5. Observe $\overline{Y}$; reject or accept H_0.

The two approaches can be represented pictorially as shown in Figure 4.26. It is easy to verify that if the confidence interval does not straddle μ_0 (as is the case in the figure), $\overline{Y}$ will fall in the rejection region, and vice versa. Will this always be the case? The answer is "yes." When we are dealing with inference about the value of a parameter, the two approaches will give the same answer. To show the equivalence algebraically, we start with the key inequality

$$P\left[-z_{1-\alpha/2} \leq \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right] = 1-\alpha$$

If we solve the inequality for $\overline{Y}$, we get

$$P\left[\mu - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}} \leq \overline{Y} \leq \mu + \frac{z_{1-\alpha/2}}{\sqrt{n}}\right] = 1-\alpha$$

Given a value $\mu = \mu_0$, the statement produces a region ($\mu_0 \pm z_{1-\alpha/2}\sigma/\sqrt{n}$) within which $100(1-\alpha)\%$ of sample means fall. If we solve the inequality for μ, we get

$$P\left[\overline{Y} - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}} \leq \mu \leq \overline{Y} + \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right] = 1-\alpha$$

This is a confidence interval for the population mean μ. In Chapter 5 we examine this approach in more detail and present a general methodology.

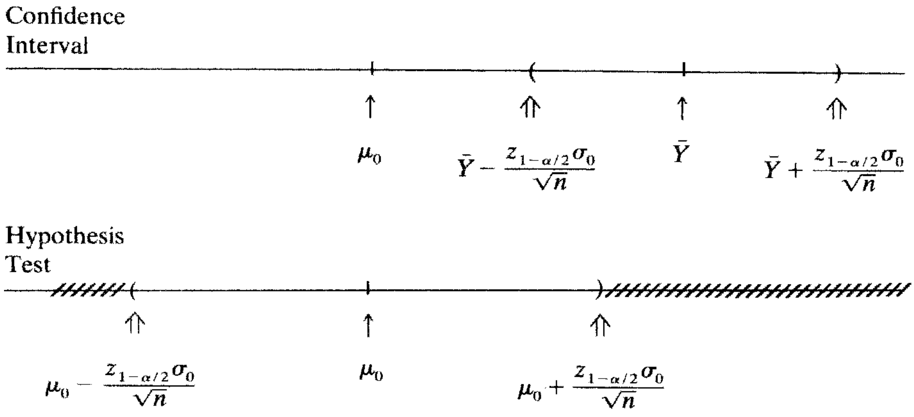

Confidence Interval
μ_0
$\bar{Y} - \frac{z_{1-\alpha/2}\sigma_0}{\sqrt{n}}$
$\bar{Y}$
$\bar{Y} + \frac{z_{1-\alpha/2}\sigma_0}{\sqrt{n}}$
Hypothesis Test
$\mu_0 - \frac{z_{1-\alpha/2}\sigma_0}{\sqrt{n}}$
μ_0
$\mu_0 + \frac{z_{1-\alpha/2}\sigma_0}{\sqrt{n}}$

Figure 4.26 Confidence intervals vs. tests of hypothesis.

If confidence intervals and hypothesis testing are but two sides of the same coin, why bother with both? The answer is (to continue the analogy) that the two sides of the coin are not the same; there is different information. The confidence interval approach emphasizes the precision of the estimate by means of the width of the interval and provides a point estimate for the parameter, regardless of any hypothesis. The hypothesis-testing approach deals with the consistency of observed (new) data with the hypothesized parameter value. It gives a probability of observing the value of the statistic or a more extreme value. In addition, it will provide a method for estimating sample sizes. Finally, by means of power calculations, we can decide beforehand whether a proposed study is feasible; that is, what is the probability that the study will demonstrate a difference if a (specified) difference exists?

You should become familiar with both approaches to statistical inference. Do not use one to the exclusion of another. In some research fields, hypothesis testing has been elevated to the only "proper" way of doing inference; all scientific questions have to be put into a hypothesis-testing framework. This is absurd and stultifying, particularly in pilot studies or investigations into uncharted fields. On the other hand, not to consider *possible* outcomes of an experiment and the chance of picking up differences is also unbalanced. Many times it will be useful to specify very carefully what is known about the parameter(s) of interest *and* to specify, in perhaps a crude way, alternative values or ranges of values for these parameters. If it is a matter of emphasis, you should stress hypothesis testing before carrying out a study and estimation after the study has been done.

4.8 INFERENCE ABOUT THE VARIANCE OF A POPULATION

4.8.1 Distribution of the Sample Variance

In previous sections we assumed that the population variance of a normal distribution was known. In this section we want to make inferences about the population variance on the basis of a sample variance. In making inferences about the population mean, we needed to know the sampling distribution of the sample mean. Similarly, we need to know the sampling distribution of the sample variance in order to make inferences about the population variance; analogous to the statement that for a normal random variable, Y, with sample mean $\overline{Y}$, the quantity

$$\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}$$

has a normal distribution with mean 0 and variance 1. We now state a result about the quantity $(n-1)s^2/\sigma^2$. The basic information is contained in the following statement:

Result 4.4. If a random variable Y is normally distributed with mean μ and variance σ^2, then for a random sample of size n the quantity $(n-1)s^2/\sigma^2$ has a chi-square distribution with $n-1$ degrees of freedom.

Each distribution is indexed by $n-1$ degrees of freedom. Recall that the sample variance is calculated by dividing $\sum(y-\overline{y})^2$ by $n-1$, the degrees of freedom.

The chi-square distribution is skewed; the amount of skewness decreases as the degrees of freedom increases. Since $(n-1)s^2/\sigma^2$ can never be negative, the sample space for the chi-square distribution is the nonnegative part of the real line. Several chi-square distributions are shown in Figure 4.27. The mean of a chi-square distribution is equal to the degrees of freedom, and

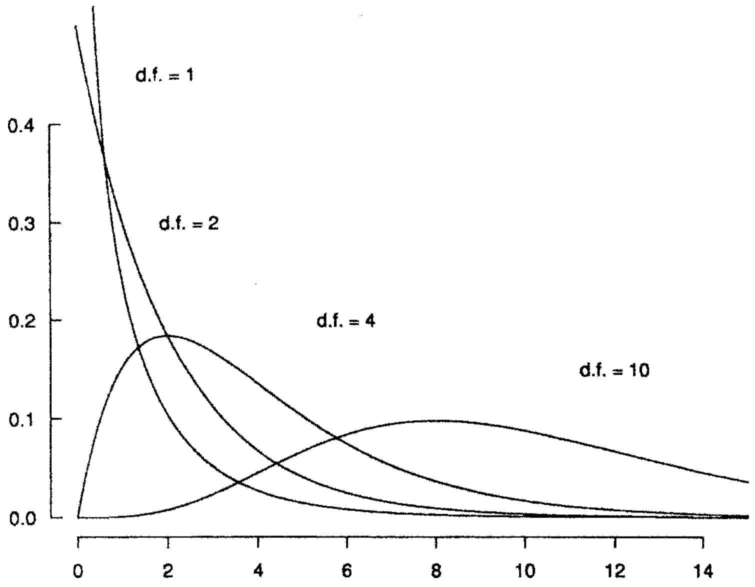


Figure 4.27 Chi-square distributions.

the variance is twice the degrees of the freedom. Formally,

$$E\left[\frac{(n-1)s^2}{\sigma^2}\right] = n-1 \tag{1}$$

$$\operatorname{var}\left[\frac{(n-1)s^2}{\sigma^2}\right] = 2(n-1) \tag{2}$$

It may seem somewhat strange to talk about the variance of the sample variance, but under repeated sampling the sample variance will vary from sample to sample, and the chi-square distribution describes this variation if the observations are from a normal distribution.

Unlike the normal distribution, a tabulation of the chi-square distribution requires a separate listing for each degree of freedom. In Table A.3, a tabulation is presented of percentiles of the chi-square distribution. For example, 95% of chi-square random variables with 10 degrees of freedom have values less than or equal to 18.31. Note that the median (50th percentile) is very close to the degrees of freedom when the number of the degrees of freedom is 10 or more.

The symbol for a chi-square random variable is χ^2, the Greek lowercase letter chi, to the power of 2. So we usually write $\chi^2 = (n-1)s^2/\sigma^2$. The degrees of freedom are usually indicated by the Greek lowercase letter ν (nu). Hence, χ^2_ν is a symbol for a chi-square random variable with ν degrees of freedom. It is not possible to maintain the notation of using a capital letter for a variable and the corresponding lowercase letter for the value of the variable.

4.8.2 Inference about a Population Variance

We begin with hypothesis testing. We have a sample of size n from a normal distribution, the sample variance s^2 has been calculated, and we want to know whether the value of s^2 observed is consistent with a hypothesized population value σ_0^2, perhaps known from previous research. Consider the quantity

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

If s^2 is very close to σ^2, the ratio s^2/σ^2 is close to 1; if s^2 differs very much from σ^2, the ratio is either very large or very close to 0: This implies that $\chi^2 = (n-1)s^2/\sigma^2$ is either very large or very small, and we would want to reject the null hypothesis. This procedure is analogous to a hypothesis test about a population mean; we measured the distance of the observed sample mean from the hypothesized value in units of standard errors; in this case we measure the "distance" in units of the hypothesized variance.

Example 4.9. The SIDS cases discussed in Section 3.3.1 were assumed to come from a normal population with variance $\sigma^2 = (800)^2$. To check this assumption, the variance, s^2, is calculated for the first 11 cases occurring in 1969. The birthweights (in grams) were

$$3374, 3515, 3572, 2977, 4111, 1899, 3544, 3912, 3515, 3232, 3289$$

The sample variance is calculated to be

$$s^2 = (574.3126 \text{ g})^2$$

The observed value of the chi-square quantity is

$$\begin{aligned}\chi^2 &= \frac{(11-1)(574.3126)^2}{(800)^2} \\ &= 5.15 \text{ with 10 degrees of freedom}\end{aligned}$$

Figure 4.14 illustrates the chi-square distribution with 10 degrees of freedom. The 2.5th and 97.5th percentiles are 3.25 and 20.48 (see Table A.3). Hence, 95% of chi-square values will fall between 3.25 and 20.48.

If we follow the usual procedure of setting our significance level at $\alpha = 0.05$, we will not reject the null hypothesis that $\sigma^2 = (800 \text{ g})^2$, since the observed value $\chi^2 = 5.15$ is less extreme than 3.25. Hence, there is not sufficient evidence for using a value of σ^2 not equal to 800 g.

As an alternative to setting up the rejection regions formally, we could have noted, using Table A.3, that the observed value of $\chi^2 = 5.15$ is between the 5th and 50th percentiles, and therefore the corresponding two-sided p-value is greater than 0.10.

A $100(1-\alpha)\%$ confidence interval is constructed using the approach of Section 4.7. The key inequality is

$$P[\chi^2_{\alpha/2} \leq \chi^2 \leq \chi^2_{1-\alpha/2}] = 1 - \alpha$$

The degrees of freedom are not indicated but assumed to be $n-1$. The values $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are chi-square values such that $1-\alpha$ of the area is between them. (In Figure 4.14, these values are 3.25 and 20.48 for $1-\alpha = 0.95$.)

The quantity χ^2 is now replaced by its equivalent, $(n-1)s^2/\sigma^2$, so that

$$P\left[\chi^2_{\alpha/2} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{1-\alpha/2}\right] = 1 - \alpha$$

If we solve for σ^2, we obtain a $100(1-\alpha)\%$ confidence interval for the population variance. A little algebra shows that this is

$$P\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right] = 1 - \alpha$$

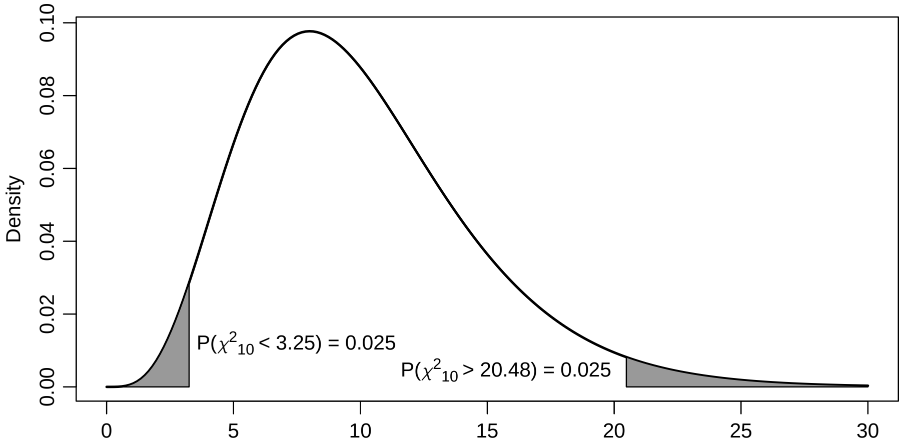


Figure 4.28 Chi-square distribution with 10 degrees of freedom.

Given an observed value of s^2, the confidence interval required can now be calculated.

To continue our example, the variance for the 11 SIDS cases above is $s^2 = (574.3126\text{ g})^2$. For $1 - \alpha = 0.95$, the values of χ^2 are (see Figure 4.28)

$$\chi^2_{0.025} = 3.25, \qquad \chi^2_{0.975} = 20.48$$

We can write the key inequality then as

$$P[3.25 \leq \chi^2 \leq 20.48] = 0.95$$

The 95% confidence interval for σ^2 can then be calculated:

$$\frac{(10)(574.3126)^2}{20.48} \leq \sigma^2 \leq \frac{(10)(574.3126)^2}{3.25}$$

and simplifying yields

$$161{,}052 \leq \sigma^2 \leq 1{,}014{,}877$$

The corresponding values for the population standard deviation are

$$\text{lower 95\% limit for } \sigma = \sqrt{161,052} = 401 \text{ g}$$

$$\text{upper 95\% limit for } \sigma = \sqrt{1{,}014{,}877} = 1007 \text{ g}$$

These are rather wide limits. Note that they include the null hypothesis value of $\sigma = 800$ g. Thus, the confidence interval approach leads to the same conclusion as the hypothesis-testing approach.

NOTES

4.1 Definition of Probability

The relative frequency definition of probability was advanced by von Mises, Fisher, and others (see Hacking [1965]). A radically different view is held by the *personal* or *subjective school*, exemplified in the work of De Finetti, Savage, and Savage. According to this school, probability reflects subjective belief and knowledge that can be quantified in terms of betting behavior. Savage [1968] states: "My probability for the event A under circumstances H is the amount of money I am indifferent to betting on A in an elementary gambling situation." What does Savage mean? Consider the thumbtack experiment discussed in Section 4.3.1. Let the event A be that the thumbtack in a single toss falls $\perp$. The other possible outcome is $\top$; call this event B. You are to bet a dollars on A and b dollars on B, such that you are indifferent to betting either on A or on B (you must bet). You clearly would not want to put all your money on A; then you would prefer outcome A. There is a split, then, in the total amount, $a + b$, to be bet so that you are indifferent to either outcome A or B. Then *your* probability of A, $P[A]$, is

$$P[A] = \frac{b}{a+b}$$

If the total amount to be bet is 1 unit, you would split it $1 - P$, P, where $0 \leq P \leq 1$, so that

$$P[A] = \frac{P}{1 - P + P} = P$$

The bet is a device to link quantitative preferences for amounts b and a of money, which are assumed to be well understood, to preferences for degrees of certainty, which we are trying to quantify. Note that Savage is very careful to require the estimate of the probability to be made under as specified circumstances. (If the thumbtack could land, say, $\top$ on a soft surface, you would clearly want to modify your probability.) Note also that betting behavior is a *definition* of personal probability rather than a guide for action. In practice, one would typically work out personal probabilities by comparison to events for which the probabilities were already established (Do I think this event is more or less likely than a coin falling heads?) rather than by considering sequences of bets.

This definition of probability is also called *personal probability*. An advantage of this view is that it can discuss more situations than the relative frequency definition, for example: the probability (rather, *my* probability) of life on Mars, or my probability that a cure for cancer will be found. You should not identify personal probability with the irrational or whimsical. Personal probabilities do utilize empirical evidence, such as the behavior of a tossed coin. In particular, if you have good reason to believe that the relative frequency of an event is P, your personal probability will also be P. It is possible to show that any self-consistent system for choosing between uncertain outcomes corresponds to a set of personal probabilities.

Although different individuals will have different personal probabilities for an event, the way in which those probabilities are updated by evidence is the same. It is possible to develop statistical analyses that summarize data in terms of how it should change one's personal probabilities. In simple analyses these *Bayesian methods* are more difficult to use than those based on relative frequencies, but the situation is reversed for some complex models. The use of Bayesian statistics is growing in scientific and clinical research, but it is still not supported by most standard software. An introductory discussion of Bayesian statistics is given by Berry [1996], and more advanced books on practical data analysis include Gelman et al. [1995] and Carlin and Louis [2000]. There are other views of probability. For a survey, see the books by Hacking [1965] and Barnett [1999] and references therein.

4.2 Probability Inequalities

For the normal distribution, approximately 68% of observations are within one standard deviation of the mean, and 95% of observations are within two standard deviations of the mean. If the distribution is not normal, a weaker statement can be made: The proportion of observations within K standard deviations of the mean is greater than or equal to $(1 - 1/K^2)$; notationally, for a variable Y,

$$P\left[-K \leq \frac{Y - E(Y)}{\sigma} \leq K\right] \leq 1 - \frac{1}{K^2}$$

where K is the number of standard deviations from the mean. This is a version of *Chebyshev's inequality*. For example, this inequality states that at least 75% of the observations fall within two standard deviations of the mean (compared to 95% for the normal distribution). This is not nearly as stringent as the first result stated, but it is more general. If the variable Y can take on only positive values and the mean of Y is μ, the following inequality holds:

$$P[Y \leq y] \leq 1 - \frac{\mu}{y}$$

This inequality is known as the *Markov inequality.*

4.3 Inference vs. Decision

The hypothesis tests discussed in Sections 4.6 and 4.7 can be thought of as decisions that are made with respect to a value of a parameter (or *state of nature*). There is a controversy in statistics as to whether the process of inference is equivalent to a decision process. It seems that a "decision" is sometimes not possible in a field of science. For example, it is not possible at this point to decide whether better control of insulin levels will reduce the risk of neuropathy in diabetes mellitus. In this case and others, the types of inferences we can make are more tenuous and cannot really be called decisions. For an interesting discussion, see Moore [2001]. This is an excellent book covering a variety of statistical topics ranging from ethical issues in experimentation to formal statistical reasoning.

4.4 Representative Samples

A random sample from a population was defined in terms of repeated independent trials or drawings of observations. We want to make a distinction between a random and a representative sample. A random sample has been defined in terms of repeated independent sampling from a population. However (see Section 4.3.2), cancer patients treated in New York are clearly not a random sample of all cancer patients in the world or even in the United States. They will differ from cancer patients in, for instance, Great Britain in many ways. Yet we do frequently make the assumption that if a cancer treatment worked in New York, patients in Great Britain can also benefit. The experiment in New York has wider applicability. We consider that with respect to the outcome of interest in the New York cancer study (e.g., increased survival time), the New York patients, although not a random sample, constitute a representative sample. That is, the survival times are a random sample from the population of survival times.

It is easier to disprove randomness than representativeness. A measure of scientific judgment is involved in determining the latter. For an interesting discussion of the use of the word *representative*, see the papers by Kruskal and Mosteller [1979a–c].

4.5 Multivariate Populations

Usually, we study more than one variable. The Winkelstein et al. [1975] study (see Example 4.1) measured diastolic and systolic blood pressures, height, weight, and cholesterol levels. In the study suggested in Example 4.2, in addition to IQ, we would measure physiological and psychological variables to obtain a more complete picture of the effect of the diet. For completeness we therefore define a *multivariate population* as the set of all possible values of a specified set of variables (measured on the objects of interest). A second category of topics then comes up:

relationships among the variables. Words such as *association* and *correlation* come up in this context. A discussion of these topics begins in Chapter 9.

4.6 Sampling without Replacement

We want to select two patients *at random* from a group of four patients. The same patient cannot be chosen twice. How can this be done? One procedure is to write each name on a slip of paper, put the four slips of paper in a hat, stir the slips of paper, and—without looking—draw out two slips. The patients whose names are on the two slips are then selected. This is known as *sampling without replacement.* (For the procedure to be *fair*, we require that the slips of paper be indistinguishable and well mixed.) The events "outcome on first draw" and "outcome on second draw" are clearly not independent. If patient A is selected in the first draw, she is no longer available for the second draw. Let the patients be labeled *A, B, C*, and *D.* Let the symbol *AB* mean "patient A is selected in the first draw and patient B in the second draw." Write down all the possible outcomes; there are 12 of them as follows:

$$\begin{array}{cccc} AB & BA & CA & DA \\ AC & BC & CB & DB \\ AD & BD & CD & DC \end{array}$$

We define the selection of two patients to be random if each of the 12 outcomes is equally likely, that is, the probability that a particular pair is chosen is 1/12. This definition has intuitive appeal: We could have prepared 12 slips of paper each with one of the 12 pairs recorded and drawnout one slip of paper. If the slip of paper is drawn randomly, the probability is 1/12 that a particular slip will be selected.

One further comment. Suppose that we only want to know which two patients have been selected (i.e., we are not interested in the order). For example, what is the probability that patients C and D are selected? This can happen in two ways: CD or DC. These events are mutually exclusive, so that the required probability is $P[CD \text{ or } DC] = P[CD] + P[DC] = 1/12 + 1/12 = 1/6$.

4.7 Pitfalls in Sampling

It is very important to define the population of interest carefully. Two illustrations of rather subtle pitfalls are Berkson's fallacy and length-biased sampling. *Berkson's fallacy* is discussed in Murphy [1979] as follows: In many studies, hospital records are reviewed or sampled to determine relationships between diseases and/or exposures. Suppose that a review of hospital records is made with respect to two diseases, A and B, which are so severe that they always lead to hospitalization. Let their frequencies in the population at large be p_1 and p_2. Then, assuming independence, the probability of the joint occurrence of the two diseases is p_1p_2. Suppose now that a healthy proportion p_3 of subjects (H) never go to the hospital; that is, $P[H] = p_3$. Now write $\overline{H}$ as that part of the population that will enter a hospital at some time; then $P[\overline{H}] = 1 - p_3$. By the rule of conditional probability, $P[A|\overline{H}] = P[A\overline{H}]/P[\overline{H}] = p_1/(1-p_3)$. Similarly, $P[B|\overline{H}] = p_2/(1-p_3)$ and $P[AB|\overline{H}] = p_1p_2/(1-p_3)$, and this is not equal to $P[A|\overline{H}]P[B|\overline{H}] = [p_1/(1-p_3)][p_2/(1-p_3)]$, which must be true in order for the two diseases to be unrelated in the hospital population. Now, you can show that $P[AB|\overline{H}] < P[AB]$, and, quoting Murphy:

> The hospital observer will find that they occur together less commonly than would be expected if they were independent. This is known as Berkson's fallacy. It has been a source of embarrassment to many an elegant theory. Thus, cirrhosis of the liver and common cancer are both reasons for admission to the hospital. *A priori*, we would expect them to be less commonly associated in the hospital than in the population at large. In fact, they have been found to be negatively correlated.

Table 4.4 Expected Composition of Visit-Based Sample in a Hypothetical Population

	Type of Patient		
Variable	Hypertensive	Other	Total
Number of patients	200	800	1000
Visits per patient per year	12	1	13
Visits contributed	2400	800	3200
Expected number of patients in a 3% sample of visits	72	24	96
Expected percent of sample	75	25	100

Source: Shepard and Neutra [1977].

(Murphy's book contains an elegant, readable exposition of probability in medicine; it will be worth your while to read it.)

A second pitfall deals with the area of *length-biased sampling*. This means that for a particular sampling scheme, some objects in the population may be more likely to be selected than others. A paper by Shepard and Neutra [1977] illustrates this phenomenon in sampling medical visits. Our discussion is based on that paper. The problem arises when we want to make a statement about a population of patients that can only be identified by a sample of patient visits. Therefore, frequent visitors will be more likely to be selected. Consider the data in Table 4.4, which illustrates that although hypertensive patients make up 20% of the total patient population, a sample based on visits would consist of 75% hypertensive patients and 25% other.

There are other areas, particularly screening procedures in chronic diseases, that are at risk for this type of problem. See Shepard and Neutra [1977] for suggested solutions as well as references to other papers.

4.8 Other Sampling Schemes

In this chapter (and almost all the remainder of the book) we are assuming *simple random sampling*, that is, sampling where every unit in the population is equally likely to end up in the sample, and sampling of different units is independent. A sufficiently large simple random sample will always be representative of the population. This intuitively plausible result is made precise in the mathematical result that the empirical cumulative distribution of the sample approaches the true cumulative distribution of the population as the sample size increases.

There are some important cases where other random sampling strategies are used, trading increased mathematical complexity for lower costs in obtaining the sample. The main techniques are as follows:

1. *Stratified sampling*. Suppose that we sampled 100 births to study low birthweight. We would expect to see about one set of twins on average, but might be unlucky and not sample any. As twins are much more likely to have low birthweight, we would prefer a sampling scheme that fixed the number of twins we observed.
2. *Unequal probability sampling*. In conjunction with stratified sampling, we might want to increase the number of twin births that we examined to more than the 1/90 in the population. We might decide to sample 10 twin births rather than just one.
3. *Cluster sampling*. In a large national survey requiring face-to-face interviews or clinical tests, it is not feasible to use a simple random sample, as this would mean that nearly every person sampled would live in a different town or city. Instead, a number of cities or counties might be sampled and simple random sampling used within the selected geographic regions.

4. *Two-phase sampling*. It is sometimes useful to take a large initial sample and then take a smaller subsample to measure more expensive or difficult variables. The probability of being included in the subsample can then depend on the values of variables measured at the first stage. For example, consider a study of genetic influences on lung cancer. Lung cancer is rare, so it would be sensible to use a stratified (case–control) sampling scheme where an equal number of people with and without lung cancer was sampled. In addition, lung cancer is extremely rare in nonsmokers. If a first-stage sample asked about smoking status it would be possible to ensure that the more expensive genetic information was obtained for a sufficient number of nonsmoker cancer cases as well as smokers with cancer.

These sampling schemes have two important features in common. The sampling scheme is fully known in advance, and the sampling is random (even if not with equal probabilities). These features mean that a valid statistical analysis of the results is possible. Although the sample is not representative of the population, it is unrepresentative in ways that are fully under the control of the analyst. Complex probability samples such as these require different analyses from simple random samples, and not all statistical software will analyze them correctly. The section on Survey Methods of the American Statistical Association maintains a list of statistical software that analyzes complex probability samples. It is linked from the Web appendix to this chapter. There are many books discussing both the statistical analysis of complex surveys and practical considerations involved in sampling, including Levy and Lemeshow [1999], Lehtonen and Pahkinen [1995], and Lohr [1999]. Similar, but more complex issues arise in environmental and ecological sampling, where measurement locations are sampled from a region.

4.9 How to Draw a Random Sample

In Note 4.6 we discussed drawing a random sample without replacement. How can we draw samples with replacement? Simply, of course, the slips could be put back in the hat. However, in some situations we cannot collect the total population to be sampled from, due to its size, for example. One way to sample populations is to use a table of random numbers. Often, these numbers are really *pseudorandom*: They have been generated by a computer. Use of such a table can be illustrated by the following problem: A random sample of 100 patient charts is to be drawn from a hospital record room containing 45,850 charts. Assume that the charts are numbered in some fashion from 1 to 45,850. (It is not necessary that they be numbered consecutively or that the numbers start with 1 and end with 45,850. All that is required is that there is some unique way of numbering each chart.) We enter the random number table randomly by selecting a page and a column on the page at random. Suppose that the first five-digit numbers are

06812, 16134, 15195, 84169, and 41316

The first three charts chosen would be chart 06812, 16134, and 15195, in that order. Now what do we do with the 84169? We can skip it and simply go to 41316, realizing that if we follow this procedure, we will have to throw out approximately half of the numbers selected.

A second example: A group of 40 animals is to be assigned at random to one of four treatments A, B, C, and D, with an equal number in each of the treatments. Again, enter the random number table randomly. The first 10-digit numbers between 1 and 40 will be the numbers of the animals assigned to treatment A, the second set of 10-digit numbers to treatment B, the third set to treatment C, and the remaining animals are assigned to treatment D. If a random number reappears in a subsequent treatment, it can simply be omitted. (Why is this reasonable?)

4.10 Algebra of Expectations

In Section 4.3.3 we discuss random variables, distributions, and expectations of random variables. We defined $E(Y) = \sum py$ for a discrete random variable. A similar definition, involving integrals rather than sums, can be made for continuous random variables. We will now state some rules for working with expectations.

1. If a is a constant, $E(aY) = aE(Y)$.
2. If a and b are constants, $E(aY + b) = aE(Y) + b$.
3. If X and Y are two random variables, $E(X + Y) = E(X) + E(Y)$.
4. If a and b are constants, $E(aX + bY) = E(aX) + E(bY) = aE(X) + bE(Y)$.

You can demonstrate the first three rules by using some simple numbers and calculating their average. For example, let $y_1 = 2$, $y_2 = 4$, and $y_3 = 12$. The average is

$$E(Y) = \frac{1}{3} \times 2 + \frac{1}{3} \times 4 + \frac{1}{3} \times 12 = 6$$

Two additional comments:

1. The second formula makes sense. Suppose that we measure temperature in °C. The average is calculated for a series of readings. The average can be transformed to °F by the formula

   $$\text{average in } {}^\circ\text{F} = \frac{9}{5} \times \text{average in } {}^\circ\text{C} + 32$$

   An alternative approach consists of transforming each original reading to °F and then taking the average. It is intuitive that the two approaches should provide the same answer.
2. It is not true that $E(Y^2) = [E(Y)]^2$. Again, a small example will verify this. Use the same three values ($y_1 = 2$, $y_2 = 4$, and $y_3 = 12$). By definition,

   $$E(Y^2) = \frac{2^2 + 4^2 + 12^2}{3} = \frac{4 + 16 + 144}{3} = \frac{164}{3} = 54.\overline{6}$$

   but

   $$[E(Y)]^2 = 6^2 = 36$$

   Can you think of a special case where the equation $E(Y^2) = [E(Y)]^2$ is true?

4.11 Bias, Precision, and Accuracy

Using the algebra of expectations, we define a statistic T to be a biased estimate of a parameter τ if $E(T) \neq \tau$. Two typical types of bias are $E(T) = \tau + a$, where a is a constant, called *location bias*; and $E(T) = b\tau$, where b is a positive constant, called *scale bias*. A simple example involves the sample variance, s^2. A more "natural" estimate of σ^2 might be

$$s_*^2 = \frac{\sum(y - \overline{y})^2}{n}$$

This statistic differs from the usual sample variance in division by n rather than $n - 1$. It can be shown (you can try it) that

$$E(s_*^2) = \frac{n-1}{n}\sigma^2$$

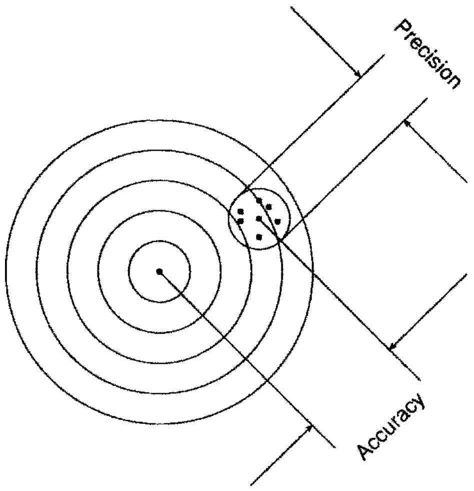


Figure 4.29 Accuracy involves the concept of bias.

Hence, s_*^2 is a biased estimate of σ^2. The statistic s_*^2 can be made unbiased by multiplying s_*^2 by $n/(n-1)$ (see rule 1 in Note 4.10); that is,

$$E\left[\frac{n}{n-1}s_*^2\right] = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2$$

But $n/(n-1)s_*^2 = s^2$, so s^2 rather than s_*^2 is an unbiased estimate of σ^2. We can now discuss precision and accuracy. *Precision* refers to the degree of closeness to each other of a set of values of a variable; *accuracy* refers to the degree of closeness of these values to the quantity (parameter) being measured. Thus, precision is an internal characteristic of a set of data, while accuracy relates the set to an external standard. For example, a thermometer that consistently reads a temperature 5 degrees too high may be very precise but will not be very accurate. A second example of the distribution of hits on a target illustrates these two concepts. Figure 4.29 shows that accuracy involves the concept of bias. Together with Note 4.10, we can now make these concepts more precise. For simplicity we will refer only to location bias.

Suppose that a statistic T estimates a quantity τ in a biased way; $E[T] = \tau + a$. The variance in this case is defined to be $E[T - E(T)]^2$. What is the quantity $E[T-\tau]^2$? This can be written as

$$E[T-\tau]^2 = E[T-(\tau+a)+a]^2 = E[T-E[T]+a]^2$$

$$\underset{\text{(mean square error)}}{E[T-\tau]^2} = \underset{\text{(variance)}}{E[T-E[T]]^2} + \underset{\text{(bias)}}{a^2}$$

The quantity $E[T-\tau]^2$ is called the *mean square error*. If the statistic is unbiased (i.e., $a = 0$), the mean square error is equal to the variance (σ^2).

4.12 Use of the Word Parameter

We have defined *parameter* as a numerical characteristic of a population of values of a variable. One of the basic tasks of statistics is to estimate values of the unknown parameter on the basis of a sample of values of a variable. There are two other uses of this word. Many clinical scientists use *parameter* for *variable*, as in: "We measured the following three parameters: blood pressure, amount of plaque, and degree of patient satisfaction." You should be aware of this pernicious use and strive valiantly to eradicate it from scientific writing. However, we are not sanguine about its ultimate success. A second incorrect use confuses *parameter* and *perimeter*, as in: "The parameters of the study did not allow us to include patients under 12 years of age." A better choice would have been to use the word *limitations*.

4.13 Significant Digits (continued)

This note continues the discussion of significant digits in Note 3.4. We discussed approximations to a quantity due to arithmetical operations, measurement rounding, and finally, sampling variability. Consider the data on SIDS cases of Example 4.11. The mean birthweight of the 78 cases was 2994 g. The probability was 95% that the interval 2994 ± 178 straddles the unknown quantity of interest: the mean birthweight of the population of SIDS cases. This interval turned out to be 2816–3172 g, although the last digits in the two numbers are not very useful. In this case we have carried enough places so that the rule mentioned in Note 3.4 is not applicable. The biggest source of approximation turns out to be due to sampling. The approximations introduced by the arithmetical operation is minimal; you can verify that if we had carried more places in the intermediate calculations, the final confidence interval would have been 2816–3171 g.

4.14 A Matter of Notation

What do we mean by 18 ± 2.6? In many journals you will find this notation. What does it mean? Is it mean plus or minus the standard deviation, or mean plus or minus the standard error? You may have to read a paper carefully to find out. Both meanings are used and thus need to be specified clearly.

4.15 Formula for the Normal Distribution

The formula for the normal probability density function for a normal random variable Y with mean μ and variance σ^2 is

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

Here, $\pi = 3.14159\ldots$, and e is the base of the natural logarithm, $e = 2.71828\ldots$. A standard normal distribution has $\mu = 0$ and $\sigma = 1$. The formula for the standard normal random variable, Z, is

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)$$

Although most statistical packages will do this for you, the heights of the curve can easily be calculated using a hand calculator. By symmetry, only one half of the range of values has to be computed [i.e., $f(z) = f(-z)$]. For completeness in Table 4.5 we give enough points to enable you to graph $f(z)$. Given any normal variable y with mean μ and variance σ^2, you can calculate $f(y)$ by using the relationships

$$Z = \frac{Y-\mu}{\sigma}$$

and plotting the corresponding heights:

$$f(y) = \frac{1}{\sigma} f(z)$$

where Z is defined by the relationship above. For example, suppose that we want to graph the curve for IQ, where we assume that IQ is normal with mean $\mu = 100$ and standard deviation

Table 4.5 Heights of the Standard Normal Curve

z	f(z)	z	f(z)	z	f(z)	z	f(z)	z	f(z)
0.0	0.3989	0.5	0.3521	1.0	0.2420	1.5	0.1295	2.0	0.0540
0.1	0.3970	0.6	0.3332	1.1	0.2179	1.6	0.1109	2.1	0.0440
0.2	0.3910	0.7	0.3123	1.2	0.1942	1.7	0.0940	2.2	0.0355
0.3	0.3814	0.8	0.2897	1.3	0.1714	1.8	0.0790	2.3	0.0283
0.4	0.3683	0.9	0.2661	1.4	0.1497	1.9	0.0656	2.4	0.0224

$\sigma = 15$. What is the height of the curve for an IQ of 109? In this case, $Z = (109-100)/15 = 0.60$ and $f(\text{IQ}) = (1/15)f(z) = (1/15)(0.3332) = 0.0222$. The height for an IQ of 91 is the same.

4.16 Null Hypothesis and Alternative Hypothesis

How do you decide which of two hypotheses is the null and which is the alternative? Sometimes the advice is to make the null hypothesis the hypothesis of "indifference." This is not helpful; indifference is a poor scientific attitude. We have three suggestions: (1) In many situations there is a prevailing view of the science that is accepted; it will continue to be accepted unless "definitive" evidence to the contrary is produced. In this instance the prevailing view would be made operational in the null hypothesis. The null hypothesis is often the "straw man" that we wish to reject. (Philosophers of science tell us that we never prove things conclusively; we can only disprove theories.) (2) An excellent guide is *Occam's razor*, which states: Do not multiply hypotheses beyond necessity. Thus, in comparing a new treatment with a standard treatment, the simpler hypothesis is that the treatments have the same effect. To postulate that the treatments are different requires an additional operation. (3) Frequently, the null hypothesis is one that allows you to calculate the p-value. Thus, if two treatments are assumed the same, we can calculate a p-value for the result observed. If we hypothesize that they are not the same, then we cannot compute a p-value without further specification.

PROBLEMS

4.1 Give examples of populations with the number of elements finite, virtually infinite, potentially infinite, and infinite. Define a sample from each population.

4.2 Give an example from a study in a research area of interest to you that clearly assumes that results are applicable to, as yet, untested subjects.

4.3 Illustrate the concepts of *population, sample, parameter*, and *statistic* by two examples from a research area of your choice.

4.4 In light of the material discussed in this chapter, now review the definitions of statistics presented at the end of Chapter 1, especially the definition by Fisher.

4.5 In Section 4.3.1, probabilities are defined as long-run relative frequencies. How would you interpret the probabilities in the following situations?

(a) The probability of a genetic defect in a child born to a mother over 40 years of age.

(b) The probability of you, the reader, dying of leukemia.

(c) The probability of life on Mars.

(d) The probability of rain tomorrow. What does the meteorologist mean?

4.6 Take a thumbtack and throw it onto a hard surface such as a tabletop. It can come to rest in two ways; label them as follows:

$$\perp = \text{up} = U$$

$$\top = \text{down} = D$$

(a) Guess the probability of U. Record your answer.

(b) Now toss the thumbtack 100 times and calculate the proportion of times the outcome is U. How does this agree with your guess? The observed proportion is an estimate of the probability of U. (Note the implied distinction between *guess* and *estimate*.)

(c) In a class situation, split the class in half. Let each member of the first half of the class toss a thumbtack 10 times and record the outcomes as a histogram: (i) the number of times that U occurs in 10 tosses; and (ii) the proportion of times that U occurs in 10 tosses. Each member of the second half of the class will toss a thumbtack 50 times. Record the outcomes in the same way. Compare the histograms. What conclusions do you draw?

4.7 The estimation of probabilities and the proper combination of probabilities present great difficulties, even to experts. The best we can do in this book is warn you and point you to some references. A good starting point is the paper by Tversky and Kahneman [1974] reprinted in Kahneman et al. [1982]. They categorize the various errors that people make in assessing and working with probabilities. Two examples from this book will test your intuition:

(a) In tossing a coin six times, is the sequence HTHHTT more likely than the sequence HHHHHH? Give your "first impression" answer, then calculate the probability of occurrence of each of the two sequences using the rules stated in the chapter.

(b) The following is taken directly from the book:

> A certain town is served by two hospitals. In the larger hospital, about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50%, sometimes lower. For a period of one year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days? The larger hospital, the smaller hospital, [or were they] about the same (that is, within 5% of each other)?

Which of the rules and results stated in this chapter have guided your answer?

4.8 This problem deals with the *gambler's fallacy*, which states, roughly, that if an event has not happened for a long time, it is "bound to come up." For example, the probability of a head on the fifth toss of a coin is assumed to be greater if the preceding four tosses all resulted in tails than if the preceding four tosses were all heads. This is incorrect.

(a) What statistical property associated with coin tosses is violated by the fallacy?

(b) Give some examples of the occurrence of the fallacy from your own area of research.

(c) Why do you suppose that the fallacy is so ingrained in people?

4.9 Human blood can be classified by the ABO blood grouping system. The four groups are A, B, AB, or O, depending on whether antigens labeled A and B are present on red blood cells. Hence, the AB blood group is one where both A and B antigens are present; the O group has none of the antigens present. For three U.S. populations, the following distributions exist:

	Blood Group				
	A	**B**	**AB**	**O**	**Total**
Caucasian	0.44	0.08	0.03	0.45	1.00
American black	0.27	0.20	0.04	0.49	1.00
Chinese	0.22	0.25	0.06	0.47	1.00

For simplicity, consider only the population of American blacks in the following question. The table shows that for a person selected randomly from this population, $P[\text{A}] = 0.27$, $P[\text{B}] = 0.20$, $P[\text{AB}] = 0.04$, and $P[\text{O}] = 0.49$.

(a) Calculate the probability that a person is *not* of blood group A.

(b) Calculate the probability that a person is either A *or* O. Are these mutually exclusive events?

(c) What is the probability that a person carries A antigens?

(d) What is the probability that in a marriage both husband and wife are of blood group O? What rule of probability did you use? (What assumption did you need to make?)

4.10 This problem continues with the discussion of ABO blood groups of Problem 4.9. We now consider the black and Caucasian population of the United States. Approximately 20% of the U.S. population is black. This produces the following two-way classification of race and blood type:

	Blood Group				
	A	**B**	**AB**	**O**	**Total**
Caucasian	0.352	0.064	0.024	0.360	0.80
American black	0.054	0.040	0.008	0.098	0.20
Total	0.406	0.104	0.032	0.458	1.00

This table specifies, for example, that the probability is 0.352 that a person selected at random is both Caucasian and blood group A.

(a) Are the events "blood group A" and "Caucasian race" statistically independent?

(b) Are the events "blood group A" and "Caucasian race" mutually exclusive?

(c) Assuming statistical independence, what is the expected probability of the event "blood group A and Caucasian race"?

(d) What is the conditional probability of "blood group A" given that the race is Caucasian?

4.11 The distribution of the Rh factor in a Caucasian population is as follows:

Rh Positive (Rh^+, Rh^+)	**Rh Positive (Rh^+, Rh^-)**	**Rh Negative**
0.35	0.48	0.17

Rh^- subjects have two Rh^- genes, while Rh^+ subjects have two Rh^+ genes or one Rh^+ gene and one Rh^- gene. A potential problem occurs when a Rh^+ male mates with an Rh^- female.

(a) Assuming random mating with respect to the Rh factor, what is the probability of an Rh^- female mating with an Rh^+ male?

(b) Since each person contributes one gene to an offspring, what is the probability of Rh incompatibility given such a mating? (Incompatibility occurs when the fetus is Rh^+ and the mother is Rh^-.)

(c) What is the probability of incompatibility in a population of such matings?

4.12 The following data for 20- to 25-year-old white males list four primary causes of death together with a catchall fifth category, and the probability of death within five years:

Cause	**Probability**
Suicide	0.00126
Homicide	0.00063
Auto accident	0.00581
Leukemia	0.00023
All other causes	0.00788

(a) What is the probability of a white male aged 20 to 25 years dying from *any* cause of death? Which rule did you use to determine this?

(b) Out of 10,000 white males in the 20 to 25 age group, how many deaths would you expect in the next five years? How many for each cause?

(c) Suppose that an insurance company sells insurance to 10,000 white male drivers in the 20 to 25 age bracket. Suppose also that each driver is insured for \$100,000 for accidental death. What annual rate would the insurance company have to charge to break even? (Assume a fatal accident rate of 0.00581.) List some reasons why your estimate will be too low or too high.

(d) Given that a white male aged 20 to 25 years has died, what is the most likely cause of death? Assume nothing else is known. Can you explain your statement?

4.13 If $Y \sim N\ (0,1)$, find

(a) $P[Y \leq 2]$

(b) $P[Y \leq -1]$

(c) $P[Y > 1.645]$

(d) $P[0.4 < Y \leq 1]$

(e) $P[Y \leq -1.96 \text{ or } Y \geq 1.96] = P[|Y| \geq 1.96]$

4.14 If $Y \sim N$ (2,4), find

(a) $P[Y \leq 2]$

(b) $P[Y \leq 0]$

(c) $P[1 \leq Y < 3]$

(d) $P[0.66 < Y \leq 2.54]$

4.15 From the paper by Winkelstein et al. [1975], glucose data for the 45 to 49 age group of California Nisei as presented by percentile are:

Percentile	90	80	70	60	50	40	30	20	10
Glucose (mg/100 mL)	218	193	176	161	148	138	128	116	104

(a) Plot these data on normal probability paper connecting the data points by straight lines. Do the data seem normal?

(b) Estimate the mean and standard deviation from the plot.

(c) Calculate the median and the interquartile range.

4.16 In a sample of size 1000 from a normal distribution, the sample mean $\overline{Y}$ was 15, and the sample variance s^2 was 100.

(a) How many values do you expect to find between 5 and 45?

(b) How many values less than 5 or greater than 45 do you expect to find?

4.17 Plot the data of Table 3.8 on probability paper. Do you think that age at death for these SIDS cases is normally distributed? Can you think of an a priori reason why this variable, age at death, is not likely to be normally distributed? Also make a QQ plot.

4.18 Plot the aflatoxin data of Section 3.2 on normal probability paper by graphing the cumulative proportions against the individual ordered values. Ignoring the last two points on the graph, draw a straight line through the remaining points and estimate the median. On the basis of the graph, would you consider the last three points in the data set *outliers*? Do you expect the arithmetic mean to be larger or smaller than the median? Why?

4.19 Plot the data of Table 3.12 (number of boys per family of eight children) on normal probability paper. Consider the endpoints of the intervals to be 0.5, 1.5, ... , 8.5. What is your conclusion about the normality of this variable? Estimate the mean and the standard deviation from the graph and compare it with the calculated values of 4.12 and 1.44, respectively.

4.20 The random variable Y has a normal distribution with mean 1.0 and variance 9.0. Samples of size 9 are taken and the sample means, $\overline{Y}$, are calculated.

(a) What is the sampling distribution of $\overline{Y}$?

(b) Calculate $P[1 < \overline{Y} \leq 2.85]$.

(c) Let $W = 4\overline{Y}$. What is the sampling distribution of W?

4.21 The sample mean and standard deviation of a set of temperature observations are 6.1°F and 3.0°F, respectively.

(a) What will be the sample mean and standard deviation of the observations expressed in °C?

(b) Suppose that the original observations are distributed with population mean μ°F and standard deviation σ°F. Suppose also that the sample mean of 6.1°F is based on 25 observations. What is the approximate sampling distribution of the mean? What are its parameters?

4.22 The frequency distributions in Figure 3.10 were based on the following eight sets of frequencies in Table 4.6.

Table 4.6 Sets of Frequencies for Figure 3.10

	Graph Number							
Y	1	2	3	4	5	6	7	8
−1	1	1	8	1	1	14	28	10
−2	2	2	8	3	5	11	14	24
−3	5	5	8	8	9	9	10	14
−4	10	9	8	11	14	6	8	10
−5	16	15	8	14	11	3	7	9
−6	20	24	8	15	8	2	6	7
−7	16	15	8	14	11	3	5	6
−8	10	9	8	11	14	6	4	4
−9	5	5	8	8	9	9	3	2
−10	2	2	8	3	5	11	2	1
−11	1	1	8	1	1	14	1	1
Total	88	88	88	88	88	88	88	88
a_4	3.03	3.20	1.78	2.38	1.97	1.36	12.1	5.78

(The numbers are used to label the graph for purposes of this exercise.) Obtain the probability plots associated with graphs 1 and 6.

4.23 Suppose that the height of male freshmen is normally distributed with mean 69 inches and standard deviation 3 inches. Suppose also (contrary to fact) that such subjects apply and are accepted at a college without regard to their physical stature.

(a) What is the probability that a randomly selected (male) freshman is 6 feet 6 inches (78 inches) or more?

(b) How many such men do you expect to see in a college freshman class of 1000 men?

(c) What is the probability that this class has at least one man who is 78 inches or more tall?

4.24 A normal distribution (e.g., IQ) has mean $\mu = 100$ and standard deviation $\sigma = 15$. Give limits within which 95% of the following would lie:

(a) Individual observations

(b) Means of 4 observations

(c) Means of 16 observations

(d) Means of 100 observations

(e) Plot the width of the interval as a function of the sample size. Join the points with an appropriate freehand line.

(f) Using the graph constructed for part (e), estimate the width of the 95% interval for means of 36 observations.

4.25 If the standard error is the measure of the precision of a sample mean, how many observations must be taken to double the precision of a mean of 10 observations?

4.26 The duration of gestation in healthy humans is approximately 280 days with a standard deviation of 10 days.

(a) What proportion of (healthy) pregnant women will be more than one week "overdue"? Two weeks?

(b) The gestation periods for a set of four women suffering from a particular condition are 240, 250, 265, and 280 days. Is this evidence that a shorter gestation period is associated with the condition?

(c) Is the sample variance consistent with the population variance of $10^2 = 100$? (We assume normality.)

(d) In view of part (c), do you want to reconsider the answer to part (b)? Why or why not?

4.27 The mean height of adult men is approximately 69 inches; the mean height of adult women is approximately 65 inches. The variance of height for both is 4^2 inches. Assume that husband–wife pairs occur without relation to height, and that heights are approximately normally distributed.

(a) What is the sampling distribution of the mean height of a couple? What are its parameters? (The variance of two statistically independent variables is the sum of the variances.)

(b) What proportion of couples is expected to have a mean height that exceeds 70 inches?

(c) In a collection of 200 couples, how many average heights would be expected to exceed 70 inches?

***(d)** In what proportion of couples do you expect the wife to be taller than the husband?

4.28 A pharmaceutical firm claims that a new analgesic drug relieves mild pain under standard conditions for 3 hours, with a standard deviation 1 hour. Sixteen patients are tested under the same conditions and have an average pain relief time of 2.5 hours. The hypothesis that the population mean of this sample is actually 3 hours is to be tested against the hypothesis that the population mean is in fact less than 3 hours; $\alpha = 0.05$.

(a) What is an appropriate test?

(b) Set up the appropriate critical region.

(c) State your conclusion.

(d) Suppose that the sample size is doubled. State precisely how the region where the null hypothesis is not rejected is changed.

***4.29** For Y, from a normal distribution with mean μ and variance σ^2, the variance of $\overline{Y}$, based on n observations, is σ^2/n. It can be shown that the sample median $\tilde{Y}$ in this situation has a variance of approximately $1.57\sigma^2/n$. Assume that the standard error of $\tilde{Y}$ equal to the standard error of $\overline{Y}$ is desired, based on $n = 10$; 20, 50, and 100 observations. Calculate the corresponding sample sizes needed for the median.

***4.30** To determine the strength of a digitalis preparation, a continuous intrajugular perfusion of a tincture is made and the dose required to kill an animal is observed. The lethal dose varies from animal to animal such that its logarithm is normally distributed. One cubic centimeter of the tincture kills 10% of all animals, 2 cm^3 kills 75%. Determine the mean and standard deviation of the distribution of the logarithm of the lethal dose.

4.31 There were 48 SIDS cases in King County, Washington, during the years 1974 and 1975. The birthweights (in grams) of these 48 cases were:

2466	3941	2807	3118	2098	3175	3515
3317	3742	3062	3033	2353	2013	3515
3260	2892	1616	4423	3572	2750	2807
2807	3005	3374	2722	2495	3459	3374
1984	2495	3062	3005	2608	2353	4394
3232	2013	2551	2977	3118	2637	1503
2438	2722	2863	2013	3232	2863	

(a) Calculate the sample mean and standard deviation for this set.

(b) Construct a 95% confidence interval for the population mean birthweight assuming that the population standard deviation is 800 g. Does this confidence interval include the mean birthweight of 3300 g for normal children?

(c) Calculate the p-value of the sample mean observed, assuming that the population mean is 3300 g and the population standard deviation is 800 g. Do the results of this part and part (b) agree?

(d) Is the sample standard deviation consistent with a population standard deviation of 800? Carry out a hypothesis test comparing the sample variance with population variance $(800)^2$. The critical values for a chi-square variable with 47 degrees of freedom are as follows:

$$\chi^2_{0.025} = 29.96, \qquad \chi^2_{0.975} = 67.82$$

(e) Set up a 95% confidence interval for the population standard deviation. Do this by first constructing a 95% confidence interval for the population variance and then taking square roots.

4.32 In a sample of 100 patients who had been hospitalized recently, the average cost of hospitalization was \$5000, the median cost was \$4000, and the modal cost was \$2500.

(a) What was the total cost of hospitalization for all 100 patients? Which statistic did you use? Why?

(b) List one practical use for *each* of the three statistics.

(c) Considering the ordering of the values of the statistics, what can you say about the distribution of the raw data? Will it be skewed or symmetric? If skewed, which way will the skewness be?

4.33 For Example 4.8, as discussed in Section 4.6.2:

(a) Calculate the probability of a Type II error and the power if α is fixed at 0.05.

(b) Calculate the power associated with a one-tailed test.

(c) What is the price paid for the increased power in part (b)?

4.34 The theory of hypothesis testing can be used to determine statistical characteristics of laboratory tests, keeping in mind the provision mentioned in connection with Example 4.6. Suppose that albumin has a normal (Gaussian) distribution in a healthy population with mean $\mu = 3.75$ mg per 100 mL and $\sigma = 0.50$ mg per 100 mL. The normal range of values will be defined as $\mu \pm 1.96\sigma$, so that values outside these limits will be classified as "abnormal." Patients with advanced chronic liver disease have reduced albumin levels; suppose that the mean for patients from this population is 2.5 mg per 100 mL and the standard deviation is the same as that of the normal population.

- **(a)** What are the critical values for the rejection region? (Here we work with an individual patient, $n = 1$.)
- **(b)** What proportion of patients with advanced chronic liver disease (ACLD) will have "normal" albumin test levels?
- **(c)** What is the probability that a patient with ACLD will be classified correctly on a test of albumin level?
- **(d)** Give an interpretation of Type I error, Type II error, and power for this example.
- **(e)** Suppose we consider only low albumin levels to be "abnormal." We want the same Type I error as above. What is the critical value now?
- **(f)** In part (e), what is the associated power?

4.35 This problem illustrates the power of probability theory.

- **(a)** Two SIDS infants are selected at random from a population of SIDS infants. We note their birthweights. What is the probability that both birthweights are (1) below the population median; (2) above the population median; (3) straddle the population median? The last interval is a nonparametric confidence interval.
- **(b)** Do the same as in part (a) for four SIDS infants. Do you see the pattern?
- **(c)** How many infants are needed to have interval 3 in part (a) have probability greater than 0.95?

REFERENCES

Barnett, V. [1999]. *Comparative Statistical Inference*. Wiley, Chichester, West Sussex, England.

Berkow, R. (ed.) [1999]. *The Merck Manual of Diagnosis and Therapy*, 17th ed. Merck, Rahway, NJ.

Berry, D. A. [1996]. *Statistics: A Bayesian Perspective*. Duxbury Press, North Scituate, MA.

Carlin, B. P., and Louis, T. A. [2000]. *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed. CRC Press, Boca Raton, FL.

Elveback, L. R., Guillier, L., and Keating, F. R., Jr. [1970]. Health, normality and the ghost of Gauss. *Journal of the American Medical Association*, **211**: 69–75.

Fisher, R. A. [1956]. *Statistical Methods and Scientific Inference*. Oliver & Boyd, London.

Fleiss, J. L., Levin, B., and Park, M. C. [2003]. *Statistical Methods for Rates and Proportions*, 3rd ed. Wiley, New York.

Galton, F. [1889]. *Natural Inheritance*. Macmillan, London.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. A. [1995]. *Bayesian Data Analysis.* CRC Press, Boca Raton, FL.

Golubjatnikov, R., Paskey, T., and Inhorn, S. L. [1972]. Serum cholesterol levels of Mexican and Wisconsin school children. *American Journal of Epidemiology*, **96**: 36–39.

Hacking, I. [1965]. *Logic of Statistical Inference*. Cambridge University Press, London.

Hagerup, L., Hansen, P. F., and Skov, F. [1972]. Serum cholesterol, serum-triglyceride and ABO blood groups in a population of 50-year-old Danish men and women. *American Journal of Epidemiology*, **95**: 99–103.

Kahneman, D., Slovic, P., and Tversky, A. (eds.) [1982]. *Judgement under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge.

Kato, H., Tillotson, J., Nichaman, M. Z., Rhoads, G. G., and Hamilton, H. B. [1973]. Epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii and California: serum lipids and diet. *American Journal of Epidemiology*, **97**: 372–385.

Kesteloot, H., and van Houte, O. [1973]. An epidemiologic study of blood pressure in a large male population. *American Journal of Epidemiology*, **99**: 14–29.

Kruskal, W., and Mosteller, F. [1979a]. Representative sampling I: non-scientific literature. *International Statistical Review*, **47**: 13–24.

Kruskal, W., and Mosteller, F. [1979b]. Representative sampling II: scientific literature excluding statistics. *International Statistical Review*, **47**: 111–127.

Kruskal, W., and Mosteller, F. [1979c]. Representative sampling III: the current statistical literature. *International Statistical Review*, **47**: 245–265.

Lehtonen, R., and Pahkinen, E. J. [1995]. *Practical Methods for Design and Analysis of Complex Surveys*. Wiley, New York.

Levy, P. S., and Lemeshow S. [1999]. *Sampling of Populations: Methods and Applications*, 3rd Ed. Wiley, New York.

Lohr, S. [1999]. *Sample: Design and Analysis*. Duxbury Press, Pacific Grove, CA.

Moore, D. S. [2001]. *Statistics: Concepts and Controversies*, 5th ed. W. H. Freeman, New York.

Murphy, E. A. [1979]. *Biostatistics in Medicine*. Johns Hopkins University Press, Baltimore.

Runes, D. D. [1959]. *Dictionary of Philosophy*. Littlefield, Adams, Ames, IA.

Rushforth, N. B., Bennet, P. H., Steinberg, A. G., Burch, T. A., and Miller, M. [1971]. Diabetes in the Pima Indians: evidence of bimodality in glucose tolerance distribution. *Diabetes*, **20**: 756–765. Copyright © 1971 by the American Diabetic Association.

Savage, I. R. [1968]. *Statistics: Uncertainty and Behavior*. Houghton Mifflin, Boston.

Shepard, D. S., and Neutra, R. [1977]. Pitfalls in sampling medical visits. *American Journal of Public Health*, **67**: 743–750. Copyright © by the American Public Health Association.

Tversky, A., and Kahneman, D. [1974]. Judgment under uncertainty: heuristics and biases. *Science*, **185**: 1124–1131. Copyright © by the AAAS.

Winkelstein, W. Jr., Kazan, A., Kato, H., and Sachs, S. T. [1975]. Epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii and California: blood pressure distributions. *American Journal of Epidemiology*, **102**: 502–513.

Zervas, M., Hamacher, H., Holmes, O., and Rieder, S. V. [1970]. Normal laboratory values. *New England Journal of Medicine*, **283**: 1276–1285.

CHAPTER 5

One- and Two-Sample Inference

5.1 INTRODUCTION

In Chapter 4 we laid the groundwork for statistical inference. The following steps were involved:

1. Define the population of interest.
2. Specify the parameter(s) of interest.
3. Take a random sample from the population.
4. Make statistical inferences about the parameter(s): (a) estimation; and (b) hypothesis testing.

A good deal of "behind-the-scenes" work was necessary, such as specifying what is meant by a *random* sample, but you will recognize that the four steps above summarize the process. In this chapter we (1) formalize the inferential process by defining pivotal quantities and their uses (Section 5.2); (2) consider normal distributions for which *both* the mean and variance are unknown, which will involve the use of the famous Student t-distribution (Sections 5.3 and 5.4); (3) extend the inferential process to a comparison of two normal populations, including comparison of the variances (Sections 5.5 to 5.7); and (4) finally begin to answer the question frequently asked of statisticians: "How many observations should I take?" (Section 5.9).

5.2 PIVOTAL VARIABLES

In Chapter 4, confidence intervals and tests of hypotheses were introduced in a somewhat ad hoc fashion as inference procedures about population parameters. To be able to make inferences, we needed the sampling distributions of the statistics that estimated the parameters. To make inferences about the mean of a normal distribution (with variance known), we needed to know that the sample mean of a random sample was normally distributed; to make inferences about the variance of a normal distribution, we used the chi-square distribution. A pattern also emerged in the development of estimation and hypothesis testing procedures. We discuss next the unifying scheme. This will greatly simplify our understanding of the statistical procedures, so that attention can be focused on the assumptions and appropriateness of such procedures rather than on understanding the mechanics.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

In Chapter 4, we used basically two quantities in making inferences:

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \quad \text{and} \quad \chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

What are some of their common features?

1. Each of these expressions involves *at least* a statistic *and* a parameter for the statistic estimated: for example, s^2 and σ^2 in the second formula.
2. The distribution of the quantity was tabulated in a standard normal table or chi-square table.
3. Distribution of the quantity was not dependent on a value of the parameter. Such a distribution is called a *fixed distribution*.
4. Both confidence intervals and tests of hypotheses were derived from a probability inequality involving either Z or χ^2.

Formally, we define:

Definition 5.1. A *pivotal variable* is a function of statistic(s) and parameter(s) having the same fixed distribution (usually tabulated) for all values of the parameter(s).

The quantities Z and χ^2 are pivotal variables. One of the objectives of theoretical statistics is to develop appropriate pivotal variables for experimental situations that cannot be modeled adequately by existing variables.

In Table 5.1 are listed eight pivotal variables and their use in statistical inference. In this chapter we introduce pivotal variables 2, 5, 6, and 8. Pivotal variables 3 and 4 are introduced in Chapter 6. For each variable, the fixed or tabulated distribution is given as well as the formula for a $100(1-\alpha)\%$ confidence interval. The corresponding test of hypothesis is obtained by replacing the statistic(s) by the hypothesized parameter value(s). The table also lists the assumptions underlying the test. Most of the time, the minimal assumption is that of normality of the underlying observations, or appeal is made to the central limit theorem.

Pivotal variables are used primarily in inferences based on the normal distribution. They provide a methodology for estimation and hypothesis testing. The aim of estimation and hypothesis testing is to make probabilistic statements about parameters. For example, confidence intervals and p-values make statements about parameters that have probabilistic aspects. In Chapters 6 to 8 we discuss inferences that do not depend as explicitly on pivotal variables; however, even in these procedures, the methodology associated with pivotal variables is used; see Figure 5.1.

5.3 WORKING WITH PIVOTAL VARIABLES

We have already introduced the manipulation of pivotal variables in Section 4.7. Table 5.1 summarizes the end result of the manipulations. In this section we again outline the process for the case of one sample from a normal population with the variance known. We have a random sample of size n from a normal population with mean μ and variance σ^2 (known). We start with the basic probabilistic inequality

$$P[z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}] = 1 - \alpha$$

Table 5.1 Pivotal Variables and Their Use in Statistical Inference

		Assumptions		100(1 − α)%	
	Pivotal Variable	Model	Other[a]	Confidence Interval[b]	Inference/Comments
1.	$\dfrac{\overline{Y}-\mu}{\sigma/\sqrt{n}}=Z$	$N(0, 1)$	(i) and (iii); or (ii)	$\overline{Y}\pm\dfrac{z_*\sigma}{\sqrt{n}}$	μ or $\mu=\mu_1-\mu_2$ based on paired data $z_*=z_{1-\alpha/2}$
2.	$\dfrac{(\overline{Y}_1-\overline{Y}_2)-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}=Z$	$N(0, 1)$	(i) and (iii); or (ii)	$(\overline{Y}_1-\overline{Y}_2)\pm z_*\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}$	$\mu_1-\mu_2$ based on independent data $z_*=z_{1-\alpha/2}$
3.	$\dfrac{p-\pi}{\sqrt{p(1-p)/n}}=Z$	$N(0, 1)$	(ii)	$p\pm z_*\sqrt{\dfrac{p(1-p)}{n}}$	π $z_*=z_{1-\alpha/2}$
4.	$\dfrac{(p_1-p_2)-(\pi_1-\pi_2)}{\sqrt{p_1q_1/n_1+p_2q_2/n_2}}=Z$	$N(0, 1)$	(ii)	$(p_1-p_2)\pm z_*\sqrt{\dfrac{p_1q_1}{n_1}+\dfrac{p_2q_2}{n_2}}$	$\pi_1-\pi_2$ based on independent data $z_*=z_{1-\alpha/2}$ $q_1=1-p_1;\, q_2=1-p_2$
5.	$\dfrac{\overline{Y}-\mu}{s/\sqrt{n}}=t$	t_{n-1}	(i)	$\overline{Y}\pm\dfrac{t_*s}{\sqrt{n}}$	μ or $\mu=\mu_1-\mu_2$ based on paired data $t_*=t_{n-1,1-\alpha/2}$
6.	$\dfrac{(\overline{Y}_1-\overline{Y}_2)-(\mu_1-\mu_2)}{s_p\sqrt{1/n_1+1/n_2}}=t$	$t_{n_1+n_2-2}$	(i) and (iv)	$(\overline{Y}_1-\overline{Y}_2)\pm t_*s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}$	$\mu_1-\mu_2$ based on independent data $t_*=t_{n_1+n_2-2,1-\alpha/2}$ $s_p^2=\dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$
7.	$\dfrac{(n-1)s^2}{\sigma^2}=\chi^2$	χ^2_{n-1}	(i)	$\dfrac{(n-1)s^2}{\chi^2_*},\ \dfrac{(n-1)s^2}{\chi^2_{**}}$	σ^2 $\chi^2_*=\chi^2_{n-1,1-\alpha/2}$ $\chi^2_{**}=\chi^2_{n-1,\alpha/2}$
8.	$\dfrac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}=F$	F_{n_1-1,n_2-1}	(i)	$\dfrac{s_1^2/s_2^2}{F_*},\ \dfrac{s_1^2/s_2^2}{F_{**}}$	$\dfrac{\sigma_1^2}{\sigma_2^2}$ $F_*=F_{n_1-1,n_2-1,1-\alpha/2}$ $F_{**}=F_{n_1-1,n_2-1,\alpha/2}$

[a] Assumptions (other): (i) Observations (for paired data, the differences) are independent, normally distributed; (ii) large-sample result; (iii) variance(s) known; (iv) population variances equal.

[b] To determine the appropriate critical region in a test of hypothesis, replace statistic(s) by hypothesized values of parameter(s).

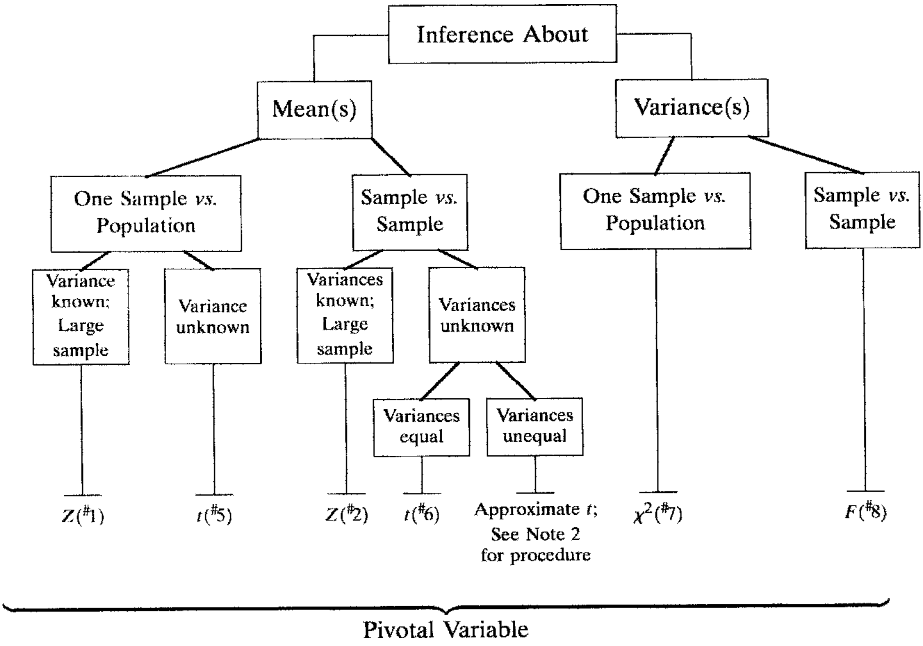


Figure 5.1 Methodology associated with pivotal variables.

We substitute $Z = (\overline{Y} - \mu)/(\sigma_0/\sqrt{n})$, writing σ_0 to indicate that the population variance is assumed to be known:

$$P\left[z_{\alpha/2} \leq \frac{\overline{Y} - \mu}{\sigma_0/\sqrt{n}} \leq z_{1-\alpha/2}\right] = 1 - \alpha$$

Solving for μ produces a $100(1-\alpha)\%$ confidence interval for μ; solving for $\overline{Y}$ and substituting a hypothesized value, μ_0, for μ produces the nonrejection region for a $100(\alpha)\%$ test of the hypothesis:

$100(1 - \alpha)\%$ confidence interval for μ:

$$[\overline{Y} + z_{\alpha/2}\sigma_0/\sqrt{n}, \quad \overline{Y} + z_{1-\alpha/2}\sigma_0/\sqrt{n}]$$

$100(\alpha)\%$ hypothesis test of $\mu = \mu_0$; reject if $\overline{Y}$ is not in

$$[\mu_0 + z_{\alpha/2}\sigma_0/\sqrt{n}, \quad \mu_0 + z_{1-\alpha/2}\sigma_0/\sqrt{n}]$$

Notice again the similarity between the two intervals. These intervals can be written in an abbreviated form using the fact that $z_{\alpha/2} = -z_{1-\alpha/2}$,

$$\overline{Y} \pm \frac{z_{1-\alpha/2}\sigma_0}{\sqrt{n}} \quad \text{and} \quad \mu_0 \pm \frac{z_{1-\alpha/2}\sigma_0}{\sqrt{n}}$$

for the confidence intervals and tests of hypothesis, respectively.

To calculate the p-value associated with a test statistic, again use is made of the pivotal variable. The null hypothesis value of the parameter is used to calculate the probability of the observed value of the statistic or an observation more extreme. As an illustration, suppose that a population variance is claimed to be $100(\sigma_0^2 = 100)$ vs. a larger value $(\sigma_0^2 > 100)$. From a random sample of size 11, we are given $s^2 = 220$. What is the p-value for this value (or more extreme)? We use the pivotal quantity $(n-1)s^2/\sigma_0^2$, which under the null hypothesis is chi-square with 10 degrees of freedom.

The one-sided p-value is the probability of a value of $s^2 \geq 220$. Using the pivotal variable, we get

$$P\left[\chi^2 \geq \frac{(11-1)(220)}{100}\right] = P[\chi^2 \geq 22.0]$$

where χ^2 has $11 - 1 = 10$ degrees of freedom, giving a one-sided p-value of 0.0151.

Additional examples in the use of pivotal variables will occur throughout this and later chapters. See Note 5.1 for some additional comments on the pivotal variable approach.

5.4 t-DISTRIBUTION

For a random sample from a normal distribution with mean μ and variance σ^2 (known), the quantity $Z = (\overline{Y} - \mu)/(\sigma/\sqrt{n})$ is a pivotal quantity that has a normal (0,1) distribution. What if the variance is unknown? Suppose that we replace the variance σ^2 by its estimate s^2 and consider the quantity $(\overline{Y} - \mu)/(s/\sqrt{n})$. What is its sampling distribution?

This problem was solved by the statistician W. S. Gossett, in 1908, who published the result under the pseudonym "Student" using the notation

$$t = \frac{\overline{Y} - \mu}{s/\sqrt{n}}$$

The distribution of this variable is now called *Student's t-distribution*. Gossett showed that the distribution of t was similar to that of the normal distribution, but somewhat more "heavy-tailed" (see below), and that for each sample size there is a different distribution. The distributions are indexed by $n - 1$, the degrees of freedom identical to that of the chi-square distribution. The t-distribution is symmetrical, and as the degrees of freedom become infinite, the standard normal distribution is reached.

A picture of the t-distribution for various degrees of freedom, as well as the limiting case of the normal distribution, is given in Figure 5.2. Note that like the standard normal distribution, the t-distribution is bell-shaped and symmetrical about zero. The t-distribution is *heavy-tailed*: The area to the right of a specified positive value is greater than for the normal distribution; in other words, the t-distribution is less "pinched." This is reasonable; unlike a standard normal deviate where only the mean ($\overline{Y}$) can vary (μ and σ are fixed), the t statistic can vary with *both* $\overline{Y}$ and s, so that t will vary even if $\overline{Y}$ is fixed.

Percentiles of the t-distribution are denoted by the symbol $t_{v,\alpha}$, where v indicates the degrees of freedom and α the 100αth percentile. This is indicated in Figure 5.3. In Table 5.1, rather than writing all the subscripts on the t variate, an asterisk is used and explained in the comment part of the table.

Table A.4 lists the percentiles of the t-distribution for each degree of freedom to 30, by fives to 100, and values for 200, 500, and ∞ degrees of freedom. This table lists the t-values such that the percent to the left is as specified by the column heading. For example, for an area of 0.975 (97.5%), the t-value for six degrees of freedom is 2.45. The last row in this column corresponds to a t with an infinite number of degrees of freedom, and the value of 1.96 is identical to the corresponding value of Z; that is, $P[Z \leq 1.96] = 0.975$. You should verify that the last row in this table corresponds precisely to the normal distribution values (i.e., $t_\infty = Z$) and that for practical purposes, t_n and Z are equivalent for $n > 30$. What are the mean and the variance of the t-distribution? The mean will be zero, and the variance is $v/(v-2)$. In the symbols used in Chapter 4, $E(t) = 0$ and $\text{Var}(t) = v/(v-2)$.

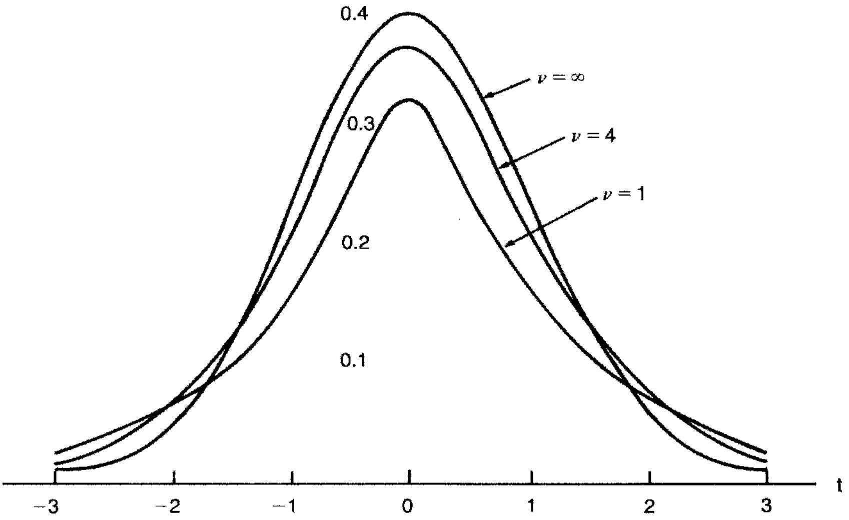


Figure 5.2 Student t-distribution with one, four, and ∞ degrees of freedom.

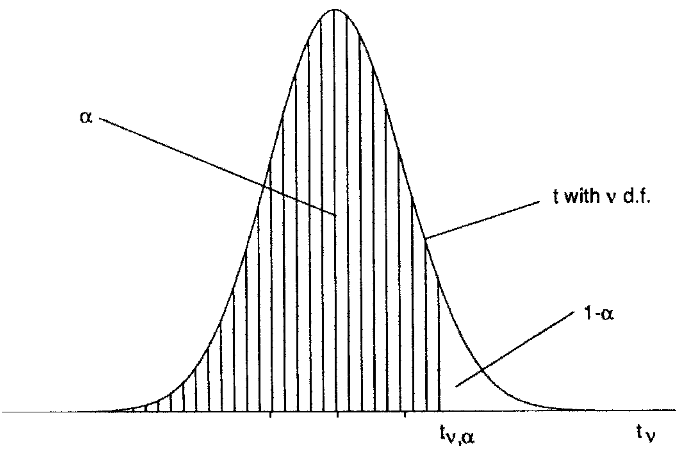


Figure 5.3 Percentiles of the t-distribution.

The converse table of percentiles for a given absolute t-value is given in the Web appendix, and most statistical software will calculate it. We find that the probability of a t-value greater than 1 in absolute value for one degree of freedom is 0.500; the corresponding areas for 7, 30, and ∞ degrees of freedom are 0.351, 0.325, and 0.317, respectively. Thus, at 30 degrees of freedom, the t-distribution is for most practical purposes, indistinguishable from a normal distribution. The term *heavy-tailed* can now be made precise: For a specified value (e.g., with an abscissa value of 1), $P[t_1 \geq 1] > P[t_7 \geq 1] > P[t_{10} \geq 1] > P[Z \geq 1]$.

5.5 ONE-SAMPLE INFERENCE: LOCATION

5.5.1 Estimation and Testing

We begin this section with an example.

Example 5.1. In Example 4.9 we considered the birthweight in grams of the first 11 SIDS cases occurring in King Country in 1969. In this example, we consider the birthweights of the first 15 cases born in 1977. The birthweights for the latter group are

2013	3827	3090	3260	4309	3374	3544	2835
3487	3289	3714	2240	2041	3629	3345	

The mean and standard deviation of this sample are 3199.8 g and 663.00 g, respectively. Without assuming that the population standard deviation is known, can we obtain an interval estimate for the population mean or test the null hypothesis that the population birthweight average of SIDS cases is 3300 g (the same as the general population)?

We can now use the t-distribution. Assuming that birthweights are normally distributed, the quantity

$$\frac{\overline{Y} - \mu}{s/\sqrt{15}}$$

has a t-distribution with $15 - 1 = 14$ degrees of freedom.

Using the estimation procedure, the point estimate of the population mean birthweight of SIDS cases is $3199.8 \doteq 3200$ g. A 95% confidence interval can be constructed on the basis of the t-distribution. For a t-distribution with $15 - 1 = 14$ degrees of freedom, the critical values are ± 2.14, that is, $P[-2.14 \leq t_{14} \leq 2.14] = 0.95$. Using Table 5.1, a 95% confidence interval is constructed using pivotal variable 5:

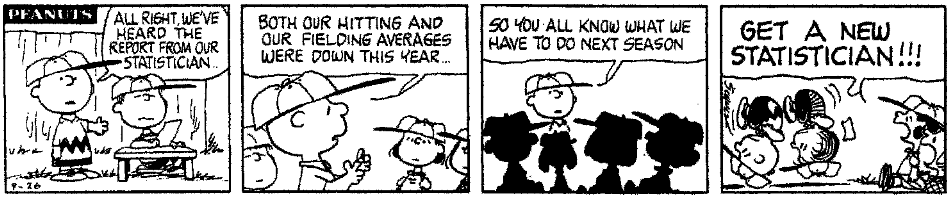
$$3200 \pm \frac{(2.14)(663.0)}{\sqrt{15}} = 3200 \pm 366, \qquad \text{lower limit : 2834 g, upper limit : 3566 g}$$

Several comments are in order:

1. This interval includes 3300 g, the average birthweight in the non-SIDS population. If the analysis had followed a hypothesis-testing procedure, we could not have rejected the null hypothesis on the basis of a two-tailed test.
2. The standard error, $633.0/\sqrt{15}$, is multiplied by 2.14 rather than the critical value 1.96 using a normal distribution. Thus, the confidence interval is wider by approximately 9%. This is the price paid for our ignorance about the value of the population standard deviation. Even in this fairly small sample, the price is modest.

5.5.2 t-Tests for Paired Data

A second example of the one-sample t-test involves its application to paired data. What are *paired data*? Typically, the term refers to repeated or multiple measurements on the same subjects. For example, we may have a measurement of the level of pain before and after administration of an analgesic drug. A somewhat different experiment might consider the level of pain in response to each of *two* drugs. One of these could be a placebo. The first experiment has the weakness that there may be a spontaneous reduction in level of pain (e.g., postoperative pain level), and thus the difference in the responses (after/before) may be made up of two effects: an effect of the drug as well as the spontaneous reduction. Some experimental design considerations are discussed further in Chapter 10. The point we want to make with these two examples is that the basic data consist of pairs, and what we want to look at is the differences within the pairs. If, in the second example, the treatments are to be compared, a common null hypothesis is that the effects are the same and therefore the differences in the treatments should be centered around zero. A natural approach then tests whether the mean of the *sample differences* could have come from a population of differences with mean zero. If we assume that the means of the sample differences are normally distributed, we can apply the t-test (under the null hypothesis),


PEANUTS
ALL RIGHT, WE'VE HEARD THE REPORT FROM OUR STATISTICIAN..
9-26
BOTH OUR HITTING AND OUR FIELDING AVERAGES WERE DOWN THIS YEAR...
SO YOU ALL KNOW WHAT WE HAVE TO DO NEXT SEASON
GET A NEW STATISTICIAN!!!

Cartoon 5.1 PEANUTS. (Reprinted by permission of UFS, Inc.)

Table 5.2 Response of 13 Patients to Aminophylline Treatment at 16 Hours Compared with 24 Hours before Treatment (Apneic Episodes per Hour)

Patient	24 h Before	16 h After	Before–After (Difference)
1	1.71	0.13	1.58
2	1.25	0.88	0.37
3	2.13	1.38	0.75
4	1.29	0.13	1.16
5	1.58	0.25	1.33
6	4.00	2.63	1.37
7	1.42	1.38	0.04
8	1.08	0.50	0.58
9	1.83	1.25	0.58
10	0.67	0.75	−0.08
11	1.13	0.00	1.13
12	2.71	2.38	0.33
13	1.96	1.13	0.83
Total	22.76	12.79	9.97
Mean	1.751	0.984	0.767
Variance	0.7316	0.6941	0.2747
Standard deviation	0.855	0.833	0.524

Source: Data from Bednarek and Roloff [1976].

and estimate the variance of the population of differences σ^2, by the variance of the *sample differences*, s^2.

Example 5.2. The procedure is illustrated with data from Bednarek and Roloff [1976] dealing with the treatment of apnea (a transient cessation of respiration) using a drug, aminophylline, in premature infants. The variable of interest, "average number of apneic episodes per hour," was measured before and after treatment with the drug. An episode was defined as the absence of spontaneous breathing for more than 20 seconds or less if associated with bradycardia or cyanosis.

Patients who had "six or more apneic episodes on each of two consecutive 8 h shifts were admitted to the study." For purposes of the study, consider only the difference between the average number of episodes 24 hours before treatment and 16 hours after. This difference is given in the fourth column of Table 5.2. The average difference for the 13 patients is 0.767 episode per hour. That is, there is a change from 1.751 episodes per hour before treatment to 0.984 episode per hour at 16 hours after treatment.

The standard deviation of the differences is $s = 0.524$. The pivotal quantity to be used is variable 5 from Table 5.1. The argument is as follows: The basic statement about the pivotal variable t with $13 - 1 = 12$ degrees of freedom is $P[-2.18 \leq t_{12} \leq 2.18] = 0.95$ using Table A.4. The form taken for this example is

$$P\left[-2.18 \leq \frac{\overline{Y} - \mu}{0.524/\sqrt{13}} \leq 2.18\right] = 0.95$$

To set up the region to test some hypothesis, we solve for $\overline{Y}$ as before. The region then is

$$P[\mu - 0.317 \leq \overline{Y} \leq \mu + 0.317] = 0.95$$

What is a "reasonable" value to hypothesize for μ? The usual procedure in this type of situation is to assume that the treatment has "no effect." That is, the average difference in the number of apneic episodes from before to after treatment represents random variation. If there is no difference in the population average number of episodes before and after treatment, we can write this as

$$H_0\colon \mu = 0$$

We can now set up the hypothesis-testing region as illustrated in Figures 5.4 and 5.5. Figure 5.4 indicates that the sample space can be partitioned without knowing the observed value of $\overline{Y}$. Figure 5.5 indicates the observed value of $\overline{Y} = 0.767$ episode per hour; it clearly falls into the rejection region. Note that the scale has been changed from Figure 5.4 to accommodate the value observed. Hence the null hypothesis is rejected and it is concluded that the average number of apneic episodes observed 16 hours after treatment differs significantly from the average number of apneic episodes observed 24 hours before treatment.

This kind of test is often used when two treatments are applied to the same experimental unit or when the experimental unit is observed over time and a treatment is administered so that it

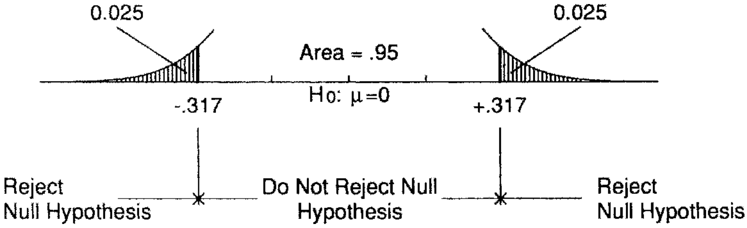


Figure 5.4 Partitioning of sample space of $\overline{Y}$ into two regions: (*a*) region where the null hypothesis is not rejected, and (*b*) region where it is rejected. (Data from Bednarek and Roloff [1976]; see Table 5.2.)

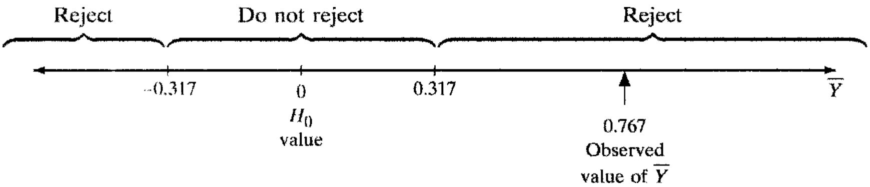


Figure 5.5 Observed value of $\overline{Y}$ and location on the sample space. (Data from Bednarek and Roloff [1976]; see Table 5.2.)

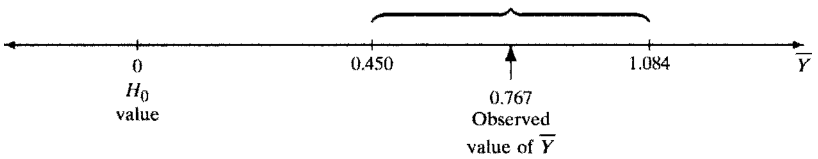


Figure 5.6 A 95% confidence interval for the difference in number of apneic episodes per hour. (Data from Bednarek and Roloff [1976]; see Table 5.2.)

is meaningful to speak of pretreatment and posttreatment situations. As mentioned before, there is the possibility that changes, if observed, are in fact due to changes over time and not related to the treatment.

To construct a confidence interval, we solve the inequality for μ so that we get

$$P[\overline{Y} - 0.317 \leq \mu \leq \overline{Y} + 0.317] = 0.95$$

Again, this interval can be set up to this point without knowing the value of $\overline{Y}$. The value of $\overline{Y}$ is observed to be 0.767 episode per hour, so that the 95% confidence interval becomes

$$[0.767 - 0.317 \leq \mu \leq 0.767 + 0.317] \text{ or } [0.450 \leq \mu \leq 1.084]$$

This interval is displayed in Figure 5.6. Two things should be noted:

1. The width of the confidence interval is the same as the width of the region where the null hypothesis is not rejected (cf. Figure 5.5).
2. The 95% confidence interval does not include zero, the null hypothesis value of μ.

5.6 TWO-SAMPLE STATISTICAL INFERENCE: LOCATION

5.6.1 Independent Random Variables

A great deal of research activity involves the comparison of two or more groups. For example, two cancer therapies may be investigated: one group of patients receives one treatment and a second group the other. The experimental situation can be thought of in two ways: (1) there is one population of subjects, and the treatments induce two subpopulations; or (2) we have two populations that are identical except in their responses to their respective treatments. If the assignment of treatment is random, the two situations are equivalent.

Before exploring this situation, we need to state a definition and a statistical result:

Definition 5.2. Two random variables Y_1 and Y_2 are *statistically independent* if for all fixed values of numbers (say, y_1 and y_2),

$$P[Y_1 \leq y_1, Y_2 \leq y_2] = P[Y_1 \leq y_1]P[Y_2 \leq y_2]$$

The notation $[Y_1 \leq y_1, Y_2 \leq y_2]$ means that Y_1 takes on a value less than or equal to y_1, and Y_2 takes on a value less than or equal to y_2. If we define an event A to have occurred when Y_1 takes on a value less than or equal to y_1, and an event B when Y_2 takes on a value less than or equal to y_2, Definition 5.2 is equivalent to the statistical independence of events $P[AB] = P[A]P[B]$ as defined in Chapter 4. So the difference between statistical independence of random variables and statistical independence of events is that the former in effect describes a relationship between many events (since the definition has to be true for *any* set of values of y_1 and y_2). A basic result can now be stated:

Result 5.1. If Y_1 and Y_2 are statistically independent random variables, then for any two constants a_1 and a_2, the random variable $W = a_1Y_1 + a_2Y_2$ has mean and variance

$$E(W) = a_1E(Y_1) + a_2E(Y_2)$$
$$\text{Var}(W) = a_1^2\text{Var}(Y_1) + a_2^2\text{Var}(Y_2)$$

The only new aspect of this result is that of the variance. In Note 4.10, the expectation of W was already derived. Before giving an example, we also state:

Result 5.2. If Y_1 and Y_2 are statistically independent random variables that are normally distributed, $W = a_1Y_1+a_2Y_2$ is normally distributed with mean and variance given by Result 5.1.

Example 5.3. Let Y_1 be normally distributed with mean $\mu_1 = 100$ and variance $\sigma_1^2 = 225$; let Y_2 be normally distributed with mean $\mu_2 = 50$ and variance $\sigma_2^2 = 175$. If Y_1 and Y_2 are statistically independent, $W = Y_1 + Y_2$ is normally distributed with mean $100 + 50 = 150$ and variance $225 + 175 = 400$. This and additional examples are given in the following summary:

$$Y_1 \sim N(100, 225), Y_2 \sim N(50, 175)$$

W	Mean of W	Variance of W
$Y_1 + Y_2$	150	400
$Y_1 - Y_2$	50	400
$2Y_1 + Y_2$	250	1075
$2Y_1 - 2Y_2$	100	1600

Note that the variance of $Y_1 - Y_2$ is the same as the variance of $Y_1 + Y_2$; this is because the coefficient of Y_1, -1, is squared in the variance formula and $(-1)^2 = (+1)^2 = 1$. In words, the variance of a sum of independent random variables is the same as the variance of a difference of independent random variables.

Example 5.4. Now we look at an example that is more interesting and indicates the usefulness of the two results stated. Heights of females and males are normally distributed with means 162 cm and 178 cm and variances (6.4 cm)^2 and (7.5 cm)^2, respectively. Let Y_1 = height of female; let Y_2 = height of male. Then we can write

$$Y_1 \sim N(162, (6.4)^2) \quad \text{and} \quad Y_2 \sim N(178, (7.5)^2)$$

Now consider husband–wife pairs. Suppose (probably somewhat contrary to societal *mores*) that husband–wife pairs are formed independent of stature. That is, we interpret this statement to mean that Y_1 and Y_2 are statistically independent. The question is: On the basis of this model, what is the probability that the wife is taller than the husband? We formulate the problem as follows: Construct the new variable $W = Y_1 - Y_2$. From Result 5.2 it follows that

$$W \sim N(-16, (6.4)^2 + (7.5)^2)$$

Now the question can be translated into a question about W; namely, if the wife is taller than the husband, $Y_1 > Y_2$, or $Y_1 - Y_2 > 0$, or $W > 0$. Thus, the question is reformulated as $P[W > 0]$.

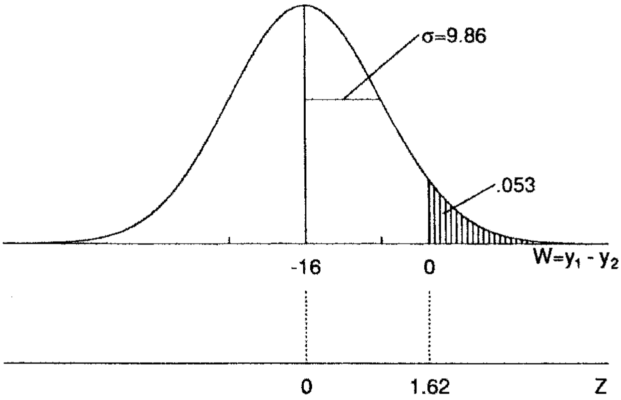


Figure 5.7 Heights of husband–wife pairs.

Hence,

$$\begin{aligned} P[W > 0] &= P\left[Z > \frac{0-(-16)}{\sqrt{(6.4)^2+(7.5)^2}}\right] \\ &\doteq P\left[Z > \frac{16}{9.86}\right] \\ &\doteq P[Z > 1.62] \\ &\doteq 0.053 \end{aligned}$$

so that under the model, in 5.3% of husband–wife pairs the wife will be taller than the husband. Figure 5.7 indicates the area of interest.

5.6.2 Estimation and Testing

The most important application of Result 5.1 involves distribution of the difference of two sample means. If $\overline{Y}_1$ and $\overline{Y}_2$ are the means from two random samples of size n_1 and n_2, respectively, and Y_1 and Y_2 are normally distributed with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, then by Result 5.2,

$$\overline{Y}_1 - \overline{Y}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

so that

$$\frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = Z$$

has a standard normal distribution. This, again, is a pivotal variable, number 2 in Table 5.1. We are now in a position to construct confidence intervals for the quantity $\mu_1 - \mu_2$ or to do hypothesis testing. In many situations, it will be reasonable to assume (null hypothesis) that $\mu_1 = \mu_2$, so that $\mu_1 - \mu_2 = 0$; although the values of the two parameters are unknown, it is reasonable for testing purposes to assume that they are equal, and hence, the difference will be zero. For example, in a study involving two treatments, we could assume that the treatments were equally effective (or ineffective) and that differences between the treatments should be centered at zero.

How do we determine whether or not random variables are statistically independent? The most common way is to note that they are causally independent in the population. That is, the value of Y for one person does not affect the value for another. As long as the observations are sampled independently (e.g., by simple random sampling), they will remain statistically independent. In some situations it is not clear a priori whether variables are independent and there are statistical procedures for testing this assumption. They are discussed in Chapter 9. For the present we will assume that the variables we are dealing with are either statistically independent or if not (as in the case of the paired t-test discussed in Section 5.5.2), use aspects of the data that can be considered statistically independent.

Example 5.5. Zelazo et al. [1972] studied the age at which children walked as related to "walking exercises" given newborn infants. They state that "if a newborn infant is held under his arms and his bare feet are permitted to touch a flat surface, he will perform well-coordinated walking movements similar to those of an adult." This reflex disappears by about eight weeks. They placed 24 white male infants into one of four "treatment" groups. For purposes of this example, we consider only two of the four groups: "active exercise group" and "eight-week control group." The active group received daily stimulation of the walking reflex for eight weeks. The control group was tested at the end of the eight-week treatment period, but there was no intervention. The age at which the child subsequently began to walk was then reported by the mother. The data and basic calculations are shown in Table 5.3.

For purposes of this example, we assume that the sample standard deviations are, in fact, population standard deviations, so that Result 5.2 can be applied. In Example 5.6 we reconsider this example using the two-sample t-test. For this example, we have

$n_1 = 6$	$n_2 = 5$
$\overline{Y}_1 = 10.125$ months	$\overline{Y}_2 = 12.350$ months
$\sigma_1 = 1.4470$ months (assumed)	$\sigma_2 = 0.9618$ month (assumed)

For purposes of this example, the quantity

$$Z = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{(1.4470)^2/6 + (0.9618)^2/5}} = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{0.7307}$$

has a standard normal distribution and is based on pivotal variable 2 of Table 5.1. Let us first set up a 95% confidence interval on the difference $(\mu_1 - \mu_2)$ in the population means. The 95% confidence interval is

$$(\overline{Y}_1 - \overline{Y}_2) \pm 1.96(0.7307)$$

with

$$\text{upper limit} = (10.125 - 12.350) + 1.4322 = -0.79 \text{ month}$$

$$\text{lower limit} = (10.125 - 12.350) - 1.4322 = -3.66 \text{ months}$$

The time line is shown in Figure 5.8.

The 95% confidence interval does not straddle zero, so we would conclude that there is a real difference in age in months when the baby first walked in the exercise group compared to the control group. The best estimate of the difference is $10.125 - 12.350 = -2.22$ months; that is, the age at first walking is about two months earlier than the control group.

Note the flow of the argument: The babies were a homogeneous group before treatment. Allocation to the various groups was on a random basis (assumed but not stated explicitly in the article); the only subsequent differences between the groups were the treatments, so significant differences between the groups must be attributable to the treatments. (Can you think of some reservations that you may want checked before accepting the conclusion?)

Table 5.3 Distribution of Ages (in Months) in Infants for Walking Alone

	Age for Walking Alone	
	Active Exercise Group	Eight-Week Control Group
	9.00	13.25
	9.50	11.50
	9.75	12.00
	10.00	13.50
	13.00	11.50
	9.50	[a]
n	6	5
Mean	10.125	12.350
Standard deviation	1.4470	0.9618

Source: Data from Zelazo et al. [1972].

[a] One observation is missing from the paper.

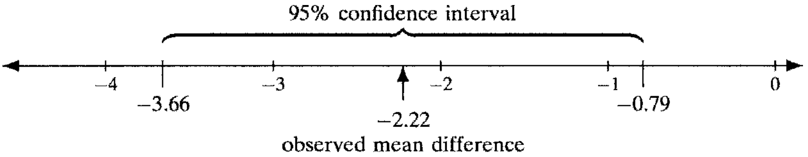


Figure 5.8 Time line for difference in time to infants walking alone.

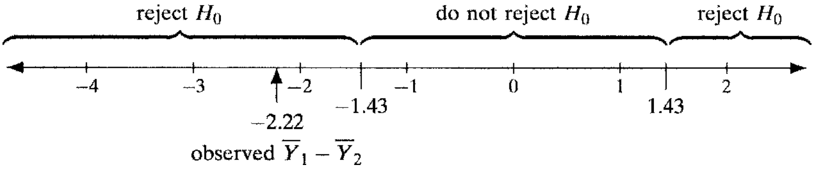


Figure 5.9 Plot showing the nonrejection region.

Formulating the problem as a hypothesis-testing problem is done as follows: A reasonable null hypothesis is that $\mu_1 - \mu_2 = 0$; in this case, the hypothesis of no effect. Comparable to the 95% confidence interval, a test at the 5% level will be carried out. Conveniently, $\mu_1 - \mu_2 = 0$, so that the nonrejection region is simply $0 \pm 1.96(0.7307)$ or 0 ± 1.4322. Plotting this on a line, we get Figure 5.9.

We would reject the null hypothesis, $H_0 : \mu_1 - \mu_2 = 0$, and accept the alternative hypothesis, $H_A : \mu_1 \neq \mu_2$; in fact, on the basis of the data, we conclude that $\mu_1 < \mu_2$.

To calculate the (one-sided) p-value associated with the difference observed, we again use the pivotal variable

$$\begin{aligned} P\left[[\overline{Y}_1 - \overline{Y}_2] \leq -2.225\right] &\doteq P\left[Z \leq \frac{-2.225 - 0}{0.7307}\right] \\ &\doteq P[Z \leq -3.05] \\ &\doteq 0.0011 \end{aligned}$$

The p-value is 0.0011, much less than 0.05, and again, we would reject the null hypothesis. To make the p-value comparable to the two-sided confidence and hypothesis testing procedure, we must multiply it by 2, to give a p-value

$$p\text{-value} = 2(0.0011) = 0.0022$$

We conclude this section by considering the two sample location problem when the population variances are not known. For this we need:

Result 5.3. If $\overline{Y}_1$ and $\overline{Y}_2$ are based on two independent random samples of size n_1 and n_2 from two normal distributions with means μ_1 and μ_2 and the same variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then

$$\frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{1/n_1 + 1/n_2}}$$

has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom. Here s_p^2 is "the pooled estimate of common variance σ^2," as defined below.

This result is summarized by pivotal variable 6 in Table 5.1. Result 5.3 assumes that the population variances are the same, $\sigma_1^2 = \sigma_2^2 = \sigma^2$. There are then two estimates of $\sigma^2 : s_1^2$ from the first sample and s_2^2 from the second sample. How can these estimates be combined to provide the best possible estimate of σ^2? If the sample sizes, n_1 and n_2, differ, the variance based on the larger sample should be given more weight; the pooled estimate of σ^2 provides this. It is defined by

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

If $n_1 = n_2$, then $s_p^2 = \frac{1}{2}(s_1^2 + s_2^2)$, just the arithmetic average of the variances. For $n_1 \neq n_2$, the variance with the larger sample size receives more weight. See Note 5.2 for a further discussion.

Example 5.5. (*continued*) Consider again the data in Table 5.3 on the age at which children first walk. We will now take the more realistic approach by treating the standard deviations as sample standard deviations, as they should be.

The pooled estimate of the (assumed) common variance is

$$s_p^2 \doteq \frac{(6-1)(1.4470)^2 + (5-1)(0.9618)^2}{6+5-2} \doteq \frac{14.1693}{9} \doteq 1.5744$$

$$s_p \doteq 1.2547 \text{ months}$$

A 95% confidence interval for the difference $\mu_1 - \mu_2$ is constructed first. From Table A.4, the critical t-value for nine degrees of freedom is $t_{9,0.975} = 2.26$. The 95% confidence interval is calculated to be

$$(10.125 - 12.350) \pm (2.26)(1.2547)\sqrt{1/6 + 1/5} \doteq -2.225 \pm 1.717$$

$$\text{lower limit} = -3.94 \text{ months and upper limit} = -0.51 \text{ month}$$

Notice that these limits are wider than the limits $(-3.66, -0.79)$ calculated on the assumption that the variances are known. The wider limits are the price for the additional uncertainty.

The same effect is observed in testing the null hypothesis that $\mu_1 - \mu_2 = 0$. The rejection region (Figure 5.10), using a 5% significance level, is outside

$$0 \pm (2.26)(2.2547)\sqrt{1/6 + 1/5} \doteq 0 \pm 1.72$$

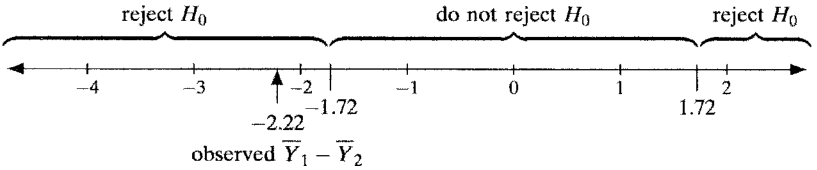


Figure 5.10 Plot showing the rejection region.

The observed value of −2.22 months also falls in the rejection region. Compared to the regions constructed when the variances were assumed known, the region where the null hypothesis is *not* rejected in this case is wider.

5.7 TWO-SAMPLE INFERENCE: SCALE

5.7.1 F-Distribution

The final inference procedure to be discussed in this chapter deals with the equality of variances of two normal populations.

Result 5.4. Given two random samples of size n_1 and n_2, with sample variances s_1^2 and s_2^2, from two normal populations with variances σ_1^2 and σ_2^2, the variable

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

has an F-distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

The F-distribution (named in honor of Sir R. A. Fisher) does not have a simple mathematical formula, but most statistical packages can compute tables of the distribution. The F-distribution is indexed by the degrees of freedom associated with s_1^2 (the numerator degrees of freedom) and the degrees of freedom associated with s_2^2 (the denominator degrees of freedom). A picture of the F-distribution is presented in Figure 5.11. The distribution is skewed; the extent of skewness depends on the degrees of freedom. As *both* increase, the distribution becomes more symmetric.

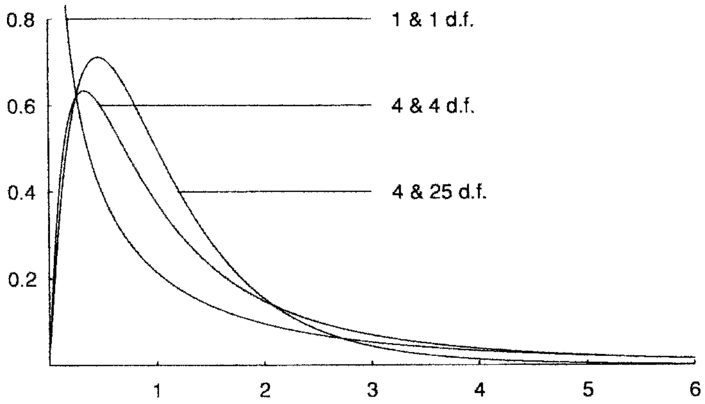


Figure 5.11 F-distribution for three sets of degrees of freedom.

We write $F_{v_1,v_2,\alpha}$ to indicate the 100αth percentile value of an F-statistic with v_1 and v_2 degrees of freedom. The mean of an F-distribution is $v_2/(v_2 - 2)$, for $v_2 > 2$; the variance is given in Note 5.3. In this note you will also find a brief discussion of the relationship between the four distributions we have now discussed: normal, chi-square, Student t, and F.

It is clear that

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

is a pivotal variable, listed as number 8 in Table 5.1. Inferences can be made on the *ratio* σ_1^2/σ_2^2. [To make inferences about σ_1^2 (or σ_2^2) by itself, we would use the chi-square distribution and the procedure outlined in Chapter 4.] Conveniently, if we want to test whether the variances are equal, that is, $\sigma_1^2 = \sigma_2^2$, the ratio σ_1^2/σ_2^2 is equal to 1 and "drops out" of the pivotal variable, which can then be written

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{s_1^2}{s_2^2}$$

We would reject the null hypothesis of equality of variances if the observed ratio s_1^2/s_1^2 is "very large" or "very small," how large or small to be determined by the F-distribution.

5.7.2 Testing and Estimation

Continuing Example 5.5, the sample variances in Table 5.3 were $s_1^2 = (1.4470)^2 = 2.0938$ and $s_2^2 = (0.9618)^2 = 0.9251$. Associated with s_1^2 are $6 - 1 = 5$ degrees of freedom, and with s_2^2, $5 - 1 = 4$ degrees of freedom. Under the null hypothesis of equality of population variances, the ratio s_1^2/s_2^2 has an F-distribution with (5, 4) degrees of freedom. For a two-tailed test at the 10% level, we need $F_{5,4,0.05}$ and $F_{5,4,0.95}$. From Table A.7, the value for $F_{5,4,0.95}$ is 6.26. Using the relationship $F_{v_1,v_2,\alpha} = 1/F_{v_2,v_1,1-\alpha}$, we obtain $F_{5,4,0.05} = 1/F_{4,5,0.95} = 0.19$. The value of F observed is $F_{5,4} = s_1^2/s_2^2 = 2.0938/0.9251 \doteq 2.26$.

From Figure 5.12 it is clear that the null hypothesis of equality of variances is not rejected. Notice that the rejection region is not symmetric about 1, due to the zero bound on the left-hand side. It is instructive to consider F-ratios for which the null hypothesis would have been rejected. On the right-hand side, $F_{5,4,0.95} = 6.26$; this implies that s_1^2 must be 6.26 times as large as s_2^2 before the 10% significance level is reached. On the left-hand side, $F_{5,4,0.05} = 0.19$, so that s_1^2 must be 0.19 times as small as s_2^2 before the 10% significance level is reached. These are reasonably wide limits (even at the 10% level).

At one time statisticians recommended performing an F-test for equality of variances before going on to the t-test. This is no longer thought to be useful. In small samples the F-test cannot reliably detect even quite large differences in variance; in large samples it will reject the hypothesis of equality of variances for differences that are completely unimportant. In addition, the F-test is extremely sensitive to the assumption of normality, even in large samples. The modern solution is to use an approximate version of the t-test that does not assume equal variances (see Note 5.2). This test can be used in all cases or only in cases where the sample variances appear substantially different. In large samples it reduces to the Z-test based on pivotal variable 2 in Table 5.1. The F-test should be restricted to the case where there is a genuine scientific interest in whether two variances are equal.

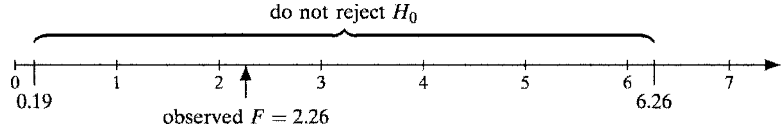


Figure 5.12 Plot showing nonrejection of the null hypothesis of equality of variances.

A few comments about terminology: Sample variances that are (effectively) the same are called *homogeneous*, and those that are not are called *heterogeneous*. A test for equality of population variances, then, is a test for homogeneity or heterogeneity. In the more technical statistical literature, you will find the equivalent terms *homoscedasticity* and *heteroscedasticity tests*.

A confidence interval on the ratios of the population variances σ_1^2/σ_2^2 can be constructed using the pivotal variable approach once more. To set up a $100(1-\alpha)\%$ confidence interval, we need the $100(\alpha/2)$ percentile and $100(1-\alpha/2)$ percentile of the F-distribution.

Continuing with Example 5.5, suppose that we want to construct a 90% confidence interval on σ_1^2/σ_2^2 on the basis of the observed sample. Values for the 5th and 95th percentiles have already been obtained: $F_{5,4,0.05} = 0.19$ and $F_{5,4,0.95} = 6.26$. A 90% confidence interval on σ_1^2/σ_2^2 is then determined by

$$\left(\frac{s_1^2/s_2^2}{F_{5,4,0.95}}, \frac{s_1^2/s_2^2}{F_{5,4,0.05}}\right)$$

For the data observed, this is

$$\left(\frac{2.0938/0.9251}{6.26}, \frac{2.0938/0.9251}{0.19}\right) = (0.36, 11.9)$$

Thus, on the basis of the data observed, we can be 90% confident that the interval (0.36, 11.9) straddles or covers the ratio σ_1^2/σ_2^2 of the population variances. This interval includes 1.0. So, also on the basis of the estimation procedure, we conclude that $\sigma_1^2/\sigma_2^2 = 1$ is not unreasonable.

A 90% confidence interval on the ratio of the standard deviations, σ_1/σ_2, can be obtained by taking square roots of the points (0.36, 11.9), producing (0.60, 3.45) for the interval.

5.8 SAMPLE-SIZE CALCULATIONS

One of the questions most frequently asked of a statistician is: How big must my n be? Stripped of its pseudojargon, a valid question is being asked: How many observations are needed in this study? Unfortunately, the question cannot be answered before additional information is supplied. We first put the requirements in words in the context of a study comparing two treatments; then we introduce the appropriate statistical terminology. To determine sample size, you need to specify or know:

1. How variable the data are
2. The chance that you are willing to tolerate concluding incorrectly that there is an effect when the treatments are equivalent
3. The magnitude of the effect to be detected
4. The certainty with which you wish to detect the effect

Each of these considerations is clearly relevant. The more variation in the data, the more observations are needed to pin down a treatment effect; when there is no difference, there is a chance that a difference will be observed, which due to sampling variability is declared significant. The more certain you want to be of detecting an effect, the more observations you will need, everything else remaining equal. Finally, if the difference in the treatments is very large, a rather economical experiment can be run; conversely, a very small difference in the treatments will require very large sample sizes to detect.

We now phrase the problem in statistical terms: The model we want to consider involves two normal populations with equal variances σ^2, differing at most in their means, μ_1 and μ_2. To determine the sample size, we must specify:

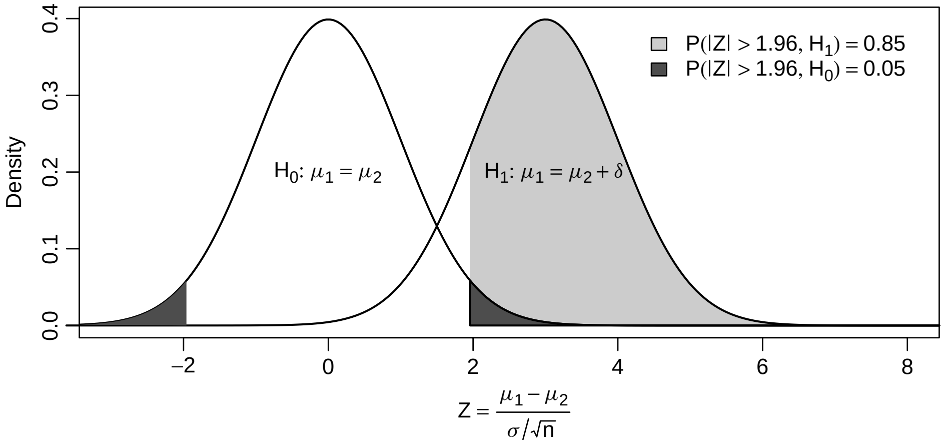


Figure 5.13 Distributions of the Z-statistic under the null and an alternative hypothesis. The probability of $Z < -1.96$ or $Z > 1.96$ under the null hypothesis (the level) is dark gray. The probability under the alternative hypothesis (the power) is light gray.

1. σ^2
2. The probability, α, of a Type I error
3. The magnitude of the difference $\mu_1 - \mu_2$ to be detected
4. The power, $1 - \beta$, or equivalently, the probability of a Type II error, β

Figure 5.13 shows an example of these quantities visually. There are two normal distributions, corresponding to the distribution of the two-sample Z-statistic under the null hypothesis that two means are equal and under the alternative hypothesis that the mean of the first sample is greater than the mean of the second. In the picture, $(\mu_1 - \mu_2)/(\sigma/\sqrt{n}) = 3$, for example, a difference in means of $\mu_1 - \mu_2 = 3$ with a standard deviation $\sigma = 10$ and a sample size of $n = 100$.

The level, or Type I error rate, is the probability of rejecting the null hypothesis if it is true. We are using a 0.05-level two-sided test. The darkly shaded regions are where $Z < -1.96$ or $Z > 1.96$ if $\mu_1 = \mu_2$, adding up to a probability (area under the curve) of 0.05. The power is the probability of rejecting the null hypothesis if it is not true. The lightly shaded region is where $Z > 1.96$ if the alternative hypothesis is true. In theory there is a second lightly shaded region where $Z < -1.96$, but this is invisibly small: There is effectively no chance of rejecting the null hypothesis "in the wrong direction." In this example the lightly shaded region adds up to a probability of 0.85, meaning that we would have 85% power.

Sample sizes are calculated as a function of

$$\Delta = \frac{|\mu_1 - \mu_2|}{\sigma}$$

which is defined to be the standardized distance between the two populations. For a two-sided test, the formula for the required sample size *per group* is

$$n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2}$$

It is instructive to contemplate this formula. The standardized difference enters as a square. Thus, to detect a treatment different *half* as small as perhaps considered initially will require *four* times as many observations per group. Decreasing the probabilities of Type I and Type II errors has the same effect on the sample size; it increases it. However, the increment is not as drastic as it is with Δ. For example, to reduce the probability of a Type I error from 0.05 to 0.025 changes the Z-value from $Z_{0.975} = 1.96$ to $Z_{0.9875} = 2.24$; even though $Z_{1-\alpha/2}$ is squared, the effect will not even be close to doubling the sample size. Finally, the formula assumes that the difference $\mu_1 - \mu_2$ can either be positive or negative. If the direction of the difference can be specified beforehand, a one-tailed value for Z can be used. This will result in a reduction of the sample sizes required for each of the groups, since $z_{1-\alpha}$ would be used.

Example 5.6. At a significance level of $1 - \alpha = 0.95$ (one tail) and power $1 - \beta = 0.80$, a difference $\Delta = 0.3$ is to be detected. The appropriate Z-values are

$$Z_{0.95} = 1.645 \text{ (a more accurate value than given in Table A.2)}$$

$$Z_{0.80} = 0.84$$

The sample size required per group is

$$n = \frac{2(1.645 + 0.84)^2}{(0.3)^2} = 137.2$$

The value is rounded up to 138, so that at least 138 observations *per group* are needed to detect the specified difference at the specified significant level and power.

Suppose that the variance σ^2 is not known, how can we estimate the sample size needed to detect a standardized difference Δ? One possibility is to have an estimate of the variance σ^2 based on a previous study or sample. Unfortunately, no explicit formulas can be given when the variance is estimated; many statistical texts suggest adding between two and four observations per group to get a reasonable approximation to the sample size (see below).

Finally, suppose that *one group*—as in a paired experiment—is to be used to determine whether a populations mean μ differs from a hypothesized mean μ_0. Using the same standardized difference $\Delta = |\mu - \mu_0|/\sigma$, it can be shown that the appropriate number in the group is

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2}$$

or one-half the number needed in one group in the two-sample case. This is why tables for sample sizes in the one-sample case tell you, in order to apply the table to the two-sample case, to (1) double the number in the table, and (2) use that number *for each group*.

Example 5.7. Consider data involving PKU children. Assume that IQ in the general population has mean $\mu = 100$ and standard deviation $= 15$. Suppose that a sample of eight PKU children whose diet has been terminated has an average IQ of 94, which is not significantly different from 100. How large would the sample have to be to detect a difference of six IQ points (i.e., the population mean is 94)? The question cannot be answered yet. (Before reading on: What else must be specified?) Additionally, we need to specify the Type I and Type II errors. Suppose that $\alpha = 0.05$ and $\beta = 0.10$. We make the test one-tailed because the alternative hypothesis is that the IQ for PKU children is less than that of children in the general population. A value of $\beta = 0.10$ implies that the power is $1 - \beta = 0.90$. We first calculate the standardized distance

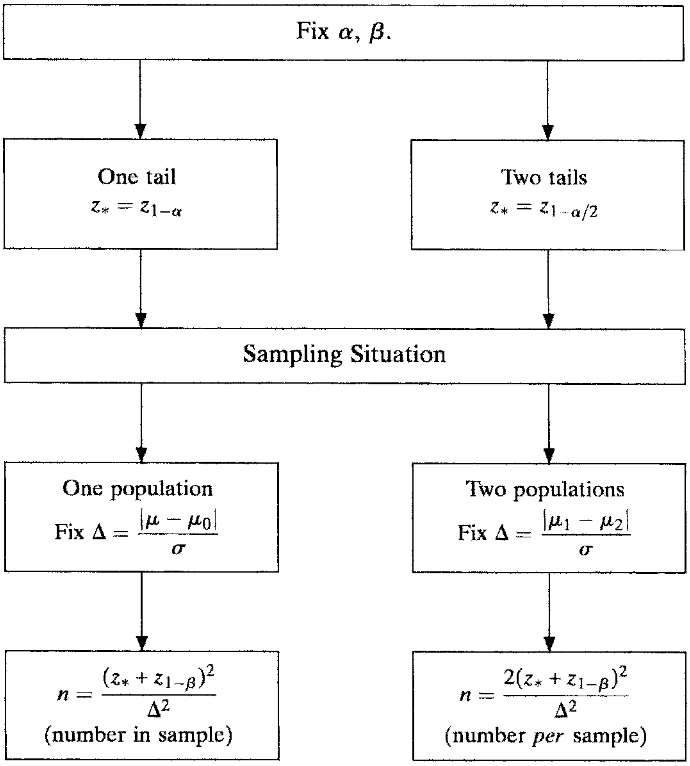
$$\Delta = \frac{|94 - 100|}{15} = \frac{6}{15} = 0.40$$

Then $z_{1-0.05} = z_{0.95} = 1.645$ and $z_{1-0.10} = z_{0.90} = 1.28$. Hence,

$$n = \frac{(1.645 + 1.28)^2}{(0.40)^2} = 53.5$$

Rounding up, we estimate that it will take a sample of 54 observations to detect a difference of $100 - 94 = 6$ IQ points (or greater) with probabilities of Type I and Type II errors as specified.

If the variance is not known, and estimated by s^2, say $s^2 = 15^2$, then statistical tables (not included in this book) indicate that the sample size is 55, not much higher than the 54 we calculated. A summary outline for calculating sample sizes is given in Figure 5.14.


Fix α, β.
One tail
$z_* = z_{1-\alpha}$
Two tails
$z_* = z_{1-\alpha/2}$
Sampling Situation
One population
Fix $\Delta = \frac{|\mu - \mu_0|}{\sigma}$
Two populations
Fix $\Delta = \frac{|\mu_1 - \mu_2|}{\sigma}$
$n = \frac{(z_* + z_{1-\beta})^2}{\Delta^2}$
(number in sample)
$n = \frac{2(z_* + z_{1-\beta})^2}{\Delta^2}$
(number *per* sample)

Comments:

1. In the case of two populations, if $\sigma_1^2 \neq \sigma_2^2$, define $\sigma^2 = (\sigma_1^2 + \sigma_2^2)/2$ and proceed as before.
2. If σ is to be estimated from the data, add to the calculated values the following values for an approximate sample size:

		One population	**Two populations**
One tail	$\alpha = 0.05$	Add 2	Add 1
	$\alpha = 0.01$	Add 4	Add 2
Two tails	$\alpha = 0.05$	Add 2	Add 1
	$\alpha = 0.01$	Add 3	Add 2

Figure 5.14 Sample-size calculations for measurement data.

There is something artificial and circular about all of these calculations. If the difference Δ is known, there is no need to perform an experiment to estimate the difference. Calculations of this type are used primarily to make the researcher aware of the kinds of differences that can be detected. Often, a calculation of this type will convince a researcher *not* to carry out a piece of research, or at least to think very carefully about the possible ways of increasing precision, perhaps even contemplating a radically different attack on the problem. In addition, the size of a sample may be limited by considerations such as cost, recruitment rate, or time constraints beyond control of the investigations. In Chapter 6 we consider questions of sample size for discrete variables.

NOTES

5.1 Inference by Means of Pivotal Variables: Some Comments

1. The problem of finding pivotal variables is a task for statisticians. Part of the problem is that such variables are not always unique. For example, when working with the normal distribution, why not use the sample median rather than the sample mean? After all, the median is admittedly more robust. However, the variance of the sample median is larger than that of the sample mean, so that a less precise probabilistic statement would result.

2. In many situations there is no exactly pivotal variable available in small samples, although a pivotal variable can typically be found in large samples.

3. The principal advantage of using the pivotal variable approach is that it gives you a unified picture of a great number of procedures that you will need.

4. There is a philosophical problem about the interpretation of a confidence interval. For example, consider the probability inequality

$$P[-1.96 \leq Z \leq 1.96] = 0.95$$

which leads to a 95% confidence interval for the mean of a normal population on the basis of a random sample of size n:

$$P\left[\overline{Y} - \frac{1.96\sigma}{\sqrt{n}} \leq \mu \leq \overline{Y} + \frac{1.96\sigma}{\sqrt{n}}\right] = 0.95$$

It is argued that once $\overline{Y}$ is observed, *this* interval either covers the mean or not; that is, P is either 0 or 1. One answer is that probabilities are not associated with a particular event—whether they have occurred or may occur at some future time—but with a population of events. For this reason we say *after the fact* that we are 95% confident that the mean is in the interval, *not* that the probability is 0.95 that the mean is in the interval.

5. Given two possible values for a parameter, which one will be designated as the null hypothesis value and which one as the alternative hypothesis value in a hypothesis testing situation? If nothing else is given, the designation will be arbitrary. Usually, there are at least four considerations in designating the *null value* of a parameter:

a. Often, the null value of the parameter permits calculation of a p-value. For example, if there are two hypotheses, $\mu = \mu_0$ and $\mu \neq \mu_0$, only under $\mu = \mu_0$ can we calculate the probability of an occurrence of the observed value or a more extreme value.

b. Past experience or previous work may suggest a specified value. The new experimentation or treatment then has a purpose: rejection of the value established previously, or assessment of the magnitude of the change.

c. Occam's razor can be appealed to. It states: "Do not multiply hypotheses beyond necessity," meaning in this context that we usually start from the value of a parameter that we would assume if no new data were available or to be produced.

d. Often, the null hypothesis is a "straw man" we hope to reject, for example, that a new drug has the same effect as a placebo.

6. Sometimes it is argued that the smaller the p-value, the stronger the treatment effect. You now will recognize that this cannot be asserted baldly. Consider the two-sample t-test. A p-value associated with this test will depend on the quantities $\mu_1 - \mu_2, s_p, n_1$, and n_2. Thus, differences in p-values between two experiments may simply reflect differences in sample size or differences in background variability (as measured by s_p).

5.2 Additional Notes on the t-Test

1. *Heterogeneous variances in the two-sample.* t*-test.* Suppose that the assumption of homogeneity of variances in the two-sample t-test is not tenable. What can be done? At least three avenues are open:

a. Use an approximation to the t procedure.

b. Transform the data.

c. Use another test, such as a nonparametric test.

With respect to the last point, alternative approaches are discussed in Chapter 8. With respect to the first point, one possibility is to rank the observations from smallest to largest (disregarding group membership) and then carry out the t-test on the ranks. This is a surprisingly good *test* but does not allow us to estimate the magnitude of the difference between the two groups. See Conover and Iman [1981] for an interesting discussion and Thompson [1991] for some precautions. Another approach adjusts the degrees of freedom of the two-sample t-test. The procedure is as follows: Let $Y_1 \sim N(\mu_1, \sigma_1^2)$ and $Y_2 \sim N(\mu_2, \sigma_2^2)$, and samples of size n_1 and n_2 are taken, respectively. The variable

$$\frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

has a standard normal distribution. However, the analogous quantity with the population variances σ_1^2 and σ_2^2 replaced by the sample variances s_1^2 and s_2^2 does not have a t-distribution. The problem of finding the distribution of this quantity is known as the *Behrens–Fisher problem.* It is of theoretical interest in statistics because there is no exact solution to such an apparently simple problem. There are, however, perfectly satisfactory practical solutions. One approach adjusts the degrees of freedom of this statistic in relation to the extent of dissimilarity of the two sample variances. The t-table is entered not with $n_1 + n_2 - 2$ degrees of freedom, but with

$$\text{degrees of freedom} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1+1) + (s_2^2/n_2)^2/(n_2+1)} - 2$$

This value need not be an integer; if you are working from tables of the t-distribution rather than software, it may be necessary to round down this number. The error in this approximation is very small and is likely to be negligible compared to the errors caused by nonnormality. For large samples (e.g., $n_1, n_2 > 30$), the statistic

$$\frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

can be treated as a standard normal deviate even if the distribution of Y is not a normal distribution.

2. *The two-sample* t*-test and design of experiments.* Given that a group has to be divided into two subgroups, the arrangement that minimizes the standard error of the difference is that of equal sample sizes in each group when there is a common σ^2. To illustrate, suppose that 10 objects are to be partitioned into two groups; consider the multiplier $\sqrt{1/n_1 + 1/n_2}$, which determines the relative size of the standard errors.

n_1	n_2	$\sqrt{1/n_1 + 1/n_2}$
5	5	0.63
6	4	0.65
7	3	0.69
8	2	0.79

This list indicates that small deviations from a 5 : 5 ratio do not affect the multiplier very much. It is sometimes claimed that sample sizes must be equal for a valid t-test: Except for giving the smallest standard error of the difference, there is no such constraint.

3. *The "wrong"* t*-test.* What is the effect of carrying out a two-sample t-test on paired data, that is, data that should have been analyzed by a paired t-test? Usually, the level of significance is reduced. On the one hand, the degrees of freedom are *increased* from $(n - 1)$, assuming n pairs of observations, to $2(n - 1)$, but at the same time, additional variation is introduced, so that the standard error of the difference is now larger. In any event the assumption of statistical independence between "groups" is usually inappropriate.

4. *Robustness of the* t*-test.* The t-test tends to be sensitive to *outliers*, unusually small or large values of a variable. We discuss other methods of analysis in Chapter 8. As a matter of routine, you should always graph the data in some way. A simple box plot or histogram will reveal much about the structure of the data. An outlier may be a perfectly legitimate value and its influence on the t-test entirely appropriate, but it is still useful to know that this influence is present.

5.3 Relationships and Characteristics of the Fixed Distributions in This Chapter

We have already suggested some relationships between the fixed distributions. The connection is more remarkable yet and illustrates the fundamental role of the normal distribution. The basic connection is between the standard normal and the chi-square distribution. Suppose that we draw randomly 10 independent values from a standard normal distribution, square each value, and sum them. This sum is a random variable. What is its sampling distribution? It turns out to be chi-square with 10 degrees of freedom. Using notation, let $Z_1, Z_2, \ldots, Z_{10}$ be the values of Z obtained in drawings 1 to 10. Then, $Z_1^2 + \cdots + Z_{10}^2$ has a chi-square distribution with 10 degrees of freedom: $\chi_{10}^2 = Z_1^2 + \cdots + Z_{10}^2$. This generalizes the special case $\chi_1^2 = Z^2$.

The second connection is between the F-distribution and the chi-square distribution. Suppose that we have two independent chi-square random variables with v_1 and v_2 degrees of freedom. The ratio

$$\frac{\chi_{v_1}^2/v_1}{\chi_{v_2}^2/v_2} = F_{v_1,v_2}$$

has an F-distribution with v_1 and v_2 degrees of freedom. Finally, the square of a t-variable with v degrees of freedom is $F_{1,v}$. Summarizing yields

$$\chi_v^2 = \sum_{i=1}^{v} Z_i^2, \qquad t_v^2 = F_{1,v} = \frac{\chi_1^2/1}{\chi_v^2/v}$$

A special case connects all four pivotal variables:

$$Z^2 = t_\infty^2 = \chi_1^2 = F_{1,\infty}$$

Thus, given the F-table, all the other tables can be generated from it. For completeness, we summarize the mean and variance of the four fixed distributions:

Distribution	**Symbol**	**Mean**	**Variance**	
Normal	Z	0	1	
Student t	t_v	0	$\frac{v}{v-2}$	$(v > 2)$
Chi-square	χ_v^2	v	$2v$	
Fisher's F	F_{v_1,v_2}	$\dfrac{v_2}{v_2-2}$	$\dfrac{2v_2^2(v_1+v_2-2)}{v_1(v_2-2)^2(v_2-4)}$	$(v_2 > 4)$

5.4 One-Sided Tests and One-Sided Confidence Intervals

Corresponding to one-sided (one-tailed) tests are one-sided confidence intervals. A one-sided confidence interval is derived from a pivotal quantity in the same was as a two-sided confidence interval. For example, in the case of a one-sample t-test, a pivotal equation is

$$P\left[-\infty \le \frac{\overline{x}-\mu}{s/\sqrt{n}} \le t_{n-1,1-\alpha}\right] = 1-\alpha$$

Solving for μ produces a $100(1-\alpha)\%$ *upper* one-sided confidence interval for μ : $(\overline{x} - t_{n-1,1-\alpha}s/\sqrt{n}, \infty)$. Similar intervals can be constructed for all the pivotal variables.

PROBLEMS

5.1 Rickman et al. [1974] made a study of changes in serum cholesterol and triglyceride levels of subjects following the Stillman diet. The diet consists primarily of protein and animal fats, restricting carbohydrate intake. The subjects followed the diet with length of time varying from 3 to 17 days. (Table 5.4). The mean cholesterol level increased significantly from 215 mg per/100 mL at baseline to 248 mg per/100 mL at the end of the diet. In this problem, we deal with the triglyceride level.

(a) Make a histogram or stem-and-leaf diagram of the *changes* in triglyceride levels.

(b) Calculate the average change in triglyceride level. Calculate the standard error of the difference.

(c) Test the significance of the average change.

(d) Construct a 90% confidence interval on the difference.

(e) The authors indicate that subjects (5,6), (7,8), (9,10), and (15,16) were "repeaters," that is, subjects who followed the diet for two sequences. Do you think it is

Table 5.4 Diet Data for Problem 5.1

		Weight (kg)		Triglyceride (mg/100 ml)	
Subject	Days on Diet	Initial	Final	Baseline	Final
1	10	54.6	49.6	159	194
2	11	56.4	52.8	93	122
3	17	58.6	55.9	130	158
4	4	55.9	54.6	174	154
5	9	60.0	56.7	148	93
6	6	57.3	55.5	148	90
7	3	62.7	59.6	85	101
8	6	63.6	59.6	180	99
9	4	71.4	69.1	92	183
10	4	72.7	70.5	89	82
11	4	49.6	47.1	204	100
12	7	78.2	75.0	182	104
13	8	55.9	53.2	110	72
14	7	71.8	68.6	88	108
15	7	71.8	66.8	134	110
16	14	70.5	66.8	84	81

reasonable to include their data the "second time around" with that of the other subjects? Supposing not, how would you now analyze the data? Carry out the analysis. Does it change your conclusions?

5.2 In data of Dobson et al. [1976], 36 patients with a confirmed diagnosis of phenylketonuria (PKU) were identified and placed on dietary therapy before reaching 121 days of age. The children were tested for IQ (Stanford–Binet test) between the ages of 4 and 6; subsequently, their normal siblings of closest age were also tested with the Stanford–Binet. The following are the first 15 pairs listed in the paper:

Pair	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
IQ of PKU case	89	98	116	67	128	81	96	116	110	90	76	71	100	108	74
IQ of sibling	77	110	94	91	122	94	121	114	88	91	99	93	104	102	82

(a) State a suitable null and an alternative hypotheses with regard to these data.

(b) Test the null hypothesis.

(c) State your conclusions.

(d) What are your assumptions?

(e) Discuss the concept of power with respect to this set of data using the fact that PKU invariably led to mental retardation until the cause was found and treatment comprising a restricted diet was instituted.

(f) The mean difference (PKU case − sibling) in IQ for the full 36 pairs was −5.25; the standard deviation of the difference was 13.18. Test the hypothesis of no difference in IQ for this entire set of data.

5.3 Data by Mazze et al. [1971] deal with the preoperative and postoperative creatinine clearance (ml/min) of six patients anesthetized by halothane:

	Patient					
	1	2	3	4	5	6
Preoperative	110	101	61	73	143	118
Postoperative	149	105	162	93	143	100

(a) Why is the paired t-test preferable to the two-sample t-test in this case?

(b) Carry out the paired t-test and test the significance of the difference.

(c) What is the model for your analysis?

(d) Set up a 99% confidence interval on the difference.

(e) Graph the data by plotting the pairs of values for each patient.

5.4 Some of the physiological effects of alcohol are well known. A paper by Squires et al. [1978] assessed the acute effects of alcohol on auditory brainstem potentials in humans. Six volunteers (including the three authors) participated in the study. The latency (delay) in response to an auditory stimulus was measured before and after an intoxicating dose of alcohol. Seven different peak responses were identified. In this exercise, we discuss only latency peak 3. Measurements of the latency of peak (in milliseconds after the stimulus onset) in the six subjects were as follows:

	Latency of Peak					
	1	2	3	4	5	6
Before alcohol	3.85	3.81	3.60	3.68	3.78	3.83
After alcohol	3.82	3.95	3.80	3.87	3.88	3.94

(a) Test the significance of the difference at the 0.05 level.

(b) Calculate the p-value associated with the result observed.

(c) Is your p-value based on a one- or two-tailed test? Why?

(d) As in Problem 5.3, graph these data and state your conclusion.

(e) Carry out an (incorrect) two-sample test and state your conclusions.

(f) Using the sample variances s_1^2 and s_2^2 associated with the set of readings observed before and after, calculate the variance of the difference, *assuming* independence (call this variance 1). How does this value compare with the variance of the difference calculated in part (a)? (Call this variance 2.) Why do you suppose variance 1 is so much bigger than variance 2? The *average* of the differences is the same as the difference in the averages. Show this. Hence, the two-sample t-test differed from the paired t-test only in the divisor. Which of the two tests in more powerful in this case, that is, declares a difference significant when in fact there is one?

5.5 The following data from Schechter et al. [1973] deal with sodium chloride preference as related to hypertension. Two groups, 12 normal and 10 hypertensive subjects, were isolated for a week and compared with respect to Na^+ intake. The following are the average daily Na^+ intakes (in milligrams):

Normal	10.2	2.2	0.0	2.6	0.0	43.1	45.8	63.6	1.8	0.0	3.7	0.0
Hypertensive	92.8	54.8	51.6	61.7	250.8	84.5	34.7	62.2	11.0	39.1		

(a) Compare the average daily Na^+ intake of the hypertensive subjects with that of the normal volunteers by means of an appropriate t-test.

(b) State your assumptions.

(c) Assuming that the population variances are not homogeneous, carry out an appropriate t-test (see Note 5.2).

5.6 Kapitulnik et al. [1976] compared the metabolism of a drug, zoxazolamine, in placentas from 13 women who smoked during pregnancy and 11 who did not. The purpose of the study was to investigate the presence of the drug as a possible proxy for the rate at which benzo[*a*]pyrene (a by-product of cigarette smoke) is metabolized. The following data were obtained in the measurement of zoxazolamine hydroxylase production (nmol $3H_2O$ formed/g per hour):

Nonsmoker	0.18	0.36	0.24	0.50	0.42	0.36	0.50	0.60	0.56	0.36	0.68		
Smoker	0.66	0.60	0.96	1.37	1.51	3.56	3.36	4.86	7.50	9.00	10.08	14.76	16.50

(a) Calculate the sample mean and standard deviation for each group.

(b) Test the assumption that the two sample variances came from a population with the same variance.

(c) Carry out the t-test using the approximation to the t-procedure discussed in Note 5.2. What are your conclusions?

(d) Suppose we agree that the variability (as measured by the standard deviations) is proportional to the level of the response. Statistical theory then suggests that the logarithms of the responses should have roughly the same variability. Take logarithms of the data and test, once more, the homogeneity of the variances.

5.7 Sometime you may be asked to do a two-sample t-test knowing only the mean, standard deviation, and sample sizes. A paper by Holtzman et al. [1975] dealing with terminating a phenylalanine-restricted diet in 4-year-old children with phenylketonuria (PKU) illustrates the problem. The purpose of the diet is to reduce the phenylalanine level. A high level is associated with mental retardation. After obtaining informed consent, eligible children of 4 years of age were randomly divided into two groups. Children in one group had their restricted diet terminated while children in the other group were continued on the restricted diet. At 6 years of age, the phenylalanine levels were tested in all children and the following data reported:

	Diet Terminated	**Diet Continued**
Number of children	5	4
Mean phenylalanine level (mg/dl)	26.9	16.7
Standard deviation	4.1	7.3

(a) State a reasonable null hypothesis and alternative hypothesis.

(b) Calculate the pooled estimate of the variance s_p^2.

(c) Test the null hypothesis of part (a). Is your test one-tailed, or two? Why?

(d) Test the hypothesis that the sample variances came from two populations with the same variance.

(e) Construct a 95% confidence interval on the difference in the population phenylalanine levels.

(f) Interpret the interval constructed in part (e).

(g) "This set of data has little power," someone says. What does this statement mean? Interpret the implications of a Type II error in this example.

(h) What is the interpretation of a Type I error in this example? Which, in your opinion, is more serious in this example: a Type I error or a Type II error?

(i) On the basis of these data, what would you recommend to a parent with a 4-year-old PKU child?

(j) Can you think of some additional information that would make the analysis more precise?

5.8 Several population studies have demonstrated an inverse correlation of sudden infant death syndrome (SIDS) rate with birthweight. The occurrence of SIDS in one of a pair of twins provides an opportunity to test the hypothesis that birthweight is a major determinant of SIDS. The data shown in Table 5.5 consist of the birthweights (in grams) of each of 22 dizygous twins and each of 19 monozygous twins.

(a) With respect to the dizygous twins, test the hypothesis given above. State the null hypothesis.

(b) Make a similar test on the monozygous twins.

(c) Discuss your conclusions.

Table 5.5 Birthweight Data for Problem 5.8

Dizygous Twins		Monozygous Twins	
SID	Non-SID	SID	Non-SID
1474	2098	1701	1956
3657	3119	2580	2438
3005	3515	2750	2807
2041	2126	1956	1843
2325	2211	1871	2041
2296	2750	2296	2183
3430	3402	2268	2495
3515	3232	2070	1673
1956	1701	1786	1843
2098	2410	3175	3572
3204	2892	2495	2778
2381	2608	1956	1588
2892	2693	2296	2183
2920	3232	3232	2778
3005	3005	1446	2268
2268	2325	1559	1304
3260	3686	2835	2892
3260	2778	2495	2353
2155	2552	1559	2466
2835	2693		
2466	1899		
3232	3714		

Source: D. R. Peterson, Department of Epidemiology, University of Washington.

5.9 A pharmaceutical firm claims that a new analgesic drug relieves mild pain under standard conditions for 3 hours with a standard deviation of 1 hour. Sixteen patients are tested under the same conditions and have an average pain relief of 2.5 hours. The hypothesis that the population mean of this sample is also 3 hours is to be tested against the hypothesis that the population mean is in fact less than 3 hours; $\alpha = 0.5$.

(a) What is an appropriate test?

(b) Set up the appropriate critical region.

(c) State your conclusion.

(d) Suppose that the sample size is doubled. State precisely how the nonrejection region for the null hypothesis is changed.

5.10 Consider Problem 3.9, dealing with the treatment of essential hypertension. Compare treatments A and B by means of an appropriate t-test. Set up a 99% confidence interval on the reduction of blood pressure under treatment B as compared to treatment A.

5.11 During July and August 1976, a large number of Legionnaires attending a convention died of mysterious and unknown cause. Epidemiologists talked of "an outbreak of Legionnaires' disease." One possible cause was thought to be toxins: nickel, in particular. Chen et al. [1977] examined the nickel levels in the lungs of nine of the cases, and selected nine controls. All specimens were coded by the Centers for Disease Control in Atlanta before being examined by the investigators. The data are as follows (μg per 100 g dry weight):

Legionnaire cases	65	24	52	86	120	82	399	87	139
Control cases	12	10	31	6	5	5	29	9	12

Note that there was no attempt to match cases and controls.

(a) State a suitable null hypothesis and test it.

(b) We now know that Legionnaires' disease is caused by a bacterium, genus *Legionella*, of which there are several species. How would you explain the "significant" results obtained in part (a)? (Chen et al. [1977] consider various explanations also.)

5.12 Review Note 5.3. Generate a few values for the normal, t, and chi-square tables from the F-table.

5.13 It is claimed that a new drug treatment can substantially reduce blood pressure. For purposes of this exercise, assume that only diastolic blood pressure is considered. A certain population of hypertensive patients has a mean blood pressure of 96 mmHg. The standard deviation of diastolic blood pressure (variability from subject to subject) is 12 mmHg. To be biologically meaningful, the new drug treatment should lower the blood pressure to at least 90 mmHg. A random sample of patients from the hypertensive population will be treated with the new drug.

(a) Assuming that $\alpha = 0.05$ and $\beta = 0.05$, calculate the sample size required to demonstrate the effect specified.

(b) Considering the labile nature of blood pressure, it might be argued that any "treatment effect" will merely be a "put-on-study effect." So the experiment is redesigned to consider two random samples from the hypertensive population, one of which will receive the new treatment, and the other, a placebo. Assuming the same specifications as above, what is the required sample size per group?

(c) Blood pressure readings are notoriously variable. Suppose that a subject's diastolic blood pressure varies randomly from measurement period to measurement period with a standard deviation of 4 mmHg. Assuming that measurement variability is independent of subject-to-subject variability, what is the overall variance or the total variability in the population? Recalculate the sample sizes for the situation described in parts (a) and (b).

(d) Suppose that the *change* in blood pressure from baseline is used. Suppose that the standard deviation of the change is 6 mmHg. How will this change the sample sizes of parts (a) and (b)?

5.14 In a paper in the *New England Journal of Medicine*, Rodeheffer et al. [1983] assessed the effect of a medication, nifedipine, on the number of painful attacks in patients with Raynaud's phenomenon. This phenomenon causes severe digital pain and functional disability, particularly in patients with underlying connective tissue disease. The drug causes "vascular smooth-muscle relaxation and relief of arterial vasospasm." In this study, 15 patients were selected and randomly assigned to one of two treatment sequences: placebo–nifedipine, or nifedipine–placebo. The data in Table 5.6 were obtained.

(a) Why were patients *randomly* assigned to one of the two sequences? What are the advantages?

(b) The data of interest are in the columns marked "placebo" and "nifedipine." State a suitable null hypothesis and alternative hypothesis for these data. Justify your choices. Test the significance of the difference in total number of attacks in two weeks on placebo with that of treatment. Use a t-test on the differences in the response. Calculate the p-value.

(c) Construct a 95% confidence interval for the difference. State your conclusions.

(d) Make a scatter plot of the placebo response (x-axis) vs. the nifedipine response (y-axis). If there was no significant difference between the treatments, about what line should the observations be scattered?

(e) Suppose that a statistician considers only the placebo readings and calculates a 95% confidence interval on the population mean. Similarly, the statistician calculates a 95% confidence interval on the nifedipine mean. A graph is made to see if the intervals overlap. Do this for these data. Compare your results with that of part (c). Is there a contradiction? Explain.

(f) One way to get rid of outliers is to carry out the following procedure: Take the differences of the data in columns 7 (placebo) and 9 (nifedipine), and rank them disregarding the signs of the differences. Put the sign of the difference on the rank. Now, carry out a paired t-test on the signed ranks. What would be an appropriate null hypothesis? What would be an appropriate alternative hypothesis? Name one advantage and one disadvantage of this procedure. (It is one form of a nonparametric test discussed in detail in Chapter 8.)

5.15 Rush et al. [1973] reported the design of a randomized controlled trial of nutritional supplementation in pregnancy. The trial was to be conducted in a poor American black population. The variable of interest was the birthweight of infants born to study participants; study design called for the random allocation of participants to one of three treatment groups. The authors then state: "The required size of the treatment groups was calculated from the following statistics: the standard deviation of birthweight ... is of the order of 500 g. An increment of 120 g in birthweight was arbitrarily taken to constitute a biologically meaningful gain. Given an expected difference between subjects and controls of 120 g, the required sample size for each group, in order to have a 5% risk of falsely rejecting, and a 20% risk of falsely accepting the null hypothesis, is about 320."

Table 5.6 Effect of Nifedipine on Patients with Raynaud's Phenomenon

						Placebo		Nifedipine	
Case	Age (yr)/ Gender	Diagnosis[a]	History of Digital Ulcer	ANA[b]	Duration of Raynaud's Phenomenon (yr)	Total Number of Attacks in 2 Weeks	Patient Assessment of Therapy[c]	Total Number of Attacks in 2 Weeks	Patient Assessment of Therapy[c]
1	49/F	R[d]	No	20	4	15	0	0	3+
2	20/F	R	No	Neg	3	3	1+	5	0
3	23/F	R	No	Neg	8	14	2+	6	2+
4	33/F	R	No	640	5	6	0	0	3+
5	31/F	R[d]	No	2560	2	12	0	2	3+
6	52/F	PSS	No	320	3	6	1+	1	0
7	45/M	PSS[d]	Yes	320	4	3	1+	2	2+
8	49/F	PSS	Yes	320	4	22	0	30	1+
9	29/M	PSS	Yes	1280	7	15	0	14	1+
10	33/F	PSS[d]	No	2560	9	11	1+	5	1+
11	36/F	PSS	Yes	2560	13	7	2+	2	3+
12	33/F	PSS[d]	Yes	2560	11	12	0	4	2+
13	39/F	PSS	No	320	6	45	0	45	0
14	39/M	PSS	Yes	80	6	14	1+	15	2+
15	32/F	SLE[d]	Yes	1280	5	35	1+	31	2+

Source: Data from Rodeheffer et al. [1983].

[a] R Raynaud's phenomenon without systemic disease; PSS, Raynaud's phenomenon with progressive systemic sclerosis; SLE, Raynaud's phenomenon with systemic lupus erythematosus (in addition, this patient had cryoglobulinemia).

[b] Reciprocal of antinuclear antibody titers.

[c] The Wilcoxon signed rank test, two-tailed, was performed on the patient assessment of placebo vs. nifedipine therapy: $p = 0.02$. Global assessment scale: 1− = worse; 0 = no change; 1+ = minimal improvement; 2+ = moderate improvement; and 3+ = marked improvement.

[d] Previous unsuccessful treatment with prazosin.

(a) What are the values for α and β?

(b) What is the estimate of Δ, the standardized difference?

(c) The wording in the paper suggests that sample size calculations are based on a two-sample test. Is the test one-tailed or two?

(d) Using a one-tailed test, verify that the sample size per group is $n = 215$. The number 320 reflects adjustments for losses and, perhaps, "multiple comparisons" since there are three groups (see Chapter 12).

5.16 This problem deals with the data of Problem 5.14. In column 4 of Table 5.6, patients are divided into those with a history of digital ulcers and those without. We want to compare these two groups. There are seven patients with a history and eight without.

(a) Consider the total number of attacks (in column 9) on the active drug. Carry out a two-sample t-test. Compare the group with a digital ulcer history with the group without this history. State your assumptions and conclusions.

(b) Rank all the observations in column 9, then separate the ranks into the two groups defined in part (a). Now carry out a two-sample t-test on the ranks. Compare your conclusions with those of part (b). Name an advantage to this approach. Name a disadvantage to this approach.

(c) We now do the following: Take the difference between the "placebo" and "nifedipine" columns and repeat the procedures of parts (a) and (b). Supposing that the conclusions of part (a) are not the same as those in this part, how would you interpret such discrepancies?

(d) The test carried out in part (c) is often called a *test for interaction*. Why do you suppose that this is so?

REFERENCES

Bednarek, F. J., and Roloff, D. W. [1976]. Treatment of apnea of prematurity with aminophylline. *Pediatrics*, **58**: 335–339. Used with permission.

Chen, J. R., Francisco, R. B., and Miller, T. E. [1977]. Legionnaires' disease: nickel levels. *Science*, **196**: 906–908. Copyright © 1977 by the AAAS.

Conover, W. J., and Iman, R. L. [1981]. Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician*, **35**: 124–129.

Dobson, J. C., Kushida, E., Williamson, M., and Friedman, E. G. [1976]. Intellectual performance of 36 phenylketonuria patients and their non-affected siblings. *Pediatrics*, **58**: 53–58. Used with permission.

Holtzman, N. A., Welcher, D. M., and Mellits, E. D. [1975]. Termination of restricted diet in children with phenylketonuria: a randomized controlled study. *New England Journal of Medicine*, **293**: 1121–1124.

Kapitulnik, J., Levin, W., Poppers, J., Tomaszewski, J. E., Jerina, D. M., and Conney, A. H. [1976]. Comparison of the hydroxylation of zoxazolamine and benzo[a]pyrene in human placenta: effect of cigarette smoking. *Clinical Pharmaceuticals and Therapeutics*, **20**: 557–564.

Mazze, R. I., Shue, G. L., and Jackson, S. H. [1971]. Renal dysfunction associated with methoxyflurane anesthesia. *Journal of the American Medical Association*, **216**: 278–288. Copyright © 1971 by the American Medical Association.

Rickman, R., Mitchell, N., Dingman, J., and Dalen, J. E. [1974]. Changes in serum cholesterol during the Stillman diet. *Journal of the American Medical Association*, **228**: 54–58. Copyright © 1974 by the American Medical Association.

Rodeheffer, R. J., Romner, J. A., Wigley, F., and Smith, C. R. [1983]. Controlled double-blind trial of Nifedipine in the treatment of Raynaud's phenomenon. *New England Journal of Medicine*, **308**: 880–883.

Rush, D., Stein, Z., and Susser, M. [1973]. The rationale for, and design of, a randomized controlled trial of nutritional supplementation in pregnancy. *Nutritional Reports International*, **7**: 547–553. Used with permission of the publisher, Butterworth-Heinemann.

Schechter, P. J., Horwitz, D., and Henkin, R. I. [1973]. Sodium chloride preference in essential hypertension. *Journal of the American Medical Association*, **225**: 1311–1315. Copyright © 1973 by The American Medical Association.

Squires, K. C., Chen, N. S., and Starr, A. [1978]. Acute effects of alcohol on auditory brainstem potentials in humans. *Science*, **201**: 174–176.

Thompson, G. L. [1991]. A unified approach to rank tests for multivariate and repeated measures designs. *Journal of the American Statistical Association*, **86**: 410–419.

Zelazo, P. R., Zelazo, N. A., and Kolb, S. [1972]. "Walking" in the newborn. *Science*, **176**: 314–315.

CHAPTER 6

Counting Data

6.1 INTRODUCTION

From previous chapters, recall the basic ideas of statistics. *Descriptive statistics* present data, usually in summary form. Appropriate *models* describe data concisely. The model *parameters* are *estimated* from the data. *Standard errors* and *confidence intervals* quantify the precision of estimates. Scientific hypotheses may be tested. A *formal hypothesis test* involves four things: (1) planning an experiment, or recognizing an opportunity, to collect appropriate data; (2) selecting a *significance level* and *critical region*; (3) collecting the data; and (4) rejecting the *null hypothesis* being tested if the value of the test statistic falls into the critical region. A less formal approach is to compute the *p-value*, a measure of how plausibly the data agree with the null hypothesis under study. The remainder of this book shows you how to apply these concepts in different situations, starting with the most basic of all data: counts.

Throughout recorded history people have been able to count. The word *statistics* comes from the Latin word for "state"; early statistics were counts used for the purposes of the state. Censuses were conducted for military and taxation purposes. Modern statistics is often dated from the 1662 comments on the Bills of Mortality in London. The Bills of Mortality counted the number of deaths due to each cause. John Graunt [1662] noticed patterns of regularity in the Bills of Mortality (see Section 3.3.1). Such vital statistics are important today for assessing the public health. In this chapter we return to the origin of statistics by dealing with data that arise by counting the number of occurrences of some event.

Count data lead to many different models. The following sections present examples of count data. The different types of count data will each be presented in three steps. First, you learn to recognize count data that fit a particular model. (This is the diagnosis phase.) Second, you examine the model to be used. (You learn about the illness.) Third, you learn the methods of analyzing data using the model. (At this stage you learn how to treat the disease.)

6.2 BINOMIAL RANDOM VARIABLES

6.2.1 Recognizing Binomial Random Variables

Four conditions characterize binomial data:

1. A response or trait takes on one and only one of two possibilities. Such a response is called a *binary response*. Examples are:

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2 Copyright © 2004 John Wiley & Sons, Inc.

- **a.** In a survey of the health system, people are asked whether or not they have hospitalization insurance.
- **b.** Blood samples are tested for the presence or absence of an antigen.
- **c.** Rats fed a potential carcinogen are examined for tumors.
- **d.** People are classified as having or not having cleft lip.
- **e.** Injection of a compound does or does not cause cardiac arrhythmia in dogs.
- **f.** Newborn children are classified as having or not having Down syndrome.

2. The response is observed a known number of times. Each observation of the response is sometimes called a *Bernoulli trial*. In condition 1(a) the number of trials is the number of people questioned. In 1(b), each blood sample is a trial. Each newborn child constitutes a trial in 1(f).

3. The chance, or probability, that a particular outcome occurs is the same for each trial. In a survey such as 1(a), people are sampled at random from the population. Since each person has the same chance of being interviewed, the probability that the person has hospitalization insurance is the same in each case. In a laboratory receiving blood samples, the samples could be considered to have the same probability of having an antigen *if* the samples arise from members of a population who submit tests when "randomly" seeking medical care. The samples would not have the same probability if batches of samples arrive from different environments: for example, from schoolchildren, a military base, and a retirement home.

4. The outcome of one trial must not be influenced by the outcome of other trials. Using the terminology of Chapter 5, the trials outcomes are independent random variables. In 1(b), the trials would not be independent if there was contamination between separate blood samples. The newborn children of 1(f) might be considered independent trials for the occurrence of Down syndrome if each child has different parents. If multiple births are in the data set, the assumption of independence would not be appropriate.

We illustrate and reinforce these ideas by examples that may be modeled by the binomial distribution.

Example 6.1. Weber et al. [1976] studied the irritating effects of cigarette smoke. Sixty subjects sat, in groups of five to six, in a 30-m^2 climatic chamber. Tobacco smoke was produced by a smoking machine. After 10 cigarettes had been smoked, 47 of the 60 subjects reported that they wished to leave the room.

Let us consider the appropriateness of the binomial model for these data. Condition 1 is satisfied. Each subject was to report whether or not he or she desired to leave the room. The answer gives one of two possibilities: yes or no. Sixty trials are observed to take place (i.e., condition 2 is satisfied).

The third condition requires that each subject have the same probability of "wishing to leave the room." The paper does not explain how the subjects were selected. Perhaps the authors advertised for volunteers. In this case, the subjects might be considered "representative" of a larger population who would volunteer. The probability would be the *unknown* probability that a person selected at random from this larger population would wish to leave the room.

As we will see below, the binomial model is often used to make inferences about the unknown probability of an outcome in the "true population." Many would say that an experiment such as this shows that cigarette smoke irritates people. The extension from the ill-defined population of this experiment to humankind in general does *not* rest on this experiment. It must be based on other formal or informal evidence that humans do have much in common; in particular, one would need to assume that if one portion of humankind is irritated by cigarette smoke, so will other segments. Do you think such inferences are reasonable?

The fourth condition needed is that the trials are independent variables. The authors report in detail that the room was cleared of all smoke between uses of the climatic chamber. There should not be a carryover effect here. Recall that subjects were tested in groups of five or six. How do you think that one person's response would be changed if another person were coughing? Rubbing the eyes? Complaining? It seems possible that condition 4 is not fulfilled; that is, it seems possible that the responses were not independent.

In summary, a binomial model might be used for these data, but with some reservation. The overwhelming majority of data collected on human populations is collected under less than ideal conditions; a subjective evaluation of the worth of an experiment often enters in.

Example 6.2. Karlowski et al. [1975] reported on a controlled clinical trial of the use of ascorbic acid (vitamin C) for the common cold. Placebo and vitamin C were randomly assigned to the subjects; the experiment was to be a double-blind trial. It turned out that some subjects were testing their capsules and claimed to know the medication. Of 64 subjects who tested the capsule and guessed at the treatment, 55 were correct. Could such a split arise by chance if testing did not help one to guess correctly?

One thinks of using a binomial model for these data since there is a binary response (correct or incorrect guess) observed on a known number of people. Assuming that people tested only their own capsules, the guesses should be statistically independent. Finally, if the guesses are "at random," each subject should have the same probability—one-half—of making a correct guess since half the participants receive vitamin C and half a placebo. This binomial model would lead to a test of the hypothesis that the probability of a correct guess was 1/2.

Example 6.3. Bucher et al. [1976] studied the occurrence of hemolytic disease in newborns resulting from ABO incompatibility between the parents. Parents are said to be incompatible if the father has antigens that the mother lacks. This provides the opportunity for production of maternal antibodies from fetal–maternal stimulation. Low-weight immune antibodies that cross the placental barrier apparently cause the disease [Cavalli-Sforza and Bodmer, 1999]. The authors reviewed 7464 consecutive infants born at North Carolina Hospital. Of 3584 "black births," 43 had ABO hemolytic disease. What can be said about the true probability that a black birth has ABO hemolytic disease?

It seems reasonable to consider the number of ABO hemolytic disease cases to be binomial. The presence of disease among the 3584 trials should be independent (assuming that no parents had more than one birth during the period of case recruitment—October 1965 to March 1973—and little or no effect from kinship of parents). The births may conceptually be thought of as a sample of the population of "potential" black births during the given time period at the hospital.

6.2.2 Binomial Model

In speaking about a Bernoulli trial, without reference to a particular example, it is customary to label one outcome as a "success" and the other outcome as a "failure." The mathematical model for the binomial distribution depends on two parameters: n, the number of trials, and π, the probability of a success in one trial. A binomial random variable, say Y, is the count of the number of successes in the n trials. Of course, Y can only take on the values $0, 1, 2, \dots, n$. If π, the probability of a success, is large (close to 1), then Y, the number of successes, will tend to be large. Conversely, if the probability of success is small (near zero), Y will tend to be small.

To do statistical analysis with binomial variables, we need the probability distribution of Y. Let k be an integer between 0 and n inclusive. We need to know $P[Y = k]$. In other words, we want the probability of k successes in n independent trials when π is the probability of success. The symbol $b(k; n, \pi)$ will be used to denote this probability. The answer involves the *binomial coefficient*. The binomial coefficient $\left(\begin{array}{c} n \\ k \end{array}\right)$ is the number of different ways that

k objects may be selected from n objects. (Problem 6.24 helps you to derive the value of $\binom{n}{k}$.) For each positive integer n, n factorial (written $n!$) is defined to be $1 \times 2 \times \cdots \times n$. So $6! = 1 \times 2 \times 3 \times 4 \times 5 \times 6 = 720$. 0!, zero factorial, is defined to be 1. With this notation the binomial coefficient may be written

$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!} = \frac{n(n-1)\cdots(k+1)}{(n-k)(n-k-1)\cdots 1} \tag{1}$$

Example 6.4. This is illustrated with the following two cases:

1. Of 10 residents, three are to be chosen to cover a hospital service on a holiday. In how many ways may the residents be chosen? The answer is

$$\binom{10}{3} = \frac{10!}{7!\,3!} = \frac{1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8 \times 9 \times 10}{(1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7)(1 \times 2 \times 3)} = 120$$

2. Of eight consecutive patients, four are to be assigned to drug A and four to drug B. In how many ways may the assignments be made? Think of the eight positions as eight objects; we need to choose four for the drug A patients. The answer is

$$\binom{8}{4} = \frac{8!}{4!\,4!} = \frac{1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8}{(1 \times 2 \times 3 \times 4)(1 \times 2 \times 3 \times 4)} = 70$$

The binomial probability, $b(k; n, \pi)$, may be written

$$b(k; n, \pi) = \binom{n}{k} \pi^k (1-\pi)^{n-k} \tag{2}$$

Example 6.5. Ten patients are treated surgically. For each person there is a 70% chance of successful surgery (i.e., $\pi = 0.7$). What is the probability of only five or fewer successful surgeries?

$$\begin{aligned} P[\text{five or fewer successful cases}] &= P[\text{five successful cases}] + P[\text{four successful cases}] \\ &\quad + P[\text{three successful cases}] + P[\text{two successful cases}] \\ &\quad + P[\text{one successful case}] + P[\text{no successful case}] \\ &= b(5; 10, 0.7) + b(4; 10, 0.7) + b(3; 10, 0.7) + b(2; 10, 0.7) \\ &\quad + b(1; 10, 0.7) + b(0; 10, 0.7) \\ &= 0.1029 + 0.0368 + 0.0090 + 0.0014 + 0.0001 + 0.0000 \\ &= 0.1502 \end{aligned}$$

(*Note*: The actual value is 0.1503; the answer 0.1502 is due to round-off error.)

The binomial probabilities may be calculated directly or found by a computer program. The mean and variance of a binomial random variable with parameters π and n are given by

$$\begin{aligned} E(Y) &= n\pi \\ \text{var}(Y) &= n\pi(1-\pi) \end{aligned} \tag{3}$$

From equation (3) it follows that Y/n has the expected value π:

$$E\left(\frac{Y}{n}\right) = \pi \tag{4}$$

In other words, the proportion of successes in n binomial trials is an unbiased estimate of the probability of success.

6.2.3 Hypothesis Testing for Binomial Variables

The hypothesis-testing framework established in Chapter 4 may be used for the binomial distribution. There is one minor complication. The binomial random variable can take on only a finite number of values. Because of this, it may not be possible to find hypothesis tests such that the significance level is precisely some fixed value. If this is the case, we construct regions so that the significance level is close to the desired significance level.

In most situations involving the binomial distribution, the number of trials (n) is known. We consider statistical tests about the true value of π. Let $p = Y/n$. If π is hypothesized to be π_0, an observed value of p close to π_0 reinforces the hypothesis; a value of p differing greatly from π_0 makes the hypothesis seem unlikely.

Procedure 1. To construct a significance test of H_0: $\pi = \pi_0$ against H_A: $\pi \neq \pi_0$, at significance level α:

1. Find the smallest c such that $P\ [|p - \pi_0| \geq c] \leq \alpha$ when H_0 is true.
2. Compute the *actual* significance level of the test; the actual significance level is $P[|p - \pi_0| \geq c]$.
3. Observe p, call it $\widehat{p}$; reject H_0 if $|\widehat{p} - \pi_0| \geq c$.

The quantity c is used to determine the critical value (see Definition 4.19); that is, determine the bounds of the rejection region, which will be $\pi_0 \pm c$. Equivalently, working in the Y scale, the region is defined by $n\pi_0 \pm nc$.

Example 6.6. For $n = 10$, we want to construct a test of the null hypothesis H_0: $\pi = 0.4$ vs. the alternative hypothesis H_A: $\pi \neq 0.4$. Thus, we want a two-sided test. The significance level is to be as close to $\alpha = 0.05$ as possible. We work in the $Y = np$ scale. Under H_0, Y has mean $n\pi = (10)(0.4) = 4$. We want to find a value C such that $P[|Y - 4| \geq C]$ is as close to $\alpha = 0.05$ (and less than α) as possible. The quantity C is the distance Y is from the null hypothesis value 4. Using the definition of the binomial distribution, we construct Table 6.1.

The closest α-value to 0.05 is $\alpha = 0.0183$; the next value is 0.1012. Hence we choose $C = 4$; we reject the null hypothesis H_0: $n\pi = 4$ if $Y = 0$ or $Y \geq 8$; equivalently, if $p = 0$ or $p \geq 0.8$, or in the original formulation, if $|p - 0.4| \geq 0.4$ since $C = 10c$.

Table 6.1 C-Values for Example 6.6

C	$4 - C$	$C + 4$	$P[\|Y - 4\| \geq C] = \alpha$
6	—	10	$0.0001 = P[Y = 10]$
5	—	9	$0.0017 = P[Y \geq 9]$
4	0	8	$0.0183 = P[Y = 0] + P[Y \geq 8]$
3	1	7	$0.1012 = P[Y \leq 1] + P[Y \geq 7]$
2	2	6	$0.3335 = P[Y \leq 2] + P[Y \geq 6]$
1	3	5	$0.7492 = P[Y \leq 3] + P[Y \geq 5]$

Procedure 2. To find the *p-value* for testing the hypothesis H_0: $\pi = \pi_0$ vs. H_A: $\pi \neq \pi_0$:

1. Observe $p : \widehat{p}$ is now fixed, where $\widehat{p} = y/n$.
2. Let $\widetilde{p}$ be a binomial random variable with parameters n and π_0. The p-value is $P[|\widetilde{p} - \pi_0| \geq |\widehat{p} - \pi_0|]$.

Example 6.7. Find the p-value for testing $\pi = 0.5$ if $n = 10$ and we observe that $p = 0.2$. $|\widetilde{p} - 0.5| \geq |0.2 - 0.5| = 0.3$ only if $\widetilde{p} = 0.0, 0.1, 0.2, 0.8, 0.9$, or 1.0. The p-value can be computed by software or by adding up the probabilities of the "more extreme" values: $0.0010 + 0.0098 + 0.0439 + 0.0439 + 0.0098 + 0.0010 = 0.1094$. Tables for this calculation are provided in the Web appendix. The appropriate one-sided hypothesis test and calculation of a one-sided p-value is given in Problem 6.25.

6.2.4 Confidence Intervals

Confidence intervals for a binomial proportion can be found by computer or by looking up the confidence limits in a table. Such tables are not included in this book, but are available in any standard handbook of statistical tables, for example, Odeh et al. [1977], Owen [1962], and Beyer [1968].

6.2.5 Large-Sample Hypothesis Testing

The central limit theorem holds for binomial random variables. If Y is binomial with parameters n and π, then for "large n,"

$$\frac{Y - E(Y)}{\sqrt{\text{var}(Y)}} = \frac{Y - n\pi}{\sqrt{n\pi(1-\pi)}}$$

has approximately the same probability distribution as an $N(0, 1)$ random variable. Equivalently, since $Y = np$, the quantity $(p - \pi)/\sqrt{\pi(1-\pi)/n}$ approaches a normal distribution. We will work interchangeably in the p scale or the Y scale. For large n, hypothesis tests and confidence intervals may be formed by using critical values of the standard normal distribution.

The closer π is to 1/2, the better the normal approximation will be. If $n \leq 50$, it is preferable to use tables for the binomial distribution and hypothesis tests as outlined above. A reasonable rule of thumb is that n is "large" if $n\pi(1-\pi) \geq 10$.

In using the central limit theorem, we are approximating the distribution of a discrete random variable by the continuous normal distribution. The approximation can be improved by using a *continuity correction.* The normal random variable with continuity correction is given by

$$Z_c = \begin{cases} \dfrac{Y - n\pi - 1/2}{\sqrt{n\pi(1-\pi)}} & \text{if } Y - n\pi > 1/2 \\[2ex] \dfrac{Y - n\pi}{\sqrt{n\pi(1-\pi)}} & \text{if } |Y - n\pi| \leq 1/2 \\[2ex] \dfrac{Y - n\pi + 1/2}{\sqrt{n\pi(1-\pi)}} & \text{if } Y - n\pi < -1/2 \end{cases}$$

For $n\pi(1-\pi) \geq 100$, or quite large, the factor of 1/2 is usually ignored.

Procedure 3. Let Y be binomial n, π, with a large n. A hypothesis test of H_0: $\pi = \pi_0$ vs. H_A: $\pi \neq \pi_0$ at significance level α is given by computing Z_c with $\pi = \pi_0$. The null hypothesis is rejected if $|Z_c| \geq z_{1-\alpha/2}$.

Example 6.8. In Example 6.2, of the 64 persons who tested their capsules, 55 guessed the treatment correctly. Could so many people have guessed the correct treatment "by chance"? In Example 6.2 we saw that chance guessing would correspond to $\pi_0 = 1/2$. At a 5% significance level, is it plausible that $\pi_0 = 1/2$?

As $n\pi_0(1 - \pi_0) = 64 \times 1/2 \times 1/2 = 16$, a large-sample approximation is reasonable. $y - n\pi_0 = 55 - 64 \times 1/2 = 23$, so that

$$Z_c = \frac{Y - n\pi_0 - 1/2}{\sqrt{n\pi_0(1 - \pi_0)}} = \frac{22.5}{\sqrt{64 \times 1/2 \times 1/2}} = 5.625$$

As $|Z_c| = 5.625 > 1.96 = z_{0.975}$, the null hypothesis that the correct guessing occurs purely by chance must be rejected.

Procedure 4. The large-sample two-sided p-value for testing H_0: $\pi = \pi_0$ vs. H_A: $\pi \neq \pi_0$ is given by $2(1 - \Phi(|Z_c|))$. $\Phi(x)$ is the probability that an $N(0, 1)$ random variable is less than x. $|Z_c|$ is the absolute value of Z_c.

6.2.6 Large-Sample Confidence Intervals

Procedure 5. For large n, say $n\widehat{p}(l - \widehat{p}) \geq 10$, an approximate $100(1-\alpha)\%$ confidence interval for π is given by

$$\left(\widehat{p} - z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1 - \widehat{p})}{n}},\quad \widehat{p} + z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1 - \widehat{p})}{n}}\right) \tag{5}$$

where $\widehat{p} = y/n$ is the observed proportion of successes.

Example 6.9. Find a 95% confidence interval for the true fraction of black children having ABO hemolytic disease in the population represented by the data of Example 6.3. Using formula (5) the confidence interval is

$$\frac{43}{3584} \pm 1.96\sqrt{\frac{(43/3584)(1 - 43/3584)}{3584}} \quad \text{or} \quad (0.0084, 0.0156)$$

6.3 COMPARING TWO PROPORTIONS

Often, one is not interested in only one proportion but wants to compare two proportions. A health services researcher may want to see whether one of two races has a higher percentage of prenatal care. A clinician may wish to discover which of two drugs has a higher proportion of cures. An epidemiologist may be interested in discovering whether women on oral contraceptives have a higher incidence of thrombophlebitis than those not on oral contraceptives. In this section we consider the statistical methods appropriate for comparing two proportions.

6.3.1 Fisher's Exact Test

Data to estimate two different proportions will arise from observations on two populations. Call the two sets of observations sample 1 and sample 2. Often, the data are presented in 2 × 2 (verbally, "two by two") tables as follows:

	Success	**Failure**
Sample 1	n_{11}	n_{12}
Sample 2	n_{21}	n_{22}

The first sample has n_{11} successes in $n_{11} + n_{12}$ trials; the second sample has n_{21} successes in $n_{21} + n_{22}$ trials. Often, the null hypothesis of interest is that the probability of success in the two populations is the same. *Fisher's exact test* is a test of this hypothesis for small samples.

The test uses the row and column totals. Let $n_{1\cdot}$ denote summation over the second index; that is, $n_{1\cdot} = n_{11} + n_{12}$. Similarly define $n_{2\cdot}$, $n_{\cdot 1}$, and $n_{\cdot 2}$. Let $n_{\cdot\cdot}$ denote summation over both indices; that is, $n_{\cdot\cdot} = n_{11} + n_{12} + n_{21} + n_{22}$. Writing the table with row and column totals gives:

	Success	**Failure**	
Sample 1	n_{11}	n_{12}	$n_{1\cdot}$
Sample 2	n_{21}	n_{22}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

Suppose that the probabilities of success in the two populations are the same. Suppose further that we are given the row and column totals but *not* n_{11}, n_{12}, n_{21}, and n_{22}. What is the probability distribution of n_{11}?

Consider the $n_{\cdot\cdot}$ trials as $n_{\cdot\cdot}$ objects; for example, $n_{1\cdot}$ purple balls and $n_{2\cdot}$ gold balls. Since each trial has the same probability of success, any subset of $n_{\cdot 1}$ trials (balls) has the same probability of being chosen as any other. Thus, the probability that n_{11} has the value k is the same as the probability that there are k purple balls among $n_{\cdot 1}$ balls chosen without replacement from an urn with $n_{1\cdot}$ purple balls and $n_{2\cdot}$ gold balls. The probability distribution of n_{11} is called the *hypergeometric distribution*.

The mathematical form of the hypergeometric probability distribution is derived in Problem 6.26.

Example 6.10. Kennedy et al. [1981] consider patients who have undergone coronary artery bypass graft surgery (CABG). CABG takes a saphenous vein from the leg and connects the vein to the aorta, where blood is pumped from the heart, and to a coronary artery, an artery that supplies the heart muscle with blood. The vein is placed beyond a narrowing, or stenosis, in the coronary artery. If the artery would close at the narrowing, the heart muscle would still receive blood. There is, however, some risk to this open heart surgery. Among patients with moderate narrowing (50 to 74%) of the left main coronary artery emergency cases have a high surgical mortality rate. The question is whether emergency cases have a surgical mortality different from that of nonemergency cases. The in-hospital mortality figures for emergency surgery and other surgery were:

	Discharge Status	
Surgical Priority	**Dead**	**Alive**
Emergency	1	19
Other	7	369

From the hypergeometric distribution, the probability of an observation this extreme is $0.3419 = P[n_{11} \geq 1] = P[n_{11} = 1] + \cdots + P[n_{11} = 8]$. (Values for $n_{\cdot\cdot}$ this large are not tabulated and need to be computed directly.) These data do not show any difference beyond that expected by chance.

Example 6.11. Sudden infant death syndrome (SIDS), or crib death, results in the unexplained death of approximately two of every 1000 infants during their first year of life. To study the genetic component of such deaths, Peterson et al. [1980] examined sets of twins with at least one SIDS child. If there is a large genetic component, the probability of both twins dying will be larger for identical twin sets than for fraternal twin sets. If there is no genetic component, but only an environmental component, the probabilities should be the same. The following table gives the data:

	SIDS Children	
Type of Twin	**One**	**Both**
Monozygous (identical)	23	1
Dizygous (fraternal)	35	2

The Fisher's exact test one-sided p-value for testing that the probability is higher for monozygous twins is $p = 0.784$. Thus, there is no evidence for a genetic component in these data.

6.3.2 Large-Sample Tests and Confidence Intervals

As mentioned above, in many situations one wishes to compare proportions as estimated by samples from two populations to see if the true population parameters might be equal or if one is larger than the other. Examples of such situations are a drug and placebo trial comparing the percentage of patients experiencing pain relief; the percentage of rats developing tumors under diets involving different doses of a food additive; and an epidemiologic study comparing the percentage of infants suffering from malnutrition in two countries.

Suppose that the first binomial variable (the sample from the first population) is of size n_1 with probability π_1, estimated by the sample proportion p_1. The second sample estimates π_2 by p_2 from a sample of size n_2.

It is natural to compare the proportions by the difference $p_1 - p_2$. The mean and variance are given by

$$E(p_1 - p_2) = \pi_1 - \pi_2,$$

$$\text{var}(p_1 - p_2) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

A version of the central limit theorem shows that for large n_1 and n_2 [say, both $n_1\pi_1(1-\pi_1)$ and $n_2\pi_2(1-\pi_2)$ greater than 10],

$$\frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} = Z$$

is an approximately normal pivotal variable. From this, hypothesis tests and confidence intervals develop in the usual manner, as illustrated below.

Example 6.12. The paper by Bucher et al. [1976] discussed in Example 6.3 examines racial differences in the incidence of ABO hemolytic disease by examining records for infants born at North Carolina Memorial Hospital. In this paper a variety of possible ways of defining hemolytic disease are considered. Using their class I definition, the samples of black and white infants have the following proportions with hemolytic disease:

$$\text{black infants, } n_1 = 3584, \qquad p_1 = \frac{43}{3584}$$

$$\text{white infants, } n_2 = 3831, \qquad p_2 = \frac{17}{3831}$$

It is desired to perform a two-sided test of the hypothesis $\pi_1 = \pi_2$ at the $\alpha = 0.05$ significance level. The test statistic is

$$Z = \frac{(43/3584) - (17/3831)}{\sqrt{[(43/3584)(1 - 43/3584)]/3584 + [(17/3831)(1 - 17/3831)]/3831}} \doteq 3.58$$

The two-sided p-value is $P[|Z| \geq 3.58] = 0.0003$ from Table A.1. As $0.0003 < 0.05$, the null hypothesis of equal rates, $\pi_1 = \pi_2$, is rejected at the significance level 0.05.

The pivotal variable may also be used to construct a confidence interval for $\pi_1 - \pi_2$. Algebraic manipulation shows that the endpoints of a symmetric (about $p_1 - p_2$) confidence interval are given by

$$p_1 - p_2 \pm z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

For a 95% confidence interval $z_{1-\alpha/2} = 1.96$ and the interval for this example is

$$0.00756 \pm 0.00414 \quad \text{or} \quad (0.00342, 0.01170)$$

A second statistic for testing for equality in two proportions is the χ^2 (chi-square) statistic. This statistic is considered in more general situations in Chapter 7. Suppose that the data are as follows:

	Sample 1	**Sample 2**	
Success	$n_1p_1 = n_{11}$	$n_2p_2 = n_{12}$	$n_{1\cdot}$
Failure	$n_1(1-p_1) = n_{21}$	$n_2(1-p_2) = n_{22}$	$n_{2\cdot}$
	$n_1 = n_{\cdot 1}$	$n_2 = n_{\cdot 2}$	$n_{\cdot\cdot}$

A statistic for testing H_0: $\pi_1 = \pi_2$ is the χ^2 statistic with one degree of freedom. It is calculated by

$$X^2 = \frac{n_{\cdot\cdot}(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}$$

For technical reasons (Note 6.2) the chi-square distribution with continuity correction, designated by X_c^2, is used by some people. The formula for X_c^2 is

$$X_c^2 = \frac{n_{\cdot\cdot}(|n_{11}n_{22} - n_{12}n_{21}| - \frac{1}{2}n_{\cdot\cdot})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}$$

For the Bucher et al. [1976] data, the values are as follows:

	Race		
ABO Hemolytic Disease	**Black**	**White**	**Total**
Yes	43	17	60
No	3541	3814	7355
Total	3584	3831	7415

$$X^2 = \frac{7415(43 \times 3814 - 17 \times 3541)^2}{60(7355)(3584)(3831)} = 13.19$$

$$X_c^2 = \frac{7415(|43 \times 3814 - 17 \times 3541| - 7415/2)^2}{60(7355)(3584)(3831)} = 12.26$$

These statistics, for large n, have a chi-square (χ^2) distribution with one degree of freedom under the null hypothesis of equal proportions. If the null hypothesis is not true, X^2 or X_c^2 will tend to be large. The null hypothesis is rejected for large values of X^2 or X_c^2. Table A.3 has χ^2 critical values. The Bucher data have $p < 0.001$ since the 0.001 critical value is 10.83 and require rejection of the null hypothesis of equal proportions.

From Note 5.3 we know that $\chi_1^2 = Z^2$. For this example the value of $Z^2 = 3.58^2 = 12.82$ is close to the value $X^2 = 13.19$. The two values would have been identical (except for rounding) if we had used in the calculation of Z an estimate of the standard error of $\sqrt{pq(1/n_1 + 1/n_2)}$, where $p = 60/7415$ is the pooled estimate of π under the null hypothesis $\pi_1 = \pi_2 = \pi$.

6.3.3 Finding Sample Sizes Needed for Testing the Difference between Proportions

Consider a study planned to test the equality of the proportions π_1 and π_2. Only studies in which both populations are sampled the same number of times, $n = n_1 = n_2$, will be considered here. There are five quantities that characterize the performance and design of the test:

1. π_1, the proportion in the first population.
2. π_2, the proportion in the second population under the alternative hypothesis.
3. n, the number of observations to be obtained from *each* of the two populations.
4. The significance level α at which the statistical test will be made. α is the probability of rejecting the null hypothesis when it is true. The null hypothesis is that $\pi_1 = \pi_2$.
5. The probability, β, of accepting the null hypothesis when it is not true, but the alternative is true. Here we will have $\pi_1 \neq \pi_2$ under the alternative hypothesis (π_1 and π_2 as specified in quantities 1 and 2 above).

These quantities are interrelated. It is not possible to change one of them without changing at least one of the others. The actual determination of sample size is usually an iterative process; the usual state of affairs is that the desire for precision and the practicality of obtaining an appropriate sample size are in conflict. In practice, one usually considers various possible combinations and arrives at a "reasonable" sample size or decides that it is not possible to perform an adequate experiment within the constraints involved.

The "classical" approach is to specify π_1, π_2 (for the alternative hypothesis), α, and β. These parameters determine the sample size n. Table A.8 gives some sample sizes for such binomial studies using one-sided hypothesis tests (see Problem 6.27). An approximation for n is

$$n = 2\left\{\frac{z_{1-\alpha} + z_{1-\beta}}{\pi_1 - \pi_2}\sqrt{\frac{1}{2}[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]}\right\}^2$$

where $\alpha = 1 - \Phi(z_{1-\alpha})$; that is, $z_{1-\alpha}$ is the value such that a $N(0, 1)$ variable Z has $P[Z > z_{1-\alpha}] = \alpha$. In words, $z_{1-\alpha}$ is the one-sided normal α critical value. Similarly, $z_{1-\beta}$ is the one-sided normal β critical value.

Figure 6.1 is a flow diagram for calculating sample sizes for discrete (binomial) as well as continuous variables. It illustrates that the format is the same for both: first, values of α and β are selected. A one- or two-sided test determines $z_{1-\alpha}$ or $z_{1-\alpha/2}$ and the quantity NUM, respectively. For discrete data, the quantities π_1 and π_2 are specified, and

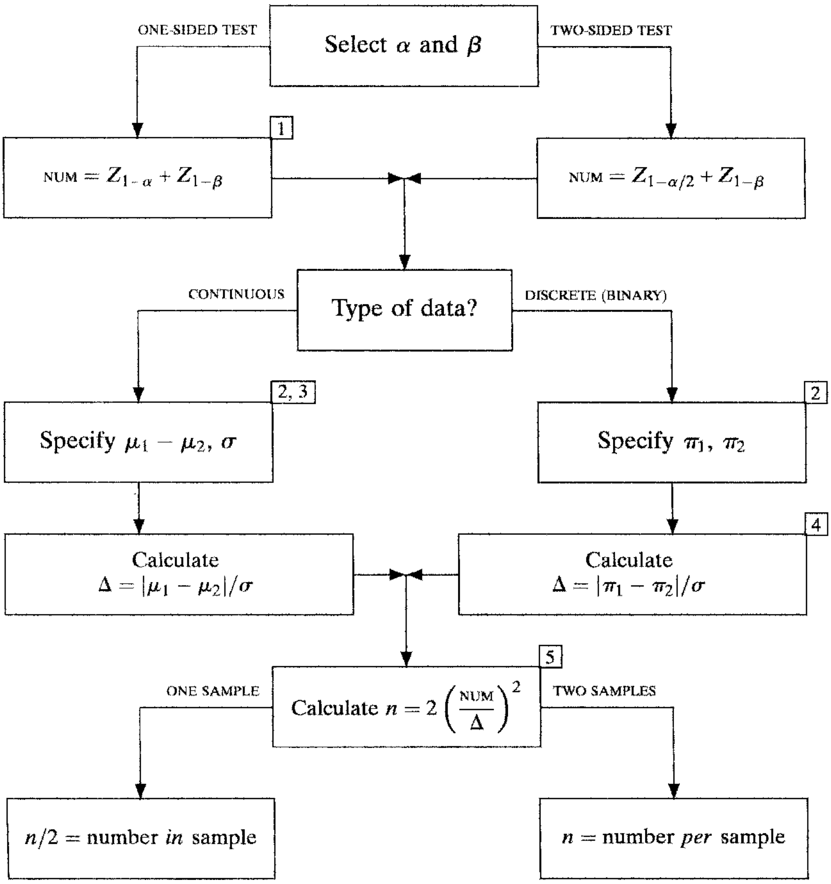

ONE-SIDED TEST
Select α and β
TWO-SIDED TEST
1
NUM $= Z_{1-\alpha} + Z_{1-\beta}$
NUM $= Z_{1-\alpha/2} + Z_{1-\beta}$
CONTINUOUS
Type of data?
DISCRETE (BINARY)
2, 3
Specify $\mu_1 - \mu_2$, σ
2
Specify π_1, π_2
4
Calculate $\Delta = |\mu_1 - \mu_2|/\sigma$
Calculate $\Delta = |\pi_1 - \pi_2|/\sigma$
5
ONE SAMPLE
Calculate $n = 2\left(\frac{\text{NUM}}{\Delta}\right)^2$
TWO SAMPLES
$n/2$ = number *in* sample
n = number *per* sample

Figure 6.1 Flowchart for sample-size calculations (continuous and discrete variables).

$\boxed{1}$ Values of Z_c for various values of c are:

c	0.500	0.800	0.900	0.950	0.975	0.990	0.995
Z_c	0.000	0.842	1.282	1.645	1.960	2.326	2.576

$\boxed{2}$ If one sample, μ_2 and π_2 are null hypothesis values.

$\boxed{3}$ If $\sigma_1^2 \neq \sigma_2^2$, calculate $\sigma^2 = \frac{1}{2}(\sigma_1^2 + \sigma_2^2)$.

$\boxed{4}$ $\sigma = \sqrt{\frac{1}{2}\,(\pi_1(1-\pi_1) + \pi_2(1-\pi_2))}$.

$\boxed{5}$ Sample size for discrete case is an approximation. For an improved estimate, use $n^* = n + 2/\Delta$.

Note: Two sample case, unequal sample sizes. Let n_1 and kn_1 be the required sample sizes. Calculate n as before. Then calculate $n_1 = n(k+1)/2k$ and $n_2 = kn_1$. (Total sample size will be larger.) If also, $\sigma_1^2 \neq \sigma_2^2$ calculate n using σ_1; then calculate $n_1 = (n/2)[1 + \sigma_2^2/(k\sigma_1^2)]$ and $n_2 = kn_1$.

$\Delta = |\pi_1 - \pi_2|/\sqrt{\frac{1}{2}(\pi_1(1-\pi_1) + \pi_2(1-\pi_2))}$ is calculated. This corresponds to the standardized differences $\Delta = |\mu_1 - \mu_2|/\sigma$ associated with normal or continuous data. The quantity $n = 2(\text{NUM}/\Delta)^2$ then produces the sample size needed for a two-sample procedure. For a one-sample procedure, the sample size is $n/2$. Hence a two-sample procedure requires a total of *four* times the number of observations of the one-sample procedure. Various refinements are available in Figure 6.1. A list of the most common Z-values is provided. If a one-sample test is wanted, the values of μ_2 and π_2 can be considered the null hypothesis values. Finally, the equation for the sample size in the discrete case is an approximation, and a more precise estimate, n^*, can be obtained from

$$n^* = n + \frac{2}{\Delta}$$

This formula is reasonably accurate.

Other approaches are possible. For example, one might specify the largest feasible sample size n, α, π_1, and π_2 and then determine the power $1 - \beta$. Figure 6.2 relates π_1, $\Delta = \pi_2 - \pi_1$, and n for two-sided tests for $\alpha = 0.05$ and $\beta = 0.10$.

Finally, we note that in certain situations where sample size is necessarily limited, for example, a rare cancer site with several competing therapies, trials with $\alpha = 0.10$ and $\beta = 0.50$ have been run.

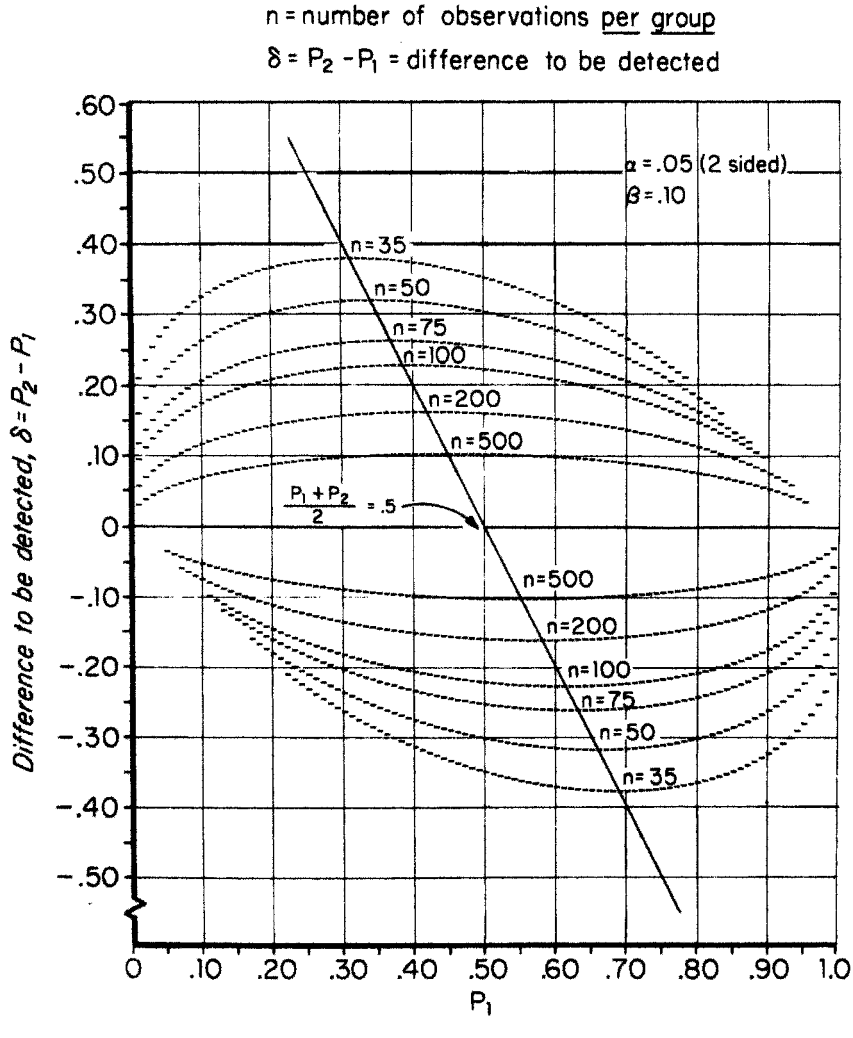


Figure 6.2 Sample sizes required for testing two proportions, π_1 and π_2 with 90% probability of obtaining a significant result at the 5% (two-sided) level. (From Feigl [1978].)

In practice, it is difficult to arrive at a sample size. To one unacquainted with statistical ideas, there is a natural tendency to expect too much from an experiment. In answer to the question, "What difference would you like to detect?" the novice will often reply, "any difference," leading to an infinite sample size!

6.3.4 Relative Risk and the Odds Ratio

In this section we consider studies looking for an association between two binary variables, that is, variables that take on only two outcomes. For definiteness we speak of the two variables as disease and exposure, since the following techniques are often used in epidemiologic and public health studies. In effect we are comparing two proportions (the proportions with disease) in two populations: those with and without exposure. In this section we consider methods of summarizing the association.

Suppose that one had a complete enumeration of the population at hand and the true proportions in the population. The probabilities may be presented in a 2 × 2 table:

	Disease	
Exposure	**+ (Yes)**	**− (No)**
+ (Yes)	π_{11}	π_{12}
− (No)	π_{21}	π_{22}

where $\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22} = 1$.

There are subtleties glossed over here. For example, by disease (for a human population), does one mean that the person develops the disease at some time before death, has the disease at a particular point in time, or develops it by some age? This ignores the problems of accurate diagnosis, a notoriously difficult problem. Similarly, exposure needs to be defined carefully as to time of exposure, length of exposure, and so on.

What might be a reasonable measure of the effect of exposure? A plausible comparison is P[disease+ |exposure+] with P[disease+ |exposure−]. In words, it makes sense to compare the probability of having disease among those exposed with the probability of having the disease among those not exposed.

Definition 6.1. A standard measure of the strength of the exposure effect is the *relative risk*. The relative risk is defined to be

$$\rho = \frac{P[\text{disease} + |\text{exposure}+]}{P[\text{disease} + |\text{exposure}-]} = \frac{\pi_{11}/(\pi_{11} + \pi_{12})}{\pi_{21}/(\pi_{21} + \pi_{22})} = \frac{\pi_{11}(\pi_{21} + \pi_{22})}{\pi_{21}(\pi_{11} + \pi_{12})}$$

Thus, a relative risk of 5 means that an exposed person is five times as likely to have the disease. The following tables of proportions or probabilities each has a relative risk of 2:

	Disease		**Disease**		**Disease**		**Disease**	
Exposure	+	−	+	−	+	−	+	−
+	0.50	0.00	0.25	0.25	0.10	0.40	0.00010	0.49990
−	0.25	0.25	0.125	0.375	0.05	0.45	0.0005	0.49995

We see that many patterns may give rise to the same relative risk. This is not surprising, as one number is being used to summarize four numbers. In particular, information on the amount of disease and/or exposure is missing.

Definition 6.2. Given that one has the exposure, the *odds* (or betting odds) of getting the disease are

$$\frac{P[\text{disease}+|\text{exposure}+]}{P[\text{disease}-|\text{exposure}+]}$$

Similarly, one may define the odds of getting the disease given no exposure. Another measure of the amount of association between the disease and exposure is the *odds ratio* defined to be

$$\begin{aligned}\omega &= \frac{P[\text{disease}+|\text{exposure}+]/P[\text{disease}-|\text{exposure}+]}{P[\text{disease}+|\text{exposure}-]/P[\text{disease}-|\text{exposure}-]}\\ &= \frac{(\pi_{11}/(\pi_{11}+\pi_{12}))/(\pi_{12}/(\pi_{11}+\pi_{12}))}{(\pi_{21}/(\pi_{21}+\pi_{22}))/(\pi_{22}/(\pi_{21}+\pi_{22}))}\\ &= \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}\end{aligned}$$

The odds ratio is also called the *cross-product ratio* on occasion; this name is suggested by the following scheme:

$$\begin{tikzcd}\pi_{11} \arrow[dr, leftrightarrow] & \pi_{12} \arrow[dl, leftrightarrow] \\ \pi_{21} & \pi_{22}\end{tikzcd}$$

Consider now how much the relative risk and odds ratio may differ by looking at the ratio of the two terms, ρ and ω,

$$\frac{\rho}{\omega} = \left(\frac{\pi_{21}+\pi_{22}}{\pi_{22}}\right)\left(\frac{\pi_{12}}{\pi_{11}+\pi_{12}}\right)$$

Suppose that the disease affects a small segment of the population. Then π_{11} is small compared to π_{12}, so that $\pi_{12}/(\pi_{11}+\pi_{12})$ is approximately equal to 1. Also, π_{21} will be small compared to π_{22}, so that $(\pi_{21}+\pi_{22})/\pi_{22}$ is approximately 1. Thus, in this case, $\rho/\omega = 1$. Restating this: If the disease affects a small fraction of the population (in both exposed and unexposed groups), the odds ratio and the relative risk are approximately equal. For this reason the odds ratio is often called the *approximate relative risk*. If the disease affects less than 5% in each group, the two quantities can be considered approximately equal.

The data for looking at the relative risk or the odds ratio usually arise in one of three ways, each of which is illustrated below. The numbers observed in each of the four cells will be denoted as follows:

	Disease	
Exposure	+	−
+	n_{11}	n_{12}
−	n_{21}	n_{22}

As before, a dot will replace a subscript when the entries for that subscript are summed. For example,

$$\begin{aligned}n_{1\cdot} &= n_{11}+n_{12}\\ n_{\cdot 2} &= n_{12}+n_{22}\\ n_{\cdot\cdot} &= n_{11}+n_{12}+n_{21}+n_{22}\end{aligned}$$

Pattern 1. (Cross-Sectional Studies: Prospective Studies of a Sample of the Population) There is a sample of size $n_{\cdot\cdot}$ from the population; both traits (exposure and disease) are measured on each subject. This is called *cross-sectional data* when the status of the two traits is measured at some fixed cross section in time. In this case the expected number in each cell is the expectation:

$$\begin{array}{cc|c}
n_{\cdot\cdot}\pi_{11} & n_{\cdot\cdot}\pi_{12} & \\
n_{\cdot\cdot}\pi_{21} & n_{\cdot\cdot}\pi_{22} & \\
\hline
 & & n_{\cdot\cdot}
\end{array}$$

Example 6.13. The following data are from Meyer et al. [1976]. This study collected information on all births in 10 Ontario (Canada) teaching hospitals during 1960–1961. A total of 51,490 births was involved, including fetal deaths and neonatal deaths (perinatal mortality). The paper considers the association of perinatal events and maternal smoking during pregnancy. Data relating perinatal mortality and smoking are as follows:

	Perinatal Mortality		
Maternal Smoking	**Yes**	**No**	**Total**
Yes	619	20,443	21,062
No	634	26,682	27,316
Total	1,253	47,125	48,378

Estimation of the relative risk and odds ratio is discussed below.

Pattern 2. (Prospective Study: Groups Based on Exposure) In a prospective study of exposure, fixed numbers—say $n_{1\cdot}$ and $n_{2\cdot}$—of people with and without the exposure are followed. The endpoints are then noted. In this case the expected number of observations in the cells are:

$$\begin{array}{cc|c}
n_{1\cdot}\dfrac{\pi_{11}}{\pi_{11}+\pi_{12}} & n_{1\cdot}\dfrac{\pi_{12}}{\pi_{11}+\pi_{12}} & n_{1\cdot} \\
n_{2\cdot}\dfrac{\pi_{21}}{\pi_{21}+\pi_{22}} & n_{2\cdot}\dfrac{\pi_{22}}{\pi_{21}+\pi_{22}} & n_{2\cdot}
\end{array}$$

Note that as the sample sizes of the exposure and nonexposure groups are determined by the experimenter, the data will not allow estimates of the proportion exposed, only the conditional probability of disease given exposure or nonexposure.

Example 6.14. As an example, consider a paper by Shapiro et al. [1974] in which they state that "by the end of this [five-year] period, there were 40 deaths in the [screened] study group of about 31,000 women as compared with 63 such deaths in a comparable group of women." Placing this in a 2×2 table and considering the screening to be the exposure, the data are:

	Breast Cancer Death		
On Study (Screened)	**Yes**	**No**	**Total**
Yes	40	30,960	31,000
No	63	30,937	31,000

Pattern 3. (Retrospective Studies) The third way of commonly collecting the data is the retrospective study. Usually, cases and an appropriate control group are identified. (Matched or paired data are *not* being discussed here.) In this case, the sizes of the disease and control groups,

$n_{\cdot 1}$ and $n_{\cdot 2}$, are specified. From such data one cannot estimate the probability of disease but rather, the probability of being exposed given that a person has the disease and the probability of exposure given that a person does not have the disease. The expected number of observations in each cell is

$$\begin{array}{cc}
n_{\cdot 1}\dfrac{\pi_{11}}{\pi_{11}+\pi_{21}} & n_{\cdot 2}\dfrac{\pi_{12}}{\pi_{12}+\pi_{22}} \\
n_{\cdot 1}\dfrac{\pi_{21}}{\pi_{11}+\pi_{21}} & n_{\cdot 2}\dfrac{\pi_{22}}{\pi_{12}+\pi_{22}} \\
\hline
n_{\cdot 1} & n_{\cdot 2}
\end{array}$$

Example 6.15. Kelsey and Hardy [1975] studied the driving of motor vehicles as a risk factor for acute herniated lumbar intervertebral disk. Their cases were people between the ages of 20 and 64; the studies were conducted in the New Haven metropolitan area at three hospitals or in the office of two private radiologists. The cases had low-back x-rays and were interviewed and given a few simple diagnostic tests. A control group was composed of those with low-back x-rays who were not classified as surgical probable or possible cases of herniated disk and who had not had their symptoms for more than one year. The in-patients, cases, and controls, of the Yale–New Haven hospital were asked if their job involved driving a motor vehicle. The data were:

	Herniated Disk?	
Motor Vehicle Job?	**Yes (Cases)**	**No (Controls)**
Yes	8	1
No	47	26
Total	55	27

Consider a two-way layout of disease and exposure to an agent thought to influence the disease:

	Disease	
Exposure	**+**	**−**
+	n_{11}	n_{12}
−	n_{21}	n_{22}

The three types of studies discussed above can be thought of as involving conditions on the marginal totals indicated in Table 6.2.

Table 6.2 Characterization of Cross-Sectional, Prospective, and Retrospective Studies and Relationship to Possible Estimation of Relative Risk and Odds Ratio

	Totals for:		Can One Estimate the:	
Type of Study	Column	Row	Relative Risk?	Odds Ratio?
Cross-sectional or prospective sample	Random	Random	Yes	Yes
Prospective on exposure	Random	Fixed	Yes	Yes
Retrospective	Fixed	Random	No	Yes

For example, a prospective study can be thought of as a situation where the totals for "exposure+" and "exposure−" are fixed by the experimenter, and the column totals will vary randomly depending on the association between the disease and the exposure.

For each of these three types of table, how might one estimate the relative risk and/or the odds ratio? From our tables of expected numbers of observations, it is seen that for tables of types 1 and 2,

$$\frac{E(n_{11})/(E(n_{11})+E(n_{12}))}{E(n_{21})/(E(n_{21})+E(n_{22}))} = \frac{E(n_{11})/E(n_{1}\cdot)}{E(n_{21})/E(n_{2}\cdot)} = \frac{\pi_{11}/(\pi_{11}+\pi_{12})}{\pi_{21}/(\pi_{21}+\pi_{22})} = \rho$$

Thus, one estimates the relative risk ρ by replacing the expected value of n_{11} by the observed value of n_{11}, etc., giving

$$\widehat{\rho} = \frac{n_{11}/n_{1}\cdot}{n_{21}/n_{2}\cdot}$$

For retrospective studies of type 3 it is not possible to estimate ρ unless the disease is rare, in which case the estimate of the odds ratio gives a reasonable estimate of the relative risk.

For all three types of tables, one sees that

$$\frac{E(n_{11})E(n_{22})}{E(n_{12})E(n_{21})} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \omega$$

Therefore, we estimate the odds ratio by

$$\widehat{\omega} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

It is clear from the definition of relative risk that if exposure has no association with the disease, $\rho = 1$. That is, both "exposed" and "nonexposed" have the same probability of disease. We verify this mathematically, and also that under the null hypothesis of no association, the odds ratio ω is also 1. Under H_0:

$$\pi_{ij} = \pi_{i}\cdot\pi\cdot_{j} \qquad \text{for} \quad i = 1, 2 \quad \text{and} \quad j = 1, 2$$

Thus,

$$\rho = \frac{\pi_{11}/\pi_{1}\cdot}{\pi_{21}/\pi_{2}\cdot} = \frac{\pi_{1}\cdot\pi\cdot_{1}/\pi_{1}\cdot}{\pi_{2}\cdot\pi\cdot_{1}/\pi_{2}\cdot} = 1 \quad \text{and} \quad \omega = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{\pi_{1}\cdot\pi\cdot_{1}\pi_{2}\cdot\pi\cdot_{2}}{\pi_{1}\cdot\pi\cdot_{2}\pi_{2}\cdot\pi\cdot_{1}} = 1$$

If ρ or ω are greater than 1, the exposed group has an increased risk of the disease. If ρ or ω are less than 1, the group not exposed has an increased risk of the disease. Note that an increased or decreased risk may, or may not, be due to a causal mechanism.

For the three examples above, let us calculate the estimated relative risk and odds ratio where appropriate. For the smoking and perinatal mortality data,

$$\widehat{\rho} = \frac{619/21{,}062}{634/27{,}316} \doteq 1.27, \qquad \widehat{\omega} = \frac{619(26{,}682)}{634(20{,}443)} \doteq 1.27$$

From these data we estimate that smoking during pregnancy is associated with an increased risk of perinatal mortality that is 1.27 times as large. (*Note*: We have not concluded that smoking causes the mortality, only that there is an association.)

The data relating screening for early detection of breast cancer and five-year breast cancer mortality gives estimates

$$\widehat{\rho} = \frac{40/31{,}000}{63/31{,}000} \doteq 0.63, \qquad \widehat{\omega} = \frac{40(30{,}937)}{63(30{,}960)} \doteq 0.63$$

Thus, in this study, screening was associated with a risk of dying of breast cancer within five years only 0.63 times as great as the risk among those not screened.

In the unmatched case–control study, only ω can be estimated:

$$\widehat{\omega} = \frac{8 \times 26}{1 \times 47} \doteq 4.43$$

It is estimated that driving jobs increase the risk of a herniated lumbar intervertebral disk by a factor of 4.43.

Might there really be no association in the tables above and the estimated $\widehat{\rho}$'s and $\widehat{\omega}$'s differ from 1 merely by chance? You may test the hypothesis of no association by using Fisher's exact test (for small samples) or the chi-squared test (for large samples).

For the three examples, using the table of χ^2 critical values with one degree of freedom, we test the statistical significance of the association by using the chi-square statistic with continuity correction.

Smoking–perinatal mortality:

$$X_c^2 = \frac{48{,}378[|619 \times 26{,}682 - 634 \times 20{,}443| - \frac{1}{2}(48{,}378)]^2}{21{,}062(27{,}316)(1253)(47{,}125)} = 17.76$$

From Table A.3, $p < 0.001$, and there is significant association. (Equivalently, for one degree of freedom, $Z = \sqrt{\chi_c^2} = 4.21$ and Table A.3 shows $p < 0.0001$.)

Breast cancer and screening:

$$X_c^2 = \frac{62{,}000[|40 \times 30{,}937 - 63 \times 30{,}960| - \frac{1}{2}(62{,}000)]^2}{31{,}000(31{,}000)(103)(61{,}897)} = 4.71$$

From the table, $0.01 < p < 0.05$ and the association is statistically significant at the 0.05 level.

Motor-vehicle job and herniated disk: $X_c^2 = 1.21$. From the χ^2 table, $p > 0.25$, and there is *not* a statistical association using only the Yale–New Haven data. In the next section we return to this data set.

If there is association, what can one say about the accuracy of the estimates? For the first two examples, where there is a statistically significant association, we turn to the construction of confidence intervals for ω. The procedure is to construct a confidence interval for ln ω, the natural log of ω, and to "exponentiate" the endpoints to find the confidence interval for ω. Our logarithms are natural logarithms, that is, to the base e. Recall e is a number; $e = 2.71828\ldots$.

The estimate of $\ln \omega$ is $\ln \widehat{\omega}$. The standard error of $\ln \widehat{\omega}$ is estimated by

$$\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

The estimate is approximately normally distributed; thus, normal critical values are used in constructing the confidence intervals. A $100(1-\alpha)\%$ confidence interval for ln ω is given by

$$\ln \widehat{\omega} \pm z_{1-\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

where an $N(0, 1)$ variable has probability $\alpha/2$ of exceeding $z_{1-\alpha/2}$.

Upon finding the endpoints of this confidence interval, we exponentiate the values of the endpoints to find the confidence interval for ω. We find a 99% confidence interval for ω with the smoking and perinatal mortality data. First we construct the confidence interval for $\ln\omega$:

$$\ln(1.27) \pm 2.576\sqrt{\frac{1}{619}+\frac{1}{20{,}443}+\frac{1}{26{,}682}+\frac{1}{634}}$$

or 0.2390 ± 0.1475 or (0.0915, 0.3865). The confidence interval for ω is

$$(e^{0.0915}, e^{0.3865}) = (1.10, 1.47)$$

To find a 95% confidence interval for the breast cancer–screening data,

$$\ln(0.63) \pm 1.96\sqrt{\frac{1}{40}+\frac{1}{30{,}960}+\frac{1}{30{,}937}+\frac{1}{63}}$$

or -0.4620 ± 0.3966 or $(-0.8586, -0.0654)$. The 95% confidence interval for the odds ratio, ω, is (0.424, 0.937).

The reason for using logarithms in constructing the confidence intervals is that $\ln\widehat{\omega}$ is more normally distributed than ω. The standard error of ω may be estimated directly by

$$\widehat{\omega}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}$$

(see Note 6.2 for the rationale). However, confidence intervals should be constructed as illustrated above.

6.3.5 Combination of 2 × 2 Tables

In this section we consider methods of combining 2 × 2 tables. The tables arise in one of two ways. In the first situation, we are interested in investigating an association between disease and exposure. There is, however, a third variable taking a finite number of values. We wish to "adjust" for the effect of the third variable. The values of the "confounding" third variable sometimes arise by taking a continuous variable and grouping by intervals; thus, the values are sometimes called *strata*. A second situation in which we will deal with several 2 × 2 tables is when the study of association and disease is made in more than one group. In some reasonable way, one would like to consider the combination of the 2 × 2 tables from each group.

Why Combine 2 × 2 *Tables?*

To see why one needs to worry about such things, suppose that there are two strata. In our first example there is no association between exposure and disease in each stratum, but if we ignore strata and "pool" our data (i.e., add it all together), an association appears. For stratum 1,

	Disease	
Exposure	**+**	**−**
+	5	50
−	10	100

$$\widehat{\omega}_1 = \frac{5(100)}{10(50)} = 1$$

and for stratum 2,

	Disease	
Exposure	**+**	**−**
+	40	60
−	40	60

$$\widehat{\omega}_2 = \frac{40(60)}{40(60)} = 1$$

In both tables the odds ratio is 1 and there is no association. Combining tables, the combined table and its odds ratio are:

	Disease	
Exposure	**+**	**−**
+	45	110
−	50	160

$$\widehat{\omega}_{\text{combined}} = \frac{45(160)}{50(110)} \doteq 1.31$$

When combining tables with no association, or odds ratios of 1, the combination may show association. For example, one would expect to find a positive relationship between breast cancer and being a homemaker. Possibly tables given separately for each gender would not show such an association. If the inference to be derived were that homemaking might be related causally to breast cancer, it is clear that one would need to adjust for gender.

On the other hand, there can be an association within each stratum that disappears in the pooled data set. The following numbers illustrate this:

Stratum 1:

	Disease	
Exposure	**+**	**−**
+	60	100
−	10	50

$$\widehat{\omega}_1 = \frac{60(50)}{10(100)} = 3$$

Stratum 2:

	Disease	
Exposure	**+**	**−**
+	50	10
−	100	60

$$\widehat{\omega}_2 = \frac{50(60)}{100(10)} = 3$$

Combined data:

	Disease	
Exposure	**+**	**−**
+	110	110
−	110	110

$$\widehat{\omega}_{\text{combined}} = 1$$

Thus, ignoring a confounding variable may “hide” an association that exists within each stratum but is not observed in the combined data.

Formally, our two situations are the same if we identify the stratum with differing groups. Also, note that there may be more than one confounding variable, that each strata of the “third” variable could correspond to a different combination of several other variables.

Questions of Interest in Multiple 2 × 2 *Tables*

In examining more than one 2 × 2 table, one or more of three questions is usually asked. This is illustrated by using the data of the study involving cases of acute herniated lumbar disk and controls (not matched) in Example 6.15, which compares the proportions with jobs driving motor vehicles. Seven different hospital services are involved, although only one of them was presented in Example 6.15. Numbering the sources from 1 to 7 and giving the data as 2 × 2 tables, the tables and the seven odds ratios are:

Source 1:

	Herniated Disk	
Motor Vehicle Job	**+**	**−**
+	8	1
−	47	26

$\widehat{\omega} = 4.43$

Source 2:

	+	−
+	5	0
−	17	21

$\widehat{\omega} = \infty$

Source 5:

	+	−
+	1	3
−	5	10

$\widehat{\omega} = 0.67$

Source 3:

	+	−
+	4	4
−	13	77

$\widehat{\omega} = 5.92$

Source 6:

	+	−
+	1	2
−	3	11

$\widehat{\omega} = 1.83$

Source 4:

	+	−
+	2	10
−	12	65

$\widehat{\omega} = 1.08$

Source 7:

	+	−
+	2	2
−	12	37

$\widehat{\omega} = 3.08$

The seven odds ratios are 4.43, ∞, 5.92, 1.08, 0.67, 1.83, and 3.08. The ratios vary so much that one might wonder whether each hospital service has the same degree of association (question 1). If they do not have the same degree of association, one might question whether the controls are appropriate, the patient populations are different, and so on.

One would also like an estimate of the overall or average association (question 2). From the previous examples it is seen that it might not be wise to sum all the tables and compute the association based on the pooled tables.

Finally, another question, related to the first two, is whether there is any evidence of any association, either overall or in some of the groups (question 3).

Two Approaches to Estimating an Overall Odds Ratio

If the seven different tables come from populations with the same odds ratio, how do we estimate the common or overall odds ratio? We will consider two approaches.

The first technique is to work with the natural logarithm, log to the base e, of the estimated odds ratio, $\widehat{\omega}$. Let $a_i = \ln \widehat{\omega}_i$, where $\widehat{\omega}_i$ is the estimated odds ratio in the ith of k 2×2 tables. The standard error of a_i is estimated by

$$s_i = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

where n_{11}, n_{12}, n_{21}, and n_{22} are the values from the ith 2×2 table. How do we investigate the problems mentioned above? To do this, one needs to understand a little of how the χ^2 distribution arises. The square of a standard normal variable has a chi-square distribution with one degree of freedom. If independent chi-square variables are added, the result is a chi-square variable whose degrees of freedom comprises the sum of the degrees of freedom of the variables that were added (see Note 5.3 also).

We now apply this to the problem at hand. Under the null hypothesis of no association in any of the tables, each a_i/s_i is approximately a standard normal value. If there is no association, $\omega = 1$ and $\ln \omega = 0$. Thus, log $\widehat{\omega}_i$ has a mean of approximately zero. Its square, $(a_i/s_i)^2$, is approximately a χ^2 variable with one degree of freedom. The sum of all k of these independent, approximately chi-square variables is approximately a chi-square variable with k degrees of freedom. The sum is

$$X^2 = \sum_{i=1}^{k} \left(\frac{a_i}{s_i}\right)^2$$

and under the null hypothesis it has approximately a χ^2-distribution with k degrees of freedom.

It is possible to partition this sum into two parts. One part tests whether the association might be the same in all k tables (i.e., it tests for homogeneity). The second part will test to see whether on the basis of all the tables there is any association.

Suppose that one wants to "average" the association from all of the 2×2 tables. It seems reasonable to give more weight to the better estimates of association; that is, one wants the estimates with higher variances to get less weight. An appropriate weighted average is

$$\overline{a} = \sum_{i=1}^{k} \frac{a_i}{s_i^2} \Bigg/ \sum_{i=1}^{k} \frac{1}{s_i^2}$$

The χ^2-statistic then is partitioned, or broken down, into two parts:

$$X^2 = \sum_{i=1}^{k} \left(\frac{a_i}{s_i}\right)^2 = \sum_{i=1}^{k} \frac{1}{s_i^2}(a_i - \overline{a})^2 + \sum_{i=1}^{k} \frac{1}{s_i^2}\overline{a}^2$$

On the right-hand side, the first sum is approximately a χ^2 random variable with $k-1$ degrees of freedom if all k groups have the same degree of association. It tests for the homogeneity of the association in the different groups. That is, if χ^2 for homogeneity is too large, we reject the null hypothesis that the degree of association (whatever it is) is the same in each group. The second term tests whether there is association on the average. This has approximately a χ^2-distribution with one degree of freedom if there is no association in each group. Thus, define

$$\chi_H^2 = \sum_{i=1}^{k} \frac{1}{s_i^2}(a_i - \overline{a})^2 = \sum_{i=1}^{k} \frac{a_i^2}{s_i^2} - \overline{a}^2 \sum_{i=1}^{k} \frac{1}{s_i^2}$$

and

$$\chi_A^2 = \overline{a}^2 \sum_{i=1}^{k} \frac{1}{s_i^2}$$

Of course, if we decide that there are different degrees of association in different groups, this means that at least one of the groups must have some association.

Consider now the data given above. A few additional points are introduced. We use the log of the odds ratio, but the second group has $\widehat{\omega} = \infty$. What shall we do about this?

With small numbers, this may happen due to a zero in a cell. The bias of the method is reduced by adding 0.5 to each cell in each table:

[1]	+	−
+	8.5	1.5
−	47.5	26.5

[2]	+	−
+	5.5	0.5
−	17.5	21.5

[3]	+	−
+	4.5	4.5
−	13.5	77.5

[4]	+	−
+	2.5	10.5
−	12.5	65.5

[5]	+	−
+	1.5	3.5
−	5.5	10.5

[6]	+	−
+	1.5	2.5
−	3.5	11.5

[7]	+	−
+	2.5	2.5
−	12.5	37.5

Now

$$\widehat{\omega}_i = \frac{(n_{11}+0.5)(n_{22}+0.5)}{(n_{12}+0.5)(n_{21}+0.5)}, \qquad s_i = \sqrt{\frac{1}{n_{11}+0.5}+\frac{1}{n_{22}+0.5}+\frac{1}{n_{12}+0.5}+\frac{1}{n_{21}+0.5}}$$

The calculations above are shown in Table 6.3.

Table 6.3 Calculations for the Seven Tables

Table i	$\widehat{\omega}_i$	$a_i = \log\widehat{\omega}_i$	s_i^2	$1/s_i^2$	a_i^2/s_i^2	a_i/s_i^2
1	3.16	1.15	0.843	1.186	1.571	1.365
2	13.51	2.60	2.285	0.438	2.966	1.139
3	5.74	1.75	0.531	1.882	5.747	3.289
4	1.25	0.22	0.591	1.693	0.083	0.375
5	0.82	−0.20	1.229	0.813	0.033	−0.163
6	1.97	0.68	1.439	0.695	0.320	0.472
7	3.00	1.10	0.907	1.103	1.331	1.212
Total				7.810	12.051	7.689

Then

$$\overline{a} = \sum_{i=1}^{k} \frac{a_i}{s_i^2} \Big/ \sum_{i=1}^{k} \frac{1}{s_i^2} = \frac{7.689}{7.810} \doteq 0.985$$

$$X_A^2 = (0.985)^2(7.810) \doteq 7.57$$

$$X_H^2 = \sum \frac{a_i^2}{s_i^2} - \chi_A^2 = 12.05 - 7.57 = 4.48$$

X_H^2 with $7 - 1 = 6$ degrees of freedom has an $\alpha = 0.05$ critical value of 12.59 from Table A.3. We do *not* conclude that the association differs between groups.

Moving to the X_A^2, we find that $7.57 > 6.63$, the χ^2 critical value with one degree of freedom at the 0.010 level. We conclude that there *is* some overall association.

The odds ratio is estimated by $\widehat{\omega} = e^{\overline{a}} = e^{0.985} = 2.68$. The standard error of $\overline{a}$ is estimated by

$$\frac{1}{\sqrt{\sum_{i=1}^{k}(1/s_i^2)}}$$

To find a confidence interval for ω, first find one for ln ω and "exponentiate" back. To find a 95% confidence interval, the calculation is

$$\overline{a} \pm \frac{z_{0.975}}{\sqrt{\sum(1/s_i^2)}} = 0.985 \pm \frac{1.96}{\sqrt{7.810}} \quad \text{or} \quad 0.985 \pm 0.701 \quad \text{or} \quad (0.284, 1.696)$$

Taking exponentials, the confidence interval for the overall odds ratio is (1.33, 5.45).

The second method of estimation is due to Mantel and Haenszel [1959]. Their estimate of the odds ratio is

$$\widehat{\omega} = \sum_{i=1}^{k} \frac{n_{11}(i)n_{22}(i)}{n_{\cdot\cdot}(i)} \Big/ \sum_{i=1}^{k} \frac{n_{12}(i)n_{21}(i)}{n_{\cdot\cdot}(i)}$$

where $n_{11}(i), n_{22}(i), n_{12}(i)$, $n_{21}(i)$, and $n_{\cdot\cdot}(i)$ are $n_{11}, n_{22}, n_{12}, n_{21}$, and $n_{\cdot\cdot}$ for the ith table.

In this problem,

$$\widehat{\omega} = \frac{\dfrac{8 \times 26}{82} + \dfrac{5 \times 21}{43} + \dfrac{4 \times 77}{98} + \dfrac{2 \times 65}{89} + \dfrac{1 \times 10}{19} + \dfrac{1 \times 11}{17} + \dfrac{2 \times 37}{53}}{\dfrac{47 \times 1}{82} + \dfrac{17 \times 10}{43} + \dfrac{13 \times 4}{98} + \dfrac{12 \times 10}{89} + \dfrac{5 \times 3}{19} + \dfrac{3 \times 2}{17} + \dfrac{12 \times 12}{53}}$$
$$\doteq \frac{12.1516}{4.0473} \doteq 3.00$$

A test of association is given by the following statistic, X_A^2, which is approximately a chi-square random variable with one degree of freedom:

$$X_A^2 = \frac{\left[\left|\sum_{i=1}^{k} n_{11}(i) - \sum_{i=1}^{k} n_{1\cdot}(i)n_{\cdot 1}(i)/n_{\cdot\cdot}(i)\right| - \frac{1}{2}\right]^2}{\sum_{i=1}^{k} n_{1\cdot}(i)n_{2\cdot}(i)n_{\cdot 1}(i)n_{\cdot 2}(i)/n_{\cdot\cdot}(i)^2[n_{\cdot\cdot}(i) - 1]}$$

The herniated disk data yield $X_A^2 = 7.92$, so that, as above, there is a significant ($p < 0.01$) association between an acute herniated lumbar intervertebral disk and whether or not a job requires driving a motor vehicle. See Schlesselman [1982] and Breslow and Day [1980] for methods of setting confidence intervals for ω using the Mantel–Haenszel estimate.

In most circumstances, combining 2×2 tables will be used to adjust for other variables that define the strata (i.e., that define the different tables). The homogeneity of the odds ratio is usually of less interest unless the odds ratio differs widely among tables. Before testing for homogeneity of the odds ratio, one should be certain that this is what is desired (see Note 6.3).

6.3.6 Screening and Diagnosis: Sensitivity, Specificity, and Bayes' Theorem

In clinical medicine, and also in epidemiology, tests are often used to screen for the presence or absence of a disease. In the simplest case the test will simply be classified as having a positive (disease likely) or negative (disease unlikely) finding. Further, suppose that there is a "gold standard" that tells us whether or not a subject actually has the disease. The definitive classification might be based on data from follow-up, invasive radiographic or surgical procedures, or autopsy results. In many cases the gold standard itself will only be relatively correct, but nevertheless the best classification available. In this section we discuss summarization of the prediction of disease (as measured by our gold standard) by the test being considered. Ideally, those with the disease should all be classified as having disease, and those without disease should be classified as nondiseased. For this reason, two indices of the performance of a test consider how often such correct classification occurs.

Definition 6.3. The *sensitivity* of a test is the percentage of people with disease who are classified as having disease. A test is sensitive to the disease if it is positive for most people having the disease. The *specificity* of a test is the percentage of people without the disease who are classified as not having the disease. A test is specific if it is positive for a small percentage of those without the disease.

Further terminology associated with screening and diagnostic tests are true positive, true negative, false positive, and false negative tests.

Definition 6.4. A test is a *true positive test* if it is positive and the subject has the disease. A test is a *true negative test* if the test is negative and the subject does not have the disease. A *false positive test* is a positive test of a person without the disease. A *false negative test* is a negative test of a person with the disease.

Definition 6.5. The *predictive value of a positive test* is the percentage of subjects with a positive test who have the disease; the *predictive value of a negative test* is the percentage of subjects with a negative test who do not have the disease.

Suppose that data are collected on a test and presented in a 2×2 table as follows:

	Disease Category	
Screening Test Result	**Disease (+)**	**Nondiseased (−)**
Positive (+) test	a (true +′ s)	b (false +′ s)
Negative (−) test	c (false −′ s)	d (true −′ s)

The sensitivity is estimated by $100a/(a+c)$, the specificity by $100d/(b+d)$. If the subjects are representative of a population, the predictive value of positive and negative tests are estimated by $100a/(a+b)$ and $100d/(c+d)$, respectively. These predictive values are useful only when the proportions with and without the disease in the study group are approximately the same as in the population where the test will be used to predict or classify (see below).

Example 6.16. Remein and Wilkerson [1961] considered a number of screening tests for diabetes. They had a group of consultants establish criteria, their gold standard, for diabetes. On each of a number of days, they recruited patients being seen in the outpatient department of the Boston City Hospital for reasons other than suspected diabetes. The table below presents results on the Folin–Wu blood test used 1 hour after a test meal and using a blood sugar level of 150 mg per 100 mL of blood sugar as a positive test.

Test	Diabetic	Nondiabetic	Total
+	56	49	105
−	14	461	475
Total	70	510	580

From this table note that there are 56 true positive tests compared to 14 false negative tests. The sensitivity is $100(56)/(56+14) = 80.0\%$. The 49 false positive tests and 461 true negative tests give a specificity of $100(461)/(49+461) = 90.4\%$. The predictive value of a positive test is $100(56)/(56+49) = 53.3\%$. The predictive value of a negative test is $100(461)/(14+461) = 97.1\%$.

If a test has a fixed value for its sensitivity and specificity, the predictive values will change depending on the prevalence of the disease in the population being tested. The values are related by *Bayes' theorem*. This theorem tells us how to update the probability of an event A: for example, the event of a subject having disease. If the subject is selected at random from some population, the probability of A is the fraction of people having the disease. Suppose that additional information becomes available; for example, the results of a diagnostic test might become available. In the light of this new information we would like to update or change our assessment of the probability that A occurs (that the subject has disease). The probability of A before receiving additional information is called the *a priori* or *prior probability*. The updated probability of A after receiving new information is called the *a posteriori* or *posterior probability*. Bayes' theorem is an explanation of how to find the posterior probability.

Bayes' theorem uses the concept of a conditional probability. We review this concept in Example 6.17.

Example 6.17. Comstock and Partridge [1972] conducted an informal census of Washington County, Maryland, in 1963. There were 127 arteriosclerotic heart disease deaths in the follow-up period. Of the deaths, 38 occurred among people whose usual frequency of church attendance was once or more per week. There were 24,245 such people as compared to 30,603 people whose usual attendance was less than once weekly. What is the probability of an arteriosclerotic heart disease death (event A) in three years given church attendance usually once or more per week (event B)?

From the data

$$P[A] = \frac{127}{24{,}245 + 30{,}603} = 0.0023$$

$$P[B] = \frac{24{,}245}{24{,}245 + 30{,}603} = 0.4420$$

$$P[A \ \& \ B] = \frac{38}{24{,}245 + 30{,}603} = 0.0007$$

$$P[A \mid B] = \frac{P[A \text{ and } B]}{P[B]} = \frac{0.0007}{0.4420} = 0.0016$$

If you knew that someone attended church once or more per week, the prior estimate of 0.0023 of the probability of an arteriosclerotic heart disease death in three years would be changed to a posterior estimate of 0.0016.

Using the conditional probability concept, Bayes' theorem may be stated.

Fact 1. (Bayes' Theorem) Let $B_1, \ldots, B_k$ be events such that one and only one of them must occur. Then for each i,

$$P[B_i|A] = \frac{P[A|B_i]P[B_i]}{P[A|B_1]P[B_1] + \cdots + P[A|B_k]P[B_k]}$$

Example 6.18. We use the data of Example 6.16 and Bayes' theorem to show that the predictive power of the test is related to the prevalence of the disease in the population. Suppose that the prevalence of the disease were not 70/580 (as in the data given), but rather, 6%. Also suppose that the sensitivity and specificity of the test were 80.0% and 90.4%, as in the example. What is the predictive value of a positive test?

We want $P[\text{disease+}|\text{test+}]$. Let B_1 be the event that the patient has disease and B_2 be the event of no disease. Let A be the occurrence of a positive test. A sensitivity of 80.0% is the same as $P[A|B_1] = 0.800$. A specificity of 90.4% is equivalent to $P[\text{not}A|B_2] = 0.904$. It is easy to see that

$$P[\text{not } A|B] + P[A|B] = 1$$

for any A and B. Thus, $P[A|B_2] = 1-0.904 = 0.096$. By assumption, $P[\text{disease+}] = P[B_1] = 0.06$, and $P[\text{disease}-] = P[B_2] = 0.94$. By Bayes' theorem,

$$P[\text{disease+}|\text{test+}] = \frac{P[\text{test} + |\text{disease+}]P[\text{disease+}]}{P[\text{test} + |\text{disease+}]P[\text{disease+}] + P[\text{test} + |\text{disease}-]P[\text{disease}-]}$$

Using our definitions of A, B_1, and B_2, this is

$$\begin{aligned} P[B_1|A] &= \frac{P[A|B_1]P[B_1]}{P[A|B_1]P[B_1] + P[A|B_2]P[B_2]} \\ &= \frac{0.800 \times 0.06}{0.800 \times 0.06 + 0.096 \times 0.94} \\ &= 0.347 \end{aligned}$$

If the disease prevalence is 6%, the predictive value of a positive test is 34.7% rather than 53.3% when the disease prevalence is 70/580 (12.1%).

Problems 6.15 and 6.28 illustrate the importance of disease prevalence in assessing the results of a test. See Note 6.8 for relationships among sensitivity, specificity, prevalence, and predictive values of a positive test. Sensitivity and specificity are discussed further in Chapter 13. See also Pepe [2003] for an excellent overview.

6.4 MATCHED OR PAIRED OBSERVATIONS

The comparisons among proportions in the preceding sections dealt with samples from different populations or from different subsets of a specified population. In many situations, the estimates of the proportions are based on the same objects or come from closely related, matched, or paired observations. You have seen matched or paired data used with a one-sample t-test.

A standard epidemiological tool is the retrospective paired case–control study. An example was given in Chapter 1. Let us recall the rationale for such studies. Suppose that one wants to see whether or not there is an association between a risk factor (say, use of oral contraceptives), and a disease (say, thromboembolism). Because the incidence of the disease is low, an extremely large prospective study would be needed to collect an adequate number of cases. One strategy is to *start* with the cases. The question then becomes one of finding appropriate controls for the cases. In a matched pair study, one control is identified for each case. The control, not having the disease, should be identical to the case in all relevant ways except, possibly, for the risk factor (see Note 6.6).

Example 6.19. This example is a retrospective matched pair case–control study by Sartwell et al. [1969] to study thromboembolism and oral contraceptive use. The cases were 175 women of reproductive age (15 to 44), discharged alive from 43 hospitals in five cities after initial attacks of idiopathic (i.e., of unknown cause) thrombophlebitis (blood clots in the veins with inflammation in the vessel walls), pulmonary embolism (a clot carried through the blood and obstructing lung blood flow), or cerebral thrombosis or embolism. The controls were matched with their cases for hospital, residence, time of hospitalization, race, age, marital status, parity, and pay status. More specifically, the controls were female patients from the same hospital during the same six-month interval. The controls were within five years of age and matched on parity (0, 1, 2, 3, or more prior pregnancies). The hospital pay status (ward, semiprivate, or private) was the same. The data for oral contraceptive use are:

	Control Use?	
Case Use?	**Yes**	**No**
Yes	10	57
No	13	95

The question of interest: Are cases more likely than controls to use oral contraceptives?

6.4.1 Matched Pair Data: McNemar's Test and Estimation of the Odds Ratio

The 2×2 table of Example 6.19 does not satisfy the assumptions of previous sections. The proportions using oral contraceptives among cases and controls cannot be considered samples from two populations since the cases and controls are paired; that is, they come together. Once a case is selected, the control for the case is constrained to be one of a small subset of people who match the case in various ways.

Suppose that there is no association between oral contraceptive use and thromboembolism after taking into account relevant factors. Suppose a case and control are such that only one of the pair uses oral contraceptives. Which one is more likely to use oral contraceptives? They may both be likely or unlikely to use oral contraceptives, depending on a variety of factors. Since the pair have the same values of such factors, neither member of the pair is more likely to have the risk factor! That is, in the case of disagreement, or discordant pairs, the probability that the case has the risk factor is 1/2. More generally, suppose that the data are

	Control Has Risk Factor?	
Case Has Risk Factor?	**Yes**	**No**
Yes	a	b
No	c	d

If there is no association between disease (i.e., case or control) and the presence or absence of the risk factor, the number b is binomial with $\pi = 1/2$ and $n = b + c$. To test for association we test $\pi = 1/2$, as shown previously. For large n, say $n \geq 30$,

$$X^2 = \frac{(b - c)^2}{b + c}$$

has a chi-square distribution with one degree of freedom if $\pi = 1/2$. For Example 6.19,

$$X^2 = \frac{(57 - 13)^2}{57 + 13} = 27.66$$

From the chi-square table, $p < 0.001$, so that there is a statistically significant association between thromboembolism and oral contraceptive use. This statistical test is called *McNemar's test.*

Procedure 6. For retrospective matched pair data, the odds ratio is estimated by

$$\widehat{\omega}_{\text{paired}} = \frac{b}{c}$$

The standard error of the estimate is estimated by

$$(1 + \widehat{\omega}_{\text{paired}})\sqrt{\frac{\widehat{\omega}_{\text{paired}}}{b + c}}$$

In Example 6.19, we estimate the odds ratio by

$$\widehat{\omega} = \frac{57}{13} \doteq 4.38$$

The standard error is estimated by

$$(1 + 4.38)\sqrt{\frac{4.38}{70}} \doteq 1.35$$

An approximate 95% confidence interval is given by

$$4.38 \pm (1.96)(1.35) \quad \text{or} \quad (1.74, 7.02)$$

More precise intervals may be based on the use of confidence intervals for a binomial proportion and the fact that $\widehat{\omega}_{\text{paired}}/(\widehat{\omega}_{\text{paired}} + 1) = b/(b + c)$ is a binomial proportion (see Fleiss [1981]). See Note 6.5 for further discussion of the chi-square analysis of paired data.

6.5 POISSON RANDOM VARIABLES

The Poisson distribution occurs primarily in two closely related situations. The first is a situation in which one counts discrete events in space or time, or some other continuous situation. For example, one might note the time of arrival (considered as a particular point in time) at an emergency medical service over a fixed time period. One may count the number of discrete occurrences of arrivals over this continuum of time. Conceptually, we may get any nonnegative integer, no matter how large, as our answer. A second example occurs when counting numbers of red blood cells that occur in a specified rectangular area marked off in the field of view. In a diluted blood sample where the distance between cells is such that they do not tend to "bump into each other," we may idealize the cells as being represented by points in the plane. Thus, within the particular area of interest, we are counting the number of points observed. A third example where one would expect to model the number of counts by a Poisson distribution would be a situation in which one is counting the number of particle emissions from a radioactive source. If the time period of observation is such that the radioactivity of the source does not decrease significantly (i.e., the time period is small compared to the half-life of a particle), the counts (which may be considered as coming at discrete time points) would again be modeled appropriately by a Poisson distribution.

The second major use of the Poisson distribution is as an approximation to the binomial distribution. If n is large and π is small in a binomial situation, the number of successes is very closely modeled by the Poisson distribution. The closeness of the approximation is specified by a mathematical theorem. As a rough rule of thumb, for most purposes the Poisson approximation will be adequate if π is less than or equal to 0.1 and n is greater than or equal to 20.

For the Poisson distribution to be an appropriate model for counting discrete points occurring in some sort of a continuum, the following two assumptions must hold:

1. The number of events occurring in one part of the continuum should be statistically independent of the number of events occurring in another part of the continuum. For example, in the emergency room, if we measure the number of arrivals during the first half hour, this event could reasonably be considered statistically independent of the number of arrivals during the second half hour. If there has been some cataclysmic event such as an earthquake, the assumption will not be valid. Similarly, in counting red blood cells in a diluted blood solution, the number of red cells in one square might reasonably be modeled as statistically independent of the number of red cells in another square.
2. The expected number of counts in a given part of the continuum should approach zero as its size approaches zero. Thus, in observing blood cells, one does not expect to find any in a very small area of a diluted specimen.

6.5.1 Examples of Poisson Data

Example 6.3 [Bucher et al., 1976] examines racial differences in the incidence of ABO hemolytic disease by examining records for infants born at the North Carolina Memorial Hospital. The samples of black and white infants gave the following estimated proportions with hemolytic disease:

$$\text{black infants},\quad n_1 = 3584,\quad p_1 = 43/3584$$

$$\text{white infants},\quad n_2 = 3831,\quad p_2 = 17/3831$$

The observed number of cases might reasonably be modeled by the Poisson distribution. (*Note:* The n is large and π is small in a binomial situation.) In this paper, studying the incidence of ABO hemolytic disease in black and white infants, the observed fractions for black and white infants of having the disease were 43/3584 and 17/3831. The 43 and 17 cases may be considered values of Poisson random variables.

A second example that would be modeled appropriately by the Poisson distribution is the number of deaths resulting from a large-scale vaccination program. In this case, n will be very large and π will be quite small. One might use the Poisson distribution in investigating the simultaneous occurrence of a disease and its association within a vaccination program. How likely is it that the particular "chance occurrence" might actually occur by chance?

Example 6.20. As a further example, a paper by Fisher et al. [1922] considers the accuracy of the plating method of estimating the density of bacterial populations. The process we are speaking about consists in making a suspension of a known mass of soil in a known volume of salt solution, and then diluting the suspension to a known degree. The bacterial numbers in the diluted suspension are estimated by plating a known volume in a nutrient gel medium and counting the number of colonies that develop from the plate. The estimate was made by a calculation that takes into account the mass of the soil taken and the degree of dilution. If we consider the colonies to be points occurring in the volume of gel, a Poisson model for the number of counts would be appropriate. Table 6.4 provides counts from seven different plates with portions of soil taken from a sample of Barnfield soil assayed in four parallel dilutions:

Example 6.21. A famous example of the Poisson distribution is data by von Bortkiewicz [1898] showing the chance of a cavalryman being killed by a horse kick in the course of a year (Table 6.5). The data are from recordings of 10 corps over a period of 20 years supplying 200 readings. A question of interest here might be whether a Poisson model is appropriate. Was the corps with four deaths an "unlucky" accident, or might there have been negligence of some kind?

Table 6.4 Counts for Seven Soil Samples

	Dilution			
Plate	I	II	III	IV
1	72	74	78	69
2	69	72	74	67
3	63	70	70	66
4	59	69	58	64
5	59	66	58	62
6	53	58	56	58
7	51	52	56	54
Mean	60.86	65.86	64.29	62.86

Table 6.5 Horse-kick Fatality Data

Number of Deaths per Corps per Year	Frequency
0	109
1	65
2	22
3	3
4	1
5	0
6	0

6.5.2 Poisson Model

The Poisson probability distribution is characterized by one parameter, λ. For each nonnegative integer k, if Y is a variable with the Poisson distribution with parameter λ,

$$P[Y = k] = \frac{e^{-\lambda}\lambda^k}{k!}$$

The parameter λ is both the mean and variance of the Poisson distribution,

$$E(Y) = \text{var}(Y) = \lambda$$

Bar graphs of the Poisson probabilities are given in Figure 6.3 for selected values of λ. As the mean (equal to the variance) increases, the distribution moves to the right and becomes more spread out and more symmetrical.

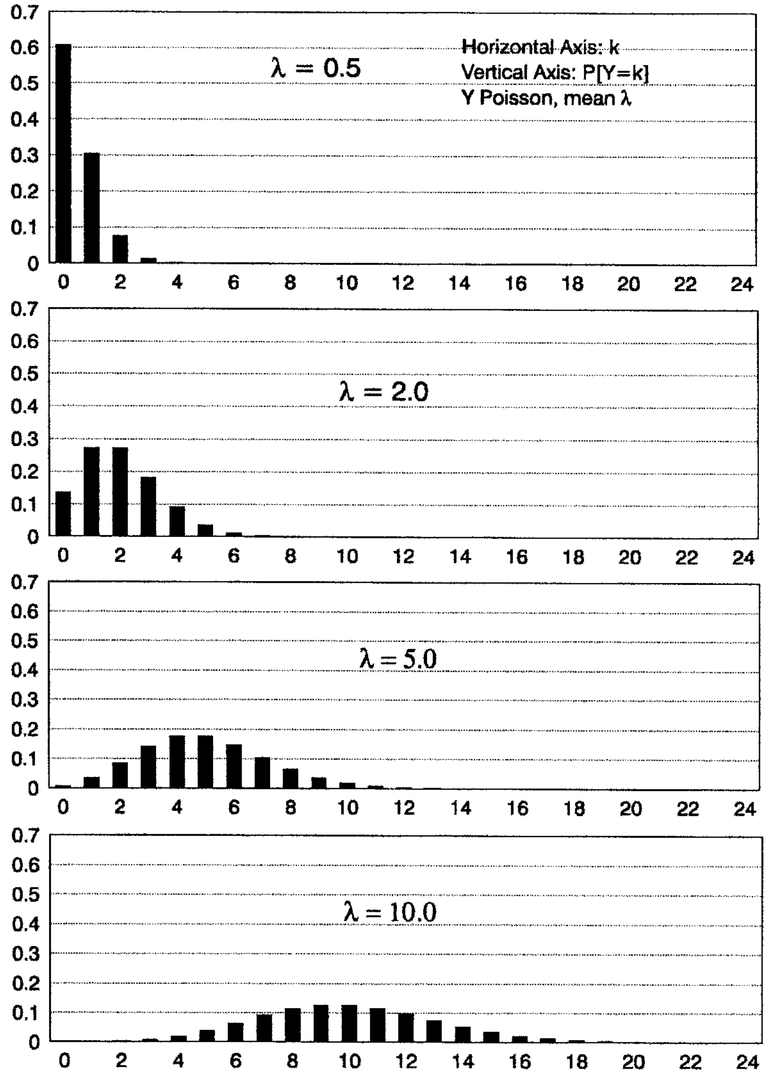


Figure 6.3 Poisson distribution.

Table 6.6 Binomial and Poisson Probabilities

	Binomial Probabilities			
	$n = 10$	$n = 20$	$n = 40$	Probabilities
k	$\pi = 0.20$	$\pi = 0.10$	$\pi = 0.05$	Poisson
0	0.1074	0.1216	0.1285	0.1353
1	0.2684	0.2702	0.2706	0.2707
2	0.3020	0.2852	0.2777	0.2707
3	0.2013	0.1901	0.1851	0.1804
4	0.0881	0.0898	0.0901	0.0902
5	0.0264	0.0319	0.0342	0.0361
6	0.0055	0.0089	0.0105	0.0120

In using the Poisson distribution to approximate the binomial distribution, the parameter λ is chosen to equal $n\pi$, the expected value of the binomial distribution. Poisson and binomial probabilities are given in Table 6.6 for comparison. This table gives an idea of the accuracy of the approximation (table entry is $P[Y = k], \lambda = 2 = n\pi$) for the first seven values of three distributions.

A fact that is often useful is that a sum of independent Poisson variables is itself a Poisson variable. The parameter for the sum is the sum of the individual parameter values. The parameter λ of the Poisson distribution is estimated by the sample mean when a sample is available. For example, the horse-kick data leads to an estimate of λ—say l—given by

$$l = \frac{0 \times 109 + 1 \times 65 + 2 \times 22 + 3 \times 3 + 4 \times 1}{109 + 65 + 22 + 3 + 1} = 0.61$$

Now, we consider the construction of confidence intervals for a Poisson parameter. Consider the case of one observation, Y, and a small result, say, $Y \leq 100$. Note 6.8 describes how confidence intervals are calculated and there is a table in the Web appendix to this chapter. From this we find a 95% confidence interval for the proportion of black infants having ABO hemolytic disease, in the Bucher et al. [1976] study. The approximate Poisson variable is the binomial variable, which in this case is equal to 43; thus, a 95% confidence interval for $\lambda = n\pi$ is (31.12, 57.92). The equation $\lambda = n\pi$ equates the mean values for the Poisson and binomial models. Now $n\pi$ is in (31.12, 57.92) if and only if π is in the interval

$$\left(\frac{31.12}{n}, \frac{57.92}{n}\right)$$

In this case, $n = 3584$, so the confidence interval is

$$\left(\frac{31.12}{3584}, \frac{57.92}{3584}\right) \quad \text{or} \quad (0.0087, 0.0162)$$

These results are comparable with the 95% binomial limits obtained in Example 6.9: (0.0084, 0.0156).

6.5.3 Large-Sample Statistical Inference for the Poisson Distribution

Normal Approximation to the Poisson Distribution

The Poisson distribution has the property that the mean and variance are equal. For the mean large, say $\geq$ 100, the normal approximation can be used. That is, let $Y \sim \text{Poisson}(\lambda)$ and $\lambda \geq 100$. Then, approximately, $Y \sim N(\lambda, \lambda)$. An approximate $100(1 - \alpha)\%$ confidence interval for λ can be formed from

$$Y \pm z_{1-\alpha/2}\sqrt{Y}$$

where $z_{1-\alpha/2}$ is a standard normal deviate at two-sided significance level α. This formula is based on the fact that Y estimates the mean as well as the variance. Consider, again, the data of Bucher et al. [1976] (Example 6.3) dealing with the incidence of ABO hemolytic disease. The observed value of Y, the number of black infants with ABO hemolytic disease, was 43. A 95% confidence interval for the mean, λ, is (31.12, 57.92). Even though $Y \leq 100$, let us use the normal approximation. The estimate of the variance, σ^2, of the normal distribution is $Y = 43$, so that the standard deviation is 6.56. An approximate 95% confidence interval is $43 \pm (1.96)(6.56)$, producing (30.1, 55.9), which is close to the values (31.12, 57.92) tabled.

Suppose that instead of one Poisson value, there is a random sample of size n, $Y_1, Y_2, \ldots, Y_n$ from a Poisson distribution with mean λ. How should one construct a confidence interval for λ based on these data? The sum $Y = Y_1 + Y_2 + \cdots + Y_n$ is Poisson with mean $n\lambda$. Construct a confidence interval for $n\lambda$ as above, say (L, U). Then, an appropriate confidence interval for λ is $(L/n, U/n)$. Consider Example 6.20, which deals with estimating the bacterial density of soil suspensions. The results for sample I were 72, 69, 63, 59, 59, 53, and 51. We want to set up a 95% confidence interval for the mean density using the seven observations. For this example, $n = 7$.

$$Y = Y_1 + Y_2 + \cdots + Y_7 = 72 + 69 + \cdots + 51 = 426$$

A 95% confidence interval for 7λ is $426 \pm 1.96\sqrt{426}$.

$$L = 385.55, \qquad \frac{L}{7} = 55.1$$
$$U = 466.45, \qquad \frac{U}{7} = 66.6$$
$$\overline{Y} = 60.9$$

The 95% confidence interval is (55.1, 66.6).

Square Root Transformation

It is often considered a disadvantage to have a distribution with a variance not "stable" but dependent on the mean in some way, as, for example, the Poisson distribution. The question is whether there is a transformation, $g(Y)$, of the variable such that the variance is no longer dependent on the mean. The answer is "yes." For the Poisson distribution, it is the square root transformation. It can be shown for "reasonably large" λ, say $\lambda \geq 30$, that if $Y \sim$ Poisson(λ), then $\text{var}(\sqrt{Y}) \doteq 0.25$.

A side benefit is that the distribution of $\sqrt{Y}$ is more "nearly normal," that is, for specified λ, the difference between the sampling distribution of $\sqrt{Y}$ and the normal distribution is smaller for most values than the difference between the distribution of Y and the normal distribution.

For the situation above, it is approximately true that

$$\sqrt{Y} \sim N(\sqrt{\lambda}, 0.25)$$

Consider Example 6.20 again. A confidence interval for $\sqrt{\lambda}$ will be constructed and then converted to an interval for λ. Let $X = \sqrt{Y}$.

Y	72	69	63	59	59	53	51
$X = \sqrt{Y}$	8.49	8.31	7.94	7.68	7.68	7.28	7.14

The sample mean and variance of X are $\overline{X} = 7.7886$ and $s_x^2 = 0.2483$. The sample variance is very close to the variance predicted by the theory $\sigma_x^2 = 0.2500$. A 95% confidence interval on $\sqrt{\lambda}$ can be set up from

$$\overline{X} \pm 1.96\frac{s_x}{\sqrt{7}} \quad \text{or} \quad 7.7886 \pm (1.96)\sqrt{\frac{0.2483}{7}}$$

producing lower and upper limits in the X scale.

$$L_x = 7.4195, \quad U_x = 8.1577$$
$$L_x^2 = 55.0, \quad U_x^2 = 66.5$$

which are remarkably close to the values given previously.

Poisson Homogeneity Test

In Chapter 4 the question of a test of normality was discussed and a graphical procedure was suggested. Fisher et al. [1922], in the paper described in Example 6.20, derived an approximate test for determining whether or not a sample of observations could have come from a Poisson distribution with the same mean. The test does not determine "Poissonness," but rather, equality of means. If the experimental situations are identical (i.e., we have a random sample), the test is a test for Poissonness.

The test, the *Poisson homogeneity test*, is based on the property that for the Poisson distribution, the mean equals the variance. The test is the following: Suppose that $Y_1, Y_2, \ldots, Y_n$ are a random sample from a Poisson distribution with mean λ. Then, for a large λ—say, $\lambda \geq 50$—the quantity

$$X^2 = \frac{(n-1)s^2}{\overline{Y}}$$

has approximately a chi-square distribution with $n - 1$ degrees of freedom, where s^2 is the sample variance.

Consider again the data in Example 6.20. The mean and standard deviation of the seven observations are

$$n = 7, \qquad \overline{Y} = 60.86, \qquad s_y = 7.7552$$

$$X^2 = \frac{(7-1)(7.7552)^2}{60.86} = 5.93$$

Under the null hypothesis that all the observations are from a Poisson distribution with the same mean, the statistic $X^2 = 5.93$ can be referred to a chi-square distribution with six degrees of freedom. What will the rejection region be? This is determined by the alternative hypothesis. In this case it is reasonable to suppose that the sample variance will be greater than expected if the null hypothesis is not true. Hence, we want to reject the null hypothesis when χ^2 is "large"; "large" in this case means $P[X^2 \geq \chi^2_{1-\alpha}] = \alpha$.

Suppose that $\alpha = 0.05$; the critical value for $\chi^2_{1-\alpha}$ with 6 degrees of freedom is 12.59. The observed value $X^2 = 5.93$ is much less than that and the null hypothesis is not rejected.

6.6 GOODNESS-OF-FIT TESTS

The use of appropriate mathematical models has made possible advances in biomedical science; the key word is *appropriate*. An inappropriate model can lead to false or inappropriate ideas.

In some situations the appropriateness of a model is clear. A random sample of a population will lead to a binomial variable for the response to a yes or no question. In other situations the issue may be in doubt. In such cases one would like to examine the data to see if the model used seems to fit the data. Tests of this type are called *goodness-of-fit tests*. In this section we examine some tests where the tests are based on count data. The count data may arise from continuous data. One may count the number of observations in different intervals of the real line; examples are given in Sections 6.6.2 and 6.6.4.

6.6.1 Multinomial Random Variables

Binomial random variables count the number of successes in n independent trials where one and only one of two possibilities must occur. *Multinomial random variables* generalize this to allow more than two possible outcomes. In a multinomial situation, outcomes are observed that take one and only one of two or more, say k, possibilities. There are n independent trials, each with the same probability of a particular outcome. Multinomial random variables count the number of occurrences of a particular outcome. Let n_i be the number of occurrences of outcome i. Thus, n_i is an integer taking a value among $0, 1, 2, \ldots, n$. There are k different n_i, which add up to n since one and only one outcome occurs on each trial:

$$n_1 + n_2 + \cdots + n_k = n$$

Let us focus on a particular outcome, say the ith. What are the mean and variance of n_i? We may classify each outcome into one of two possibilities, the ith outcome or anything else. There are then n independent trials with two outcomes. We see that n_i is a binomial random variable when considered alone. Let π_i, where $i = 1, \ldots, k$, be the probability that the ith outcome occurs. Then

$$E(n_i) = n\pi_i, \qquad \operatorname{var}(n_i) = n\pi_i(1 - \pi_i) \tag{6}$$

for $i = 1, 2, \ldots, k$.

Often, multinomial outcomes are visualized as placing the outcome of each of the n trials into a separate *cell* or box. The probability π_i is then the probability that an outcome lands in the ith cell.

The remainder of this section deals with multinomial observations. Tests are presented to see if a specified multinomial model holds.

6.6.2 Known Cell Probabilities

In this section, the cell probabilities $\pi_1, \ldots, \pi_k$ are specified. We use the specified values as a null hypothesis to be compared with the data $n_1, \ldots, n_k$. Since $E(n_i) = n\pi_i$, it is reasonable to examine the differences $n_i - n\pi_i$. The statistical test is given by the following fact.

Fact 2. Let n_i, where $i = 1, \ldots, k$, be multinomial. Under $H_0 : \pi_i = \pi_i^0$,

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - n\pi_i^0)^2}{n\pi_i^0}$$

has approximately a chi-square distribution with $k - 1$ degrees of freedom. If some π_i are not equal to π_i^0, X^2 will tend to be too large.

The distribution of X^2 is well approximated by the chi-square distribution if all of the expected values, $n\pi_i^0$, are at least five, except possibly for one or two of the values. When the null hypothesis is not true, the null hypothesis is rejected for X^2 too large. At significance level

α, reject H_0 if $X^2 \geq \chi^2_{1-\alpha,k-1}$, where $\chi^2_{1-\alpha,k-1}$ is the $1-\alpha$ percentage point for a χ^2 random variable with $k-1$ degrees of freedom.

Since there are k cells, one might expect the labeling of the degrees of freedom to be k instead of $k-1$. However, since the n_i add up to n we only need to know $k-1$ of them to know all k values. There are really only $k-1$ quantities that may vary at a time; the last quantity is specified by the other $k-1$ values.

The form of X^2 may be kept in mind by noting that we are comparing the observed values, n_i, and expected values, $n\pi_i^0$. Thus,

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Example 6.22. Are births spread uniformly throughout the year? The data in Table 6.7 give the number of births in King County, Washington, from 1968 through 1979 by month. The estimated probability of a birth in a given month is found by taking the number of days in that month and dividing by the total number of days (leap years are included in Table 6.7).

Testing the null hypothesis using Table A.3, we see that $163.15 > 31.26 = \chi^2_{0.001,11}$, so that $p < 0.001$. We reject the null hypothesis that births occur uniformly throughout the year. With this large sample size ($n = 160{,}654$) it is not surprising that the null hypothesis can be rejected. We can examine the magnitude of the effect by comparing the ratio of observed to expected numbers of births, with the results shown in Table 6.8. There is an excess of births in the spring (March and April) and a deficit in the late fall and winter (October through January). Note that the difference from expected values is small. The maximum "excess" of births occurred

Table 6.7 Births in King County, Washington, 1968–1979

Month	Births	Days	π_i^0	$n\pi_i^0$	$(n_i - n\pi_i^0)^2/n\pi_i^0$
January	13,016	310	0.08486	13,633	27.92
February	12,398	283	0.07747	12,446	0.19
March	14,341	310	0.08486	13,633	36.77
April	13,744	300	0.08212	13,193	23.01
May	13,894	310	0.08486	13,633	5.00
June	13,433	300	0.08212	13,193	4.37
July	13,787	310	0.08486	13,633	1.74
August	13,537	310	0.08486	13,633	0.68
September	13,459	300	0.08212	13,193	5.36
October	13,144	310	0.08486	13,633	17.54
November	12,497	300	0.08212	13,193	36.72
December	13,404	310	0.08486	13,633	3.85
Total	160,654 (n)	3653	0.99997		$163.15 = X^2$

Table 6.8 Ratios of Observed to Expected Births

Month	Observed/Expected Births	Month	Observed/Expected Births
January	0.955	July	1.011
February	0.996	August	0.993
March	1.052	September	1.020
April	1.042	October	0.964
May	1.019	November	0.947
June	1.018	December	0.983

in March and was only 5.2% above the number expected. A plot of the ratio vs. month would show a distinct sinusoidal pattern.

Example 6.23. Mendel [1866] is justly famous for his theory and experiments on the principles of heredity. Sir R. A. Fisher [1936] reviewed Mendel's work and found a surprisingly good fit to the data. Consider two parents heterozygous for a dominant–recessive trait. That is, each parent has one dominant gene and one recessive gene. Mendel hypothesized that all four combinations of genes would be equally likely in the offspring. Let A denote the dominant gene and a denote the recessive gene. The two parents are Aa. The offspring should be

Genotype	Probability
AA	1/4
Aa	1/2
aa	1/4

The Aa combination has probability 1/2 since one cannot distinguish between the two cases where the dominant gene comes from one parent and the recessive gene from the other parent. In one of Mendel's experiments he examined whether a seed was wrinkled, denoted by a, or smooth, denoted by A. By looking at offspring of these seeds, Mendel classified the seeds as aa, Aa, or AA. The results were

	AA	Aa	aa	Total
Number	159	321	159	639

as presented in Table II of Fisher [1936]. Do these data support the hypothesized 1 : 2 : 1 ratio? The chi-square statistic is

$$X^2 = \frac{(159 - 159.75)^2}{159.75} + \frac{(321 - 319.5)^2}{319.5} + \frac{(159 - 159.75)^2}{159.75} = 0.014$$

For the χ^2 distribution with two degrees of freedom, $p > 0.95$ from Table A.3 (in fact $p = 0.993$), so that the result has more agreement than would be expected by chance. We return to these data in Example 6.24.

6.6.3 Addition of Independent Chi-Square Variables: Mean and Variance of the Chi-Square Distribution

Chi-square random variables occur so often in statistical analysis that it will be useful to know more facts about chi-square variables. In this section facts are presented and then applied to an example (see also Note 5.3).

Fact 3. Chi-square variables have the following properties:

1. Let X^2 be a chi-square random variable with m degrees of freedom. Then

$$E(X^2) = m \quad \text{and} \quad \text{var}(X^2) = 2m$$

2. Let $X_1^2, \dots, X_n^2$ be independent chi-square variables with $m_1, \dots, m_n$ degrees of freedom. Then $X^2 = X_1^2 + \cdots + X_n^2$ is a chi-square random variable with $m = m_1 + m_2 + \cdots + m_n$ degrees of freedom.

Table 6.9 Chi-Square Values for Mendel's Experiments

Experiments	X^2	Degrees of Freedom
3 : 1 Ratios	2.14	7
2 : 1 Ratios	5.17	8
Bifactorial experiments	2.81	8
Gametic ratios	3.67	15
Trifactorial experiments	15.32	26
Total	29.11	64

3. Let X^2 be a chi-square random variable with m degrees of freedom. If m is large, say $m \geq 30$,

$$\frac{X^2 - m}{\sqrt{2m}}$$

is approximately a $N(0, 1)$ random variable.

Example 6.24. We considered Mendel's data, reported by Fisher [1936], in Example 6.23. As Fisher examined the data, he became convinced that the data fit the hypothesis too well [Box, 1978, pp. 195, 300]. Fisher comments: "Although no explanation can be expected to be satisfactory, it remains a possibility among others that Mendel was deceived by some assistant who knew too well what was expected."

One reason Fisher arrived at his conclusion was by combining χ^2 values from different experiments by Mendel. Table 6.9 presents the data.

If all the null hypotheses are true, by the facts above, $X^2 = 29.11$ should look like a χ^2 with 64 degrees of freedom. An approximate normal variable,

$$Z = \frac{29.11 - 64}{\sqrt{128}} = -3.08$$

has less than 1 chance in 1000 of being this small ($p = 0.99995$). One can only conclude that something peculiar occurred in the collection and reporting of Mendel's data.

6.6.4 Chi-Square Tests for Unknown Cell Probabilities

Above, we considered tests of the goodness of fit of multinomial data when the probability of being in an individual cell was specified precisely: for example, by a genetic model of how traits are inherited. In other situations, the cell probabilities are not known but may be estimated. First, we motivate the techniques by presenting a possible use; next, we present the techniques, and finally, we illustrate the use of the techniques by example.

Consider a sample of n numbers that may come from a normal distribution. How might we check the assumption of normality? One approach is to divide the real number line into a finite number of intervals. The number of points observed in each interval may then be counted. The numbers in the various intervals or cells are multinomial random variables. If the sample were normal with known mean μ and known standard deviation σ, the probability, π_i, that a point falls between the endpoints of the ith interval—say Y_1 and Y_2—is known to be

$$\pi_i = \Phi\left(\frac{Y_2 - \mu}{\sigma}\right) - \Phi\left(\frac{Y_1 - \mu}{\sigma}\right)$$

where Φ is the distribution function of a standard normal random variable. In most cases, μ and σ are not known, so μ and σ, and thus π_i, must be estimated. Now π_i depends on two variables, μ and $\sigma : \pi_i = \pi_i(\mu, \sigma)$ where the notation $\pi_i(\mu, \sigma)$ means that π_i is a function of μ and σ. It is natural if we estimate μ and σ by, say, $\widehat{\mu}$ and $\widehat{\sigma}$, to estimate π_i by $p_i(\widehat{\mu}, \widehat{\omega})$. That is,

$$p_i(\widehat{\mu}, \widehat{\sigma}) = \Phi\left(\frac{Y_2 - \widehat{\mu}}{\widehat{\sigma}}\right) - \Phi\left(\frac{Y_1 - \widehat{\mu}}{\widehat{\sigma}}\right)$$

From this, a statistic (X^2) can be formed as above. If there are k cells,

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_{i=1}^{k} \frac{[n_i - np_i(\widehat{\mu}, \widehat{\sigma})]^2}{np_i(\widehat{\mu}, \widehat{\sigma})}$$

Does X^2 now have a chi-square distribution? The following facts describe the situation.

Fact 4. Suppose that n observations are grouped or placed into k categories or cells such that the probability of being in cell i is $\pi_i = \pi_i(\Theta_1, \ldots, \Theta_s)$, where π_i depends on s parameters Θ_j and where $s < k-1$. Suppose that none of the s parameters are determined by the remaining $s - 1$ parameters. Then:

1. If $\widehat{\Theta}_1, \ldots, \widehat{\Theta}_s$, the parameter estimates, are chosen to minimize X^2, the distribution of X^2 is approximately a chi-square random variable with $k - s - 1$ degrees of freedom for large n. Estimates chosen to minimize the value of X^2 are called *minimum chi-square estimates*.
2. If estimates of $\Theta_1, \ldots, \Theta_s$ other than the minimum chi-square estimates are used, then for large n the distribution function of X^2 lies between the distribution functions of chi-square variables with $k - s - 1$ degrees of freedom and $k - 1$ degrees of freedom. More specifically, let $X^2_{1-\alpha,m}$ denote the α-significance-level critical value for a chi-square distribution with m degrees of freedom. The significance-level-α critical value of X^2 is less than or equal to $X^2_{1-\alpha,k-1}$. A conservative test of the multinomial model is to reject the null hypothesis that the model is correct if $X^2 \geq \chi^2_{1-\alpha,k-1}$.

These complex statements are best understood by applying them to an example.

Example 6.25. Table 3.4 in Section 3.3.1 gives the age in days at death of 78 SIDS cases. Test for normality at the 5% significance level using a χ^2-test.

Before performing the test, we need to divide the real number line into intervals or cells. The usual approach is to:

1. Estimate the parameters involved. In this case the unknown parameters are μ and σ. We estimate by $\overline{Y}$ and s.
2. Decide on k, the number of intervals. Let there be n observations. A good approach is to choose k as follows:

   a. For $20 \leq n \leq 100, k \doteq n/5$.

   b. For $n > 300, k \doteq 3.5n^{2/5}$ (here, $n^{2/5}$ is n raised to the 2/5 power).

3. Find the endpoints of the k intervals so that each interval has probability $1/k$. The k intervals are

$$\begin{array}{ll} (-\infty, a_1] & \text{interval 1} \\ (a_1, a_2] & \text{interval 2} \\ \vdots & \vdots \\ (a_{k-2}, a_{k-1}] & \text{interval } (k-1) \\ (a_{k-1}, \infty) & \text{interval k} \end{array}$$

Let Z_i be a value such that a standard normal random variable takes a value less than Z_i with probability i/k. Then

$$a_i = \overline{X} + sZ_i$$

(In testing for a distribution other than the normal distribution, other methods of finding cells of approximately equal probability need to be used.)

4. Compute the statistic

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - n/k)^2}{n/k}$$

where n_i is the number of data points in cell i.

To apply steps 1 to 4 to the data at hand, one computes $n = 78$, $\overline{X} = 97.85$, and $s = 55.66$. As $78/5 = 15.6$, we will use $k = 15$ intervals. From tables of the normal distribution, we find Z_i, $i = 1, 2, \ldots, 14$, so that a standard normal random variable has probability $i/15$ of being less than Z_i. The values of Z_i and a_i are given in Table 6.10.

The number of observations observed in the 15 cells, from left to right, are 0, 8, 7, 5, 7, 9, 7, 5, 6, 6, 2, 2, 3, 5, and 6. In each cell, the number of observations expected is $np_i = n/k$ or $78/15 = 5.2$. Then

$$X^2 = \frac{(0-5.2)^2}{5.2} + \frac{(8-5.2)^2}{5.2} + \cdots + \frac{(6-5.2)^2}{5.2} = 16.62$$

We know that the 0.05 critical values are between the chi-square critical values with 12 and 14 degrees of freedom. The two values are 21.03 and 23.68. Thus, we do not reject the hypothesis of normality. (If the X^2 value had been greater than 23.68, we would have rejected the null hypothesis of normality. If X^2 were between 21.03 and 23.68, the answer would be in doubt. In that case, it would be advisable to compute the minimum chi-square estimates so that a known distribution results.)

Note that the largest observation, 307, is $(307 - 97.85)/55.6 = 3.76$ sample standard deviations from the sample mean. In using a chi-square goodness-of-fit test, all large observations are placed into a single cell. The magnitude of the value is lost. If one is worried about large outlying values, there are better tests of the fit to normality.

Table 6.10 Z_i and a_i Values

i	Z_i	a_i	i	Z_i	a_i	i	Z_i	a_i
1	−1.50	12.8	6	−0.25	84.9	11	0.62	135.0
2	−1.11	35.3	7	−0.08	94.7	12	0.84	147.7
3	−0.84	50.9	8	0.08	103.9	13	1.11	163.3
4	−0.62	63.5	9	0.25	113.7	14	1.50	185.8
5	−0.43	74.5	10	0.43	124.1			

NOTES

6.1 Continuity Correction for 2 × 2 *Table Chi-Square Values*

There has been controversy about the appropriateness of the continuity correction for 2 × 2 tables [Conover, 1974]. The continuity correction makes the *actual* significance levels under the null hypothesis closer to the hypergeometric (Fisher's exact test) actual significance levels. When compared to the chi-square distribution, the *actual* significance levels are too low [Conover, 1974; Starmer et al., 1974; Grizzle, 1967]. The *uncorrected* "chi-square" value referred to chi-square critical values gives actual and nominal significance levels that are close. For this reason, the authors recommend that the continuity correction *not* be used. Use of the continuity correction would be correct but overconservative. For arguments on the opposite side, see Mantel and Greenhouse [1968]. A good summary can be found in Little [1989].

6.2 Standard Error of $\widehat{\omega}$ as Related to the Standard Error of $\log\widehat{\omega}$

Let X be a positive variate with mean μ_x and standard deviation σ_x. Let $Y = \log_e X$. Let the mean and standard deviation of Y be μ_y and σ_y, respectively. It can be shown that under certain conditions

$$\frac{\sigma_x}{\mu_x} \doteq \sigma_y$$

The quantity σ_x/μ_x is known as the *coefficient of variation.* Another way of writing this is

$$\sigma_x \doteq \mu_x \sigma_y$$

If the parameters are replaced by the appropriate statistics, the expression becomes

$$s_x \doteq \overline{x} s_y$$

and the standard deviation of $\widehat{\omega}$ then follows from this relationship.

6.3 Some Limitations of the Odds Ratio

The odds ratio uses one number to summarize four numbers, and some information about the relationship is necessarily lost. The following example shows one of the limitations. Fleiss [1981] discusses the limitations of the odds ratio as a measure for public health. He presents the mortality rates per 100,000 person-years from lung cancer and coronary artery disease for smokers and nonsmokers of cigarettes [U.S. Department of Health, Education and Welfare, 1964]:

	Smokers	**Nonsmokers**	**Odds Ratio**	**Difference**
Cancer of the lung	48.33	4.49	10.8	43.84
Coronary artery disease	294.67	169.54	1.7	125.13

The point is that although the risk ω is increased much more for cancer, the added number dying of coronary artery disease is higher, and in some sense smoking has a greater effect in this case.

6.4 Mantel–Haenszel Test for Association

The chi-square test of association given in conjunction with the Mantel–Haenszel test discussed in Section 6.3.5 arises from the approach of the section by choosing a_i and s_i appropriately

[Fleiss, 1981]. The corresponding chi-square test for homogeneity does *not* make sense and should not be used. Mantel et al. [1977] give the problems associated with using this approach to look at homogeneity.

6.5 Matched Pair Studies

One of the difficult aspects in the design and execution of matched pair studies is to decide on the matching variables, and then to find matches to the degree desired. In practice, many decisions are made for logistic and monetary reasons; these factors are not discussed here. The primary purpose of matching is to have a *valid* comparison. Variables are matched to increase the validity of the comparison. Inappropriate matching can hurt the statistical power of the comparison. Breslow and Day [1980] and Miettinen [1970] give some fundamental background. Fisher and Patil [1974] further elucidate the matter (see also Problem 6.30).

6.6 More on the Chi-Square Goodness-of-Fit Test

The goodness-of-fit test as presented in this chapter did not mention some of the subtleties associated with the subject. A few arcane points, with appropriate references, are given in this note.

1. In Fact 4, the estimate used should be maximum likelihood estimates or equivalent estimates [Chernoff and Lehmann, 1954].
2. The initial chi-square limit theorems were proved for fixed cell boundaries. Limiting theorems where the boundaries were random (depending on the data) were proved later [Kendall and Stuart, 1967, Secs. 30.20 and 30.21].
3. The number of cells to be used (as a function of the sample size) has its own literature. More detail is given in Kendall and Stuart [1967, Secs. 30.28 to 30.30]. The recommendations for k in the present book are based on this material.

6.7 Predictive Value of a Positive Test

The predictive value of a positive test, PV$^+$, is related to the prevalence (PREV), sensitivity (SENS), and specificity (SPEC) of a test by the following equation:

$$\text{PV}^+ = \frac{1}{1 + \left[(1 - \text{SPEC})/\text{SENS}\right]\left[(1 - \text{PREV})/\text{PREV}\right]}$$

Here PREV, SENS, and SPEC, are on a scale of 0 to 1 of proportions instead of percentages.

If we define $\text{logit}(p) = \log[p/(1 - p)]$, the predictive value of a positive test is related very simply to the prevalence as follows:

$$\text{logit}[\text{PV}^+] = \log\left(\frac{\text{SENS}}{1 - \text{SPEC}}\right) + \text{logit}(\text{PREV})$$

This is a very informative formula. For rare diseases (i.e., low prevalence), the term "logit (PREV)" will dominate the predictive value of a positive test. So no matter what the sensitivity or specificity of a test, the predictive value will be low.

6.8 Confidence Intervals for a Poisson Mean

Many software packages now provide confidence intervals for the mean of a Poisson distribution. There are two formulas: an approximate one that can be done by hand, and a more complex exact formula. The approximate formula uses the following steps. Given a Poisson variable Y:

1. Take $\sqrt{Y}$.

2. Add and subtract 1.

3. Square the result $[(\sqrt{Y}-1)^2,\ (\sqrt{Y}+1)^2]$.

This formula is reasonably accurate for $Y \geq 5$. See also Note 6.9 for a simple confidence interval when $Y = 0$. The exact formula uses the relationship between the Poisson and χ^2 distributions to give the confidence interval

$$\left[\frac{1}{2}\chi^2_{\alpha/2}(2x),\ \frac{1}{2}\chi^2_{1-\alpha/2}(2x+2)\right]$$

where $\chi^2_{\alpha/2}(2x)$ is the $\alpha/2$ percentile of the χ^2 distribution with $2x$ degrees of freedom.

6.9 Rule of Threes

An upper 90% confidence bound for a Poisson random variable with observed values 0 is, to a very good approximation, 3. This has led to the *rule of threes*, which states that if in n trials zero events of interest are observed, a 95% confidence bound on the underlying rate is $3/n$. For a fuller discussion, see Hanley and Lippman-Hard [1983]. See also Problem 6.29.

PROBLEMS

6.1 In a randomized trial of surgical and medical treatment a clinic finds eight of nine patients randomized to medicine. They complain that the randomization must not be working; that is, π cannot be 1/2.

(a) Is their argument reasonable from their point of view?

***(b)** With 15 clinics in the trial, what is the probability that *all* 15 clinics have fewer than eight people randomized to each treatment, of the first nine people randomized? Assume independent binomial distributions with $\pi = 1/2$ at each site.

6.2 In a dietary study, 14 of 20 subjects lost weight. If weight is assumed to fluctuate by chance, with probability 1/2 of losing weight, what is the exact two-sided p-value for testing the null hypothesis $\pi = 1/2$?

6.3 Edwards and Fraccaro [1960] present Swedish data about the gender of a child and the parity. These data are:

	Order of Birth							
Gender	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
Males	2846	2554	2162	1667	1341	987	666	12,223
Females	2631	2361	1996	1676	1230	914	668	11,476
Total	5477	4915	4158	3343	2571	1901	1334	23,699

(a) Find the p-value for testing the hypothesis that a birth is equally likely to be of either gender using the combined data and binomial assumptions.

(b) Construct a 90% confidence interval for the probability that a birth is a female child.

(c) Repeat parts (a) and (b) using only the data for birth order 6.

6.4 Ounsted [1953] presents data about cases with convulsive disorders. Among the cases there were 82 females and 118 males. At the 5% significance level, test the hypothesis that a case is equally likely to be of either gender. The siblings of the cases were 121 females and 156 males. Test at the 10% significance level the hypothesis that the siblings represent 53% or more male births.

6.5 Smith et al. [1976] report data on ovarian carcinoma (cancer of the ovaries). People had different numbers of courses of chemotherapy. The five-year survival data for those with 1–4 and 10 or more courses of chemotherapy are:

	Five-Year Status	
Courses	**Dead**	**Alive**
1–4	21	2
$\geq$10	2	8

Using Fisher's exact test, is there a statistically significant association ($p \leq 0.05$) in this table? (In this problem and the next, you will need to compute the hypergeometric probabilities using the results of Problem 6.26.)

6.6 Borer et al. [1980] study 45 patients following an acute myocardial infarction (heart attack). They measure the *ejection fraction* (EF), the percent of the blood pumped from the left ventricle (the pumping chamber of the heart) during a heart beat. A low EF indicates damaged or dead heart muscle (myocardium). During follow-up, four patients died. Dividing EF into low (<35%) and high ($\geq$35%) EF groups gave the following table:

	Vital Status	
EF	**Dead**	**Alive**
<35%	4	9
$\geq$35%	0	32

Is there reason to suspect, at a 0.05 significance level, that death is more likely in the low EF group? Use a one-sided p-value for your answer, since biological plausibility (and prior literature) indicates that low EF is a risk factor for mortality.

6.7 Using the data of Problem 6.4, test the hypothesis that the proportions of male births among those with convulsive disorders and among their siblings are the same.

6.8 Lawson and Jick [1976] compare drug prescription in the United States and Scotland.

(a) In patients with congestive heart failure, two or more drugs were prescribed in 257 of 437 U.S. patients. In Scotland, 39 of 179 patients had two or more drugs prescribed. Test the null hypothesis of equal proportions giving the resulting p-value. Construct a 95% confidence interval for the difference in proportions.

(b) Patients with dehydration received two or more drugs in 55 of 74 Scottish cases as compared to 255 of 536 in the United States. Answer the questions of part (a).

6.9 A randomized study among patients with angina (heart chest pain) is to be conducted with five-year follow-up. Patients are to be randomized to medical and surgical treatment. Suppose that the estimated five-year medical mortality is 10% and it is hoped that the surgical mortality will be only half as much (5%) or less. If a test of binomial proportions at the 5% significance level is to be performed, and we want to be 90% certain of detecting a difference of 5% or more, what sample sizes are needed for the two (equal-sized) groups?

6.10 A cancer with poor prognosis, a three-year mortality of 85%, is studied. A new mode of chemotherapy is to be evaluated. Suppose that when testing at the 0.10 significance level, one wishes to be 95% certain of detecting a difference if survival has been increased to 50% or more. The randomized clinical trial will have equal numbers of people in each group. How many patients should be randomized?

6.11 Comstock and Partridge [1972] show data giving an association between church attendance and health. From the data of Example 6.17, which were collected from a prospective study:

(a) Compute the relative risk of an arteriosclerotic death in the three-year follow-up period if one usually attends church less than once a week as compared to once a week or more.

(b) Compute the odds ratio and a 95% confidence interval.

(c) Find the percent error of the odds ratio as an approximation to the relative risk; that is, compute 100(OR − RR)/RR.

(d) The data in this population on deaths from cirrhosis of the liver are:

Usual Church	**Cirrhosis Fatality?**	
Attendance	**Yes**	**No**
≥1 per week	5	24,240
<1 per week	25	30,578

Repeat parts (a), (b), and (c) for these data.

6.12 Peterson et al. [1979] studied the patterns of infant deaths (especially SIDS) in King County, Washington during the years 1969–1977. They compared the SIDS deaths with a 1% sample of all births during the time period specified. Tables relating the occurrence of SIDS with maternal age less than or equal to 19 years of age, and to birth order greater than 1, follow for those with single births.

	Child			**Child**	
Birth Order	**SIDS**	**Control**	**Maternal Age**	**SIDS**	**Control**
>1	201	689	≤19	76	164
=1	92	626	>19	217	1151

	Child	
	SIDS	Control
Birth order >1 *and* maternal age ≤19	26	17
Birth order =1 *or* maternal age >19	267	1298
Birth order >1 *and* maternal age ≤19	26	17
Birth order =1 *and* maternal age >19	42	479

(a) Compute the odds ratios and 95% confidence intervals for the data in these tables.

(b) Which pair of entries in the second table do you think best reflects the risk of both risk factors at once? Why? (There is not a definitely correct answer.)

***(c)** The control data represent a 1% sample of the population data. Knowing this, how would you estimate the relative risk directly?

6.13 Rosenberg et al. [1980] studied the relationship between coffee drinking and myocardial infarction in young women aged 30–49 years. This retrospective study included 487 cases hospitalized for the occurrence of a myocardial infarction (MI). Nine hundred eighty controls hospitalized for an acute condition (trauma, acute cholecystitis, acute respiratory diseases, and appendicitis) were selected. Data for consumption of five or more cups of coffee containing caffeine were:

Cups per Day	MI	Control
≥5	152	183
<5	335	797

Compute the odds ratio of a MI for heavy (≥5 cups per day) coffee drinkers vs. nonheavy coffee drinkers. Find the 90% confidence interval for the odds ratio.

6.14 The data of Problem 6.13 were considered to be possibly confounded with smoking. The 2 × 2 tables by smoking status, in cigarettes per day, are displayed in Table 6.11.

(a) Compute the Mantel–Haenszel estimate of the odds ratio and the chi-square statistic for association. Would you reject the null hypothesis of no association between coffee drinking and myocardial infarction at the 5% significance level?

(b) Using the log odds ratio as the measure of association in each table, compute the chi-square statistic for association. Find the estimated overall odds ratio and a 95% confidence interval for this quantity.

6.15 The paper of Remein and Wilkerson [1961] considers screening tests for diabetes. The Somogyi–Nelson (venous) blood test (data at 1 hour after a test meal and using 130 mg per 100 mL as the blood sugar cutoff level) gives the following table:

Test	Diabetic	Nondiabetic	Total
+	59	48	107
−	11	462	473
Total	70	510	580

Table 6.11 2 x 2 Tables for Problem 6.14

Cups per Day	MI	Control
Never smoked		
≥5	7	31
<5	55	269
Former smoker		
≥5	7	18
<5	20	112
1–14 cigarettes per day		
≥5	7	24
<5	33	114
15–24 cigarettes per day		
≥5	40	45
<5	88	172
25–34 cigarettes per day		
≥5	34	24
<5	50	55
35–44 cigarettes per day		
≥5	27	24
<5	55	58
45+ cigarettes per day		
≥5	30	17
<5	34	17

(a) Compute the sensitivity, specificity, predictive value of a positive test, and predictive value of a negative test.

(b) Using the sensitivity and specificity of the test as given in part (a), plot curves of the predictive values of the test vs. the percent of the population with diabetes (0 to 100%). The first curve will give the probability of diabetes given a positive test. The second curve will give the probability of diabetes given a negative test.

6.16 Remein and Wilkerson [1961] present tables showing the trade-off between sensitivity and specificity that arises by changing the cutoff value for a positive test. For blood samples collected 1 hour after a test meal, three different blood tests gave the data given in Table 6.12.

(a) Plot three curves, one for each testing method, on the same graph. Let the vertical axis be the sensitivity and the horizontal axis be (1 − specificity) of the test. The curves are generated by the changing cutoff values.

(b) Which test, if any, looks most promising? Why? (See also Note 6.7)

6.17 Data of Sartwell et al. [1969] that examine the relationship between thromboembolism and oral contraceptive use are presented below for several subsets of the population. For each subset:

(a) Perform McNemar's test for a case–control difference (5% significance level).

(b) Estimate the relative risk.

(c) Find an appropriate 90% confidence interval for the relative risk.

Table 6.12 Blood Sugar Data for Problem 6.16

Blood Sugar	Type of Test					
	Somogyi–Nelson		Folin–Wu		Anthrone	
(mg/100 mL)	SENS	SPEC	SENS	SPEC	SENS	SPEC
70	—	—	100.0	8.2	100.0	2.7
80	—	1.6	97.1	22.4	100.0	9.4
90	100.0	8.8	97.1	39.0	100.0	22.4
100	98.6	21.4	95.7	57.3	98.6	37.3
110	98.6	38.4	92.9	70.6	94.3	54.3
120	97.1	55.9	88.6	83.3	88.6	67.1
130	92.9	70.2	78.6	90.6	81.4	80.6
140	85.7	81.4	68.6	95.1	74.3	88.2
150	80.0	90.4	57.1	97.8	64.3	92.7
160	74.3	94.3	52.9	99.4	58.6	96.3
170	61.4	97.8	47.1	99.6	51.4	98.6
180	52.9	99.0	40.0	99.8	45.7	99.2
190	44.3	99.8	34.3	100.0	40.0	99.8
200	40.0	99.8	28.6	100.0	35.7	99.8

For nonwhites:

	Case	
Control	**Yes**	**No**
Yes	3	3
No	11	9

For married:

	Case	
Control	**Yes**	**No**
Yes	8	10
No	41	46

and for ages 15–29:

	Case	
Control	**Yes**	**No**
Yes	5	33
No	7	57

6.18 Janerich et al. [1980] compared oral contraceptive use among mothers of malformed infants and matched controls who gave birth to healthy children. The controls were matched for maternal age and race of the mother. For each of the following, estimate the odds ratio and form a 90% confidence interval for the odds ratio.

(a) Women who conceived while using the pill or immediately following pill use.

	Case	
Control	**Yes**	**No**
Yes	1	33
No	49	632

(b) Women who experienced at least one complete pill-free menstrual period prior to conception.

	Case	
Control	**Yes**	**No**
Yes	38	105
No	105	467

(c) Cases restricted to major structural anatomical malformations; use of oral contraceptives after the last menstrual period or in the menstrual cycle prior to conception.

	Case	
Control	**Yes**	**No**
Yes	0	21
No	45	470

(d) As in part (c) but restricted to mothers of age 30 or older.

	Case	
Control	**Yes**	**No**
Yes	0	1
No	6	103

6.19 Robinette et al. [1980] studied the effects on health of occupational exposure to microwave radiation (radar). The study looked at groups of enlisted naval personnel who were enrolled during the Korean War period. Find 95% confidence intervals for the percent of men dying of various causes, as given in the data below. Deaths were recorded that occurred during 1950–1974.

(a) Eight of 1412 aviation electronics technicians died of malignant neoplasms.

(b) Six of the 1412 aviation electronics technicians died of suicide, homicide, or other trauma.

(c) Nineteen of 10,116 radarmen died by suicide.

(d) Sixteen of 3298 fire control technicians died of malignant neoplasms.

(e) Three of 9253 radiomen died of infective and parasitic disease.

(f) None of 1412 aviation electronics technicians died of infective and parasitic disease.

6.20 The following data are also from Robinette et al. [1980]. Find 95% confidence intervals for the population percent dying based on these data: (1) 199 of 13,078 electronics technicians died of disease; (2) 100 of 13,078 electronics technicians died of circulatory disease; (3) 308 of 10,116 radarmen died (of any cause); (4) 441 of 13,078 electronics technicians died (of any cause); (5) 103 of 10,116 radarmen died of an accidental death.

(a) Use the normal approximation to the Poisson distribution (which is approximating a binomial distribution).

(b) Use the large-sample binomial confidence intervals (of Section 6.2.6). Do you think the intervals are similar to those calculated in part (a)?

6.21 Infant deaths in King County, Washington were grouped by season of the year. The number of deaths by season, for selected causes of death, are listed in Table 6.13.

Table 6.13 Death Data for Problem 6.21

	Season			
	Winter	Spring	Summer	Autumn
Asphyxia	50	48	46	34
Immaturity	30	40	36	35
Congenital malformations	95	93	88	83
Infection	40	19	40	43
Sudden infant death syndrome	78	71	87	86

(a) At the 5% significance level, test the hypothesis that SIDS deaths are uniformly ($p = 1/4$) spread among the seasons.

(b) At the 10% significance level, test the hypothesis that the deaths due to infection are uniformly spread among the seasons.

(c) What can you say about the p-value for testing that asphyxia deaths are spread uniformly among seasons? Immaturity deaths?

6.22 Fisher [1958] (after [Carver, 1927]) provided the following data on 3839 seedlings that were progeny of self-fertilized heterozygotes (each seedling can be classified as either starchy or sugary and as either green or white):

Number of Seedlings	**Green**	**White**	**Total**
Starchy	1997	906	2903
Surgary	904	32	936
Total	2901	938	3839

(a) On the assumption that the green and starchy genes are dominant and that the factors are independent, show that by Mendel's law that the ratio of expected frequencies (starchy green, starchy white, sugary green, sugary white) should be 9 : 3: 3 : 1.

(b) Calculate the expected frequencies under the hypothesis that Mendel's law holds and assuming 3839 seedlings.

(c) The data are multinomial with parameters π_1, π_2, π_3, and π_4, say. What does Mendel's law imply about the relationships among the parameters?

(d) Test the goodness of fit.

6.23 Fisher [1958] presented data of Geissler [1889] on the number of male births in German families with eight offspring. One model that might be considered for these data is the binomial distribution. This problem requires a goodness-of-fit test.

(a) Estimate π, the probability that a birth is male. This is done by using the estimate $p =$ (total number of male births)/(total number of births). The data are given in Table 3.10.

(b) Using the p of part (a), find the binomial probabilities for number of boys = 0, 1, 2, 3, 4, 5, 6, 7, and 8. Estimate the expected number of observations in each cell if the binomial distribution is correct.

(c) Compute the X^2 value.

(d) The X^2 distribution lies between chi-square distributions with what two degrees of freedom? (Refer to Section 6.6.4)

***(e)** Test the goodness of fit by finding the two critical values of part (d). What can you say about the p-value for the goodness-of-fit test?

***6.24** **(a)** Let $R(n)$ be the number of ways to arrange n distinct objects in a row. Show that $R(n) = n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$. By definition, $R(0) = 1$. *Hint:* Clearly, $R(1) = 1$. Use *mathematical induction*. That is, show that if $R(n-1) = (n-1)!$, then $R(n) = n!$. This would show that for all positive integers n, $R(n) = n!$. Why? [To show that $R(n) = n!$, suppose that $R(n-1) = (n-1)!$. Argue that you may choose any of the n objects for the first position. For each such choice, the remaining $n-1$ objects may be arranged in $R(n-1) = (n-1)!$ different ways.]

(b) Show that the number of ways to select k objects from n objects, denoted by $\binom{n}{k}$ (the binomial coefficient), is $n!/((n-k)!\,k!)$. *Hint*: We will choose the k objects by arranging the n objects in a row; the first k objects will be the ones we select. There are $R(n)$ ways to do this. When we do this, we get the *same* k objects many times. There are $R(k)$ ways to arrange the *same* k objects in the first k positions. For each such arrangement, the other $n-k$ objects may be arranged in $R(n-k)$ ways. The number of ways to arrange these objects is $R(k)R(n-k)$. Since each of the k objects is counted $R(k)R(n-k)$ times in the $R(n)$ arrangements, the number of different ways to select k objects is

$$\frac{R(n)}{R(k)R(n-k)} = \frac{n!}{k!\,(n-k)!}$$

from part (a). Then check that

$$\binom{n}{n} = \binom{n}{0} = 1$$

(c) Consider the binomial situation: n independent trials each with probability π of success. Show that the probability of k successes

$$b(k; n, \pi) = \binom{n}{k} \pi^k (1-\pi)^{n-k}$$

Hint: Think of the n trials as ordered. There are $\binom{n}{k}$ ways to choose the k trials that give a success. Using the independence of the trials, argue that the probability of the k trials being a success is $\pi^k(1-\pi)^{n-k}$.

(d) Compute from the definition of $b(k; n, \pi)$: (i) $b(3; 5, 0.5)$; (ii) $b(3; 3, 0.3)$; (iii) $b(2; 4, 0.2)$; (iv) $b(1; 3, 0.7)$; (v) $b(4; 6, 0.1)$.

6.25 In Section 6.2.3 we presented procedures for two-sided hypothesis tests with the binomial distribution. This problem deals with one-sided tests. We present the procedures for a test of $H_0 : \pi \geq \pi_0$ vs. $H_A : \pi < \pi_0$. [The same procedures would be used for $H_0 : \pi = \pi_0$ vs. $H_A : \pi < \pi_0$. For $H_0 : \pi \leq \pi_0$ vs. $H_A : \pi > \pi_0$, the procedure would be modified (see below).]

Procedure A: To construct a significance test of $H_0 : \pi \geq \pi_0$ vs. $H_a : \pi < \pi_0$ at significance level α:

(a) Let Y be binomial n, π_0, and $p = Y/n$. Find the largest c such that $P[p \leq c] \leq \alpha$.

(b) Compute the actual significance level of the test as $P[p \leq c]$.

(c) Observe p. Reject H_0 if $p \leq c$.

Procedure B: The p-value for the test if we observe p is $P[\widetilde{p} \leq p]$, where p is the *fixed* observed value and $\widetilde{p}$ equals $\widetilde{Y}/n$, where $\widetilde{Y}$ is binomial n, π_0.

(a) In Problem 6.2, let π be the probability of losing weight. (i) Find the critical value c for testing $H_0 : \pi \geq 1/2$ vs. $H_A : \pi < 1/2$ at the 10% significance level. (ii) Find the one-sided p-value for the data of Problem 6.2.

(b) Modify procedures A and B for the hypotheses $H_0 : \pi \leq \pi_0$ vs. $H_A : \pi > \pi_0$.

***6.26** Using the terminology and notation of Section 6.3.1, we consider proportions of success from two samples of size $n_{1\cdot}$ and $n_{2\cdot}$. Suppose that we are told that there are $n_{\cdot 1}$ total successes. That is, we observe the following:

	Success	**Failures**	
Sample 1	?		$n_{1\cdot}$
Sample 2			$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

If both populations are equally likely to have a success, what can we say about n_{11}, the number of successes in population 1, which goes in the cell with the question mark?

Show that

$$P[n_{11} = k] = \binom{n_{1\cdot}}{k}\binom{n_{2\cdot}}{n_{\cdot 1} - k} \bigg/ \binom{n_{\cdot\cdot}}{n_{\cdot 1}}$$

for $k \leq n_{1\cdot}$, $k \leq n_{\cdot 1}$, and $n_{\cdot 1} - k \leq n_{2\cdot}$. *Note*: $P[n_{11} = k]$, which has the parameters $n_{1\cdot}, n_{2\cdot}$, and $n_{\cdot 1}$, is called a *hypergeometric probability. Hint*: As suggested in Section 6.3.1, think of each trial (in sample 1 or 2) as a ball [purple ($n_{1\cdot}$) or gold ($n_{2\cdot}$)].

Since successes are equally likely in either population, any ball is as likely as any other to be drawn in the $n_{\cdot 1}$ successes. All subsets of size $n_{\cdot 1}$ are equally likely, so the probability of k successes is the number of subsets with k purple balls divided by the total number of subsets of size $n_{\cdot 1}$. Argue that the first number is $\binom{n_{1\cdot}}{k}\binom{n_{2\cdot}}{n_{\cdot 1}-k}$ and the second is $\binom{n_{\cdot\cdot}}{n_{\cdot 1}}$.

6.27 This problem gives more practice in finding the sample sizes needed to test for a difference in two binomial populations.

(a) Use Figure 6.2 to find *approximate* two-sided sample sizes *per group* for $\alpha = 0.05$ and $\beta = 0.10$ when (i) $P_1 = 0.5, P_2 = 0.6$; (ii) $P_1 = 0.20, P_2 = 0.10$; (iii) $P_1 = 0.70, P_2 = 0.90$.

(b) For each of the following, find one-sided sample sizes *per group* as needed from the formula of Section 6.3.3. (i) $\alpha = 0.05, \beta = 0.10, P_1 = 0.25, P_2 = 0.10$; (ii) $\alpha = 0.05, \beta = 0.05, P_1 = 0.60, P_2 = 0.50$; (iii) $\alpha = 0.01, \beta = 0.01, P_1 = 0.15, P_2 = 0.05$; (iv) $\alpha = 0.01, \beta = 0.05, P_1 = 0.85, P_2 = 0.75$. To test π_1 vs. π_2, we need the same sample size as we would to test $1 - \pi_1$ vs. $1 - \pi_2$. Why?

6.28 You are examined by an excellent screening test (sensitivity and specificity of 99%) for a rare disease (0.1% or 1/1000 of the population). Unfortunately, the test is positive. What is the probability that you have the disease?

***6.29** **(a)** Derive the rule of threes defined in Note 6.9.

(b) Can you find a similar constant to set up a 99% confidence interval?

***6.30** Consider the matched pair data of Problem 6.17: What null hypothesis does the usual chi-square test for a 2 × 2 table test on these data? What would you decide about the matching if this chi-square was not significant (e.g., the "married" table)?

REFERENCES

Beyer, W. H. (ed.) [1968]. *CRC Handbook of Tables for Probability and Statistics*. CRC Press, Cleveland, OH.

Borer, J. S., Rosing, D. R., Miller, R. H., Stark, R. M., Kent, K. M., Bacharach, S. L., Green, M. V., Lake, C. R., Cohen, H., Holmes, D., Donahue, D., Baker, W., and Epstein, S. E. [1980]. Natural history of left ventricular function during 1 year after acute myocardial infarction: comparison with clinical, electrocardiographic and biochemical determinations. *American Journal of Cardiology*, **46**: 1–12.

Box, J. F. [1978]. *R. A. Fisher: The Life of a Scientist*. Wiley, New York.

Breslow, N. E., and Day, N. E. [1980]. *Statistical Methods in Cancer Research*, Vol. 1, *The Analysis of Case–Control Studies*, IARC Publication 32. International Agency for Research in Cancer, Lyon, France.

Bucher, K. A., Patterson, A. M., Elston, R. C., Jones, C. A., and Kirkman, H. N., Jr. [1976]. Racial difference in incidence of ABO hemolytic disease. *American Journal of Public Health*, **66**: 854–858. Copyright © 1976 by the American Public Health Association.

Carver, W. A. [1927]. A genetic study of certain chlorophyll deficiencies in maize. *Genetics*, **12**: 415–440.

Cavalli-Sforza, L. L., and Bodmer, W. F. [1999]. *The Genetics of Human Populations*. Dover Publications, New York.

Chernoff, H., and Lehmann, E. L. [1954]. The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Annals of Mathematical Statistics*, **25**: 579–586.

Comstock, G. W., and Partridge, K. B. [1972]. Church attendance and health. *Journal of Chronic Diseases*, **25**: 665–672. Used with permission of Pergamon Press, Inc.

Conover, W. J. [1974]. Some reasons for not using the Yates continuity correction on 2 × 2 contingency tables (with discussion). *Journal of the American Statistical Association*, **69**: 374–382.

Edwards, A. W. F., and Fraccaro, M. [1960]. Distribution and sequences of sex in a selected sample of Swedish families. *Annals of Human Genetics, London*, **24**: 245–252.

Feigl, P. [1978]. A graphical aid for determining sample size when comparing two independent proportions. *Biometrics*, **34**: 111–122.

Fisher, L. D., and Patil, K. [1974]. Matching and unrelatedness. *American Journal of Epidemiology*, **100**: 347–349.

Fisher, R. A. [1936]. Has Mendel's work been rediscovered? *Annals of Science*, **1**: 115–137.

Fisher, R. A. [1958]. *Statistical Methods for Research Workers*, 13th ed. Oliver & Boyd, London.

Fisher, R. A., Thornton, H. G., and MacKenzie, W. A. [1922]. The accuracy of the plating method of estimating the density of bacterial populations. *Annals of Applied Biology*, **9**: 325–359.

Fleiss, J. L., Levin, B., and Park, M. C. [2003]. *Statistical Methods for Rates and Proportions*, 3rd ed. Wiley, New York.

Geissler, A. [1889]. Beiträge zur Frage des Geschlechts Verhältnisses der Geborenen. *Zeitschrift des K. Sachsischen Statistischen Bureaus*.

Graunt, J. [1662]. *Natural and Political Observations Mentioned in a Following Index and Made Upon the Bills of Mortality*. Given in part in Newman, J. R. (ed.) [1956]. *The World of Mathematics*, Vol. 3. Simon & Schuster, New York.

Grizzle, J. E. [1967]. Continuity correction in the χ^2-test for 2 × 2 tables. *American Statistician*, **21**: 28–32.

Hanley, J. A., and Lippman-Hand, A. [1983]. If nothing goes wrong, is everything alright? *Journal of the American Medical Association*, **249**: 1743–1745.

Janerich, D. T., Piper, J. M., and Glebatis, D. M. [1980]. Oral contraceptives and birth defects. *American Journal of Epidemiology*, **112**: 73–79.

Karlowski, T. R., Chalmers, T. C., Frenkel, L. D., Zapikian, A. Z., Lewis, T. L., and Lynch, J. M. [1975]. Ascorbic acid for the common cold: a prophylactic and therapeutic trial. *Journal of the American Medical Association*, **231**: 1038–1042.

Kelsey, J. L., and Hardy, R. J. [1975]. Driving of motor vehicles as a risk factor for acute herniated lumbar intervertebral disc. *American Journal of Epidemiology*, **102**: 63–73.

Kendall, M. G., and Stuart, A. [1967]. *The Advanced Theory of Statistics*, Vol. 2, *Inference and Relationship*. Hafner, New York.

Kennedy, J. W., Kaiser, G. W., Fisher, L. D., Fritz, J. K., Myers, W., Mudd, J. G., and Ryan, T. J. [1981]. Clinical and angiographic predictors of operative mortality from the collaborative study in coronary artery surgery (CASS). *Circulation*, **63**: 793–802.

Lawson, D. H., and Jick, H. [1976]. Drug prescribing in hospitals: an international comparison. *American Journal of Public Health*, **66**: 644–648.

Little, R. J. A. [1989]. Testing the equality of two independent binomial proportions. *American Statistician*, **43**: 283–288.

Mantel, N., and Greenhouse, S. W. [1968]. What is the continuity correction? *American Statistician*, **22**: 27–30.

Mantel, N., and Haenszel, W. [1959]. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, **22**: 719–748.

Mantel, N., Brown, C., and Byar, D. P. [1977]. Tests for homogeneity of effect in an epidemiologic investigation. *American Journal of Epidemiology*, **106**: 125–129.

Mendel, G. [1866]. Versuche über Pflanzenhybriden. *Verhandlungen Naturforschender Vereines in Brunn*, **10**: 1.

Meyer, M. B., Jonas, B. S., and Tonascia, J. A. [1976]. Perinatal events associated with maternal smoking during pregnancy. *American Journal of Epidemiology*, **103**: 464–476.

Miettinen, O. S. [1970]. Matching and design efficiency in retrospective studies. *American Journal of Epidemiology*, **91**: 111–118.

Odeh, R. E., Owen, D. B., Birnbaum, Z. W., and Fisher, L. D. [1977]. *Pocket Book of Statistical Tables*. Marcel Dekker, New York.

Ounsted, C. [1953]. The sex ratio in convulsive disorders with a note on single-sex sibships. *Journal of Neurology, Neurosurgery and Psychiatry*, **16**: 267–274.

Owen, D. B. [1962]. *Handbook of Statistical Tables*. Addison-Wesley, Reading, MA.

Peterson, D. R., van Belle, G., and Chinn, N. M. [1979]. Epidemiologic comparisons of the sudden infant death syndrome with other major components of infant mortality. *American Journal of Epidemiology*, **110**: 699–707.

Peterson, D. R., Chinn, N. M., and Fisher, L. D. [1980]. The sudden infant death syndrome: repetitions in families. *Journal of Pediatrics*, **97**: 265–267.

Pepe, M. S. [2003]. *The Statistical Evaluation of Medical Tests for Clarification and Prediction*. Oxford University Press, Oxford.

Remein, Q. R., and Wilkerson, H. L. C. [1961]. The efficiency of screening tests for diabetes. *Journal of Chronic Diseases*, **13**: 6–21. Used with permission of Pergamon Press, Inc.

Robinette, C. D., Silverman, C., and Jablon, S. [1980]. Effects upon health of occupational exposure to microwave radiation (radar). *American Journal of Epidemiology*, **112**: 39–53.

Rosenberg, L., Slone, D., Shapiro, S., Kaufman, D. W., Stolley, P. D., and Miettinen, O. S. [1980]. Coffee drinking and myocardial infarction in young women. *American Journal of Epidemiology*, **111**: 675–681.

Sartwell, P. E., Masi, A. T., Arthes, F. G., Greene, G. R., and Smith, H. E. [1969]. Thromboembolism and oral contraceptives: an epidemiologic case–control study. *American Journal of Epidemiology*, **90**: 365–380.

Schlesselman, J. J. [1982]. *Case–Control Studies: Design, Conduct, Analysis*. Monographs in Epidemiology and Biostatistics. Oxford University Press, New York.

Shapiro, S., Goldberg, J. D., and Hutchinson, G. B. [1974]. Lead time in breast cancer detection and implications for periodicity of screening. *American Journal of Epidemiology*, **100**: 357–366.

Smith, J. P., Delgado, G., and Rutledge, F. [1976]. Second-look operation in ovarian cancer. Cancer, **38**: 1438–1442. Used with permission from J. B. Lippincott Company.

Starmer, C. F., Grizzle, J. E., and Sen, P. K. [1974]. Comment. *Journal of the American Statistical Association*, **69**: 376–378.

U.S. Department of Health, Education, and Welfare [1964]. *Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service*. U.S. Government Printing Office, Washington, DC.

von Bortkiewicz, L. [1898]. *Das Gesetz der Kleinen Zahlen*. Teubner, Leipzig.

Weber, A., Jermini, C., and Grandjean, E. [1976]. Irritating effects on man of air pollution due to cigarette smoke. *American Journal of Public Health*, **66**: 672–676.

CHAPTER 7

Categorical Data: Contingency Tables

7.1 INTRODUCTION

In Chapter 6, *discrete variables* came up by counting the number of times that specific outcomes occurred. In looking at the presence or absence of a risk factor and a disease, *odds ratio* and *relative risk* were introduced. In doing this, we looked at the relationship between two discrete variables; each variable took on one of two possible states (i.e., risk factor present or absent and disease present or absent). In this chapter we show how to analyze more general discrete data. Two types of generality are presented.

The first generalization considers two jointly distributed discrete variables. Each variable may take on more than two possible values. Some examples of discrete variables with three or more possible values might be: smoking status (which might take on the values "never smoked," "former smoker," and "current smoker"); employment status (which could be coded as "full-time," "part-time," "unemployed," "unable to work due to medical reason," "retired," "quit," and "other"); and clinical judgment of improvement (classified into categories of "considerable improvement," "slight improvement," "no change," "slight worsening," "considerable worsening," and "death").

The second generalization allows us to consider three or more discrete variables (rather than just two) at the same time. For example, method of treatment, gender, and employment status may be analyzed jointly. With three or more variables to investigate, it becomes difficult to obtain a "feeling" for the interrelationships among the variables. If the data fit a relatively simple mathematical model, our understanding of the data may be greatly increased.

In this chapter, our first *multivariate statistical model* is encountered. The model is the *log-linear model* for multivariate discrete data. The remainder of the book depends on a variety of models for analyzing data; this chapter is an exciting, important, and challenging introduction to such models!

7.2 TWO-WAY CONTINGENCY TABLES

Let two or more discrete variables be measured on each unit in an experiment or observational study. In this chapter, methods of examining the relationship among the variables are studied. In most of the chapter we study the relationship of two discrete variables. In this case we count the number of occurrences of each pair of possibilities and enter them in a table. Such tables are called *contingency tables*. Example 7.1 presents two contingency tables.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2 Copyright © 2004 John Wiley & Sons, Inc.

Example 7.1. In 1962, Wangensteen et al., published a paper in the *Journal of the American Medical Association* advocating gastric freezing. A balloon was lowered into a subject's stomach, and coolant at a temperature of −17 to −20°C was introduced through tubing connected to the balloon. Freezing was continued for approximately 1 hour. The rationale was that gastric digestion could be interrupted and it was thought that a duodenal ulcer might heal if treatment could be continued over a period of time. The authors advanced three reasons for the interruption of gastric digestion: (1) interruption of vagal secretory responses; (2) "rendering of the central mucosa nonresponsive to food ingestion . . . "; and (3) "impairing the capacity of the parietal cells to secrete acid and the chief cells to secrete pepsin." Table 7.1 was presented as evidence for the effectiveness of gastric freezing. It shows a decrease in acid secretion.

On the basis of this table and other data, the authors state: "These data provide convincing objective evidence of significant decreases in gastric secretory responses attending effective gastric freezing" and conclude: "When profound gastric hypothermia is employed with resultant freezing of the gastric mucosa, the method becomes a useful agent in the control of many of the manifestations of peptic ulcer diathesis. Symptomatic relief is the rule, followed quite regularly by x-ray evidence of healing of duodenal ulcer craters and evidence of effective depression of gastric secretory responses." *Time* [1962] reported that "all [the patients'] ulcers healed within two to six weeks."

However, careful studies attempting to confirm the foregoing conclusion failed. Two studies in particular failed to confirm the evidence, one by Hitchcock et al. [1966], the other by Ruffin et al. [1969]. The latter study used an elaborate sham procedure (control) to simulate gastric freezing, to the extent that the tube entering the patient's mouth was cooled to the same temperature as in the actual procedure, but the coolant entering the stomach was at room temperature, so that no freezing took place. The authors defined an endpoint to have occurred if one of the following criteria was met: "perforation; ulcer pain requiring hospitalization for relief; obstruction, partial or complete, two or more weeks after hyperthermia; hemorrhage, surgery for ulcer; repeat hypothermia; or x-ray therapy to the stomach."

Several institutions cooperated in the study, and to ensure objectivity and equal numbers, random allocations to treatment and sham were balanced within groups of eight. At the termination of the study, patients were classified as in Table 7.2. The authors conclude: "The results of

Table 7.1 Gastric Response of 10 Patients with Duodenal Ulcer Whose Stomachs Were Frozen at −17 to −20°C for 1 Hour

		Average Percent Decrease in HCl after Gastric Freezing		
Patients	Patients with Decrease in Free HCl	Overnight Secretion	Peptone Stimulation	Insulin
10	10[a]	87	51	71

Source: Data from Wangensteen et al. [1962].

[a] All patients, except one, had at least a 50% decrease in free HCl in overnight secretion.

Table 7.2 Causes of Endpoints

Group	Patients	With Hemorrhage	With Operation	With Hospitalization	Not Reaching Endpoint
F (freeze)	69	9	17	9	34
S (sham)	68	9	14	7	38

Source: Data from Ruffin et al. [1969].

Table 7.3 Contingency Table for Gastric Freezing Data

	j			
i	1	2	$\cdots$	c
1	n_{11}	n_{12}	$\cdots$	n_{1c}
2	n_{21}	n_{22}	$\cdots$	n_{2c}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rc}

this study demonstrate conclusively that the 'freezing' procedure was not better than the sham in the treatment of duodenal ulcer, confirming the work of others. ... It is reasonable to assume that the relief of pain and subjective improvement reported by early investigators was probably due to the psychological effect of the procedure."

Contingency tables set up from two variables are called *two-way tables*. Let the variable corresponding to rows have r (for "row") possible outcomes, which we index by i $(i = 1, 2, \ldots, r)$. Let the variables corresponding to the column headings have c (for "column") possible states indexed by j $(j = 1, 2, \ldots, c)$. One speaks of an *$r \times c$ contingency table*. Let n_{ij} be the number of observations corresponding to the ith state of the row variable and the jth state of the column variable. In the example above, $n_{11} = 9, n_{12} = 17, n_{13} = 9, n_{14} = 34, n_{21} = 9, n_{22} = 14, n_{23} = 7$, and $n_{24} = 38$. In general, the data are presented as shown in Table 7.3. Such tables usually arise in one of two ways:

1. A sample of observations is taken. On each unit we observe the values of two traits. Let π_{ij} be the probability that the row variable takes on level i and the column variable takes on level j. Since one of the combinations must occur,

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\pi_{ij} = 1 \qquad (1)$$

2. Each row corresponds to a sample from a different population. In this case, let π_{ij} be the probability the column variable takes on state j when sampling from the ith population. Thus, for each i,

$$\sum_{j=1}^{c}\pi_{ij} = 1 \qquad (2)$$

   If the samples correspond to the column variable, the π_{ij} are the probabilities that the row variable takes on state i when sampling from population j. In this circumstance, for each j,

$$\sum_{i=1}^{r}\pi_{ij} = 1 \qquad (3)$$

Table 7.2 comes from the second model since the treatment is assigned by the experimenter; it is not a trait of the experimental unit. Examples for the first model are given below.

The usual null hypothesis in a model 1 situation is that of independence of row and column variables. That is (assuming row variable $= i$ and column variable $= j$), $P[i \text{ and } j] = P[i]P[j]$,

$$H_0\colon \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$$

In the model 2 situation, suppose that the row variable identifies the population. The usual null hypothesis is that all r populations have the same probabilities of taking on each value of the column variable. That is, for any two rows, denoted by i and i', say, and all j,

$$H_0\colon \pi_{ij} = \pi_{i'j}$$

If one of these hypotheses holds, we say that there is *no association*; otherwise, the table is said to have *association* between the categorical variables.

We will use the following notation for the sum over the elements of a row and/or column: $n_{i\cdot}$ is the sum of the elements of the ith row; $n_{\cdot j}$ is the sum of the elements of the jth column:

$$n_{i\cdot} = \sum_{j=1}^{c} n_{ij}, \qquad n_{\cdot j} = \sum_{i=1}^{r} n_{ij}, \qquad n_{\cdot\cdot} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}$$

It is shown in Note 7.1 that under either model 1 or model 2, the null hypothesis is reasonably tested by comparing n_{ij} with

$$\frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}$$

The latter is the value expected in the ijth cell given the observed marginal configuration and assuming either of the null hypotheses under model 1 or model 2. This is shown as

$n_{11} = 9$	$n_{12} = 17$	$n_{13} = 9$	$n_{14} = 34$	$n_{1\cdot} = 69$
$n_{21} = 9$	$n_{22} = 14$	$n_{23} = 7$	$n_{24} = 38$	$n_{2\cdot} = 68$
$n_{\cdot 1} = 18$	$n_{\cdot 2} = 31$	$n_{\cdot 3} = 16$	$n_{\cdot 4} = 72$	$n_{\cdot\cdot} = 137$

Under the null hypothesis, the table of expected values $n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}$ is

$69 \times 18/137$	$69 \times 31/137$	$69 \times 16/137$	$69 \times 72/137$
$68 \times 18/137$	$68 \times 31/137$	$68 \times 16/137$	$68 \times 72/137$

or

9.07	15.61	8.06	36.26
8.93	15.39	7.94	35.74

It is a remarkable fact that both null hypotheses above may be tested by the χ^2 statistic,

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot})^2}{n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}}$$

Note that n_{ij} is the observed cell entry; $n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}$ is the expected cell entry, so this statistic may be remembered as

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

For example, the array above gives

$$\begin{aligned}X^2 = {} & \frac{(9-9.07)^2}{9.07} + \frac{(17-15.61)^2}{15.61} + \frac{(9-8.06)^2}{8.06} \\ & + \frac{(34-36.26)^2}{36.26} + \frac{(9-8.93)^2}{8.93} + \frac{(14-15.39)^2}{15.39} \\ & + \frac{(7-7.94)^2}{7.94} + \frac{(38-35.76)^2}{35.76} = 0.752\end{aligned}$$

Under the null hypothesis, the X^2 statistic has approximately a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom. This approximation is for large samples and is appropriate when all of the *expected* values, $n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}$, are 5 or greater. There is some evidence to indicate that the approximation is valid if all the expected values, except possibly one, are 5 or greater.

For our example, the degrees of freedom for the example are $(2-1)(4-1) = 3$. The rejection region is for X^2 too large. The 0.05 critical value is 7.81. As $0.752 < 7.81$, we do *not* reject the null hypothesis at the 0.05 significance level.

Example 7.2. Robertson [1975] examined seat belt use in automobiles with starter interlock and buzzer/light systems. The use or nonuse of safety belts by drivers in their vehicles was observed at 138 sites in Baltimore, Maryland; Houston, Texas; Los Angeles, California; the New Jersey suburbs; New York City; Richmond, Virginia; and Washington, DC during late 1973 and early 1974. The sites were such that observers could see whether or not seat belts were being used. The sites were freeway entrances and exits, traffic-jam areas, and other points where vehicles usually slowed to less than 15 miles per hour. The observers dictated onto tape the gender, estimated age, and racial appearance of the driver of the approaching car; as the vehicles slowed alongside, the observer recorded whether or not the lap belt and/or shoulder belt was in use, not in use, or could not be seen. The license plate numbers were subsequently sent to the appropriate motor vehicle administration, where they were matched to records from which the manufacturer and year were determined. In the 1973 models, a buzzer/light system came on when the seat belt was not being used. The buzzer was activated for at least 1 minute when the driver's seat was occupied, the ignition switch was on, the transmission gear selector was in a forward position, and the driver's lap belt was not extended at least 4 inches from its normal resting position. Effective on August 15, 1973, a federal standard required that the automobile could be started only under certain conditions. In this case, when the driver was seated, the belts had to be extended more than 4 inches from their normally stored position and/or latched. Robertson states that as a result of the strong negative public reaction to the interlock system, federal law has banned the interlock system. Data on the buzzer/light-equipped models and interlock-equipped models are given in Table 7.4. As can be seen from the table, column percentages were presented to aid assimilation of the information in the table.

Table 7.4 Robertson [1975] Seat Belt Data

	1973 Models (Buzzer/Light)		1974 Models (Interlock)		
Belt Use	%	Number	%	Number	Total
Lap and shoulder	7	432	48	1007	1439
Lap only	21	1262	11	227	1489
None	72	4257	41	867	5124
Total	100	5951	100	2101	8052

Percentages in two-way contingency tables are useful in aiding visual comprehension of the contents. There are three types of percent tables:

1. *Column percent tables* give the percentages for each column (the columns add to 100%, except possibly for rounding errors). This is best for comparing the distributions of different columns.
2. *Row percent tables* give the percentages for each row (the rows add to 100%). This is best for comparing the distributions of different rows.
3. The *total percent table* gives percentages, so that all the entries in a table add to 100%. This aids investigation of the proportions in each combination.

The column percentages in Table 7.4 facilitate comparison of seat belt use in the 1973 buzzer/light models and the 1974 interlock models. They illustrate that there are strategies for getting around the interlock system, such as disabling it, connecting the seat belt and leaving it connected on the seat, as well as possible other strategies, so that even with an interlock system, not everyone uses it. The computed value of the chi-square statistic for this table is 1751.6 with two degrees of freedom. The p-value is effectively zero, as shown in Table A.3 in the Appendix.

Given that we have a statistically significant association, the next question that arises is: To what may we attribute this association? To determine why the association occurs, it is useful to have an idea of which entries in the table differ more than would be expected by chance from their value under the null hypothesis of no association. Under the null hypothesis, for each entry in the table, the following *adjusted residual value* is approximately distributed as a standard normal distribution. The term *residual* is used since it looks at the difference between the observed value and the value expected under the null hypothesis. This difference is then standardized by its standard error,

$$z_{ij} = \frac{n_{ij} - (n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot})}{\sqrt{n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}\,(1 - n_{i\cdot}/n_{\cdot\cdot})\left(1 - n_{\cdot j}/n_{\cdot\cdot}\right)}} \tag{4}$$

For example, for the (1, 1) entry in the table, a standardized residual, is given by

$$\frac{(432 - 1439 \times 5951/8052)}{\sqrt{\dfrac{1439(5951)}{8052}\left(1 - \dfrac{1439}{8052}\right)\left(1 - \dfrac{5951}{8052}\right)}} = 41.83$$

The matrix of the residual values observed with the corresponding normal probability p-values is given in Table 7.5. Note that the values add to zero for the residuals across each row. This occurs because there are only two columns. The adjusted residual values observed are so far from zero that the normal p-values are miniscule.

In general, there is a problem in looking at a contingency table with many cells. Because there are a large number of residual values in the table, it may be that one or more of them differs by chance from zero at the 5% significance level. Even *under the null hypothesis*, because of the many possibilities examined, *this would occur much more than 5% of the time*. One conservative way to deal with this problem is to multiply the p-values by the number of rows minus one and the number of columns minus one. If the corresponding p-value is less than 0.05, one can conclude that the entry is different from that expected by the null hypothesis at the 5% significance level *even after looking at all of the different entries*. (This problem of looking at many possibilities, called the *multiple comparison problem*, is dealt with in considerable detail in Chapter 12.) For this example, even after multiplying by the number of rows minus one and the number of columns minus one, all of the entries differ from those expected under the null hypothesis. Thus, one can conclude, using the sign of the residual to tell us whether the

Table 7.5 Adjusted Residual Values (Example 7.2)

i	j	Residual (Z_{ij})	p-value	p-value $\times (r-1) \times (c-1)$
1	1	−41.83	0+	0+
1	2	41.83	0+	0+
2	1	10.56	3×10^{-22}	6×10^{-22}
2	2	−10.56	3×10^{-22}	6×10^{-22}
3	1	24.79	9×10^{-53}	2×10^{-52}
3	2	−24.79	9×10^{-53}	2×10^{-52}

percentage is too high or too low, that in the 1973 models there is less lap and shoulder belt use than in the 1974 models. Further, if we look at the "none" category, there are fewer people without any belt use in the 1974 interlock models than in the 1973 buzzer/light-equipped models. One would conclude that the interlock system, although a system disliked by the public, was successful as a public health measure in increasing the amount of seat belt use.

Suppose that we decide there is an association in a contingency table. We can interpret the table by using residuals (as we have done above) to help to find out whether particular entries differ more than expected by chance. Another approach to interpretation is to characterize numerically the amount of association between the variables, or proportions in different groups, in the contingency table. To date, no single measure of the amount of association in contingency tables has gained widespread acceptance. There have been many proposals, all of which have some merit. Note 7.2 presents some measures of the amount of association.

7.3 CHI-SQUARE TEST FOR TREND IN 2 × k TABLES

There are a variety of techniques for improving the statistical power of χ^2 tests. Recall that power is a function of the alternative hypothesis. One weakness of the chi-square test is that it is an "omnibus" test; it tests for independence vs. dependence without specifying the nature of the latter. In some cases, a small subset of alternative hypotheses may be specified to increase the power of the chi-square test by defining a special test. One such situation occurs in $2 \times k$ tables when the alternative hypothesis is that there is an ordering in the variable producing the k categories. For example, exposure categories can be ordered, and the alternative hypothesis may be that the probability of disease *increases* with increasing exposure.

In this case the row variable takes on one of two states (say + or − for definiteness). For each state of the column variable ($j = 1, 2, \ldots, k$), let π_j be the conditional probability of a positive response. The test for trend is designed to have statistical power against the alternatives:

$$H_1 : \pi_1 \leq \pi_2 \leq \cdots \leq \pi_k, \qquad \text{with at least one strict inequality}$$

$$H_2 : \pi_1 \geq \pi_2 \geq \cdots \geq \pi_k, \qquad \text{with at least one strict inequality}$$

That is, the alternatives of interest are that the proportion of + responses increases or decreases with the column variable. For these alternatives to be of interest, the column variable will have a "natural" ordering. To compute the statistic, a score needs to be assigned to each state j of the column variable. The scores x_j are assigned so that they increase or decrease. Often, the x_j are consecutive integers. The data are laid out as shown in Table 7.6.

Table 7.6 Scores Assigned to State *j*

	j				
i	1	2	$\cdots$	k	Total
1+	n_{11}	n_{12}	$\cdots$	n_{1k}	$n_{1\cdot}$
2−	n_{21}	n_{22}	$\cdots$	n_{2k}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot k}$	$n_{\cdot\cdot}$
Score	x_1	x_2	$\cdots$	x_k	

Before stating the test, we define some notation. Let

$$[n_1 x] = \sum_{j=1}^{k} n_{1j} x_j - \frac{n_{1\cdot} \sum n_{\cdot j} x_j}{n_{\cdot\cdot}}$$

and

$$[x^2] = \sum_{j=1}^{k} n_{\cdot j} x_j^2 - \frac{(\sum n_{\cdot j} x_j)^2}{n_{\cdot\cdot}}$$

and

$$p = \frac{n_{1\cdot}}{n_{\cdot\cdot}}$$

Then the chi-square test for trend is defined to be

$$X^2_{\text{trend}} = \frac{[n_1 x]^2}{[x^2] p(1-p)}$$

and when there is no association, this quantity has approximately a chi-square distribution with one degree of freedom. [In the terminology of Chapter 9, this is a chi-square test for the slope of a weighted regression line with dependent variable $p_j = n_{1j}/n_{\cdot j}$, predictor variable x_j, and weights $n_{1j}/p(1-p)$, where $j = 1, 2, \ldots, k$.]

Example 7.3. For an example of this test, we use data of Maki et al. [1977], relating risk of catheter-related infection to the duration of catheterization. An infection was considered to be present if there were 15 or more colonies of microorganisms present in a culture associated with the withdrawn catheter. A part of the data dealing with the number of positive cultures as related to duration of catheterization is given in Table 7.7. A somewhat natural set of values of the scores x_i is the duration of catheterization in days. The designation ≥ 4 is, somewhat arbitrarily, scored 4.

Before carrying out the analysis, note that a graph of the proportion of positive cultures vs. duration such as in the one shown in Figure 7.1 clearly suggests a trend. The general chi-square test on the 2 × 4 table produces a value of $X^2 = 6.99$ with three degrees of freedom and a significance level of 0.072.

Table 7.7 Relations of Results of Semiquantitative Culture and Catheterization

	Duration of Catheterization (days)				
Culture	1	2	3	≥ 4	Total
Positive[a]	1[b]	5	5	14	25
Negative	46	64	39	76	225
Total	47	69	44	90	250

Source: Data from Maki et al. [1977].

[a] Culture is positive if 15 or more colonies on the primary plate.

[b] Numbers in the body of the table are the numbers of catheters.

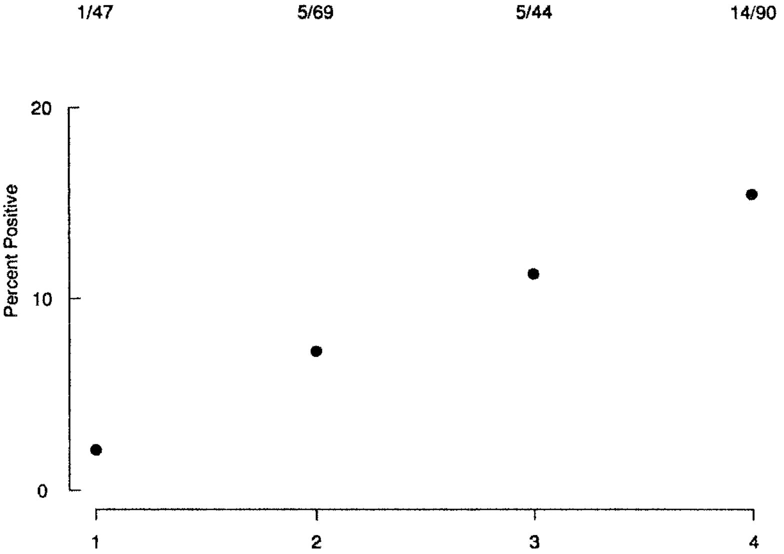


Figure 7.1 Graph of percentage of cultures positive vs. duration of catheterization. The fractions 1/47, etc., are the number of positive cultures to the total number of cultures for a particular day. (Data from Maki et al. [1977]; see Table 7.7.)

To calculate the chi-square test for trend, we calculate the quantities $[n_1x]$, $[x^2]$, and p as defined above.

$$[n_1x] = 82 - \frac{(25)(677)}{250} = 14.3$$

$$[x^2] = 2159 - \frac{677^2}{250} \doteq 325.6840$$

$$p = \frac{25}{250} = 0.1, \qquad (1-p) = 0.9$$

$$X^2_{\text{trend}} = \frac{[n_1x]^2}{[x^2]p(1-p)} \doteq \frac{14.3^2}{325.6840(0.1)(0.9)} \doteq 6.98$$

This statistic has one degree of freedom associated with it, and from the chi-square Table A.3, it can be seen that $0.005 < p < 0.01$; hence there is a significant linear trend.

Note two things about the chi-square test for trend. First, the degrees of freedom are one, *regardless* of how large the value k. Second, the values of the scores chosen (x_j) are not too crucial, and evenly spaced scores will give more statistical power against a trend than will the usual χ^2 test. The example above indicates one type of contingency table in which ordering is clear: when the categories result from grouping a continuous variable.

7.4 KAPPA: MEASURING AGREEMENT

It often happens in measuring or categorizing objects that the variability of the measurement or categorization is investigated. For example, one might have two physicians independently judge a patient's status as "improved," "remained the same," or "worsened." A study of psychiatric patients might have two psychiatrists independently classifying patients into diagnostic categories. When we have two discrete classifications of the same object, we may put the entries into a two-way *square* $(r = c)$ contingency table. The chi-square test of this chapter may then be used to test for association. Usually, when two measurements are taken of the same objects, there is not much trouble showing association; rather, the concern is to study the degree or amount of agreement in the association. This section deals with a statistic, *kappa* (κ), designed for such situations. We will see that the statistic has a nice interpretation; the value of the statistic can be taken as a measure of the degree of agreement. As we develop this statistic, we shall illustrate it with the following example.

Example 7.4. Fisher et al. [1982] studied the reproducibility of coronary arteriography. In the coronary artery surgery study (CASS), coronary arteriography is the key diagnostic procedure. In this procedure, a tube is inserted into the heart and fluid injected that is opaque to x-rays. By taking x-ray motion pictures, the coronary arteries may be examined for possible narrowing, or *stenosis*. The three major arterial systems of the heart were judged with respect to narrowing. Narrowing was significant if it was 70% or more of the diameter of the artery. Because the angiographic films are a key diagnostic tool and are important in the decision about the appropriateness of bypass surgery, the quality of the arteriography was monitored and the amount of agreement was ascertained.

Table 7.8 presents the results for randomly selected films with two readings. One reading was that of the patient's clinical site and was used for therapeutic decisions. The angiographic film was then sent to another clinical site designated as a quality control site. The quality control site read the films blindly, that is, without knowledge of the clinical site's reading. From these readings, the amount of disease was classified as "none" (entirely normal), "zero-vessel disease but some disease," and one-, two-, and three-vessel disease.

We wish to study the amount of agreement. One possible measure of this is the proportion of the pairs of readings that are the same. This quantity is estimated by adding up the numbers

Table 7.8 Agreement with Respect to Number of Diseased Vessels

Quality Control	Clinical Site Reading					
Site Reading	Normal	Some	One	Two	Three	Total
Normal	13	8	1	0	0	22
Some	6	43	19	4	5	77
One	1	9	155	54	24	243
Two	0	2	18	162	68	250
Three	0	0	11	27	240	278
Total	20	62	204	247	337	870

on the diagonal of the table; those are the numbers where both the clinical and quality control sites read the same quantity. In such a situation, the contingency table will be square. Let r be the number of categories (in the table of this example, $r = 5$). The proportion of cases with agreement is given by

$$P_A = \frac{n_{11} + n_{22} + \cdots + n_{rr}}{n_{..}} = \sum_{i=1}^{r} \frac{n_{ii}}{n_{..}}$$

For this table, the proportion with agreement is given by $P_A = (13{+}43{+}155{+}162{+}240)/870 = 613/870 \doteq 0.7046$.

The proportion of agreement is limited because it is determined heavily by the proportions of people in the various categories. Consider, for example, a situation where each of two judges places 90% of the measurements in one category and 10% in the second category, such as in the following array:

81	9	90
9	1	10
90	10	100

Here there is no association whatsoever between the two measurements. In fact, the chi-square value is precisely zero by design; there is no more agreement between the patients than that expected by chance. Nevertheless, because both judges have a large proportion of the cases in the first category, in 82% of the cases there is agreement; that is, $P_A = 0.82$. We have a paradox: On the one hand, the agreement seems good (there is an agreement 82% of the time); on the other hand, the agreement is no more than can be expected by chance. To have a more useful measure of the amount of agreement, the *kappa statistic* was developed to adjust for the amount of agreement that one expects purely by chance.

If one knows the totals of the different rows and columns, the proportion of observations expected to agree by chance is given by the following equation:

$$P_C = \frac{n_{1\cdot}n_{\cdot 1} + \cdots + n_{r\cdot}n_{\cdot r}}{n_{..}^2} = \sum_{i=1}^{r} \frac{n_{i\cdot}n_{\cdot i}}{n_{..}^2}$$

For the angiography example, the proportion of agreement expected by chance is given by

$$P_C = \frac{22 \times 20 + 77 \times 62 + 243 \times 204 + 250 \times 247 + 278 \times 337}{870^2} \doteq 0.2777$$

The kappa statistic uses the fact that the best possible agreement is 1 and that, by chance, one expects an agreement P_C. A reasonable measure of the amount of agreement is the proportion of difference between 1 and P_C that can be accounted for by actual observed agreement. That is, kappa is the ratio of the agreement actually observed minus the agreement expected by chance, divided by 1 (which corresponds to perfect agreement), minus the agreement expected by chance:

$$\kappa = \frac{P_A - P_C}{1 - P_C}$$

For our example, the computed value of kappa is

$$\kappa = \frac{0.7046 - 0.2777}{1 - 0.2777} \doteq 0.59$$

The kappa statistic runs from $-P_C/(1 - P_C)$ to 1. If the agreement is totally by chance, the expected value is zero. Kappa is equal to 1 if and only if there is complete agreement between the two categorizations [Cohen, 1968; Fleiss, 1981].

Since the kappa statistic is generally used where it is clear that there will be statistically significant agreement, the real issue is the amount of agreement. κ is a measure of the amount of agreement. In our example, one can state that 59% of the difference between perfect agreement and the agreement expected by chance is accounted for by the agreement between the clinical and quality control reading sites.

Now that we have a parameter to measure the amount of agreement, we need to consider the effect of the sample size. For small samples, the estimation of κ will be quite variable; for larger samples it should be quite good. For relatively large samples, when there is no association, the variance of the estimate is estimated as follows:

$$\text{var}_0(\kappa) = \frac{P_C + P_C^2 - \sum_{i=1}^{r} (n_{i\cdot}^2 n_{\cdot i} + n_{i\cdot} n_{\cdot i}^2)/n_{\cdot\cdot}^3}{n_{\cdot\cdot}(1 - P_C)^2}$$

The subscript on $\text{var}_0(\kappa)$ indicates that it is the variance under the null hypothesis. The standard error of the estimate is the square root of this quantity. κ divided by the standard error is approximately a standard normal variable when there is no association between the quantities. This may be used as a statistical test for association in lieu of the chi-square test [Fleiss et al., 1969].

A more useful function of the general standard error is construction of a confidence interval for the true κ. A $100(1-\alpha)\%$ confidence interval for the population value of κ for large samples is given by

$$(\kappa - z_{1-\alpha/2}\sqrt{\text{var}(\kappa)}, \kappa + z_{1-\alpha/2}\sqrt{\text{var}(\kappa)})$$

The estimated standard error, allowing for association, is the square root of

$$\text{var}(\kappa) = \frac{\sum_{i=1} \frac{n_{ii}}{n_{\cdot\cdot}} \left[1 - \left(\frac{n_{i\cdot} + n_{\cdot i}}{n_{\cdot\cdot}}\right)(1-\kappa)\right]^2 + \sum_{i \neq j} \sum \frac{n_{ij}}{n_{\cdot\cdot}} \left[\left(\frac{n_{\cdot i} + n_{j\cdot}}{n_{\cdot\cdot}}\right)(1-\kappa)\right]^2 - [\kappa - P_C(1-\kappa)]^2}{n_{\cdot\cdot}(1 - P_C)^2}$$

For our particular example, the estimated variance of κ is

$$\text{var}(\kappa) = 0.000449$$

The standard error of κ is approximately 0.0212. The 95% confidence interval is

$$(0.57 - 1.96 \times 0.0212, 0.57 + 1.96 \times 0.0212) \doteq (0.55, 0.63)$$

A very comprehensive discussion of the use of κ in medical research can be found in Kraemer et al. [2002], and a discussion in the context of other ways to measure agreement is given by Nelson and Pepe [2000].

The kappa statistic has drawbacks. First, as indicated, the small sample variance is quite complicated. Second, while the statistic is supposed to adjust for marginal agreement is does not really do so (see, e.g., Agresti [2002, p. 453]). Third, κ ignores the ordering of the categories (see Maclure and Willett [1987]). Finally, it is difficult to embed κ in a statistical model: as, for example, a function of the odds ratio or correlation coefficient. Be sure to consider alternatives to kappa when measuring agreement; for example, the odds ratio and logistic regression as in Chapter 6 or the log-linear models discussed in the next section.

*7.5 LOG-LINEAR MODELS

For the first time we will examine statistical methods that deal with more than two variables at one time. Such methods are important for the following reasons: In one dimension, we have been able to summarize data with the normal distribution and its two parameters, the mean and the variance, or equivalently, the mean and the standard deviation. Even when the data did not appear normally distributed, we could get a feeling for our data by histograms and other graphical methods in one dimension. When we observe two numbers at the same time, or are working with two-dimensional data, we can plot the points and examine the data visually. (This is discussed further in Chapter 9. Even in the case of two variables, we shall see that it is useful to have models summarizing the data.) When we move to three variables, however, it is much harder to get a "feeling" for the data. Possibly, in three dimensions, we could construct visual methods of examining the data, although this would be difficult. With more than three variables, such physical plots cannot be obtained; although mathematicians may think of space and time as being a four-dimensional space, we, living in a three-dimensional world, cannot readily grasp what the points mean. In this case it becomes very important to simplify our understanding of the data by fitting a model to the data. *If* the model fits, it may summarize the complex situation very succinctly. In addition, the model may point out relationships that may reasonably be understood in a simple way. The fitting of probability models or distributions to many variables at one time is an important topic.

The models are necessarily mathematically complex; thus, the reader needs discipline and perseverance to work through and understand the methods. It is a very worthwhile task. Such methods are especially useful in the analysis of observational biomedical data. We now proceed to our first model for multiple variables, the log-linear model.

Before beginning the details of the actual model, we define some terms that we will be using. The models we investigate are for *multivariate categorical data*. We already know the meaning of *categorical data*: values of a variable or variables that put subjects into one of a finite number of categories. The term *multivariate* comes from the prefix *multi*-, meaning "many," and *variate*, referring to variables; the term refers to multiple variables at one time.

Definition 7.1. *Multivariate data* are data for which each observation consists of values for more than one random variable on each experimental unit. *Multivariate statistical analysis* consists of data analysis of multivariate data.

The majority of data collected are, in fact, multivariate data. If one measures systolic and diastolic blood pressure on each subject, there are two variables—thus, multivariate data. If we administer a questionnaire on the specifics of brushing teeth, flossing, and so on, the response of a person to each question is a separate variable, and thus one has multivariate data. Strictly speaking, some of the two-way contingency table data we have looked at are multivariate data since they cross-classify by two variables. On the other hand, tables that arose from looking at one quantity in different subgroups are not multivariate when the group was not observed on experimental units picked from a population but was part of a data collection or experimental procedure.

Additional terminology is included in the term *log-linear models*. We already have an idea of the meaning of a model. Let us consider the two terms *log* and *linear*. The logarithm was discussed in connection with the likelihood ratio chi-square statistics. (In this section, and indeed throughout this book, the logarithm will be to the base e.) Recall, briefly some of the properties of the logarithm. Of most importance to us is that the log of the product of terms is the sum of the individual logs. For example, if we have three numbers, a, b, and c (all positive), then

$$\ln(abc) = \ln a + \ln b + \ln c$$

Here, "ln" represents the *natural logarithm*, the log to the base e. Recall that by the definition of natural log, if one exponentiates the logarithm—that is, takes the number e to the power represented by the logarithm—one gets the original number back:

$$e^{\ln a} = a$$

Inexpensive hand calculators compute both the logarithm and the exponential of a number. If you are rusty with such manipulations, Problem 7.24 will give you practice in the use of logarithms and exponentials.

The second term we have used is the term *linear*. It is associated with a straight line or a linear relationship. For two variables x and y, y is a linear function of x if $y = a + bx$, where a and b are constants. For three variables, x, y, and z, z is a linear function of x and y if $z = a + bx + cy$, where a, b, and c are constant. In general, in a linear relationship, one *adds* a constant multiple for each of the variables involved. The linear models we use will look like the following: Let

$$g_{ij}^{IJ}$$

be the logarithm of the probability that an observation falls into the ijth cell in the two-dimensional contingency table. Let there be I rows and J columns. One possible model would be

$$g_{ij}^{IJ} = u + u_i^I + u_j^J$$

(For more detail on why the term *linear* is used for such models, see Note 7.4.)

We first consider the case of two-way tables. Suppose that we want to fit a model for independence. We know that independence in terms of the cell probabilities π_{ij} is equivalent to the following equation:

$$\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$$

If we take logarithms of this equation and use the notation g_{ij} for the natural log of the cell probability, the following results:

$$g_{ij} = \ln \pi_{ij} = \ln \pi_{i\cdot} + \ln \pi_{\cdot j}$$

When we denote the natural logs of $\pi_{i\cdot}$ and $\pi_{\cdot j}$ by the quantities h_i^I and h_j^J, we have

$$g_{ij} = h_i^I + h_j^J$$

The quantities h_i^I and h_j^J are not all independent. They come from the marginal probabilities for the I row variables and the J column variables. For example, the h_i^I's satisfy the equation

$$e^{h_1^I} + e^{h_2^I} + \cdots + e^{h_I^I} = 1$$

This equation is rather awkward and unwieldy to work with; in particular, given $I - 1$ of the h_i's, determination of the other coefficient takes a bit of work. It is possible to choose a different normalization of the parameters if we add a constant. Rewrite the equation above as follows:

$$g_{ij} = \left(\sum_{i'=1}^{I} \frac{h_{i'}^I}{I}\right) + \left(\sum_{j'=1}^{J} \frac{h_{j'}^J}{J}\right) + \left(h_i^I - \sum_{i'=1}^{I} \frac{h_{i'}^I}{I}\right) + \left(h_j^J - \sum_{j'=1}^{J} \frac{h_{j'}^J}{J}\right)$$

The two quantities in parentheses farthest to the right both add to zero when we sum over the indices i and j, respectively. In fact, that is why those terms were added and subtracted. Thus, we can rewrite the equation for g_{ij} as follows:

$$g_{ij} = u + u_i^I + u_j^J, \qquad i = 1, \dots, I, \quad j = 1, \dots, J$$

where

$$\sum_{i=1}^{I} u_i^I = 0, \ \sum_{j=1}^{J} u_j^J = 0$$

It is easier to work with this normalization. Note that this is a linear model for the log of the cell probability π_{ij}; that is, this is a log-linear model.

Recall that estimates for the $\pi_{i\cdot}$ and $\pi_{\cdot j}$ were $n_{i\cdot}/n_{\cdot\cdot}$ and $n_{\cdot j}/n_{\cdot\cdot}$, respectively. If one follows through all of the mathematics involved, estimates for the parameters in the log-linear model result. At this point, we shall slightly abuse our notation by using the same notation for both the population parameter values and the estimated parameter values from the sample at hand. The estimates are

$$u = \frac{1}{I}\sum_{i=1}^{I} \ln \frac{n_{i\cdot}}{n_{\cdot\cdot}} + \frac{1}{J}\sum_{j=1}^{J} \ln \frac{n_{\cdot j}}{n_{\cdot\cdot}}$$

$$u_i^I = \ln \frac{n_{i\cdot}}{n_{\cdot\cdot}} - \frac{1}{I}\sum_{i'=1}^{I} \ln \frac{n_{i'\cdot}}{n_{\cdot\cdot}}$$

$$u_j^J = \ln \frac{n_{\cdot j}}{n_{\cdot\cdot}} - \frac{1}{J}\sum_{j'=1}^{I} \ln \frac{n_{\cdot j'}}{n_{\cdot\cdot}}$$

From these estimates we get fitted values for the number of observations in each cell. This is done as follows: By inserting the estimated parameters from the log-linear model and then taking the exponential, we have an estimate of the probability that an observation falls into the ijth cell. Multiplying this by $n_{\cdot\cdot}$, we have an estimate of the number of observations we should see in the cell if the model is correct. In this particular case, the fitted value for the ijth cell turns out to be the expected value from the chi-square test presented earlier in this chapter, that is, $n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}$.

Let us illustrate these complex formulas by finding the estimates for one of the examples above.

Example 7.1. (*continued*) We know that for the 2 × 4 table, we have the following values:

$$n_{\cdot 1} = 18, \quad n_{\cdot 2} = 31, \quad n_{\cdot 3} = 16, \quad n_{\cdot 4} = 72, \quad n_{1\cdot} = 69, \quad n_{2\cdot} = 68, \quad n_{\cdot\cdot} = 137$$

$$\ln(n_{1\cdot}/n_{\cdot\cdot}) \doteq -0.6859, \qquad \ln(n_{2\cdot}/n_{\cdot\cdot}) \doteq -0.7005$$

$$\ln(n_{\cdot 1}/n_{\cdot\cdot}) \doteq -2.0296, \qquad \ln(n_{\cdot 2}/n_{\cdot\cdot}) \doteq -1.4860$$

$$\ln(n_{\cdot 3}/n_{\cdot\cdot}) \doteq -2.1474, \qquad \ln(n_{\cdot 4}/n_{\cdot\cdot}) \doteq -0.6433$$

With these numbers, we may compute the parameters for the log-linear model. They are

$$u \doteq \frac{-0.6859 - 0.7005}{2} + \frac{-2.0296 - 1.4860 - 2.1474 - 0.6433}{4}$$

$$\doteq -0.6932 - 1.5766 = -2.2698$$

$$u_1^J \doteq -2.0296 - (-1.5766) \doteq -0.4530$$

$$u_1^I \doteq -0.6859 - (-0.6932) \doteq 0.0073 \qquad u_2^J \doteq -1.4860 - (-1.5766) \doteq 0.0906$$

$$u_2^I \doteq -0.7004 - (-0.6932) \doteq -0.0073 \qquad u_3^J \doteq -2.1474 - (-1.5766) \doteq -0.5708$$

$$u_4^J \doteq -0.6433 - (-1.5766) \doteq 0.9333$$

The larger the value of the coefficient, the larger will be the cell probability. For example, looking at the two values indexed by i, the second state having a minus sign will lead to a slightly smaller contribution to the cell probability than the term with the plus sign. (This is also clear from the marginal probabilities, which are 68/137 and 69/137.) The small magnitude of the term means that the difference between the two I state values has very little effect on the cell probability. We see that of all the contributions for the j variable values, $j = 4$ has the biggest effect, 1 and 3 have fairly large effects (tending to make the cell probability small), while 2 is intermediate.

The chi-square goodness of fit and the likelihood ratio chi-square statistics that may be applied to this setting are

$$X^2 = \sum \frac{(\text{observed} - \text{fitted})^2}{\text{fitted}}$$

$$\text{LRX}^2 = 2\sum \left(\text{observed} \ln \frac{\text{observed}}{\text{fitted}}\right)$$

Finally, if the model for independence does not hold, we may add more parameters. We can find a log-linear model that will fit any possible pattern of cell probabilities. The equation for the log of the cell probabilities is given by the following:

$$g_{ij} = u + u_i^I + u_j^J + u_{ij}^{IJ}, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, J$$

where

$$\sum_{i=1}^{I} u_i^I = 0, \qquad \sum_{j=1}^{J} u_j^J = 0, \qquad \sum_{i=1}^{I} u_{ij}^{IJ} = 0, \qquad \sum_{j=1}^{J} u_{ij}^{IJ} = 0$$

It seems rather paradoxical, or at least needlessly confusing, to take a value indexed by i and j and to set it equal to the sum of four values, including some indexed by i and j; the right-hand side is much more complex than the left-hand side. The reason for doing this is that, usually, the full (or saturated) model, which can give any possible pattern of cell probabilities, is not desirable. It is hoped during the modeling effort that the data will allow a simpler model, which would allow a simpler interpretation of the data. In the case at hand, we examine the possibility of the simpler interpretation that the two variables are independent. If they are not, the particular model is not too useful.

Note two properties of the fitted values. First, in order to fit the independence model, where each term depends on at most one factor or one variable, we only needed to know the marginal values of the frequencies, the $n_{i\cdot}$ and $n_{\cdot j}$. We did not need to know the complete distribution of the frequencies to find our fitted values. Second, when we had fit values to the frequency table, the fitted values summed to the marginal value used in the estimation; that is, if we sum across i or j, the sum of the expected values is equal to the sum actually observed.

At this point it seems that we have needlessly confused a relatively easy matter: the analysis of two-way contingency tables. If only two-way contingency tables were involved, this would be a telling criticism; however, the strength of log-linear models appears when we have more than two cross-classified categorical variables. We shall now discuss the situation for three cross-classified categorical variables. The analyses may be extended to any number of variables, but such extensions are not done in this book.

Suppose that the three variables are labeled X, Y, and Z, where the index i is used for the X variable, j for the Y variable, and k for the Z variable. (This is to say that X will take values $1, \dots, I$, Y will take on $1, \dots, J$, and so on.) The methods of this section are illustrated by the following example.

Example 7.5. The study of Weiner et al. [1979] is used in this example. The study involves exercise treadmill tests for men and women. Among men with chest pain thought probably to be angina, a three-way classification of the data is as follows: One variable looks at the resting electrocardiogram and tells whether or not certain parts of the electrocardiogram (the ST- and T-waves) are normal or abnormal. Thus, $J = 2$. A second variable considers whether or not the exercise test was positive or negative ($I = 2$). A positive exercise test shows evidence of an ischemic response (i.e., lack of appropriate oxygen to the heart muscles for the effort being exerted). A positive test is thought to be an indicator of coronary artery disease. The third variable was an evaluation of the coronary artery disease as determined by coronary arteriography. The disease is classified as normal or minimal disease, called zero-vessel disease, one-vessel disease, and multiple-vessel disease ($K = 3$). The data are presented in Table 7.9.

The most general log-linear model for the three factors is given by the following extension of the two-factor work:

$$g_{ijk} = u + u_i^I + u_j^J + u_k^K + u_{ij}^{IJ} + u_{ik}^{IK} + u_{jk}^{JK} + u_{ijk}^{IJK}$$

where

$$\sum_{i=1}^{I} u_i^I = \sum_{j=1}^{J} u_j^J = \sum_{k=1}^{K} u_k^K = 0$$

$$\sum_{i=1}^{I} u_{ij}^{IJ} = \sum_{j=1}^{J} u_{ij}^{IJ} = \sum_{i=1}^{I} u_{ik}^{IK} = \sum_{k=1}^{K} u_{ik}^{IK} = \sum_{j=1}^{J} u_{jk}^{JK} = \sum_{k=1}^{K} u_{jk}^{JK} = 0$$

$$\sum_{i=1}^{I} u_{ijk}^{IJK} = \sum_{j=1}^{J} u_{ijk}^{IJK} = \sum_{k=1}^{K} u_{ijk}^{IJK} = 0$$

Table 7.9 Exercise Test Data

		Number of Vessels Diseased (K)		
Exercise Test Response (I)	Resting Electrocardiogram ST- and T-Waves (J)	0 ($k = 1$)	1 ($k = 2$)	2 or 3 ($k = 3$)
+	Normal ($j = 1$)	30	64	147
($i = 1$)	Abnormal ($j = 2$)	17	22	80
−	Normal ($j = 1$)	118	46	38
($i = 2$)	Abnormal ($j = 2$)	14	7	11

Source: Weiner et al. [1979].

In other words, there is a u term for every possible combination of the variables, including no variables at all. For each term involving one or more variables, if we sum over any one variable, the sum is equal to zero. The term involving I, J, and K is called a *three-factor term*, or a *second-order interaction term*; in general, if a coefficient involves M variables, it is called an *M-factor term* or an $(M - 1)$th-*order interaction term*.

With this notation we may now formulate a variety of simpler models for our three-way contingency table. For example, the model might be any one of the following simpler models:

$$H_1 : g_{ijk} = u + u_i^I + u_j^J + u_k^K$$

$$H_2 : g_{ijk} = u + u_i^I + u_j^J + u_k^K + u_{ij}^{IJ}$$

$$H_3 : g_{ijk} = u + u_i^I + u_j^J + u_k^K + u_{ij}^{IJ} + u_{ik}^{IK} + u_{jk}^{JK}$$

The notation has become so formidable that it is useful to introduce a shorthand notation for the hypotheses. One or more capitalized indices contained in brackets will indicate a hypothesis where the terms involving that particular set of indices as well as any terms involving subsets of the indices are to be included in the model. Any terms not specified in this form are assumed not to be in the model. For example,

$$[IJ] \longrightarrow u + u_i^I + u_j^J + u_{ij}^{IJ}$$

$$[K] \longrightarrow u + u_k^K$$

$$[IJK] \longrightarrow u + u_i^I + u_j^J + u_k^K + u_{ij}^{IJ} + u_{ik}^{IK} + u_{jk}^{JK} + u_{ijk}^{IJK}$$

The formulation of the three hypotheses given above in this notation would be simplified as follows:

$$H_1 : [I][J][K]$$

$$H_2 : [IJ][K]$$

$$H_3 : [IJ][IK][JK]$$

This notation describes a *hierarchical hypothesis*; that is, if we have two factor terms containing, say, variables I and J, we also have the one-factor terms for the same variables. The hypothesis would not be written $[IJ][I][J]$, for example, because the last two parts would be redundant, as already implied by the first. Using this bracket notation for the three-factor model, there are eight possible hypotheses of interest. All except the most complex one have a simple interpretation in terms of the probability relationships among the factors X, Y, and Z. This is given in Table 7.10.

Hypotheses 5, 6, and 7 are of particular interest. Take, for example, hypothesis 5. This hypothesis states that if you take into account the X variable, there is no association between Y and Z. In particular, if one only looks at the two-way table of Y and Z, an association may be seen, because in fact they are associated. However, if hypothesis 5 holds, one could then conclude that the association is due to interaction with the variable X and could be "explained away" by taking into account the values of X.

There is a relationship between hypotheses involving the bracket notation and the corresponding tables that one gets from the higher-dimensional contingency table. For example, consider the term $[IJ]$. This is related to the contingency table one gets by summing over K (i.e., over the Z variable). In general, a contingency table that results from summing over the cells for one or more variables in a higher-dimensional contingency table is called a *marginal table*. Very simple examples of marginal tables are the marginal total column and the marginal total row along the bottom of the two-way table.

Table 7.10 Three-Factor Hypotheses and their Interpretation

Hypothesis	Meaning in Words	Hypothesis Restated in Terms of the π_{ijk}'s
1. $[I][J][K]$	X, Y, and Z are independent	$\pi_{ijk} = \pi_{i\cdot\cdot}\pi_{\cdot j\cdot}\pi_{\cdot\cdot k}$
2. $[IJ][K]$	Z is independent of X and Y	$\pi_{ijk} = \pi_{ij\cdot}\pi_{\cdot\cdot k}$
3. $[IK][J]$	Y is independent of X and Z	$\pi_{ijk} = \pi_{i\cdot k}\pi_{\cdot j\cdot}$
4. $[I][JK]$	X is independent of Y and Z	$\pi_{ijk} = \pi_{i\cdot\cdot}\pi_{\cdot jk}$
5. $[IJ][IK]$	For X known, Y and Z are independent; that is, Y and Z are conditionally independent given X	$\pi_{ijk} = \pi_{ij.}\pi_{i.k}/\pi_{i\cdot\cdot}$
6. $[IJ][JK]$	X and Z are conditionally independent given Y	$\pi_{ijk} = \pi_{ij.}\pi_{.jk}/\pi_{.j.}$
7. $[IK][JK]$	X and Y are conditionally independent given Z	$\pi_{ijk} = \pi_{i.k}\pi_{.jk}/\pi_{\cdot\cdot k}$
8. $[IJ][IK][JK]$	No three-factor interaction	No simple form

Using the idea of marginal tables, we can discuss some properties of fits of the various hierarchical hypotheses for log-linear models. Three facts are important:

1. The fit is estimated using only the marginal tables associated with the bracket terms that state the hypothesis. For example, consider hypothesis 1, the independence of the X, Y, and Z variables. To compute the estimated fit, one only needs the one-dimensional frequency counts for the X, Y, and Z variables individually and does not need to know the joint relationship between them.
2. Suppose that one looks at the fitted estimates for the frequencies and sums the *fitted* values to give marginal tables. The marginal sum for the fit is equal to the marginal table for the actual data set when the marginal table is involved in the fitting.
3. The chi-square and likelihood ratio chi-square tests discussed above using the observed and fitted values still hold.

We consider fitting hypothesis 5 to the data of Example 7.5. The hypothesis stated that if one knows the response to the maximal treadmill test, the resting electrocardiogram ST- and T-wave abnormalities are independent of the number of vessels diseased. The observed frequencies and the fitted frequencies, as well as the values of the u-parameters for this model, are given in Table 7.11.

The relationship between the fitted parameter values and the expected, or fitted, number of observations in a cell is given by the following equations:

$$\widehat{\pi}_{ijk} = e^{u+u_i^I+u_j^J+u_k^K+u_{ij}^{IJ}+u_{ik}^{IK}}$$

The fitted value $= n_{\ldots}\widehat{\pi}_{ijk}$, where $n_{\ldots}$ is the total number of observations.
For these data, we compute the right-hand side of the first equation for the (1,1,1) cell. In this case,

$$\begin{aligned}\widehat{\pi}_{111} &= \exp(-2.885 + 0.321 + 0.637 - 0.046 - 0.284 - 0.680)\\ &= e^{-2.937} \doteq 0.053\end{aligned}$$

$$\text{fitted value} \doteq 594 \times 0.053 \doteq 31.48$$

Table 7.11 Fitted Model for the Hypothesis That the Resting Electrocardiogram ST- and T-Wave (Normal or Abnormal) Is Independent of the Number of Vessels Diseased (0, 1, and 2–3) Conditionally upon Knowing the Exercise Response (+ or −)

Cell $(i,\ j,\ k)$	Observed	Fitted	u-Parameters
(1,1,1)	30	31.46	$u = -2.885$
(1,1,2)	64	57.57	$u_1^I = -u_2^I = 0.321$
(1,1,3)	147	151.97	$u_1^J = -u_2^J = 0.637$
(1,2,1)	17	15.54	$u_1^K = -0.046, u_2^K = -0.200$
(1,2,2)	22	28.43	$u_3^K = 0.246$
(1,2,3)	80	75.04	$u_{1,1}^{IJ} = -0.284, u_{1,2}^{IJ} = 0.284$
(2,1,1)	118	113.95	$u_{2,1}^{IJ} = 0.284, u_{2,2}^{IJ} = -0.284$
(2,1,2)	46	45.75	$u_{1,1}^{IK} = -0.680, u_{1,2}^{IK} = 0.078$
(2,1,3)	38	42.30	$u_{1,3}^{IK} = 0.602$
(2,2,1)	14	18.05	$u_{2,1}^{IK} = 0.680, u_{2,2}^{IK} = -0.078$
(2,2,2)	7	7.25	$u_{2,3}^{IK} = -0.602$
(2,2,3)	11	6.70	

where exp(argument) is equal to the number e raised to a power equal to the argument. The computed value of 31.48 differs slightly from the tabulated value, because the tabulated value came from computer output that carried more accuracy than the accuracy used in this computation.

We may test whether the hypothesis is a reasonable fit by computing the chi-square value under this hypothesis. The likelihood ratio chi-square value is computed as follows:

$$\mathrm{LRX}^2 = 2(30 \ln \frac{30}{31.46} + \cdots + 11 \ln \frac{11}{6.70}) \doteq 6.86$$

To assess the statistical significance we need the degrees of freedom to examine the chi-square value. For the log-linear model the degrees of freedom is given by the following rule:

Rule 1. The chi-square statistic for model fit of a log-linear model has degrees of freedom equal to the total number of cells in the table $(I \times J \times K)$ minus the number of independent parameters fitted. By *independent parameters* we mean the following: The number of parameters fitted for the X variable is $I - 1$ since the u_i^I terms sum to zero. For each of the possible terms in the model, the number of independent parameters is given in Table 7.12.

For the particular model at hand, the number of independent parameters fitted is the sum of the last column in Table 7.13. There are 12 cells in the table, so that the number of degrees of freedom is 12 − 8, or 4. The p-value for a chi-square of 6.86 for four degrees of freedom is 0.14, so that we cannot reject the hypothesis that this particular model fits the data.

We are now faced with a new consideration. Just because this model fits the data, there may be other models that fit the data as well, including some simpler model. In general, one would like as simple a model as possible (Occam's razor); however, models with more parameters generally give a better fit. In particular, a simpler model may have a p-value much closer to the significance level that one is using. For example, if one model has a p of 0.06 and is simple, and a slightly more complicated model has a p of 0.78, which is to be preferred? If the sample size is small, the p of 0.06 may correspond to estimated cell values that differ considerably from the actual values. For a very large sample, the fit may be excellent. There is no hard-and-fast rule in the trade-off between the simplicity of the model and the goodness of the fit. To understand the data, we are happy with the simple model that fits fairly well, although presumably, it is not precisely the probability model that would fit the entirety of the population values. Here we would hope for considerable scientific understanding from the simple model.

Table 7.12 Degrees of Freedom for Log-Linear Model Chi-Square

Term	Number of Parameters
u	1
u_i^I	$I - 1$
u_j^J	$J - 1$
u_k^K	$K - 1$
u_{ij}^{IJ}	$(I - 1)(J - 1)$
u_{ik}^{IK}	$(I - 1)(K - 1)$
u_{jk}^{JK}	$(J - 1)(K - 1)$
u_{ijk}^{IJK}	$(I - 1)(J - 1)(K - 1)$

Table 7.13 Parameters for Example 7.5

	Number of Parameters	
Model Terms	General	Example 7.5
u	1	1
u_i^I	$I - 1$	1
u_j^J	$J - 1$	1
u_k^K	$K - 1$	2
u_{ij}^{IJ}	$(I - 1)(J - 1)$	1
u_{ik}^{IK}	$(I - 1)(K - 1)$	2

Table 7.14 Chi-Square Goodness-of-Fit Statistics for Example 7.5 Data

Model	d.f.	LRX^2	p-Value	X^2
$[I][J][K]$	7	184.21	< 0.0001	192.35
[IJ][K]	6	154.35	< 0.0001	149.08
[IK][J]	5	36.71	< 0.0001	34.09
[I][JK]	5	168.05	< 0.0001	160.35
[IJ][IK]	4	6.86	0.14	7.13
[IJ][JK]	4	138.19	< 0.0001	132.30
[IK][JK]	3	20.56	0.0001	21.84
[IJ][IK][JK]	2	2.96	0.23	3.03

For this example, Table 7.14 shows for each of the eight possible models the degrees of freedom (d.f.), the LRX^2 value (with its corresponding p-value for reference), and the "usual" goodness-of-fit chi-square value. We see that there are only two possible models if we are to simplify at all rather than using the entire data set as representative. They are the model fit above and the model that contains each of the three two-factor interactions. The model fit above is simpler, while the other model below has a larger p-value, possibly indicating a better fit. One way of approaching this is through what are called *nested hypotheses*.

Definition 7.2. One hypothesis is *nested* within another if it is the special case of the other hypothesis. That is, whenever the nested hypothesis holds it necessarily implies that the hypothesis it is nested in also holds.

If nested hypotheses are considered, one takes the difference between the likelihood ratio chi-square statistic for the more restrictive hypothesis, minus the likelihood ratio chi-square statistic for the more general hypothesis. This difference will itself be a chi-square statistic if the special case holds. The degrees of freedom of the difference is equal to the difference of freedom for the two hypotheses. In this case, the chi-square statistic for the difference is $6.86 - 2.96 = 3.90$. The degrees of freedom are $4 - 2 = 2$. This corresponds to a p-value of more than 0.10. At the 5% significance level, there is marginal evidence that the more general hypothesis does fit the data better than the restrictive hypothesis. In this case, however, because of the greater simplicity of the restrictive hypothesis, one might choose it to fit the data. Once again, there is no hard and fast answer to the payoff between fit of the data and simplicity of interpretation of a hypothesis.

This material is an introduction to log-linear models. There are many extensions, some of which are mentioned briefly in the Notes at the end of the chapter. An excellent introduction to log-linear models is given in Fienberg [1977]. Other elementary books on log-linear models are those by Everitt [1992] and Reynolds [1977]. A more advanced and thorough treatment is given by Haberman [1978, 1979]. A text touching on this subject and many others is Bishop et al. [1975].

NOTES

7.1 Testing Independence in Model 1 and Model 2 Tables

This note refers to Section 7.2.

1. *Model 1.* The usual null hypothesis is that the results are statistically independent. That is (assuming row variable $= i$ and column variable $= j$):

$$P[i \text{ and } j] = P[i]P[j]$$

The probability on the left-hand side of the equation is π_{ij}. From Section 7.2, the marginal probabilities are found to be

$$\pi_{i\cdot} = \sum_{j=1}^{c} \pi_{ij} \quad \text{and} \quad \pi_{\cdot j} = \sum_{i=1}^{r} \pi_{ij}$$

The null hypothesis of statistical independence of the variables is

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$$

Consider how one might estimate these probabilities under two circumstances:

a. Without assuming the variables are independent.
b. Assuming the variables are independent.

In the first instance we are in a binomial situation. Let a success be the occurrence of the ijth pair. Let

$$n_{\cdot\cdot} = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$$

The binomial estimate for π_{ij} is the number of successes divided by the number of trials:

$$p_{ij} = \frac{n_{ij}}{n_{\cdot\cdot}}$$

If we assume independence, the natural approach is to estimate $\pi_{i\cdot}$ and $\pi_{\cdot j}$. But the occurrence of state i for the row variable is also a binomial event. The estimate of $\pi_{i\cdot}$ is the number of occurrences of state i for the row variable ($n_{i\cdot}$) divided by the sample size ($n_{\cdot\cdot}$). Thus,

$$p_{i\cdot} = \frac{n_{i\cdot}}{n_{\cdot\cdot}}$$

Similarly, $\pi_{\cdot j}$ is estimated by

$$p_{\cdot j} = \frac{n_{\cdot j}}{n_{\cdot\cdot}}$$

Under the hypothesis of statistical independence, the estimate of $\pi_{i\cdot}\pi_{\cdot j} = \pi_{ij}$ is

$$\frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}^2}$$

The chi-square test will involve comparing estimates of the expected number of observations with and without assuming independence. With independence, we expect to observe $n_{\cdot\cdot}\pi_{ij}$ entries in the ijth cell. This is estimated by

$$n_{\cdot\cdot}p_{i\cdot}p_{\cdot j} = \frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}$$

2. *Model 2*. Suppose that the row variable identifies the population. The null hypothesis is that all r populations have the same probabilities of taking on each value of the column variable. That is, for any two rows, denoted by i and i', say, and all j,

$$H_0 : \pi_{ij} = \pi_{i'j}$$

As in the first part above, we want to estimate these probabilities in two cases:

a. Without assuming anything about the probabilities.

b. Under H_0, that is, assuming that each population has the same distribution of the column variable.

Under (a), if no assumptions are made, π_{ij} is the probability of obtaining state j for the column variable in the $n_{i\cdot}$ trials from the ith population. Again the binomial estimate holds:

$$p_{ij} = \frac{n_{ij}}{n_{i\cdot}}$$

If the null hypothesis holds, we may "pool" all our $n_{\cdot\cdot}$ trials to get a more accurate estimate of the probabilities. Then the proportion of times the column variable takes on state j is

$$p_j = \frac{n_{\cdot j}}{n_{\cdot\cdot\cdot}}$$

As in the first part, let us calculate the numbers we expect in the cells under (a) and (b). If (a) holds, the expected number of successes in the $n_{i\cdot}$ trials of the ith population is $n_{i\cdot}\pi_{ij}$. We estimate this by

$$n_{i\cdot}(\frac{n_{ij}}{n_{i\cdot}}) = n_{ij}$$

Under the null hypothesis, the expected number $n_{i\cdot}\pi_{ij}$ is estimated by

$$n_{i\cdot}p_j = \frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}$$

In summary, under either model 1 or model 2, the null hypothesis is reasonably tested by comparing n_{ij} with $n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}$.

7.2 Measures of Association in Contingency Tables

Suppose that we reject the null hypothesis of no association between the row and column categories in a contingency table. It is useful then to have a measure of the degree of association. In a series of papers, Goodman and Kruskal [1979] argue that no single measure of association for contingency tables is best for all purposes. Measures must be chosen to help with the problem at hand. Among the measures they discuss are the following:

1. *Measure* λ_C. Call the row variable or row categorization R and the column variable or column categorization C. Suppose that we wish to use the value of R to predict the value of C. The measure λ_C is an estimate of the proportion of the errors made in classification if we do not know R that can be eliminated by knowing R before making a prediction. From the data, λ_C is given by

$$\lambda_C = \frac{\left(\sum_{i=1}^{r} \max_j n_{ij}\right) - \max_j n_{\cdot j}}{n_{\cdot\cdot} - \max_j n_{\cdot j}}$$

λ_R is defined analogously.

2. *Symmetric measure* λ. λ_C does not treat the row and column classifications symmetrically. A symmetric measure may be found by assuming that the chances are 1/2 and 1/2 of needing to predict the row and column variables, respectively. The proportion of the errors in classification that may be reduced by knowing the other (row or column variable) when predicting is estimated by λ:

$$\lambda = \frac{\left(\sum_{i=1}^{r} \max_j n_{ij}\right) + \left(\sum_{j=1}^{c} \max_i n_{ij}\right) - \max_i n_{i\cdot} - \max_j n_{\cdot j}}{2n_{\cdot\cdot} - (\max_i n_{i\cdot} + \max_j n_{\cdot j})}$$

3. *Measure* γ *for ordered categories*. In many applications of contingency tables the categories have a natural order: for example, last grade in school, age categories, number of weeks hospitalized. Suppose that the orderings of the variables correspond to the indices i and j for the rows and columns. The γ measure is the difference in the proportion of the time that the two measures have the same ordering minus the proportion of the time that they have the opposite ordering, when there are no ties. Suppose that the indices for the two observations are i, j and i, j. The indices have the same ordering if

$$(1) i < i \text{ and } j < j \quad \text{or} \quad (2) i > i \text{ and } j > j$$

They have the opposite ordering if

$$(1)\ i < \boldsymbol{i} \text{ and } j > \boldsymbol{j} \quad \text{or} \quad (2)\ i > \boldsymbol{i} \text{ and } j < \boldsymbol{j}$$

There are ties if $i = \boldsymbol{i}$ and/or $j = \boldsymbol{j}$. The index is

$$\gamma = \frac{2S - 1 + T}{1 - T}$$

where

$$S = 2\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{n_{ij}\sum_{\boldsymbol{i}>i}\sum_{\boldsymbol{j}>j} n_{\boldsymbol{ij}}}{n_{\cdot\cdot}^2}$$

and

$$T = \frac{\sum_{i=1}^{r}\left(\sum_{j=1}^{c} n_{ij}\right)^2 + \sum_{j=1}^{c}\left(\sum_{i=1}^{r} n_{ij}\right)^2 - \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}^2}{n_{\cdot\cdot}^2}$$

4. *Karl Pearson's contingency coefficient*, C. Since the chi-square statistic (X^2) is based on the square of the difference between the values observed in the contingency table and the values estimated, if association does not hold, it is reasonable to base a measure of association on X^2. However, chi-square increases as the sample size increases. One would like a measure of association that estimated a property of the total population. For this reason, $X^2/n_{\cdot\cdot}$ is used in the next three measures. Karl Pearson proposed the measure C.

$$C = \sqrt{\frac{X^2/n_{\cdot\cdot}}{1 + X^2/n_{\cdot\cdot}}}$$

5. *Cramer's* V. Harold Cramer proposed a statistic with values between 0 and 1. The coefficient can actually attain both values.

$$V = \sqrt{\frac{X^2/n_{\cdot\cdot}}{\text{minimum}(r-1, c-1)}}$$

6. *Tshuprow's* T, *and the* Φ^2 *coefficient*. The two final coefficients based on X^2 are

$$T = \sqrt{\frac{X^2/n_{\cdot\cdot}}{\sqrt{(r-1)(c-1)}}} \quad \text{and} \quad \Phi = \sqrt{X^2/n_{\cdot\cdot}}$$

We compute these measures of association for two contingency tables. The first table comes from the Robertson [1975] seat belt paper discussed in the text. The data are taken for 1974 cars with the interlock system. They relate age to seat belt use. The data and the column percents are given in Table 7.15. Although the chi-square value is 14.06 with $p = 0.007$, we can see from the column percentages that the relationship is weak. The coefficients of association are

$$\begin{array}{llll} \lambda_C = 0, & \lambda = 0.003, & C = 0.08, & T = 0.04 \\ \lambda_R = 0.006, & \gamma = -0.03, & V = 0.06, & \Phi = 0.08 \end{array}$$

Table 7.15 Seat Belt Data by Age

	Age (Years)			Column Percents (Age)		
Belt Use	<30	30–49	≥ 50	< 30	30–49	≥ 50
Lap and shoulder	206	580	213	45	50	45
Lap only	36	125	65	8	11	14
None	213	459	192	47	39	41
				100	100	100

In general, all these coefficients lie between -1 or 0, and $+1$. They are zero if the variables are not associated at all. These values are small, indicating little association.

Consider the following data from Weiner et al. [1979], relating clinical diagnosis of chest pain to the results of angiographic examination of the coronary arteries:

	Frequency (Vessels Diseased)			**Row Percents (Vessels Diseased)**			
Chest Pain	**0**	**1**	**2 or 3**	**0**	**1**	**2 or 3**	**Total**
Definite angina	66	135	419	11	22	68	101
Probable angina	179	139	276	30	23	46	99
Nonischemic	197	39	15	78	16	6	100

The chi-square statistic is 418.48 with a p-value of effectively zero. Note that those with definite angina were very likely (89%) to have disease, and even the probability of having multivessel disease was 68%. Chest pain thought to be nonischemic was associated with "no disease" 78% of the time. Thus, there is a strong relationship. The measures of association are

$$\begin{array}{llll} \lambda_C = 0.24, & \lambda = \ \ 0.20, & C = 0.47, & T = 0.38 \\ \lambda_R = 0.16, & \gamma = -0.64, & V = 0.38, & \Phi = 0.53 \end{array}$$

More information on these measures of association and other potentially useful measures is available in Reynolds [1977] and in Goodman and Kruskal [1979].

7.3 Testing for Symmetry in a Contingency Table

In a square table, one sometimes wants to test the table for symmetry. For example, when examining two alternative means of classification, one may be interested not only in the amount of agreement (κ), but also in seeing that the pattern of misclassification is the same. In this case, estimate the expected value in the ijth cell by $(n_{ij} + n_{ji})/2$. The usual chi-square value is appropriate with $r(r-1)/2$ degrees of freedom, where r is the number of rows (and columns). See van Belle and Cornell [1971].

7.4 Use of the Term Linear in Log-Linear Models

Linear equations are equations of the form $y = c + a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$ for some variables $X_1, \ldots, X_n$ and constants c and $a_1, \ldots, a_n$. The log-linear model equations can be put into this form. For concreteness, consider the model $[IJ][K]$, where $i = 1, 2$, $j = 1, 2$, and $k = 1, 2$. Define new variables as follows:

$$X_1 = \begin{cases} 1 & \text{if } i = 1, \\ 0 & \text{if } i = 2; \end{cases} \quad X_2 = \begin{cases} 1 & \text{if } i = 2, \\ 0 & \text{if } i = 1; \end{cases} \quad X_3 = \begin{cases} 1 & \text{if } j = 1, \\ 0 & \text{if } j = 2; \end{cases}$$

$$X_4 = \begin{cases} 1 & \text{if } j = 2, \\ 0 & \text{if } j = 1 \end{cases} \quad X_5 = \begin{cases} 1 & \text{if } k = 1, \\ 0 & \text{if } k = 2 \end{cases} \quad X_6 = \begin{cases} 1 & \text{if } k = 2, \\ 0 & \text{if } k = 1, \end{cases}$$

$$X_7 = \begin{cases} 1 & \text{if } i = 1, j = 1, \\ 0 & \text{otherwise}; \end{cases} \quad X_8 = \begin{cases} 1 & \text{if } i = 1, j = 2, \\ 0 & \text{otherwise}; \end{cases}$$

$$X_9 = \begin{cases} 1 & \text{if } i = 2, j = 1, \\ 0 & \text{otherwise} \end{cases} \quad X_{10} = \begin{cases} 1 & \text{if } i = 2, j = 2, \\ 0 & \text{otherwise} \end{cases}$$

Then the model is

$$\begin{aligned} \log \pi_{ijk} &= u + u_1^I X_1 + u_2^I X_2 + u_1^J X_3 + u_2^J X_4 + u_1^K X_5 + u_2^K X_6 \\ &\quad + u_{1,1}^{IJ} X_7 + u_{1,2}^{IJ} X_8 + u_{2,1}^{IJ} X_9 + u_{2,2}^{IJ} X_{10} \end{aligned}$$

Thus the log-linear model is a linear equation of the same form as $y = c + a_1X_1 + a_2X_2 + \cdots + a_nX_n$. We discuss such equations in Chapter 11. Variables created to pick out a certain state (e.g., $i = 2$) by taking the value 1 when the state occurs, and taking the value 0 otherwise, are called *indicator* or *dummy variables*.

7.5 Variables of Constant Probability in Log-Linear Models

Consider the three-factor X, Y, and Z log-linear model. Suppose that Z terms are entirely "omitted" from the model, for example, $[IJ]$ or

$$\log \pi_{ijk} = u + u_i^I + u_j^J + u_{ij}^{IJ}$$

The model then fits the situation where Z is uniform on its state; that is,

$$P[Z = k] = \frac{1}{k}, \qquad k = 1, \ldots, K$$

7.6 Log-Linear Models with Zero Cell Entries

Zero values in the contingency tables used for log-linear models are of two types. Some arise as *sampling zeros* (values could have been observed, but were not in the sample). In this case, if zeros occur in marginal tables used in the estimation:

- Only certain u-parameters may be estimated.
- The chi-square goodness-of-fit statistic has reduced degrees of freedom.

Some zeros are necessarily *fixed*; for example, some genetic combinations are fatal to offspring and will not be observed in a population. Log-linear models can be used in the analysis (see Bishop et al., [1975]; Haberman [1979]; Fienberg [1977]).

7.7 GSK Approach to Higher-Dimensional Contingency Tables

The second major method of analyzing multivariate contingency tables is due to Grizzle et al. [1969]. They present an analysis method closely related to multiple regression (Chapter 11). References in which this method are considered are Reynolds [1977] and Kleinbaum et al. [1988].

PROBLEMS

In Problems 7.1–7.9, perform the following tasks as well as any other work requested. Problems 7.1–7.5 are taken from the seat belt paper of Robertson [1975].

- **(a)** If a table of expected values is given with one or more missing values, compute the missing values.
- **(b)** If the chi-square value is not given, compute the value of the chi-square statistic.
- **(c)** State the degrees of freedom.
- **(d)** State whether the chi-square p-value is less than or greater than 0.01, 0.05, and 0.10 .
- **(e)** When tables are given with missing values for the adjusted residual values, p-values and $(r - 1) \times (c - 1) \times p$-values, fill in the missing values.
- **(f)** When percent tables are given with missing values, fill in the missing percentages for the row percent table, column percent table, and total percent table, as applicable.
- **(g)** Using the 0.05 significance level, interpret the findings. (Exponential notation is used for some numbers, e.g., $34,000 = 3.4 \times 10^4 = 3.4\text{E}4$; $0.0021 = 2.1 \times 10^{-3} = 2.1\text{E} - 3$.)
- **(h)** Describe verbally what the row and column percents mean. That is, "of those with zero vessels diseased ... ," and so on.

7.1 In 1974 vehicles, seat belt use was considered in association with the ownership of the vehicle. ("L/S" means "both lap and shoulder belt.")

	Ownership			
Belt Use	**Individuals**	**Rental**	**Lease**	**Other Corporate**
L/S	583	145	86	182
Lap Only	139	24	24	31
None	524	59	74	145

Expected			
615.6	112.6	90.9	176.9
134.7	24.7	19.9	38.7
495.7	90.7	?	?

Adjusted Residuals			
−2.99	?	−0.76	0.60
0.63	−0.15	1.02	−1.44
2.65	−4.55	?	0.31

$p-$Values			
0.0028	5E − 6	0.4481	0.5497
0.5291	0.8821	?	?
0.0080	5.3E − 6	0.8992	0.7586

$(r - 1) \times (c - 1) \times p-$ Values			
0.017	3E − 5	1+	1+
1+	1+	1+	0.8869
0.048	3E − 5	1+	1+

Column Percents			
47	?	47	51
11	11	13	9
42	?	?	41

$d.f. =?$

$X^2 = 26.72$

7.2 In 1974 cars, belt use and manufacturer were also examined. One hundred eighty-nine cars from “other” manufacturers are not entered into the table.

	Manufacturer					
Belt Use	**GM**	**Toyota**	**AMC**	**Chrysler**	**Ford**	**VW**
L/S	498	25	36	74	285	33
Lap only	102	5	12	29	43	11
None	334	18	30	67	259	51

Adjusted Residuals					
3.06	0.33	−0.65	−1.70	−0.69	−3.00
0.49	−0.03	?	2.89	−3.06	0.33
−3.43	−0.32	−0.23	−0.08	2.63	?

***p*−Values**					
0.0022	0.7421	0.5180	0.0898	0.4898	0.0027
0.6208	0.9730	?	0.0039	0.0022	0.7415
0.0006	0.7527	?	0.9366	0.0085	0.0043

Column Percents					
53	52	46	44	49	?
11	10	15	17	7	?
36	38	38	39	?	?

d.f. =?
$X^2 = 34.30$

7.3 The relationship between belt use and racial appearance in the 1974 models is given here. Thirty-four cases whose racial appearance was “other” are excluded from this table.

	Racial Appearance	
Belt Use	**White**	**Black**
L/S	866	116
Lap only	206	20
None	757	102

Expected	
868.9	113.1
?	26.0
?	98.9

Adjusted Residuals	
−0.40	0.40
1.33	−1.33
?	?

***p*−Values**	
0.69	0.69
?	?
0.67	0.67

d.f. =?
X^2 =?

7.4 The following data are given as the first example in Note 7.2. In the 1974 cars, belt use and age were cross-tabulated.

Expected		
217.59	556.64	?
49.22	125.93	?
?	481.42	194.39

Adjusted Residuals		
?	2.06	−1.23
−2.26	−0.13	?
2.67	−2.00	−0.25

p−Values		
0.219	?	0.217
0.024	0.895	0.017
0.007	?	0.799

$(r-1) \times (c-1) \times p$−Values		
0.88	0.16	0.87
?	?	?
0.03	0.18	1+

Column %s		
45	?	45
?	?	14
47	39	41

Row %s		
?	?	?
16	55	29
25	53	22

d.f. =?
$X^2 = 14.06$

7.5 In the 1974 cars, seat belt use and gender of the driver were related as follows:

	Gender	
Belt Use	Female	Male
L/S	267	739
Lap only	85	142
None	261	606

Expected	
?	?
66.3	160.7
253.1	613.9

Adjusted Residuals	
?	?
2.90	−2.90
0.77	−0.77

p−Values	
0.0104	0.0104
0.0038	0.0038
?	?

$(r-1) \times (c-1) \times p$−Values	
0.02	0.02
0.01	0.01
?	?

Column %s	
44	50
14	?
43	?

Total %s	
13	35
4	?
?	?

d.f. =?
X^2 =?

7.6 The data are given in the second example of Note 7.2. The association of chest pain classification and amount of coronary artery disease was examined.

Adjusted Residuals			$(r-1)\times(c-1)\times p$−Values		
−13.95	0.33	12.54	1.0E − 30	1+	4.3E − 27
?	1.57	−1.27	1+	0.47	0.82
18.32	−2.47	−14.80	1.4E − 40	0.05	8.8E − 33

Row %s			Column %s			
11	22	68	?	43	59	d.f. =?
30	23	46	?	44	39	$X^2 = 418.48$
?	?	?	?	12	2	

7.7 Peterson et al. [1979] studied the age at death of children who died from sudden infant death syndrome (SIDS). The deaths from a variety of causes, including SIDS, were cross-classified by the age at death, as in Table 7.16, taken from death records in King County, Washington, over the years 1969–1977.

Table 7.16 Death Data for Problem 7.7[a]

	Age at Death				
Cause	0 Days	1–6 Days	2–4 Weeks	5–26 Weeks	27–51 Weeks
Hyaline membrane disease	19	51	7	0	0
Respiratory distress syndrome	68	191	46	0	3
Asphyxia of the newborn	105	60	7	4	2
Immaturity	104	34	3	0	0
Birth injury	115	105	17	2	0
Congenital malformation	79	101	72	75	32
Infection	7	38	36	43	18
SIDS	0	0	24	274	24
All other	60	51	28	58	35

[a] d.f. =?; X^2 = 1504.18.

(a) The values of $(r-1)\times(c-1)\times p$-value for the adjusted residual are given here multiplied by −1 if the adjusted residual is negative and multiplied by +1 if the adjusted residual is positive.

−1+	1.4E − 9	−1+	−3.8E − 5	−0.89
−0.43	4.6E − 26	1+	−3.3E − 20	−3.2E − 3
3.0E − 18	1+	−0.02	−4.5E − 10	−0.18
2.3E − 26	−1+	−5.8E − 3	−1.2E − 9	−0.08
1.1E − 11	3.9E − 4	−0.42	−3.8E − 15	−1.6E − 3
−0.20	−1+	7.7E − 6	−1+	0.12
−1.1E − 8	−1+	1.3E − 5	0.90	6.5E − 3
−31.2E − 25	−3.4E − 28	−0.19	1.7E − 57	1+
−1+	−0.03	1+	1+	2.9E − 9

What is the distribution of SIDS cases under the null hypothesis that all causes have the same distribution?

(b) What percent display (row, column, or total) would best emphasize the difference?

7.8 Morehead [1975] studied the relationship between the retention of intrauterine devices (IUDs) and other factors. The study participants were from New Orleans, Louisiana. Tables relating retention to the subjects' age and to parity (the number of pregnancies) are studied in this problem (one patient had a missing age).

(a) Was age related to IUD retention?

Age	**Continuers**	**Terminators**
19–24	41	48
25–29	50	40
30+	63	27

Expected	
50.95	?
51.52	38.5
51.52	38.5

Adjusted Residuals	
−2.61	2.61
−0.40	0.40
?	?

p−Values	
0.0091	0.0091
?	?
0.0027	0.0027

Column %s	
26.6	41.7
?	34.8
?	23.5

Row %s	
46.1	53.9
?	?
70.0	30.0

d.f. =?
X^2 =?

(b) The relationship of parity and IUD retention gave these data:

Parity	**Continuers**	**Terminators**
1–2	59	53
3–4	39	34
5+	57	28

Adjusted Residuals	
−1.32	1.32
−0.81	0.81
?	?

Total %s	
?	19.6
14.4	?
21.1	?

d.f. =?
$X^2 = 4.74$

7.9 McKeown et al. [1952] investigate evidence that the environment is involved in infantile pyloric stenosis. The relationship between the age at onset of the symptoms in days, and the rank of the birth (first child, second child, etc.) was given as follows:

	Age at Onset of Symptoms (Days)						
Birth Rank	**0–6**	**7–13**	**14–20**	**21–27**	**28–34**	**35–41**	**≥ 42**
1	42	41	116	140	99	45	58
2	28	35	63	53	49	23	31
≥ 3	26	21	39	48	39	14	23

(a) Find the expected value (under independence) for cell ($i = 2, j = 3$). For this cell compute (observed - expected)^2/ expected.

(b) The chi-square statistic is 13.91. What are the degrees of freedom? What can you say about the p-value?

(c) In the paper, the authors present, the column percents, not the frequencies, as above. Fill in the missing values in both arrays below. The arrangement is the same as the first table.

44	42	53	58	53	55	52
29	36	29	?	26	28	28
?	?	18	20	21	17	21

The adjusted residual p-values are

0.076	0.036	0.780	0.042	0.863	0.636	0.041
0.667	0.041	0.551	0.035	0.710	0.874	0.734
0.084	0.734	?	0.856	0.843	0.445	0.954

What can you conclude?

(d) The authors note that the first two weeks appear to have different patterns. They also present the data as:

	Age at Onset (Days)	
Birth Rank	**0–13**	**≥ 14**
1	83	458
2	63	219
≥ 3	47	163

For this table, $X^2 = 8.35$. What are the degrees of freedom? What can you say about the p-value?

(e) Fill in the missing values in the adjusted residual table, p-value table, and column percent table. Interpret the data.

Adjusted Residuals		p–Values		Column %s	
−2.89	2.89	0.0039	0.0039	43	55
?	?	0.065	0.065	33	?
1.54	−1.54	?	?	24	?

(f) Why is it crucial to know whether prior to seeing these data the investigators had hypothesized a difference in the parity distribution between the first two weeks and the remainder of the time period?

Problems 7.10–7.16 deal with the chi-square test for trend. The data are from a paper by Kennedy et al. [1981] relating operative mortality during coronary bypass operations to various risk factors. For each of the tables, let the scores for the chi-square test for trend be consecutive integers. For each of the tables:

a. Compute the chi-square statistic for trend. Using Table A.3, give the strongest possible statement about the p-value.

b. Compute, where not given, the percentage of operative mortality, and plot the percentage for the different categories using equally spaced intervals.

c. The usual chi-square statistic (with $k - 1$ degrees of freedom) is given with its p-value. When possible, from Table A.3 or the chi-square values, tell which statistic is more highly significant (has the smallest p-value). Does your figure in (b) suggest why?

7.10 The amount of anginal (coronary artery disease) chest pain is categorized by the Canadian Heart Classification from mild (class I) to severe (class IV).

	Anginal Pain Classification			
Surgical Mortality	**I**	**II**	**III**	**IV**
Yes	6	19	47	59
No	242	1371	2494	1314
% surgical mortality	2.4	1.4	1.8	?

Usual
$X^2 = 31.19$
$p = 7.7\text{E} - 7$

7.11 Congestive heart failure occurs when the heart is not pumping sufficient blood. A heart damaged by a myocardial infarction, heart attack, can incur congestive heart failure. A score from 0 (good) to 4 (bad) for congestive heart failure is related to operative mortality.

	Congestive Heart Failure Score				
Operative Mortality	**0**	**1**	**2**	**3**	**4**
Yes	73	50	13	12	4
No	4480	1394	404	164	36
% operative mortality	1.6	3.4	?	6.8	10.0

Usual
$X^2 = 46.45$
$p = 1.8\text{E} - 9$

7.12 A measure of left ventricular performance, or the pumping action of the heart, is the *ejection fraction*, which is the percentage of the blood in the left ventricle that is pumped during the beat. A high number indicates a more efficient performance.

	Ejection Fraction (%)				
Operative Mortality	**< 19**	**20–29**	**30–39**	**40–49**	**≥ 50**
Yes	1	4	5	22	74
No	14	88	292	685	3839
% operative mortality	6.7	?	?	3.1	1.9

Usual
$X^2 = 8.34$
$p = 0.080$

7.13 A score was derived from looking at how the wall of the left ventricle moved while the heart was beating (details in CASS [1981]). A score of 5 was normal; the larger the score, the worse the motion of the left ventricular wall looked. The relationship to operative mortality is given here.

	Wall Motion Score				
Operative Mortality	**5–7**	**8–11**	**12–15**	**16–19**	**≥ 20**
Yes	65	36	32	10	2
No	3664	1605	746	185	20
% operative mortality	1.7	2.2	?	5.1	9.1

Usual
$X^2 = 28.32$
$p = 1.1\text{E} - 5$

What do you conclude about the relationship? That is, if you were writing a paragraph to describe this finding in a medical journal, what would you say?

7.14 After the blood has been pumped from the heart, and the pressure is at its lowest point, a low blood pressure in the left ventricle is desirable. This left ventricular end diastolic pressure [LVEDP] is measured in millimeters of mercury (mmHg).

	LVEDP			
Operative Mortality	**0–12**	**13–18**	**19–24**	**≥24**
Yes	56	43	22	26
No	3452	1692	762	416
% operative mortality	?	2.5	2.8	5.9

Usual
$X^2 = 34.49$
$p = 1.6\text{E} - 7$

7.15 The number of diseased vessels and operative mortality are given by:

	Diseased Vessels		
Operative Mortality	**1**	**2**	**3**
Yes	17	43	91
No	1196	2018	3199
% operative mortality	1.4	2.1	?

Usual
$X^2 = 7.95$
$p = 0.019$

7.16 The left main coronary artery, if occluded (i.e., totally blocked), blocks two of the three major arterial vessels to the heart. Such an event almost always leads to death. Thus, people with much narrowing of the left main coronary artery usually receive surgical therapy. Is this narrowing also associated with higher surgical mortality?

	Percentage Narrowing			
Operative Mortality	**0–49**	**50–74**	**75–89**	**≥ 90**
Yes	116	8	10	19
No	5497	486	268	222
% operative mortality	2.1	1.6	?	7.9

Usual
$X^2 = 37.75$
$p = 3.2\text{E} - 8$

7.17 In Robertson's [1975] seat belt study, the observers (unknown to them) were checked by sending cars through with a known seat belt status. The agreement numbers between the observers and the known status were:

	Belt Use in Vehicles Sent		
Belt Use Reported	**S/L**	**Lap Only**	**No Belt**
Shoulder and lap	28	2	0
Lap only	3	33	6
No belt	0	15	103

(a) Compute P_A, P_C, and κ.

(b) Construct a 95% confidence interval for κ.

(c) Find the two-sided p-value for testing $\kappa = 0$ (for the entire population) by using $Z = \kappa/\text{SE}_0(\kappa)$.

7.18 The following table is from [Fisher et al., 1982]. The coronary artery tree has considerable biological variability. If the right coronary artery is normal-sized and supplies its usual share of blood to the heart, the circulation of blood is called *right dominant*. As the right coronary artery becomes less important, the blood supply is characterized as balanced and then *left dominant*. The data for the clinical site and quality control site joint readings of angiographic films are given here.

	Dominance (Clinical Site)		
Dominance (QC Site)	**Left**	**Balanced**	**Right**
Left	64	7	4
Balanced	4	35	32
Right	8	21	607

(a) Compute P_A, P_C, and κ (Section 7.4).

(b) Find var(κ) and construct a 90% confidence interval for the population value of κ.

7.19 Example 7.4 discusses the quality control data for the CASS arteriography (films of the arteries). A separate paper by Wexler et al. [1982] examines the study of the left ventricle. Problem 7.12 describes the ejection fraction. Clinical site and quality control site readings of ejection gave the following table:

	Ejection Fraction (QC Site)		
Ejection Fraction (Clinical Site)	**$\geq$ 50%**	**30–49%**	**< 30%**
$\geq$ 50%	302	27	5
30–49%	40	55	9
< 30%	1	9	18

(a) Compute P_A, P_C, and κ.

(b) Find SE(κ) and construct a 99% confidence interval for the population value of κ.

7.20 The value of κ depends on how we construct our categories. Suppose that in Example 7.4 we combine normal and other zero-vessel disease to create a zero-vessel disease category. Suppose also that we combine two- and three-vessel disease into a multivessel-disease category. Then the table becomes:

Vessels Diseased (QC Site)	Vessels Diseased (Clinical Site) 0	1	Multi-
0	70	20	9
1	10	155	78
Multi-	2	29	497

(a) Compute P_A, P_C, and κ.

(b) Is this kappa value greater than or less than the value in Example 7.4? Will this always occur? Why?

(c) Construct a 95% confidence interval for the population value of κ.

7.21 Zeiner-Henriksen [1972a] compared personal interview and postal inquiry methods of assessing infarction. His introduction follows:

> The questionnaire developed at the London School of Hygiene and Tropical Medicine and later recommended by the World Health Organization for use in field studies of cardiovascular disease has been extensively used in various populations. While originally developed for personal interviews, this questionnaire has also been employed for postal inquiries. The postal inquiry method is of course much cheaper than personal interviewing and is without interviewer error.
>
> A Finnish–Norwegian lung cancer study offered an opportunity to evaluate the repeatability at interview of the cardiac pain questionnaire, and to compare the interview symptom results with those of a similar postal inquiry. The last project, confined to a postal inquiry of the chest pain questions in a sub-sample of the 4092 men interviewed, was launched in April 1965, $2\frac{1}{2}$ to 3 years after the original interviews.
>
> The objective was to compare the postal inquiry method with the personal interview method as a means of assessing the prevalence of angina and possible infarction ....

The data are given in Table 7.17.

(a) Compute P_A, P_C, and κ.

(b) Construct a 90% confidence interval for the population value of $\kappa(\sqrt{\text{var}(\kappa)} = 0.0231)$.

(c) Group the data in three categories by:

(**i**) combining PI + AP, PI only, and AP only; (**ii**) combining the two PI/AP negatives categories; (**iii**) leaving "incomplete" as a third category. Recompute P_A, P_C, and κ. (This new grouping has the categories "cardiovascular symptoms," "no symptoms," and "incomplete.")

Table 7.17 Interview Data for Problem 7.21

	Interview						
				PI/AP Negative			
Postal Inquiry	PI[a] + AP[a]	PI Only	AP Only	Nonspecific	Other	Incomplete	Total
PI + AP	23	15	9	6	—	1	54
PI only	14	18	14	24	8	—	78
AP only	3	5	20	12	17	3	60
PI/AP negative							
Nonspecific	2	8	8	54	24	5	101
Other	2	3	5	62	279	1	352
Incomplete	—	2	—	22	37	—	61
Total	44	51	56	180	365	10	706

[a] PI, possible infarction; AP, angina pectoris.

Table 7.18 Interview Results for Problem 7.22

	Interview					
				I− A−		
Postal Inquiry[a]	I+ A+	I+ A−	I− A+	Nonspecific	Other	Total
I+ A+	11	3	1	1	—	16
I+ A−	2	14	—	4	—	20
I− A+	5	2	7	1	1	16
I− A−						
Nonspecific	1	4	5	39	9	58
Other	1	8	6	40	72	127
Total	20	31	19	85	82	237

[a] I+, positive infarction; I−, negative infarction; A+ and A−, positive or negative indication of angina.

7.22 In a follow-up study, Zeiner-Henriksen [1972b] evaluated the reproducibility of their method using reinterviews. Table 7.18 shows the results.

(a) Compute P_A, P_C, and κ for these data.

(b) Construct a 95% confidence interval for the population value of kappa. $\text{SE}(\kappa) = 0.043$.

(c) What is the value of the Z-statistic for testing no association that is computed from kappa and its estimated standard error $\sqrt{\text{var}_0(\kappa)} = 0.037$?

7.23 Weiner et al. [1979] studied men and women with suspected coronary disease. They were studied by a maximal exercise treadmill test. A positive test ($\geq$ 1 mm of ST-wave depression on the exercise electrocardiogram) is thought to be indicative of coronary artery disease. Disease was classified into zero-, one- (or single-), and multivessel disease. Among people with chest pain thought probably anginal (i.e., due to coronary artery disease), the following data are found.

	Vessels Diseased		
Category	**0**	**1**	**Multi-**
Males, + test	47	86	227
Males, − test	132	53	49
Females, + test	62	28	44
Females, − test	83	14	9

The disease prevalence is expected to be significantly different in men and women. We want to see whether the exercise test is related to disease separately for men and women.

(a) For males, the relationship of + or − test and disease give the data below. Fill in the missing values, interpret these data, and answer the questions.

	Vessels Diseased		
Exercise Test	**0**	**1**	**Multi-**
+	47	86	?
−	132	?	49

Expected		
108.5	84.2	167.3
70.5	54.8	?

Adjusted Residuals		
0+	0.73	0+
0+	0.73	0+

Row Percents		
?	?	63.1
56.4	22.6	20.9

Column Percents		
26.3	61.9	?
73.7	38.1	?

Formulate a question for which the row percents would be a good method of presenting the data. Formulate a question where the column percents would be more appropriate.

***7.24** **(a)** Find the natural logarithms, ln x, of the following x: 1.24, 0.63, 0.78, 2.41, 2.7182818, 1.00, 0.10. For what values do you think ln x is positive? For what values do you think ln x is negative? (A plot of the values may help.)

(b) Find the exponential, e^x, of the following x: −2.73, 5.62, 0.00, −0.11, 17.3, 2.45. When is e^x less than 1? When is e^x greater than 1?

(c) $\ln(a \times b) = \ln a + \ln b$. Verify this for the following pairs of a and b:

a :	2.00	0.36	0.11	0.62
b :	0.50	1.42	0.89	0.77

(d) $e^{a+b} = e^a \cdot e^b$. Verify this for the following pairs of numbers:

a:	−2.11	0.36	0.88	−1.31
b:	2.11	1.59	−2.67	−0.45

Table 7.19 Angina Data for Problem 7.25

Model[a]	d.f.	LRX^2	$p-$Value
$[I][J][K]$	7	114.41	0+
$[I][K]$	6	103.17	0+
$[IK][J]$	5	26.32	0+
$[I][JK]$	5	94.89	0+
$[IJ][IK]$	4	15.08	0.0045
$[IJ][JK]$	4	83.65	0+
$[IK][JK]$	3	6.80	0.079
$[IJ][IK][JK]$	2	2.50	0.286

[a] I, J, and K refer to variables as in Example 7.5.

Table 7.20 Hypothesis Data for Problem 7.25

Cell (i, j, k)	Observed	r Fitted	u-Parameters
(1,1,1)	17	18.74	$u = -3.37$
(1,1,2)	86	85.01	$u_1^I = -u_2^I = 0.503$
(1,1,3)	244	243.25	$u_1^J = -u_2^J = 0.886$
(1,2,1)	5	3.26	$u_1^K = -0.775, u_2^K = -0.128, u_3^K = 0.903$
(1,2,2)	14	14.99	$u_{1,1}^{IJ} = -u_{1,2}^{IJ} = -u_{2,1}^{IJ} = u_{2,2}^{IJ} = -0.157$
(1,2,3)	99	99.75	$u_{1,1}^{IK} = -u_{2,1}^{IK} = -0.728$
(2,1,1)	42	40.26	$u_{1,2}^{IK} = -u_{2,2}^{IK} = 0.143$
(2,1,2)	31	31.99	$u_{1,3}^{IK} = -u_{2,3}^{IK} = 0.586$
(2,1,3)	37	37.75	$u_{1,1}^{JK} = -u_{2,1}^{JK} = 0.145$
(2,2,1)	2	3.74	$u_{1,2}^{JK} = -u_{2,2}^{JK} = 0.138$
(2,2,2)	4	3.01	$u_{1,3}^{JK} = -u_{2,3}^{JK} = -0.283$
(2,2,3)	9	8.25	

***7.25** Example 7.5 uses Weiner et al. [1979] data for cases with probable angina. The results for the cases with definite angina are given in Table 7.19.

(a) Which models are at all plausible?

(b) The data for the fit of the $[IJ][IK][JK]$ hypothesis are given in Table 7.20.

Using the u-parameters, compute the fitted value for the (1,2,3) cell, showing that it is (approximately) equal to 99.75 as given.

(c) Using the fact that hypothesis 7 is nested within hypothesis 8, compute the chi-square statistic for the additional gain in fit between the models. What is the p-value (as best as you can tell from the tables)?

***7.26** As in Problem 7.25, the cases of Example 7.5, but with chest pain thought not to be due to heart disease (nonischemic), gave the goodness-of-fit likelihood ratio chi-square statistics shown in Table 7.21.

(a) Which model would you prefer? Why?

(b) For model $[IJ]$ $[IK]$, the information on the fit is given in Table 7.22.

Using the u-parameter values, verify the fitted value for the (2,1,1) cell.

(c) Interpret the probabilistic meaning of the model in words for the variables of this problem.

Table 7.21 Goodness-of-Fit Data for Problem 7.23

Model	d.f.	LRX^2	$p-$Value
$[I][J][K]$	7	35.26	0+
$[IJ][K]$	6	28.45	0+
$[IK][J]$	5	11.68	0.039
$[I][JK]$	5	32.46	0+
$[IJ][IK]$	4	4.87	0.30
$[IJ][JK]$	4	25.65	0+
$[IK][JK]$	3	8.89	0.031
$[IJ][IK][JK]$	2	2.47	0.29

Table 7.22 Fit Data for Problem 7.23

Cell (i, j, k)	Observed	r Fitted	u-Parameters
(1,1,1)	33	32.51	$u = -3.378$
(1,1,2)	13	12.01	$u_1^I = -u_2^I = 0.115$
(1,1,3)	7	8.48	$u_1^J = -u_2^J = 0.658$
(1,2,1)	13	13.49	$u_1^K = 1.364, u_2^K = -0.097, u_3^K = -1.267$
(1,2,2)	4	4.99	$u_{1,1}^{IJ} = -u_{1,2}^{IJ} = -u_{2,1}^{IJ} = u_{2,2}^{IJ} = -0.218$
(1,2,3)	5	3.52	$u_{1,1}^{IK} = -u_{2,1}^{IK} = -0.584$
(2,1,1)	126	128.69	$u_{1,2}^{IK} = -u_{2,2}^{IK} = -0.119$
(2,1,2)	21	18.75	$u_{1,3}^{IK} = -u_{2,3}^{IK} = 0.703$
(2,1,3)	3	2.56	
(2,2,1)	25	22.31	
(2,2,2)	1	3.25	
(2,2,3)	0	0.44	

***7.27** Willkens et al. [1981] study possible diagnostic criteria for Reiter's syndrome. This rheumatic disease was considered in the context of other rheumatic diseases. Eighty-three Reiter's syndrome cases were compared with 136 cases with one of the following four diagnoses: ankylosing spondylitis, seronegative definite rheumatoid arthritis, psoriatic arthritis, and gonococcal arthritis. A large number of potential diagnostic criteria were considered. Here we consider two factors: the presence or absence of urethritis and/or cervicitis (for females); and the duration of the initial attack evaluated as greater than or equal to one month or less than one month. The data are given in Table 7.23, and the goodness-of-fit statistics are given in Table 7.24.

(a) Fill in the question marks in Table 7.24.

(b) Which model(s) seem plausible (at the 0.05 significance level)?

(c) Since we are looking for criteria to differentiate between Reiter's syndrome and the other diseases, one strategy that makes sense is to assume independence of the disease category ($[K]$) and then look for the largest departures from the observed and fitted cells. The model we want is then $[IJ][K]$. The fit is given in Table 7.25.

Which cell of Reiter's syndrome cases has the largest excess of observed minus fitted?

(d) If you use the cell found in part (c) as your criteria for Reiter's syndrome, what are the specificity and sensitivity of this diagnostic criteria for these cases?

Table 7.23 Reiter's Syndrome Data for Problem 7.27

Urethritis and/or		Initial Attack [J]	
Cervicitis [I]	1 Disease [K]	<1 Month	≥1 Month
Yes	Reiter's	2	70
	Other	11	3
No	Reiter's	1	10
	Other	20	132

Table 7.24 Goodness-of-Fit Data for Problem 7.27

Model	d.f.	LRX^2	$p-$Value
$[I][J][K]$	?	200.65	?
$[IJ][K]$	?	200.41	?
$[IK][J]$	?	40.63	?
$[I][JK]$	?	187.78	?
$[IJ][IK]$	?	40.39	?
$[IJ][JK]$	?	187.55	?
$[IK][JK]$	?	27.76	?
$[IJ][IK][JK]$	?	5.94	?

Table 7.25 Goodness-of-Fit Data for Problem 7.27

Cell (i, j, k)	Observed	Fitted
(1,1,1)	70	24.33
(1,1,2)	3	48.67
(1,2,1)	2	4.33
(1,2,2)	11	8.67
(2,1,1)	10	47.33
(2,1,2)	132	94.67
(2,2,1)	1	7.00
(2,2,2)	20	14.00

***7.28** We claim in the text that the three-factor log-linear model $[IJ][IK]$ means that the J and K variables are independent conditionally upon the I variable. Prove this by showing the following steps:

(a) By definition, Y and Z are independent conditionally upon X if

$$P[Y = j \text{ and } Z = k|X = i] = P[Y = j|X = i]P[Z = k|X = i]$$

Using the probabilities π_{ijk}, show that this is equivalent to

$$\frac{\pi_{ijk}}{\pi_{i\cdot\cdot}} = \left(\frac{\pi_{ij\cdot}}{\pi_{i\cdot\cdot}}\right)\left(\frac{\pi_{i\cdot k}}{\pi_{i\cdot\cdot}}\right)$$

(b) If the equation above holds true, show that

$$\ln \pi_{ijk} = u + u_i^I + u_j^J + u_k^K + u_{ij}^{IJ} + u_{ik}^{IK}$$

where

$$u_{ij}^{IJ} = \ln(\pi_{ij\cdot}) - \frac{1}{I}\sum_{\boldsymbol{i}=1}^{I} \ln(\pi_{\boldsymbol{i}j\cdot}) - \frac{1}{J}\sum_{\boldsymbol{j}=1}^{J} \ln(\pi_{i\boldsymbol{j}\cdot}) + \frac{1}{IJ}\sum_{\boldsymbol{i}=1}^{I}\sum_{\boldsymbol{j}=1}^{J} \ln(\pi_{\boldsymbol{i}\boldsymbol{j}\cdot})$$

$$u_{ik}^{IK} = \ln(\pi_{i\cdot k}) - \frac{1}{I}\sum_{\boldsymbol{i}=1}^{I} \ln(\pi_{\boldsymbol{i}\cdot k}) - \frac{1}{K}\sum_{k=\mathbf{1}}^{K} \ln(\pi_{i\cdot\boldsymbol{k}}) + \frac{1}{IK}\sum_{\boldsymbol{i}=1}^{I}\sum_{k=\mathbf{1}}^{K} \ln(\pi_{\boldsymbol{i}\cdot\boldsymbol{k}})$$

$$u_i^I = \frac{1}{J}\sum_{\boldsymbol{j}=1}^{J} \ln(\pi_{i\boldsymbol{j}\cdot}) + \frac{1}{K}\sum_{\boldsymbol{k}=1}^{K} \ln(\pi_{i\cdot\boldsymbol{k}}) - \ln(\pi_{i\cdot\cdot}) + \frac{1}{IJ}\sum_{\boldsymbol{i}=1}^{I}\sum_{\boldsymbol{j}=1}^{J} \ln(\pi_{\boldsymbol{i}\boldsymbol{j}\cdot})$$

$$+\frac{1}{IK}\sum_{\boldsymbol{i}=1}^{I}\sum_{\boldsymbol{k}=1}^{K} \ln(\pi_{\boldsymbol{i}\cdot\boldsymbol{k}}) - \frac{1}{I}\sum_{\boldsymbol{i}=1}^{I} \ln(\pi_{\boldsymbol{i}\cdot\cdot})$$

$$u_j^J = \frac{1}{I}\sum_{\boldsymbol{i}=1}^{I} \ln(\pi_{\boldsymbol{i}j\cdot}) - \frac{1}{IJ}\sum_{\boldsymbol{j}=1}^{J}\sum_{\boldsymbol{i}=1}^{I} \ln(\pi_{\boldsymbol{i}\boldsymbol{j}\cdot})$$

$$u_k^J = \frac{1}{I}\sum_{\boldsymbol{i}=1}^{I} \ln(\pi_{\boldsymbol{i}\cdot k}) - \frac{1}{IK}\sum_{\boldsymbol{k}=1}^{K}\sum_{\boldsymbol{i}=1}^{I} \ln(\pi_{\boldsymbol{i}\cdot\boldsymbol{k}})$$

$$u = -\frac{1}{IJ}\sum_{\boldsymbol{i}=1}^{I}\sum_{\boldsymbol{j}=1}^{J} \ln(\pi_{\boldsymbol{i}\boldsymbol{j}\cdot}) - \frac{1}{IK}\sum_{\boldsymbol{i}=1}^{I}\sum_{\boldsymbol{k}=1}^{K} \ln(\pi_{\boldsymbol{i}\cdot\boldsymbol{k}}) - \frac{1}{I}\sum_{\boldsymbol{i}=1}^{I} \ln(\pi_{\boldsymbol{i}\cdot\cdot})$$

(c) If the equation above holds, use $\pi_{ijk} = e^{\ln \pi_{ijk}}$ to show that the first equation then holds.

***7.29** The notation and models for the three-factor log-linear model extend to larger numbers of factors. For example, for variables W, X, Y, and Z (denoted by the indices i, j, k, and l, respectively), the following notation and model correspond:

$$[IJK][L] = u + u_i^I + u_j^J + u_k^K + u_l^L + u_{ij}^{IJ} + u_{ik}^{IK} + u_{jk}^{JK} + u_{ijk}^{IJK}$$

(a) For the four-factor model, write the log-linear u-terms corresponding to the following model notations: **(i)** $[IJ][KL]$; **(ii)** $[IJK][IJL][JKL]$; **(iii)** $[IJ][IK][JK][L]$.

(b) Give the bracket notation for the models corresponding to the u-parameters: **(i)** $u + u_i^I + u_j^J + u_k^K + u_l^L$; **(ii)** $u + u_i^I + u_j^J + u_k^K + u_l^L + u_{ij}^{IJ} + u_{kl}^{KL}$; **(iii)** $u + u_i^I + u_j^J + u_k^K + u_l^L + u_{ij}^{IJ} + u_{ik}^{IK} + u_{il}^{IL} + u_{jk}^{JK} + u_{ijk}^{IJK}$.

***7.30** Verify the values of the contingency coefficients, or measures of association, given in the first example of Note 7.2.

***7.31** Verify the values of the measures of association given in the second example of Note 7.2.

***7.32** Prove the following properties of some of the measures of association, or contingency coefficients, presented in Note 7.2.

(a) $0 \leq \lambda_C \leq 1$. Show by example that 0 and 1 are possible values.

(b) $0 \leq \lambda \leq 1$. Show by example that 0 and 1 are possible values. What happens if the two traits are independent in the sample $n_{ij} = n_{i}.n._{j}/n..$?

(c) $-1 \leq \gamma \leq 1$. Can γ be -1 or $+1$? If the traits are independent in the sample, show that $\gamma = 0$. Can $\gamma = 0$ otherwise? If yes, give an example.

(d) $0 < C < 1$.

(e) $0 \leq V \leq 1$.

(f) $0 \leq T \leq 1$ [use part (e) to show this].

(g) Show by example that ϕ^2 can be larger than 1.

***7.33** Compute the contingency coefficients of Note 7.2, omitting γ, for the data of:

(a) Problem 7.1.

(b) Problem 7.5.

REFERENCES

Agresti, A. [2002]. *Categorical Data Analysis*, 2nd ed. Wiley, New York.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. [1975]. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA.

CASS [1981]. (Principal investigators of CASS and their associates); Killip, T. (ed.); Fisher, L., and Mock, M. (assoc. eds.) National Heart, Lung and Blood Institute Coronary Artery Surgery Study. *Circulation*, **63**: part II, I-1 to I-81.

Cohen, J. [1968]. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**: 213–220.

Everitt, B. S. [1992]. *The Analysis of Contingency Tables*, 2nd ed. Halstead Press, New York.

Fienberg, S. E. [1977]. *The Analysis of Cross-Classified Categorical Data*. MIT Press, Cambridge, MA.

Fisher, L. D., Judkins, M. P., Lesperance, J., Cameron, A., Swaye, P., Ryan, T. J., Maynard, C., Bourassa, M., Kennedy, J. W., Gosselin, A., Kemp, H., Faxon, D., Wexler, L., and Davis, K. [1982]. Reproducibility of coronary arteriographic reading in the Coronary Artery Surgery Study (CASS). *Catheterization and Cardiovascular Diagnosis*, **8**: 565–575. Copyright © 1982 by Wiley-Liss.

Fleiss, J. L., Levin, B., and Park, M. C. [2003]. *Statistical Methods for Rates and Proportions*, 3rd ed. Wiley, New York.

Fleiss, J. L., Cohen, J., and Everitt, B. S. [1969]. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, **72**: 323–327.

Goodman, L. A., and Kruskal, W. H. [1979]. *Measures of Association for Cross-Classifications*. Springer-Verlag, New York.

Grizzle, J. E., Starmer, C. F., and Koch, G. G. [1969]. Analysis of categorical data by linear models. *Biometrics*, **25**: 489–504.

Haberman, S. J. [1978]. *Analysis of Qualitative Data, Vol. 1, Introductory Topics*. Elsevier, New York.

Haberman, S. J. [1979]. *Analysis of Qualitative Data, Vol. 2, New Developments*. Elsevier, New York.

Hitchcock, C. R., Ruiz, E., Sutherland, D., and Bitter, J. E. [1966]. Eighteen-month follow-up of gastric freezing in 173 patients with duodenal ulcer. *Journal of the American Medical Association*, **195**: 115–119.

Kennedy, J. W., Kaiser, G. C., Fisher, L. D., Fritz, J. K., Myers, W., Mudd, J. G., and Ryan, T. J. [1981]. Clinical and angiographic predictors of operative mortality from the collaborative study in coronary artery surgery (CASS). *Circulation*, **63**: 793–802.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E., and Nizam, A. [1997]. *Applied Regression Analysis and Multivariable Methods*, 3rd ed. Brooks/Cole, Pacific Grove, California.

Kraemer, H. C., Periyakoil, V. S., and Noda, A. [2002]. Tutorial in biostatistics: kappa coefficient in medical research. *Statistics in Medicine*, **21**: 2109–2119.

Maclure, M., and Willett, W. C. [1987]. Misinterpretation and misuses of the kappa statistic. *American Journal of Epidemiology*, **126**: 161–169.

Maki, D. G., Weise, C. E., and Sarafin, H. W. [1977]. A semi-quantitative culture method for identifying intravenous-catheter-related infection. *New England Journal of Medicine*, **296**: 1305–1309.

McKeown, T., MacMahon, B., and Record, R. G. [1952]. Evidence of post-natal environmental influence in the aetiology of infantile pyloric stenosis. *Archives of Diseases in Children*, **58**: 386–390.

Morehead, J. E. [1975]. Intrauterine device retention: a study of selected social-psychological aspects. *American Journal of Public Health*, **65**: 720–730.

Nelson, J. C., and Pepe, M. S. [2000]. Statistical description of interrater reliability in ordinal ratings. *Statistical Methods in Medical Research*, **9**: 475–496.

Peterson, D. R., van Belle, G., and Chinn, N. M. [1979]. Epidemiologic comparisons of the sudden infant death syndrome with other major components of infant mortality. *American Journal of Epidemiology*, **110**: 699–707.

Reynolds, H. T. [1977]. *The Analysis of Cross-Classifications*. Free Press, New York.

Robertson, L. S. [1975]. Safety belt use in automobiles with starter-interlock and buzzer-light reminder systems. *American Journal of Public Health*, **65**: 1319–1325. Copyright © 1975 by the American Health Association.

Ruffin, J. M., Grizzle, J. E., Hightower, N. C., McHarcy, G., Shull, H., and Kirsner, J. B. [1969]. A cooperative double-blind evaluation of gastric "freezing" in the treatment of duodenal ulcer. *New England Journal of Medicine*, **281**: 16–19.

Time [1962]. Frozen ulcers. *Time*, May 18, pp. 45–47.

van Belle, G., and Cornell, R. G. [1971]. Strengthening tests of symmetry in contingency tables. *Biometrics*, **27**: 1074–1078.

Wangensteen, C. H., Peter, E. T., Nicoloff, M., Walder, A. I., Sosin, H., and Bernstein, E. F. [1962]. Achieving "physiologic gastrectomy" by gastric freezing. *Journal of the American Medical Association*, **180**: 439–444. Copyright © 1962 by the American Medical Association.

Weiner, D. A., Ryan, T. J., McCabe, C. H., Kennedy, J. W., Schloss, M., Tristani, F., Chaitman, B. R., and Fisher, L. D. [1979]. Correlations among history of angina, ST-segment response and prevalence of coronary-artery disease in the Coronary Artery Surgery Study (CASS). *New England Journal of Medicine*, **301**: 230–235.

Wexler, L., Lesperance, J., Ryan, T. J., Bourassa, M. G., Fisher, L. D., Maynard, C., Kemp, H. G., Cameron, A., Gosselin, A. J., and Judkins, M. P. [1982]. Interobserver variability in interpreting contrast left ventriculograms (CASS). *Catheterization and Cardiovascular Diagnosis*, **8**: 341–355.

Willkens, R. F., Arnett, F. C., Bitter, T., Calin, A., Fisher, L., Ford, D. K., Good, A. E., and Masi, A. T. [1981]. Reiter's syndrome: evaluation of preliminary criteria. *Arthritis and Rheumatism*, **24**: 844–849. Used with permission from J. B. Lippincott Company.

Zeiner-Henriksen, T. [1972a]. Comparison of personal interview and inquiry methods for assessing prevalences of angina and possible infarction. *Journal of Chronic Diseases*, **25**: 433–440. Used with permission of Pergamon Press, Inc.

Zeiner-Henriksen, T. [1972b]. The repeatability at interview of symptoms of angina and possible infarction. *Journal of Chronic Diseases*, **25**: 407–414. Used with permission of Pergamon Press, Inc.

CHAPTER 8

Nonparametric, Distribution-Free, and Permutation Models: Robust Procedures

8.1 INTRODUCTION

In Chapter 4 we worked with the normal distribution, noting the fact that many populations have distributions that are approximately normal. In Chapter 5 we presented elegant one- and two-sample methods for estimating the mean of a normal distribution, or the difference of the means, and constructing confidence intervals. We also examined the corresponding tests about the mean(s) from normally distributed populations. The techniques that we learned are very useful. Suppose, however, that the population under consideration is not normal. What should we do? If the population is not normal, is it appropriate to use the same t-statistic that applies when the sample comes from a normally distributed population? If not, is there some other approach that can be used to analyze such data?

In this chapter we consider such questions. In Section 8.2 we introduce terminology associated with statistical procedures needing few assumptions and in Section 8.3 we note that some of the statistical methods that we have already looked at require very few assumptions.

The majority of this chapter is devoted to specific statistical methods that require weaker assumptions than that of normality. Statistical methods are presented that apply to a wide range of situations. Methods of constructing statistical tests for specific situations, including computer simulation, are also discussed. We conclude with

1. An indication of newer research in the topics of this chapter
2. Suggestions for additional reading if you wish to learn more about the subject matter

8.2 ROBUSTNESS: NONPARAMETRIC AND DISTRIBUTION-FREE PROCEDURES

In this section we present terminology associated with statistical procedures that require few assumptions for their validity.

The first idea we consider is *robustness*:

Definition 8.1. A statistical procedure is *robust* if it performs well when the needed assumptions are not violated "too badly" or if the procedure performs well for a large family of probability distributions.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2 Copyright © 2004 John Wiley & Sons, Inc.

By a *procedure* we mean an estimate, a statistical test, or a method of constructing a confidence interval. We elaborate on this definition to give the reader a better idea of the meaning of the term. The first thing to note is that the definition is *not* a mathematical definition. We have talked about a procedure performing "well" but have not given a precise mathematical definition of what "well" means. The term *robust* is analogous to beauty: Things may be considered more or less beautiful. Depending on the specific criteria for beauty, there may be greater or lesser agreement about the beauty of an object. Similarly, different statisticians may disagree about the robustness of a particular statistical procedure depending on the probability distributions of concern and use of the procedure. Nevertheless, as the concept of beauty is useful, the concept of robustness also proves to be useful conceptually and in discussing the range of applicability of statistical procedures.

We discuss some of the ways that a statistical test may be robust. Suppose that we have a test statistic whose distribution is derived for some family of distributions (e.g., normal distributions). Suppose also that the test is to be applied at a particular significance level, which we designate the *nominal* significance level. When other distributions are considered, the *actual* probability of rejecting the null hypothesis when it holds may differ from the *nominal* significance level if the distribution is not one of those used to derive the statistical test. For example, in testing for a specific value of the mean with a normally distributed sample, the t-test may be used. Suppose, however, that the distribution considered is not normal. Then, if testing at the 5% significance level, the actual significance level (the true probability of rejecting under the *null* hypothesis that the population mean has the hypothesized value) may not be 5%; it may vary. A statistical test would be robust over a larger family of distributions if the true significance level and nominal significance level were close to each other. Also, a statistical test is robust if under specific alternatives, the probability of rejecting the null hypothesis tends to be large even when the alternatives are in a more extensive family of probability distributions.

A statistical test may be robust in a particular way for large samples, but not for small samples. For example, for most distributions, if one uses the t-test for the mean when the sample size becomes quite large, the central limit theory shows that the nominal significance level is approximately the same as the true significance level when the null hypothesis holds. On the other hand, if the samples come from a skewed distribution and the sample size is small, the t-test can perform quite badly. Lumley et al., [2002] reviewed this issue and reported that in most cases the t-test performs acceptably even with 30 or so observations, and even in a very extreme example the performance was excellent with 250 observations.

A technique of constructing confidence intervals is robust to the extent that the nominal confidence level is maintained over a larger family of distributions. For example, return to the t-test. If we construct 95% confidence intervals for the mean, the method is robust to the extent that samples from a nonnormal distribution straddle the mean about 95% of the time. Alternatively, a method of constructing confidence intervals is nonrobust if the confidence with which the parameters are in the interval differs greatly from the nominal confidence level. An estimate of a parameter is robust to the extent that the estimate is close to the true parameter value over a large class of probability distributions.

Turning to a new topic, the normal distribution model is useful for summarizing data, because two parameters (in this case, the mean and variance, or equivalently, the mean and the standard deviation) describe the entire distribution. Such a set or family of distribution functions with each member described (or indexed) by a few parameters is called a *parametric family*. The distributions used for test statistics are also parametric families. For example, the t-distribution, the F-distribution, and the χ^2-distribution depend on one or two integer parameters: the degrees of freedom. Other examples of parametric families are the binomial distribution, with its two parameters n and π, and the Poisson distribution, with its parameter λ.

By contrast, *semiparametric families* and *nonparametric families* of distributions are families that cannot be conveniently characterized, or indexed, by a few parameters. For example, if one looked at all possible continuous distributions, it is not possible to find a few parameters that characterize all these distributions.

Definition 8.2. A family of probability distributions that can be characterized by a few parameters is a *parametric family*. A family is *nonparametric* if it can closely approximate any arbitrary probability distribution. A family of probability distributions that is neither parametric nor nonparametric is *semiparametric*.

In small samples the t-test holds for the family of normal distributions, that is, for a parametric family. It would be nice to have a test statistic whose distribution was valid for a larger family of distributions. In large samples the t-test qualifies, but in small samples it does not.

Definition 8.3. Statistical procedures that hold, or are valid for a nonparametric family of distributions, are called *nonparametric statistical procedures*.

The definition of nonparametric here can be made precise in a number of nonequivalent ways, and no single definition is in universal use. See also Note 8.1. The usefulness of the t-distribution in small samples results from the fact that samples from a normal distribution give the same t-distribution for all normal distributions under the null hypothesis. More generally, it is very useful to construct a test statistic whose distribution is the same for all members of some family of distributions. That is, assuming that the sample comes from some member of the family, and the null hypothesis holds, the statistic has a known distribution; in other words, the distribution does not depend upon, or is *free* of, which member of the underlying family of distributions is sampled. This leads to our next definition.

Definition 8.4. A statistical procedure is *distribution-free* over a specified family of distributions if the statistical properties of the procedure do not depend on (or are free of) the underlying distribution being sampled.

A test statistic is distribution-free if under the null hypothesis, it has the same distribution for all members of the family. A method of constructing confidence intervals is distribution-free if the nominal confidence level holds for all members of the underlying family of distributions.

The usefulness of the (unequal variances) t-test in large samples results from the fact that samples from any distribution give the same large-sample normal distribution under the null hypothesis that the means are equal. That is, the t-statistic becomes free of any information about the shape of the distribution as the sample size increases. This leads to a definition:

Definition 8.5. A statistical procedure is *asymptotically distribution-free* over a specified family of distributions if the statistical properties of the procedure do not depend on (or are free of) the underlying distribution being sampled for sufficiently large sample sizes.

In practice, one selects statistical procedures that hold over a wide class of distributions. Often, the wide class of distributions is nonparametric, and the resulting statistical procedure is distribution-free for the family. The procedure would then be both nonparametric and distribution-free. The terms *nonparametric* and *distribution-free* are used somewhat loosely and are often considered interchangeable. The term *nonparametric* is used much more often than the term *distribution-free*.

One would expect that a nonparametric procedure would not have as much statistical power as a parametric procedure *if* the sample observed comes from the parametric family. This is frequently, but not necessarily, true. One method of comparing procedures is to look at their relative efficiency. *Relative efficiency* is a complex term when defined precisely (see Note 8.2), but the essence is contained in the following definition:

Definition 8.6. The *relative efficiency* of statistical procedure A to statistical procedure B is the ratio of the sample size needed for B to the sample size needed for A in order that both procedures have the same statistical power.

For example, if the relative efficiency of A to B is 1.5, then B needs 50% more observations than A to get the same amount of statistical power.

8.3 SIGN TEST

Suppose that we are testing a drug to reduce blood pressure using a crossover design with a placebo. We might analyze the data by taking the blood pressure while not on the drug and subtracting it from the blood pressure while on the drug. These differences resulting from the matched or paired data will have an expected mean of zero if the drug under consideration had no more effect than the placebo effect. If we want to assume normality, a one-sample t-test with a hypothesized mean of zero is appropriate. Suppose, however, that we knew from past experience that there were occasional large fluctuations in blood pressure due to biological variability. If the sample size were small enough that only one or two such fluctuations were expected, we would be hesitant to use the t-test because of the known fact that one or two large observations, or outliers, destroyed the probability distribution of the test (see Problem 8.20). What should we do?

An alternative nonparametric way of analyzing the data is the following. Suppose that there is no treatment effect. All of the difference between the blood pressures measured on-drug and on-placebo will be due to biological variability. Thus, the difference between the two measurements will be due to symmetric random variability; the number is equally likely to be positive or negative. The *sign test* is appropriate for the null hypothesis that observed values have the same probability of being positive or negative: If we look at the number of positive numbers among the differences (and exclude values equal to zero), under the null hypothesis of no drug effect, this number has a binomial distribution, with $\pi = \frac{1}{2}$. *A test of the null hypothesis could be a test of the binomial parameter* $\pi = \frac{1}{2}$. This was discussed in Chapter 6 when we considered McNemar's test. Such tests are called *sign tests*, since we are looking at the sign of the difference.

Definition 8.7. Tests based on the sign of an observation (i.e., plus or minus), and which test the hypothesis that the observation is equally likely to be a plus or minus, are called *sign test procedures*.

Note that it is possible to use a sign test in situations where numbers are not observed, but there is only a rating. For example, one could have a blinded evaluation of patients as worse on-drug than on-placebo, the same on-drug as on-placebo, and better on-drug than on-placebo. By considering only those who were better or worse on the drug, the null hypothesis of no effect is equivalent to testing that each outcome is equally likely; that is, the binomial probability is $1/2$, the sign test may be used. Ratings of this type are useful in evaluating drugs when numerical quantification is not available. As tests of $\pi = \frac{1}{2}$ for binomial random variables were discussed in Chapter 6, we will not elaborate here. Problems 8.1 to 8.3 use the sign test.

Suppose that the distribution of blood pressures *did* follow a normal distribution: How much would be lost in the way of efficiency by using the sign test? We can answer this question mathematically in large sample sizes. The relative efficiency of the sign test with respect to the t-test when the normal assumptions are satisfied is 0.64; that is, compared to analyzing data using the t-test, 36% of the samples are effectively thrown away. Alternatively, one needs 1/0.64, or 1.56 times as many observations for the sign test as one would need using the t-test to have the same statistical power in a normal distribution. On the other hand, if the data came from a different mathematical distribution, the Laplace or double exponential distribution, the sign test would be more efficient than the t-test.

In some cases a more serious price paid by switching to the sign test is that a different scientific question is being answered. With the t-test we are asking whether the average blood pressure is lower on drug than on placebo; with the sign test we are asking whether the majority of patients have lower blood pressure on drug than on placebo. The answers may be different and it is important to consider which is the more important question.

The sign test is useful in many situations. It is a "quick-and-dirty" test that one may compute mentally without the use of computational equipment; provided that statistical tables are available, you can get a quick estimate of the statistical significance of an appropriate null hypothesis.

8.4 RANKS

Many of the nonparametric, distribution-free tests are based on one simple and brilliant idea. The approach is motivated by an example.

Example 8.1. The following data are for people who are exercised on a treadmill to their maximum capacity. There were five people in a group that underwent heavy distance-running training and five control subjects who were sedentary and not trained. The maximum oxygen intake rate adjusted for body weight is measured in mL/kg per minute. The quantity is called $\mathrm{VO}_{2\mathrm{MAX}}$. The values for the untrained subjects were 45, 38, 48, 49, and 51. The values for the trained subjects were 63, 55, 59, 65, and 77. Because of the larger spread among the trained subjects, especially one extremely large $\mathrm{VO}_{2\mathrm{MAX}}$ (as can be seen from Figure 8.1), the values do not look like they are normally distributed. On the other hand, it certainly appears that the training has some benefits, since the five trained persons all exceed the treadmill times of the five sedentary persons. Although we do not want to assume that the values are normally distributed, we should somehow use the fact that the larger observations come from one group and the smaller observations come from the other group. We desire a statistical test whose distribution can be tabulated under the null hypothesis that the probability distributions are the same in the two groups.

The crucial idea is the rank of the observation, which is the position of the observation among the other observations when they are arranged in order.

Definition 8.8. The *rank* of an observation, among a set of observations, is its position when the observations are arranged from smallest to largest. The smallest observation has rank 1, the next smallest has rank 2, and so on. If observations are tied, the rank assigned is the average of the ranks appropriate to the equal numbers.

For example, the ranks of the 10 observations given above would be found as follows: first, order the observations from the smallest to largest; then number them from left to right, beginning at 1.

Observation	38	45	48	49	51	55	59	63	65	77
Rank	1	2	3	4	5	6	7	8	9	10

We now consider several of the benefits of using ranks. In the example above, suppose there was no difference in the $\mathrm{VO}_{2\ \mathrm{MAX}}$ value between the two populations. Then we have 10 independent samples (five from each population). Since there would be nothing to distinguish between observations, the five observations from the set of people who experienced training would be equally likely to be any five of the given observations. That is, if we consider the

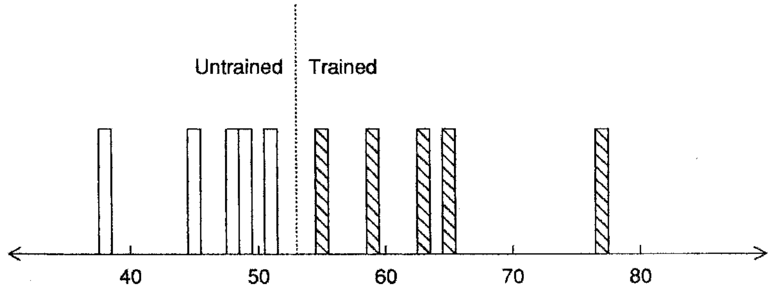


Figure 8.1 $\mathrm{VO}_{2\ \mathrm{MAX}}$ in trained and untrained persons.

ranks from 1 to 10, all subsets of size 5 would be equally likely to represent the ranks of the five trained subjects. This is true regardless of the underlying distribution of the 10 observations.

We repeat for emphasis: *If we consider continuous probability distributions (so that there are no ties) under the null hypothesis that two groups of observations come from the same distribution, the ranks have the same distribution*! Thus, tests based on the ranks will be nonparametric tests over the family of continuous probability distributions. Another way of making the same point: Any test that results from using the ranks will be distribution-free, because the distribution of the ranks does not depend on the underlying probability distribution under the null hypothesis.

There is a price to be paid in using rank tests. If we have a small number of observations, say two in each group, even if the two observations in one group are larger than both observations in the other group, a rank test will not allow rejection of the null hypothesis that the distributions are the same. On the other hand, if one knows that the data are approximately normally distributed if the two large observations are considerably larger than the smaller observations, the t-test would allow one to reject the null hypothesis that the distributions are the same. However, this increased statistical power in tiny samples *critically* depends on the normality assumptions. With small sample sizes, one cannot check the adequacy of the assumptions. One may reject the null hypothesis incorrectly (when, in fact, the two distributions are the same) because a large outlying value is observed. This price is specific to small samples: In large samples a particular rank-based test may be more or less powerful than the t-test. Note 8.6 describes another disadvantage of rank tests.

Many nonparametric statistical tests can be devised using the simple idea of ranks. In the next three sections of this chapter we present specific rank tests of certain hypotheses.

8.5 WILCOXON SIGNED RANK TEST

In this section we consider our first rank test. The test is an alternative to the one-sample t-test. Whenever the one-sample t-test of Chapter 5 is appropriate, this test may also be used, as its assumptions will be satisfied. However, since the test is a nonparametric test, its assumptions will be satisfied much more generally than under the assumptions needed for the one-sample t-test. In this section we first discuss the needed assumptions and null hypothesis for this test. The test itself is then presented and illustrated by an example. For large sample sizes, the value of the test statistic may be approximated by a standard normal distribution; the appropriate procedure for this is also presented.

8.5.1 Assumptions and Null Hypotheses

The signed rank test is appropriate for statistically independent observations. The null hypothesis to be tested is that each observation comes from a distribution that is symmetric with a mean of zero. That is, for any particular observation, the value is equally likely to be positive or negative.

For the one-sample t-test, we have independent observations from a normal distribution; suppose that the null hypothesis to be tested has a mean of zero. When the mean is zero, the distribution is symmetric about zero, and positive or negative values are equally likely. Thus, the signed rank test may be used wherever the one-sample t-test of mean zero is appropriate. For large sample sizes, the signed rank test has an efficiency of 0.955 relative to the t-test; the price paid for using this nonparametric test is equivalent to losing only 4.5% of the observations. In addition, when the normal assumptions for the t-test hold and the mean is not zero, the signed rank test has equivalent statistical power.

An example where the signed rank test is appropriate is a crossover experiment with a drug and a placebo. Suppose that subjects have the sequence “placebo, then drug” or “drug, then placebo,” each assigned at random, with a probability of 0.5. The null hypothesis of interest is that the drug has the same effect as the placebo. If one takes the difference between measurements taken on the drug and on the placebo, and if the treatment has no effect, the distribution of the difference will not depend on whether the drug was given first or second. The probability is one-half that the placebo was given first and that the observation being looked at is the second observation minus the first observation. The probability is also 1/2 that the observation being examined came from a person who took the drug first. In this case, the observation being used in the signed rank test would be the first observation minus the second observation. Since under the null hypothesis, these two differences have the same distribution except for a minus sign, the distribution of observations under the null hypothesis of "no treatment effect" is symmetric about zero.

8.5.2 Alternative Hypotheses Tested with Power

To use the test, we need to know what type of alternative hypotheses may be detected with some statistical power. For example, suppose that one is measuring blood pressure, and the drug supposedly lowers the blood pressure compared to a placebo. The difference between the measurements on the drug and the blood pressure will tend to be negative. If we look at the observations, two things will occur. First, there will tend to be more observations that have a negative value (i.e., a minus sign) than expected by chance. Second, if we look at the values of the data, the largest absolute values will tend to be negative values. The differences that are positive will usually have smaller absolute values. The signed rank test is designed to use both sorts of information. The signed rank statistic is designed to have power where the alternatives of interest correspond roughly to a shift of the distribution (e.g., the median, rather than being zero, is positive or negative).

8.5.3 Computation of the Test Statistic

We compute the signed rank statistic as follows:

1. Rank the absolute values of the observations from smallest to largest. Note that we do *not* rank the observations themselves, but rather, the absolute values; that is, we ignore minus signs. Drop observations equal to zero.
2. Add up the values of the ranks assigned to the positive observations. Do the same to the negative observations. The smaller of the two values is the value of the Wilcoxon signed rank statistic used in Table A.9 in the Appendix.

The procedure is illustrated in the following example.

Example 8.2. Brown and Hurlock [1975] investigated three methods of preparing the breasts for breastfeeding. The methods were:

1. Toughening the skin of the nipple by nipple friction or rolling
2. Creams to soften and lubricate the nipple
3. Prenatal expression of the first milk secreted before or after birth (colostrum)

Each subject had one randomly chosen treated breast and one untreated breast. Nineteen different subjects were randomized to each of three treatment groups; that is, each subject received the three treatments in random order. The purpose of the study was to evaluate methods of preventing postnatal nipple pain and trauma. The effects were evaluated by the mothers filling out a subjective questionnaire rating nipple sensitivity from "comfortable" (1) to "painful" (2) after each feeding. The data are presented in Table 8.1.

We use the signed rank test to examine the statistical significance of the nipple-rolling data. The first step is to rank the absolute values of the observations, omitting zero values. The observations ranked by absolute value and their ranks are given in Table 8.2.

Note the tied absolute values corresponding to ranks 4 and 5. The average rank 4.5 is used for both observations. Also note that two zero observations were dropped.

Table 8.1 Mean Subjective Difference between Treated and Untreated Breasts

Nipple Rolling	Masse Cream	Expression of Colostrum
−0.525	0.026	−0.006
0.172	0.739	0.000
−0.577	−0.095	−0.257
0.200	−0.040	−0.070
0.040	0.006	0.107
−0.143	−0.600	0.362
0.043	0.007	−0.263
0.010	0.008	0.010
0.000	0.000	−0.080
−0.522	−0.100	−0.010
0.007	0.000	0.048
−0.122	0.000	0.300
−0.040	0.060	0.182
0.000	−0.180	−0.378
−0.100	0.000	−0.075
0.050	0.040	−0.040
−0.575	0.080	−0.080
0.031	−0.450	−0.100
−0.060	0.000	−0.020

Source: Data from Brown and Hurlock [1975].

Table 8.2 Ranked Observation Data

Observation	Rank	Observation	Rank
0.007	1	−0.122	10
0.010	2	−0.143	11
0.031	3	0.172	12
0.040	4.5	0.200	13
−0.040	4.5	−0.522	14
0.043	6	−0.525	15
0.050	7	−0.575	16
−0.060	8	−0.577	17
−0.100	9		

The sum of the ranks of the positive numbers is $S = 1+2+3+4.5+6+7+12+13 = 48.5$. This is less than the sum of the negative ranks. For a sample size of 17, Table A.9 shows that the two-sided p-value is ≥ 0.10. If there are no ties, Owen [1962] shows that $P[S \geq 48.5] = 0.1$ and the two-sided p-value is 0.2. No treatment effect has been shown.

8.5.4 Large Samples

When the number of observations is moderate to large, we may compute a statistic that has approximately a standard normal distribution under the null hypothesis. We do this by subtracting the mean under the null hypothesis from the observed signed rank statistic, and dividing by the standard deviation under the null hypothesis. Here we do not take the minimum of the sums of positive and negative ranks; the usual one- and two-sided normal procedures can be used. The mean and variance under the null hypothesis are given in the following two equations:

$$E(S) = \frac{n(n+1)}{4} \tag{1}$$

$$\operatorname{var}(S) = \frac{n(n+1)(2n+1)}{24} \tag{2}$$

From this, one gets the following statistic, which is approximately normally distributed for large sample sizes:

$$Z = \frac{S - E(S)}{\sqrt{\operatorname{var}(S)}} \tag{3}$$

Sometimes, data are recorded on such a scale that ties can occur for the absolute values. In this case, tables for the signed rank test are conservative; that is, the probability of rejecting the null hypothesis when it is true is *less* than the nominal significance level. The asymptotic statistic may be adjusted for the presence of ties. The effect of ties is to reduce the variance in the statistic. The rank of a term involved in a tie is replaced by the average of the ranks of those tied observations. Consider, for example, the following data:

$$6, -6, -2, 0, 1, 2, 5, 6, 6, -3, -3, -2, 0$$

Note that there are not only some ties, but zeros. In the case of zeros, the zero observations are omitted from the computation as noted before. These data, ranked by absolute value, with average ranks replacing the given rank when the absolute values are tied, are shown below. The first row (A) represents the data ranked by absolute value, omitting zero values; the second row (B) gives the ranks; and the third row (C) gives the ranks, with ties averaged (in this row, ranks of positive numbers are shown in bold type):

A	1	−2	2	−2	−3	−3	5	6	−6	6	6
B	1	2	3	4	5	6	**7**	8	9	10	11
C	**1**	3	**3**	3	5.5	5.5	**7**	**9.5**	9.5	**9.5**	**9.5**

Note that the ties are with respect to the absolute value (without regard to sign). Thus the three ranks corresponding to observations of −2 and +2 are 2, 3, and 4, the average of which is 3. The S-statistic is computed by adding the ranks for the positive values. In this case,

$$S = 1 + 3 + 7 + 9.5 + 9.5 + 9.5 = 39.5$$

Before computing the asymptotic statistic, the variance of S must be adjusted because of the ties. To make this adjustment, we need to know the number of groups that have ties and the number of ties in each group. In looking at the data above, we see that there are three sets of ties, corresponding to absolute values 2, 3, and 6. The number of ties corresponding to observations of absolute value 2 (the "2 group") is 3; the number of ties in the "3 group" is 2; and the number of ties in the "6 group" is 4. In general, let q be the number of groups of ties, and let t_i, where i goes from 1 to q, be the number of observations involved in the particular group. In this case,

$$t_1 = 3, \qquad t_2 = 2, \qquad t_3 = 4, \qquad q = 3$$

In general, the variance of S is reduced according to the equation:

$$\text{var}(S) = \frac{n(n+1)(2n+1) - \frac{1}{2}\sum_{i=1}^{q} t_i(t_i-1)(t_i+1)}{24} \tag{4}$$

For the data that we are working with, we started with 13 observations, but the n used for the test statistic is 11, since two zeros were eliminated. In this case, the expected mean and variance are

$$E(S) = 11 \times \frac{12}{4} = 33$$

$$\text{var}(S) = \frac{11 \times 12 \times 23 - \frac{1}{2}(3 \times 2 \times 4 + 2 \times 1 \times 3 + 4 \times 3 \times 5)}{24} \doteq 135.6$$

Using test statistic S gives

$$Z = \frac{S - E(S)}{\sqrt{\text{var}(S)}} = \frac{39.5 - 33}{\sqrt{135.6}} \doteq 0.56$$

With a Z-value of only 0.56, one would not reject the null hypothesis for commonly used values of the significance level. For testing at a 0.05 significance level, if n is 15 or larger with few ties, the normal approximation may reasonably be used. Note 8.4 and Problem 8.22 have more information about the distribution of the signed-rank test.

Example 8.2. (*continued*) We compute the asymptotic Z-statistic for the signed rank test using the data given. In this case, $n = 17$ after eliminating zero values. We have one set of two tied values, so that $q = 1$ and $t_1 = 2$. The null hypothesis mean is $17 \times 18/4 = 76.5$. This variance is $[17 \times 18 \times 35 - (1/2) \times 2 \times 1 \times 3]/24 = 446.125$. Therefore, $Z = (48.5 - 76.5)/21.12 \doteq -1.326$. Table A.9 shows that a two-sided p is about 0.186. This agrees with $p = 0.2$ as given above from tables for the distribution of S.

8.6 WILCOXON (MANN–WHITNEY) TWO-SAMPLE TEST

Our second example of a rank test is designed for use in the two-sample problem. Given samples from two different populations, the statistic tests the hypothesis that the distributions of the two populations are the same. The test may be used whenever the two-sample t-test is appropriate. Since the test given depends upon the ranks, it is nonparametric and may be used more generally. In this section, we discuss the null hypothesis to be tested, and the efficiency of the test relative to the two-sample t-test. The test statistic is presented and illustrated by two examples. The large-sample approximation to the statistic is given. Finally, the relationship between two equivalent statistics, the Wilcoxon statistic and the Mann–Whitney statistic, is discussed.

8.6.1 Null Hypothesis, Alternatives, and Power

The null hypothesis tested is that each of two independent samples has the same probability distribution. Table A.10 for the Mann–Whitney two-sample statistic assumes that there are no ties. Whenever the two-sample t-test may be used, the *Wilcoxon statistic* may also be used. The statistic is designed to have statistical power in situations where the alternative of interest has one population with generally larger values than the other. This occurs, for example, when the two distributions are normally distributed, but the means differ. For normal distributions with a shift in the mean, the efficiency of the Wilcoxon test relative to the two-sample t-test is 0.955.

For other distributions with a shift in the mean, the Wilcoxon test will have relative efficiency near 1 if the distribution is *light-tailed* and greater than 1 if the distribution is *heavy-tailed.*

However, as the Wilcoxon test is designed to be less sensitive to extreme values, it will have less power against an alternative that adds a few extreme values to the data. For example, a pollutant that generally had a normally distributed concentration might have occasional very high values, indicating an illegal release by a factory. The Wilcoxon test would be a poor choice if this were the alternative hypothesis. Johnson et al. [1987] shows that a *quantile test* (see Note 8.5) is more powerful than the Wilcoxon test against the alternative of a shift in the extreme values, and the U.S. EPA [1994] has recommended using this test. In large samples a t-test might also be more powerful than the Wilcoxon test for this alternative.

8.6.2 Test Statistic

The test statistic itself is easy to compute. The combined sample of observations from both populations are ordered from the smallest observation to the largest. The sum of the ranks of the population with the smaller sample size (or in the case of equal sample sizes, an arbitrarily designated first population) gives the value of the Wilcoxon statistic.

To evaluate the statistic, we use some notation. Let m be the number of observations for the smaller sample, and n the number of observations in the larger sample. The Wilcoxon statistic W is the sum of the ranks of the m observations when both sets of observations are ranked together.

The computation is illustrated in the following example:

Example 8.3. This example deals with a small subset of data from the Coronary Artery Surgery Study [CASS, 1981]. Patients were studied for suspected or proven coronary artery disease. The disease was diagnosed by coronary angiography. In coronary angiography, a tube is placed into the aorta (where the blood leaves the heart) and a dye is injected into the arteries of the heart, allowing x-ray motion pictures (angiograms) of the arteries. If an artery is narrowed by 70% or more, the artery is considered significantly diseased. The heart has three major arterial systems, so the disease (or lack thereof) is classified as zero-, one-, two-, or three-vessel disease (abbreviated 0VD, 1VD, 2VD, and 3VD). Narrowed vessels do not allow as much blood to give oxygen and nutrients to the heart. This leads to chest pain (angina) and total blockage of arteries, killing a portion of the heart (called a *heart attack* or *myocardial infarction*). For those reasons, one does not expect people with disease to be able to exercise vigorously. Some subjects in CASS were evaluated by running on a treadmill to their maximal exercise performance. The treadmill increases in speed and slope according to a set schedule. The total time on the treadmill is a measure of exercise capacity. The data that follow present treadmill time in seconds for men with normal arteries (but suspected coronary artery disease) and men with three-vessel disease are as follows:

Normal	1014	684	810	990	840	978	1002	1111		
3VD	864	636	638	708	786	600	1320	750	594	750

Note that $m = 8$ (normal arteries) and $n = 10$ (three-vessel disease). The first step is to rank the combined sample and assign ranks, as in Table 8.3. The sum of the ranks of the smaller normal group is 101. Table A.10, for the closely related Mann–Whitney statistic of Section 8.6.4, shows that we reject the null hypothesis of equal population distributions at a 5% significance level.

Under the null hypothesis, the expected value of the Wilcoxon statistic is

$$E(W) = \frac{m(m+n+1)}{2} \tag{5}$$

Table 8.3 Ranking Data for Example 8.3

Value	Rank	Group	Value	Rank	Group	Value	Rank	Group
594	1	3VD	750	7.5	3VD	978	13	Normal
600	2	3VD	750	7.5	3VD	990	14	Normal
636	3	3VD	786	9	3VD	1002	15	Normal
638	4	3VD	810	10	Normal	1014	16	Normal
684	5	Normal	840	11	Normal	1111	17	Normal
708	6	3VD	864	12	3VD	1320	18	3VD

In this case, the expected value is 76. As we conjectured (*before* seeing the data) that the normal persons would exercise longer (i.e., W would be large), a one-sided test that rejects the null hypothesis if W is too large might have been used. Table A.10 shows that at the 5% significance level, we would have rejected the null hypothesis using the one-sided test. (This is also clear, since the more-stringent two-sided test rejected the null hypothesis.)

8.6.3 Large-Sample Approximation

There is a large-sample approximation to the Wilcoxon statistic (W) under the null hypothesis that the two samples come from the same distribution. The approximation may fail to hold if the distributions are different, even if neither has systematically larger or smaller values. The mean and variance of W, with or without ties, is given by equations (5) through (7). In these equations, m is the size of the smaller group (the number of ranks being added to give W), n the number of observations in the larger group, q the number of groups of tied observations (as discussed in Section 8.6.2), and t_i the number of ranks that are tied in the ith set of ties. First, without ties,

$$\text{var}(W) = \frac{mn(m+n+1)}{12} \tag{6}$$

and with ties,

$$\text{var}(W) = \frac{mn(m+n+1)}{12} - \left[\sum_{i=1}^{q} t_i(t_i-1)(t_i+1)\right] \frac{mn}{12(m+n)(m+n-1)} \tag{7}$$

Using these values, an asymptotic statistic with an approximately standard normal distribution is

$$Z = \frac{W - E(W)}{\sqrt{\text{var}(W)}} \tag{8}$$

Example 8.3. (*continued*) The normal approximation is best used when $n \geq 15$ and $m \geq 15$. Here, however, we compute the asymptotic statistic for the data of Example 8.3.

$$E(W) = \frac{8(10+8+1)}{2} = 76$$

$$\text{var}(W) = \frac{8 \cdot 10(8+10+1)}{12} - 2(2-1)(2+1)\left[\frac{8 \cdot 10}{12(8+10)(8+10+1)}\right]$$

$$= 126.67 - 0.12 = 126.55$$

$$Z = \frac{101-76}{\sqrt{126.55}} \doteq 2.22$$

The one-sided p-value is 0.013, and the two-sided p-value is $2(0.013) = 0.026$. In fact, the exact one-sided p-value is 0.013. Note that the correction for ties leaves the variance virtually unchanged.

Example 8.4. The Wilcoxon test may be used for data that are ordered and ordinal. Consider the angiographic findings from the CASS [1981] study for men and women in Table 8.4. Let us test whether the distribution of disease is the same in the men and women studied in the CASS registry.

You probably recognize that this is a contingency table, and the χ^2-test may be applied. If we want to examine the possibility of a trend in the proportions, the χ^2-test for trend could be used. That test assumes that the proportion of females changes in a linear fashion between categories. Another approach is to use the Wilcoxon test as described here.

The observations may be ranked by the six categories (none, mild, moderate, 1VD, 2VD, and 3VD). There are many ties: 4517 ties for the lowest rank, 1396 ties for the next rank, and so on. We need to compute the average rank for each of the six categories. If J observations have come before a category with K tied observations, the average rank for the k tied observations is

$$\text{average rank} = \frac{2J + K + 1}{2} \tag{9}$$

For these data, the average ranks are computed as follows:

K	*J*	Average	*K*	*J*	Average
4,517	0	2,259	4,907	6,860	9,314
1,396	4,517	5,215.5	5,339	11,767	14,437
947	5,913	6,387	6,997	17,106	20,605

Now our smaller sample of females has 2360 observations with rank 2259, 572 observations with rank 5215.5, and so on. Thus, the sum of the ranks is

$$\begin{aligned} W &= 2360(2259) + 572(5215.5) + 291(6387) + 1020(9314) + 835(14{,}437) + 882(20{,}605) \\ &= 49{,}901{,}908 \end{aligned}$$

The expected value from equation (5) is

$$E(W) = \frac{5960(5960 + 18{,}143 + 1)}{2} = 71{,}829{,}920$$

Table 8.4 Extent of Coronary Artery Disease by Gender

Extent of Disease	Male	Female	Total
None	2,157	2,360	4,517
Mild	824	572	1,396
Moderate	656	291	947
Significant			
1VD	3,887	1,020	4,907
2VD	4,504	835	5,339
3VD	6,115	882	6,997
Total	18,143	5,960	24,103

Source: Data from CASS [1981].

From equation (7), the variance, taking into account ties, is

$$\begin{aligned}\text{var}(W) &= 5960 \times 18{,}143 \times \frac{5960 + 18{,}143 + 1}{12} \\ &\quad -(4517 \times 4516 \times 4518 + \cdots + 6997 \times 6996 \times 6998)\frac{5960 \times 18{,}143}{12 \times 20{,}103 \times 20{,}102} \\ &= 2.06 \times 10^{11}\end{aligned}$$

From this,

$$z = \frac{W - E(W)}{\sqrt{\text{var}(W)}} \doteq -48.29$$

The p-value is extremely small and the population distributions clearly differ.

8.6.4 Mann–Whitney Statistic

Mann and Whitney developed a test statistic that is equivalent to the Wilcoxon test statistic. To obtain the value for the Mann–Whitney test, which we denote by U, one arranges the observations from the smallest to the largest. The statistic U is obtained by counting the number of times an observation from the group with the smallest number of observations precedes an observation from the second group. With no ties, the statistics U and W are related by the following equation:

$$U + W = \frac{m(m + 2n + 1)}{2} \tag{10}$$

Since the two statistics add to a constant, using one of them is equivalent to using the other. We have used the Wilcoxon statistic because it is easier to compute by hand. The values of the two statistics are so closely related that books of statistical tables contain tables for only one of the two statistics, since the transformation from one to the other is almost immediate. Table A.10 is for the Mann–Whitney statistic.

To use the table for Example 8.3, the Mann–Whitney statistic would be

$$U = \frac{8[8 + 2(10) + 1]}{2} - 101 = 116 - 101 = 15$$

From Table A.10, the two-sided 5% significance levels are given by the tabulated values and mn minus the tabulated value. The tabulated two-sided value is 63, and $8 \times 10 - 63 = 17$. We do reject for a two-sided 5% test. For a one-sided test, the upper critical value is 60; we want the lower critical value of $8 \times 10 - 60 = 20$. Clearly, again we reject at the 5% significance level.

8.7 KOLMOGOROV–SMIRNOV TWO-SAMPLE TEST

Definition 3.9 showed one method of describing the distributions of values from a population: the *empirical cumulative distribution*. For each value on the real line, the empirical cumulative distribution gives the proportion of observations less than or equal to that value. One visual way of comparing two population samples would be a graph of the two empirical cumulative distributions. If the two empirical cumulative distributions differ greatly, one would suspect that the populations being sampled were not the same. If the two curves were quite close, it would be reasonable to assume that the underlying population distributions were essentially the same.

The *Kolmogorov–Smirnov statistic* is based on this observation. The value of the statistic is the maximum absolute difference between the two empirical cumulative distribution functions. Note 8.7 discusses the fact that the Kolmogorov–Smirnov statistic is a rank test. Consequently, the test is a nonparametric test of the null hypothesis that the two distributions are the same. When the two distributions have the same shape but different locations, the Kolmogorov–Smirnov statistic is far less powerful than the Wilcoxon rank-sum test (or the t-test if it applies), but the Kolmogorov–Smirnov test can pick up any differences between distributions, whatever their form.

The procedure is illustrated in the following example:

Example 8.4. (*continued*) The data of Example 8.3 are used to illustrate the statistic. Using the method of Chapter 3, Figure 8.2 was constructed with both distribution functions.

From Figure 8.2 we see that the maximum difference is 0.675 between 786 and 810. Tables of the statistic are usually tabulated not in terms of the maximum absolute difference D, but in terms of $(mn/d)D$ or mnD, where m and n are the two sample sizes and d is the lowest common denominator of m and n. The benefit of this is that $(mn/d)D$ or mnD is always an integer. In this case, $m = 8$, $n = 10$, and $d = 2$. Thus, $(mn/d)D = (8)(10/2)(0.675) = 27$ and $mnD = 54$. Table 44 of Odeh et al. [1977] gives the 0.05 critical value for mnD as 48. Since $54 > 48$, we reject the null hypothesis at the 5% significance level. Tables of critical values are not given in this book but are available in standard tables (e.g., Odeh et al. [1977]; Owen [1962]; Beyer [1990]) and most statistics packages. The tables are designed for the case with no ties. If there are ties, the test is conservative; that is, the probability of rejecting the null hypothesis when it is true is even less than the nominal significance level.

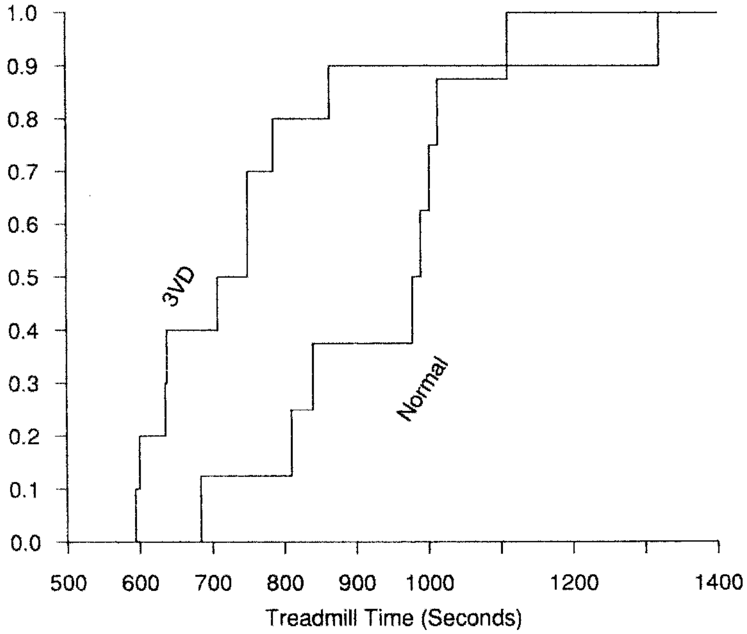


Figure 8.2 Empirical cumulative distributions for the data of Example 8.3.

The large-sample distribution of D is known. Let n and m both be large, say, both 40 or more. The large-sample test rejects the null hypothesis according to the following table:

Significance Level	**Reject the Null Hypothesis if:**
0.001	KS $\geq$ 1.95
0.01	KS $\geq$ 1.63
0.05	KS $\geq$ 1.36
0.10	KS $\geq$ 1.22

KS is defined as

$$\text{KS} = \max_x \sqrt{\frac{nm}{n+m}} |F_n(x) - G_m(x)| = \sqrt{\frac{nm}{n+m}} D \tag{11}$$

where F_n and G_m are the two empirical cumulative distributions.

8.8 NONPARAMETRIC ESTIMATION AND CONFIDENCE INTERVALS

Many nonparametric tests have associated estimates of parameters. Confidence intervals for these estimates are also often available. In this section we present two estimates associated with the Wilcoxon (or Mann–Whitney) two-sample test statistic. We also show how to construct a confidence interval for the median of a distribution.

In considering the Mann–Whitney test statistic described in Section 8.6, let us suppose that the sample from the first population was denoted by X's, and the sample from the second population by Y's. Suppose that we observe mX's and nY's. The Mann–Whitney test statistic U is the number of times an X was less than a Y among the nmX and Y pairs. As shown in equation (12), the Mann–Whitney test statistic U, when divided by mn, gives an unbiased estimate of the probability that X is less than Y.

$$E\left(\frac{U}{mn}\right) = P[X < Y] \tag{12}$$

Further, an approximate $100(1-\alpha)\%$ confidence interval for the probability that X is less than Y may be constructed using the asymptotic normality of the Mann–Whitney test statistic. The confidence interval is given by the following equation:

$$\frac{U}{mn} \pm Z_{1-\alpha/2}\sqrt{\frac{1}{\min(m,n)}\frac{U}{mn}\left(1-\frac{U}{mn}\right)} \tag{13}$$

In large samples this interval tends to be too long, but in small samples it can be too short if U/mn is close to 0 or 1 [Church and Harris, 1970]. In Section 8.10.2 we show another way to estimate a confidence interval.

Example 8.5. This example illustrates use of the Mann–Whitney test statistic to estimate the probability that X is less than Y and to find a 95% confidence interval for $P[X < Y]$.

Examine the normal/3VD data in Example 8.3. We shall estimate the probability that the treadmill time of a randomly chosen person with normal arteries is less than that of a three-vessel disease patient.

Note that 1014 is less than one three-vessel treadmill time; 684 is less than 6 of the three-vessel treadmill times, and so on. Thus,

$$U = 1 + 6 + 2 + 1 + 2 + 1 + 1 + 1 = 15$$

We also could have found U by using equation (9) and $W = 101$ from Example 8.3. Our estimate of $P[X < Y]$ is $15/(8 \times 10) = 0.1875$. The confidence interval is

$$0.1875 \pm (1.96)\sqrt{\frac{1}{8}(0.1875)(1 - 0.1875)} = 0.1875 \pm 0.2704$$

We see that the lower limit of the confidence interval is below zero. As zero is the minimum possible value for $P[X < Y]$, the confidence interval could be rounded off to $[0, 0.458]$.

If it is known that the underlying population distributions of X and Y are the same shape and differ only by a shift in means, it is possible to use the Wilcoxon test (or any other rank test) to construct a confidence interval. This is an example of a *semiparametric* procedure: it does not require the underlying distributions to be known up to a few parameters, but it does impose strong assumptions on them and so is not *nonparametric*. The procedure is to perform Wilcoxon tests of $X + \delta$ vs. Y to find values of δ at which the p-value is exactly 0.05. These values of δ give a 95% confidence interval for the difference in locations.

Many statistical packages will compute this confidence interval and may not warn the user about the assumption that the distributions have the same shape but a different location. In the data from Example 8.5, the assumption does not look plausible: The treadmill times for patients with three-vessel disease are generally lower but with one outlier that is higher than the times for all the normal subjects.

In Chapter 3 we saw how to estimate the median of a distribution. We now show how to construct a confidence interval for the median that will hold for any distribution. To do this, we use *order statistics*.

Definition 8.9. Suppose that one observes a sample. Arrange the sample from the smallest to the largest number. The smallest number is the *first-order statistic*, the second smallest is the *second-order statistic*, and so on; in general, the ith-*order statistic* is the ith number in line.

The notation used for an order statistic is to put the subscript corresponding to the particular order statistic in parentheses. That is,

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

To find a $100(1 - \alpha)\%$ confidence interval for the median, we first find from tables of the binomial distribution with $\pi = 0.5$, the largest value of k such that the probability of k or fewer successes is less than or equal to $\alpha/2$. That is, we choose k to be the largest value of k such that

$$P[\text{number of heads in } n \text{ flips of a fair coin} = 0 \text{ or } 1 \text{ or} \ldots \text{or } k] \leq \frac{\alpha}{2}$$

Given the value of k, the confidence interval for the median is the interval between the $(k + 1)$- and $(n - k)$-order statistics. That is, the interval is

$$(X_{(k+1)}, X_{(n-k)})$$

Example 8.6. The treadmill times of 20 females with normal or minimal coronary artery disease in the CASS study are

$$570, 618, 30, 780, 630, 738, 900, 750, 750, 540, 660,$$

$$780, 720, 750, 936, 900, 762, 840, 816, 690$$

We estimate the median time and construct a 90% confidence interval for the median of this population distribution. The order statistics (ordered observations) from 1 to 20 are

$$30, 540, 570, 618, 630, 660, 690, 720, 738, 750, 750,$$

$$750, 762, 780, 780, 816, 840, 900, 900, 936$$

Since we have an odd number of observations,

$$\text{median} = \frac{X_{(10)} + X_{(11)}}{2} = \frac{750 + 750}{2} = 750$$

If X is binomial, $n = 20$ and $\pi = 0.5$, $P[X \leq 5] = 0.0207$ and $P[X \leq 6] = 0.0577$. Thus, $k = 5$. Now, $X_{(6)} = 690$ and $X_{(15)} = 780$. Hence, the confidence interval is (690, 780). The actual confidence is $100(1 - 2 \times 0.0207)\% \doteq 95.9\%$. Because of the discrete nature of the data, the nominal 90% confidence interval is also a 95.9% confidence interval.

*8.9 PERMUTATION AND RANDOMIZATION TESTS

In this section we present a method that may be used to generate a wide variety of statistical procedures. The arguments involved are subtle; you need to pay careful attention to understand the logic. We illustrate the idea by working from an example.

Suppose that one had two samples, one of size n and one of size m. Consider the null hypothesis that the distributions of the two populations are the same. Let us suppose that, in fact, this null hypothesis is true; the combined $n+m$ observations are independent and sampled from the same population. Suppose now that you are told that one of the $n + m$ observations is equal to 10. Which of the $n + m$ observations is most likely to have taken the value 10? There is really nothing to distinguish the observations, since they are all taken from the same distribution or population. Thus, any of the $n + m$ observations is equally likely to be the one that was equal to 10. More generally, suppose that our samples are taken in a known order; for example, the first n observations come from the first population and the next m from the second. Let us suppose that the null hypothesis still holds. Suppose that you are now given the observed values in the sample, all $n + m$ of them, but not told which value was obtained from which ordered observation. Which arrangement is most likely? Since all the observations come from the same distribution, and the observations are independent, there is nothing that would tend to associate any one sequence or arrangement of the numbers with a higher probability than any other sequence. In other words, every assignment of the observed numbers to the $n + m$ observations is equally likely. This is the idea underlying a class of tests called *permutation tests*. To understand why they are called this, we need the definition of a permutation:

Definition 8.10. Given a set of $(n + m)$ objects arranged or numbered in a sequence, a *permutation* of the objects is a rearrangement of the objects into the same or a different order. The number of permutations is $(n + m)!$.

What we said above is that if the null hypothesis holds in the two-sample problem, all permutations of the numbers observed are equally likely. Let us illustrate this with a small example. Suppose that we have two observations from the first group and two observations from the second group. Suppose that we know that the four observations take on the values 3,

Table 8.5 Permutations of Four Observations

x		y		$\overline{x}-\overline{y}$	x		y		$\overline{x}-\overline{y}$
3	7	8	10	-4	7	8	3	10	1
3	7	10	8		7	8	10	3	
7	3	8	10		8	7	3	10	
7	3	10	8		8	7	10	3	
3	8	7	10	-3	7	10	3	8	3
3	8	10	7		7	10	8	3	
8	3	7	10		10	7	3	8	
8	3	10	7		10	7	8	3	
3	10	7	8	-1	8	10	3	7	4
3	10	8	7		8	10	7	3	
10	3	7	8		10	8	3	7	
10	3	8	7		10	8	7	3	

7, 8, and 10. Listed in Table 8.5 are the possible permutations where the first two observations would be considered to come from the first group and the second two from the second group. (Note that x represents the first group and y represents the second.)

If we only know the four values 3, 7, 8, and 10 but do not know in which order they came, any of the 24 possible arrangements listed above are equally likely. If we wanted to perform a two-sample test, we could generate a statistic and calculate its value for each of the 24 arrangements. We could then order the values of the statistic according to some alternative hypothesis so that the more extreme values were more likely under the alternative hypothesis. By looking at what sequence *actually occurred*, we can get a p-value for this set of data. The p-value is determined by the position of the statistic among the possible values. The p-value is the number of possibilities as extreme or more extreme than that observed divided by the number of possibilities.

Suppose, for example, that with the data above, we decided to use the difference in means between the two groups, $\overline{x}-\overline{y}$, as our test statistic. Suppose also that our alternative hypothesis is that group 1 has a larger mean than group 2. Then, if any of the last four rows of Table had occurred, the one-sided p-value would be 4/24, or 1/6. Note that this would be the most extreme finding possible. On the other hand, if the data had been 8, 7, 3, and 10, with an $\overline{x}-\overline{y}=1$, the p-value would be 12/24, or 1/2.

The tests we have been discussing are called *permutation tests*. They are possible when a permutation of all or some subset of the data is considered equally likely under the null hypothesis; the test is based on this fact. These tests are sometimes also called *conditional tests*, because the test takes some portion of the data as fixed or known. In the case above, we assume that we know the actual observed values, although we do not know in which order they occurred. We have seen an example of a conditional test before: Fisher's exact test in Chapter 6 treated the row and column totals as known; conditionally, upon that information, the test considered what happened to the entries in the table. The permutation test can be used to calculate appropriate p-values for tests such as the t-test when, in fact, normal assumptions do not hold. To do this, proceed as in the next example.

Example 8.7. Given two samples, a sample of size n of X observations and a sample of size m of Y observations, it can be shown (Problem 8.24) that the two-sample t-test is a monotone function of $\overline{x}-\overline{y}$; that is, as $\overline{x}-\overline{y}$ increases, t also increases. Thus, if we perform a permutation test on $\overline{x}-\overline{y}$, we are in fact basing our test on extreme values of the t-statistic. The illustration above is equivalent to a t-test on the four values given. Consider now the data

$$x_1 = 1.3, \qquad x_2 = 2.3, \qquad x_3 = 1.9, \qquad y_1 = 2.8, \qquad y_2 = 3.9$$

The 120 permutations $(3+2)!$ fall into 10 groups of 12 permutations with the same value of $\overline{x}-\overline{y}$ (a complete table is included in the Web appendix). The observed value of $\overline{x}-\overline{y}$ is -1.52, the lowest possible value. A one-sided test of $E(Y) < E(X)$ would have $p = 0.1 = 12/120$. The two-sided p-value is 0.2.

The Wilcoxon test may be considered a permutation test, where the values used are the ranks and not the observed values. For the Wilcoxon test we know what the values of the ranks will be; thus, one set of statistical tables may be generated that may be used for the entire sample. For the general permutation test, since the computation depends on the numbers actually observed, it cannot be calculated until we have the sample in hand. Further, the computations for large sample sizes are very time consuming. If n is equal to 20, there are over 2×10^{18} possible permutations. Thus, the computational work for permutation tests becomes large rapidly. This would appear to limit their use, but as we discuss in the next section, it is possible to sample permutations rather than evaluating every one.

We now turn to *randomization tests*. Randomization tests proceed in a similar manner to permutation tests. In general, one assumes that some aspects of the data are known. If certain aspects of the data are known (e.g., we might know the numbers that were observed, but not which group they are in), one can calculate a number of equally likely outcomes for the complete data. For example, in the permutation test, if we know the actual values, all possible permutations of the values are equally likely under the null hypothesis. In other words, it is as if a permutation were to be selected at random; the permutation tests are examples of randomization tests.

Here we consider another example. This idea is the same as that used in the signed rank test. Suppose that under the null hypothesis, the numbers observed are independent and symmetric about zero. Suppose also that we are given the absolute values of the numbers observed but not whether they are positive or negative. Take a particular number a. Is it more likely to be positive or negative? Because the distribution is symmetric about zero, it is not more likely to be either one. It is equally likely to be $+a$ or $-a$. Extending this to all the observations, every pattern of assigning pluses or minuses to our absolute values is equally likely to occur under the null hypothesis that all observations are symmetric about zero. We can then calculate the value of a test statistic for all the different patterns for pluses and minuses. A test basing the p-value on these values would be called a *randomization test*.

Example 8.8. One can perform a randomization one-sample t-test, taking advantage of the absolute values observed rather than introducing the ranks. For example, consider the first four paired observations of Example 8.2. The values are -0.0525, 0.172, 0.577, and 0.200. Assign all 16 patterns of pluses and minuses to the four absolute values (0.0525, 0.172, 0.577, and 0.200) and calculate the values of the paired or one-sample t-test. The 16 computed values, in increasing order, are -3.47, -1.63, -1.49, $\mathbf{-0.86}$, -0.46, -0.34, -0.08, -0.02, 0.02, 0.08, 0.34, 0.46, 0.86, 1.48, 1.63, and 3.47. The observed t-value (in bold type) is -0.86. It is the fourth of 16 values. The two-sided p-value is $2(4/16) = 0.5$.

*8.10 MONTE CARLO OR SIMULATION TECHNIQUES

*8.10.1 Evaluation of Statistical Significance

To compute statistical significance, we need to compare the observed values with something else. In the case of symmetry about the origin, we have seen it is possible to compare the observed value to the distribution where the plus and minus signs are independent with probability 1/2. In cases where we do not know a prior appropriate comparison distribution, as in a drug trial, the distribution without the drug is found by either using the same subjects in a crossover trial or forming a control group by a separate sample of people who are not treated with the drug. There are cases where one can conceptually write down the probability structure that would generate the distribution under the null hypothesis, but in practice could not calculate the distribution. One example of this would be the permutation test. As we mentioned previously, if there are 20 different values in the sample, there are more than 2×10^{18} different permutations. To generate them all would not be feasible, even with modern electronic computers. However, one could evaluate the particular value of the test statistic by generating a second sample from the null distribution with all permutations being equally likely. If there were some way to generate permutations randomly and compute the value of the statistic, one could take the observed statistic (thinking of this as a sample of size 1) and compare it to the randomly generated value under the null hypothesis, the second sample. One would then order the observed and generated values of the statistic and decide which values are more extreme; this would lead to a rejection region for the null hypothesis. From this, a p-value could be computed. These abstract ideas are illustrated by the following examples.

Example 8.9. As mentioned above, for fixed observed values, the two-sample t-test is a monotone function of the value of $\overline{x}-\overline{y}$, the difference in the means of the two samples. Suppose that we have the $\overline{x} - \overline{y}$ observed. One might then generate random permutations and compute the values of $\overline{x} - \overline{y}$. Suppose that we generate n such values. For a two-sided test, let us order the *absolute* values of the statistic, including both our random sample under the null hypothesis and the actual observation, giving us $n + 1$ values. Suppose that the actual observed value of the statistic from the data is the kth-order statistic, where we have ordered the absolute values from smallest to largest. Larger values tend to give more evidence against the null hypothesis of equal means. Suppose that we would reject for all observations as large as the kth-order statistic or larger. This corresponds to a p-value of $(n + 2 - k)/(n + 1)$.

One problem that we have not discussed yet is the method for generating the random permutation and $\overline{x} - \overline{y}$ values. This is usually done by computer. The computer generates random permutations by using what are called *random number generators* (see Note 8.10). A study using the generation of random quantities by computer is called a *Monte Carlo study*, for the gambling establishment at Monte Carlo with its random gambling devices and games. Note that by using Monte Carlo permutations, we can avoid the need to generate all possible permutations! This makes permutation tests feasible for large numbers of observations.

Another type of example comes about when one does not know how to compute the distribution theoretically under the null hypothesis.

Example 8.10. This example will not give all the data but will describe how a Monte Carlo test was used. In the Coronary Artery Surgery Study (CASS [1981], Alderman et al. [1982]), a study was made of the reasons people that were treated by coronary bypass surgery or medical therapy. Among 15 different institutions, it was found that many characteristics affected the assignments of patients to surgical therapy. A multivariate statistical analysis of a type described later in this book (linear discriminant analysis) was used to identify factors related to choice of therapy and to estimate the probability that someone would have surgery. It was clear that the sites differed in the percentage of people assigned to surgery, but it was also clear that the clinical sites had patient populations with different characteristics. Thus, one could not immediately conclude that the clinics had different philosophies of assignment to therapy merely by running a χ^2 test. Conceivably, the differences between clinics could be accounted for by the different characteristics of the patient populations. Using the estimated probability that each patient would or would not have surgery, the total number of surgical cases was distributed among the clinics using a Monte Carlo technique. The corresponding χ^2 test for the observed and expected values was computed for each of these randomly generated assignments under the null hypothesis of no clinical difference. This was done 1000 times. The actual observed value for the statistic turned out to be larger than any of the 1000 simulations. Thus, the estimated p-value for the significance of the conjecture that the clinics had different methods of assigning people to therapy was less than 1/1001. It was thus concluded that the clinics had different philosophies by which they assigned people to medical or surgical therapy.

We now turn to other possible uses of the Monte Carlo technique.

8.10.2 The Bootstrap

The motivation for distribution-free statistical procedures is that we need to know the distribution of a statistic when the frequency distribution F of the data is not known a priori. A very ingenious way around this problem is given by the *bootstrap*, a procedure due in its full maturity to Efron [1979], although special cases and related ideas had been around for many years.

The idea behind the bootstrap is that although we do not know F, we have a good estimate of it in the empirical frequency distribution F_n. If we can estimate the distribution of our statistic when data are sampled from F_n, we should have a good approximation to the distribution of the statistic when data are sampled from the true, unknown F. We can create data sets sampled from F_n simply by resampling the observed data: We take a sample of size n from our data set of size n (replacing the sampled observation each time). Some observations appear once, others twice, others not at all.

The bootstrap appears to be too good to be true (the name emphasizes this, coming from the concept of "lifting yourself by your bootstraps"), but both empirical and theoretical analysis confirm that it works in a fairly wide range of cases. The two main limitations are that it works only for independent observations and that it fails for certain extremely nonrobust statistics (the only simple examples being the maximum and minimum). In both cases there are more sophisticated variants of the bootstrap that relax these conditions.

Because it relies on approximating F by F_n the bootstrap is a large-sample method that is only asymptotically distribution-free, although it is successful in smaller samples than, for example, the t-test for nonnormal data. Efron and Tibshirani [1986, 1993] are excellent references; much of the latter is accessible to the nonstatistician. Davison and Hinckley [1997] is a more advanced book covering many variants on the idea of resampling. The Web appendix to this chapter links to more demonstrations and examples of the bootstrap.

Example 8.11. We illustrate the bootstrap by reexamining the confidence interval for $P[X < Y]$ generated in Example 8.5. Recall that we were comparing treadmill times for normal subjects and those with three-vessel disease. The observed $P[X < Y]$ was $15/80 = 0.1875$. In constructing a bootstrap sample we sample 8 observations from the normal and 10 from the three-vessel disease data and compute U/mn for the sample. Repeating this 1000 times gives an estimate of the distribution of $P[X < Y]$. Taking the upper and lower $\alpha/2$ percentage points of the distribution gives an approximate 95% confidence interval. In this case the confidence interval is $[0, 0.41]$. Figure 8.3 shows a histogram of the bootstrap distribution with the normal approximation from Example 8.5 overlaid on it.

Comparing this to the interval generated from the normal approximation, we see that both endpoints of the bootstrap interval are slightly higher, and the bootstrap interval is not quite symmetric about the observed value, but the two intervals are otherwise very similar. The bootstrap technique requires more computer power but is more widely applicable: It is less conservative in large samples and may be less liberal in small samples.

Related resampling ideas appear elsewhere in the book. The idea of splitting a sample to estimate the effect of a model in an unbiased manner is discussed in Chapters 11 and 13 and elsewhere. Systematically omitting part of a sample, estimating values, and testing on the omitted part is used; if one does this, say for all subsets of a certain size, a *jackknife* procedure is being used (see Efron [1982]; Efron and Tibshirani [1993]).

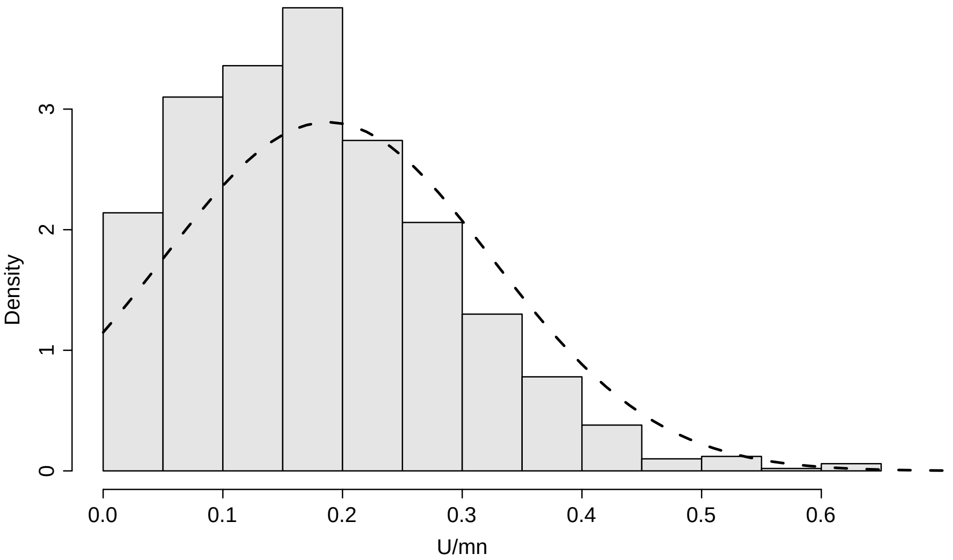


Figure 8.3 Histogram of bootstrap distribution of U/mn and positive part of normal approximation (dashed line). (Data from CASS [1981]; see Example 8.5.)

8.10.3 Empirical Evaluation of the Behavior of Statistics: Modeling and Evaluation

Monte Carlo generation on a computer is also useful for studying the behavior of statistics. For example, we know that the χ^2-statistic for contingency tables, as discussed in Chapter 7, has approximately a χ^2-distribution for large samples. But is the distribution approximately χ^2 for smaller samples? In other words, is the statistic fairly robust with respect to sample size? What happens when there are small numbers of observations in the cells? One way to evaluate small-sample behavior is a Monte Carlo study (also called a *simulation study*). One can generate multinomial samples with the two traits independent, compute the χ^2-statistic, and observe, for example, how often one would reject at the 5% significance level. The Monte Carlo simulation would allow evaluation of how large the sample needs to be for the asymptotic χ^2 critical value to be useful.

Monte Carlo simulation also provides a general method for estimating power and sample size. When designing a study one usually wishes to calculate the probability of obtaining statistically significant results under the proposed alternative hypothesis. This can be done by simulating data from the alternative hypothesis distribution and performing the planned test. Repeating this many times allows the power to be estimated. For example, if 910 of 1000 simulations give a statistically significant result, the power is estimated to be 91%. In addition to being useful when no simple formula exists for the power, the simulation approach is helpful in concentrating the mind on the important design factors. Having to simulate the possible results of a study makes it very clear what assumptions go into the power calculation.

Another use of the Monte Carlo method is to model very complex situations. For example, you might need to design a hospital communications network with many independent inputs. If you knew roughly the distribution of calls from the possible inputs, you could simulate by Monte Carlo techniques the activity of a proposed network if it were built. In this manner, you could see whether or not the network was often overloaded. As another example, you could model the hospital system of an area under the assumption of new hospitals being added and various assumptions about the case load. You could also model what might happen in catastrophic circumstances (*provided* that realistic assumptions could be made). In general, the modeling and simulation approach gives one method of evaluating how changes in an environment might affect other factors without going through the expensive and potentially catastrophic exercise of actually building whatever is to be simulated. Of course, such modeling depends *heavily* on the skill of the people constructing the model, the realism of the assumptions they make, and whether or not the probabilistic assumptions used correspond approximately to the real-life situation.

A starting reference for learning about Monte Carlo ideas is a small booklet by Hoffman [1979]. More theoretical texts are Edgington [1987] and Ripley [1987] .

*8.11 ROBUST TECHNIQUES

Robust techniques cover more than the field of nonparametric and distribution-free statistics. In general, distribution-free statistics give robust techniques, but it is possible to make more classical methods robust against certain violations of assumptions.

We illustrate with three approaches to making the sample mean robust. Another approach discussed earlier, which we shall not discuss again here, is to use the sample median as a measure of location. The three approaches are modifications of the traditional mean statistic $\overline{x}$. Of concern in computing the sample mean is the effect that an outlier will have. An observation far away from the main data set can have an enormous effect on the sample mean. One would like to eliminate or lessen the effect of such outlying and possibly spurious observations.

An approach that has been suggested is the α-trimmed mean. With the α-trimmed mean, we take some of the largest and smallest observations and drop them from each end. We then compute the usual sample mean on the data remaining.

Definition 8.11. The *α-trimmed mean* of n observations is computed as follows: Let k be the smallest integer greater than or equal to αn. Let $X_{(i)}$ be the order statistics of the sample. The α-trimmed mean drops approximately a proportion α of the observations from both ends of the distribution. That is,

$$\alpha\text{-trimmed mean} = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} X_{(i)}$$

We move on to the two other ways of modifying the mean, and then illustrate all three with a data set. The second method of modifying the mean is called *Winsorization*. The α-trimmed mean drops the largest and smallest observations from the samples. In the Winsorized mean, such observations are included, but the large effect is reduced. The approach is to shrink the smallest and largest observations to the next remaining observations, and count them as if they had those values. This will become clearer with the example below.

Definition 8.12. The *α-Winsorized mean* is computed as follows. Let k be the smallest integer greater than or equal to αn. The α-Winsorized mean is

$$\alpha\text{-Winsorized mean} = \frac{1}{n}\left[(k+1)(X_{(k+1)} + X_{(n-k)}) + \sum_{i=k+2}^{n-k-1} X_{(i)}\right]$$

The third method is to weight observations differentially. In general, we would want to weight the observations at the ends or tails less and those in the middle more. Thus, we will base the weights on the order statistics where the weights for the first few order statistics and the last few order statistics are typically small. In particular, we define the weighted mean to be

$$\text{weighted mean} = \frac{\sum_{i=1}^{n} W_i X_{(i)}}{\sum_{i=1}^{n} W_i}, \qquad \text{where } W_i \geq 0$$

Problem 8.26 shows that the α-trimmed mean and the α-Winsorized mean are examples of weighted means with appropriately chosen weights.

Example 8.12. We compute the mean, median, 0.1-trimmed mean, and 0.1-Winsorized mean for the female treadmill data of Example 8.6.

$$\text{mean} = \overline{x} = \frac{30 + \cdots + 936}{20} = 708$$

$$\text{median} = \frac{X_{(10)} + X_{(11)}}{2} = 750$$

Now $0.1 \times 20 = 2$, so $k = 2$.

$$\alpha\text{-trimmed mean} = \frac{570 + \cdots + 900}{16} = 734.6$$

$$\alpha\text{-Winsorized mean} = \frac{1}{20}(3(579 + 900) + 618 + \cdots + 840) = 734.7$$

Note that the median and both robust mean estimates are considerably higher than the sample mean $\overline{x}$. This is because of the small outlier of 30.

The Winsorized mean was intended to give outlying observations the same influence on the estimate as the most extreme of the interior estimates. In fact, the trimmed mean does this and the Winsorized mean gives outlying observations rather more influence. This, combined with the simplicity of the trimmed mean, makes it more attractive.

Robust techniques apply in a much more general context than shown here, and indeed are more useful in other situations. In particular, for regression and multiple regression (subjects of subsequent chapters in this book), a large amount of statistical theory has been developed for making the procedures more robust [Huber, 1981].

*8.12 FURTHER READING AND DIRECTIONS

There are several books dealing with nonparametric statistics. Among these are Lehmann and D'Abrera [1998] and Kraft and van Eeden [1968]. Other books deal exclusively with nonparametric statistical techniques. Three that are accessible on a mathematical level suitable for readers of this book are Marascuilo and McSweeney [1977], Bradley [1968], and Siegel and Castellan [1990].

A book that gives more of a feeling for the mathematics involved at a level above this text but which does not require calculus is Hajek [1969]. Another very comprehensive text that outlines much of the theory of statistical tests but is on a somewhat more advanced mathematical level, is Hollander and Wolfe [1999]. Finally, a comprehensive text on robust methods, written at a very advanced mathematical level, is Huber [2003].

In other sections of this book we give nonparametric and robust techniques in more general settings. They may be identified by one of the words *nonparametric, distribution-free*, or *robust* in the title of the section.

NOTES

8.1 Definitions of Nonparametric and Distribution-Free

The definitions given in this chapter are close to those of Huber [2003]. Bradley [1968] states that "roughly speaking, a nonparametric test is a test which makes no hypothesis about the value of a parameter in a statistical density function, whereas a distribution-free test is one which makes no assumptions about the precise form of the sampled population."

8.2 Relative Efficiency

The statements about relative efficiency in this chapter refer to asymptotic relative efficiency [Bradley, 1968; Hollander and Wolfe, 1999; Marascuilo and McSweeney, 1977]. For two possible *estimates*, the asymptotic relative efficiency of A to B is the limit of the ratio of the variance of B to the variance of A as the sample size increases. For two possible *tests*, first select a sequence of alternatives such that as n becomes large, the power (probability of rejecting the null hypothesis) for test A converges to a fixed number greater than zero and less than 1. Let this number be C. For each member of the sequence, find sample sizes n_A and n_B such that both tests have (almost) power C. The limit of the ratio n_B to n_A is the asymptotic relative efficiency. Since the definition is for large sample sizes (asymptotic), for smaller sample sizes the efficiency may be more or less than the figures we have given. Both Bradley [1968] and Hollander and Wolfe [1999] have considerable information on the topic.

8.3 Crossover Designs for Drugs

These are subject to a variety of subtle differences. There may be carryover effects from the drugs. Changes over time—for example, extreme weather changes—may make the second part of the crossover design different than the first. Some drugs may permanently change the subjects in some way. Peterson and Fisher [1980] give many references germane to randomized clinical trials.

8.4 Signed Rank Test

The values of the ranks are known; for n observations, they are the integers $1 - n$. The only question is the sign of the observation associated with each rank. Under the null hypothesis, the sign is equally likely to be plus or minus. Further, knowing the rank of an observation based on the absolute values does not predict the sign, which is still equally likely to be plus or minus independently of the other observations. Thus, all 2^n patterns of plus and minus signs are equally likely. For $n = 2$, the four patterns are:

Ranks	1	2	1	2	1	2	1	2
Signs	−	−	+	−	−	+	+	+
S	0		1		2		3	

So $P[S \leq 0] = 1/4$, $P[S \leq 1] = 1/2$, $P[S \leq 2] = 3/4$, and $P[S \leq 3] = 1$.

8.5 Quantile Test

If the alternative hypothesis of interest is an increase in extreme values of the outcome variable, a more powerful rank test can be based on the number of values above a given threshold. That is, the outcome value X_i is recoded to 1 if it is above the threshold and 0 if it is below the threshold. This recoding reduces the data to a 2×2 table, and Fisher's exact test can be used to make the comparison (see Section 6.3). Rather than prespecifying a threshold, one could specify that the threshold was to be, say, the 90th percentile of the combined sample. Again the data would be recoded to 1 for an observation in the top 10%, 0 for other observations, giving a 2 × 2 table. It is important that either a threshold or a percentile be specified in advance. Selecting the threshold that gives the largest difference in proportions gives a test related to the Kolmogorov–Smirnov test, and when proper control of the implicit multiple comparisons is made, this test is not particularly powerful.

8.6 Transitivity

One disadvantage of the rank tests is that they are not necessarily *transitive*. Suppose that we conclude from the Mann–Whitney test that group A has larger values than group B, and group B has larger values than group C. It would be natural to assume that group A has larger values than group C, but the Mann–Whitney test could conclude the reverse—that C was larger than A. This fact is important in the theory of elections, where different ways of running elections are generally equivalent to different rank tests. It implies that candidate A could beat B, B could beat C, and C could beat A in fair two-way runoff elections, a problem noted in the late eighteenth century by Condorcet. Many interesting issues related to nontransitivity were discussed in Martin Gardner's famous "Mathematical Games" column in *Scientific American* of December 1970, October 1974, and November 1997.

The practical importance of nontransitivity is unclear. It is rare in real data, so may largely be a philosophical issue. On the other hand, it does provide a reminder that the rank-based tests are not just a statistical garbage disposal that can be used for any data whose distribution is unattractive.

8.7 Kolmogorov–Smirnov Statistic Is a Rank Statistic

We illustrate one technique used to show that the Kolmogorov–Smirnov statistic is a rank test. Looking at Figure 8.2, we could slide both curves along the x-axis without changing the value of the maximum difference, D. Since the curves are horizontal, we can stretch them along the axis (as long as the order of the jumps does not change) and not change the value of D. Place the first jump at 1, the second at 2, and so on. We have placed the jumps then at the ranks! The height of the jumps depends on the sample size. Thus, we can compute D from the ranks (and knowing which group have the rank) and the sample sizes. Thus, D is nonparametric and distribution-free.

8.8 One-Sample Kolmogorov–Smirnov Tests and One-Sided Kolmogorov–Smirnov Tests

It is possible to compare one sample to a hypothesized distribution. Let F be the empirical cumulative distribution function of a sample. Let H be a hypothesized distribution function. The statistic

$$D = \max_x |F(x) - H(x)|$$

is the one-sample statistic. If H is continuous, critical values are tabulated for this nonparametric test in the tables already cited in this chapter. An approximation to the p-value for the one-sample Kolmogorov–Smirnov test is

$$P(D > d) \leq 2e^{-2d^2/n}$$

This is conservative regardless of sample size, the value of d, the presence or absence of ties, and the true underlying distribution F, and is increasingly accurate as the p-value decreases. This approximation has been known for a long time, but the fact that it is guaranteed to be conservative is a recent, very difficult mathematical result [Massart, 1990].

The Kolmogorov–Smirnov two-sample statistic was based on the largest difference between two empirical cumulative distribution functions; that is,

$$D = \max_x |F(x) - G(x)|$$

where F and G are the two empirical cumulative distribution functions. Since the absolute value is involved, we are not differentiating between F being larger and G being larger. If we had hypothesized as an alternative that the F population took on larger values in general, F would tend to be less than G, and we could use

$$D^+ = \max_x (G(x) - F(x))$$

Such one-sided Kolmogorov–Smirnov statistics are used and tabulated. They also are nonparametric rank tests for use with one-sided alternatives.

8.9 More General Rank Tests

The theory of tests based on ranks is well developed [Hajek, 1969; Hajek and Sidak, 1999; Huber, 2003]. Consider the two-sample problem with groups of size n and m, respectively. Let $R_i (i = 1, 2, \ldots, n)$ be the ranks of the first sample. Statistics of the following form, with a a function of R_i, have been studied extensively.

$$S = \frac{1}{n} \sum_{i=1}^{n} a(R_i)$$

The $a(R_i)$ may be chosen to be efficient in particular situations. For example, let $a(R_i)$ be such that a standard normal variable has probability $R_i/(n+m+1)$ of being less than or equal to this value. Then, when the usual two-sample t-test normal assumptions hold, the relative efficiency is 1. That is, this rank test is as efficient as the t-test for large samples. This test is called the *normal scores test* or *van der Waerden test.*

8.10 Monte Carlo Technique and Pseudorandom Number Generators

The term *Monte Carlo technique* was introduced by the mathematician Stanislaw Ulam [1976] while working on the Manhattan atomic bomb project.

Computers typically do not generate random numbers; rather, the numbers are generated in a sequence by a specific computer algorithm. Thus, the numbers are called *pseudorandom numbers*. Although not random, the sequence of numbers need to appear random. Thus, they are tested in part by statistical tests. For example, a program to generate random integers from zero to nine may have a sequence of generated integers tested by the χ^2 goodness-of-fit test to see that the "probability" of each outcome is 1/10. A generator of uniform numbers on the interval (0, 1) can have its empirical distribution compared to the uniform distribution by the one-sample Kolmogorov–Smirnov test (Note 8.8). The subject of pseudorandom number generators is very deep both philosophically and mathematically. See Chaitin [1975] and Dennett [1984, Chaps. 5 and 6] for discussions of some of the philosophical issues, the former from a mathematical viewpoint.

Computer and video games use pseudorandom number generation extensively, as do computer security systems. A number of computer security failures have resulted from poor-quality pseudorandom number generators being used in encryption algorithms. One can generally assume that the generators provided in statistical packages are adequate for statistical (not cryptographic) purposes, but it is still useful to repeat complex simulation experiments with a different generator if possible. A few computer systems now have "genuine" random number generators that collect and process randomness from sources such as keyboard and disk timings.

PROBLEMS

8.1 The following data deal with the treatment of essential hypertension (*essential* is a technical term meaning that the cause is unknown; a synonym is *idiopathic*) and is from a paper by Vlachakis and Mendlowitz [1976]. Seventeen patients received treatments C, A, and B, where C is the control period, A is propranolol+phenoxybenzamine, and B is propranolol + phenoxybenzamine + hydrochlorothiazide. Each patient received C first, then either A or B, and finally, B or A. The data in Table 8.6 consist of the systolic blood pressure in the recumbent position.

Table 8.6 Blood Pressure Data for Problem 8.1

Patient	C	A	B	Patient	C	A	B
1	185	148	132	10	180	132	136
2	160	128	120	11	176	140	135
3	190	144	118	12	200	165	144
4	192	158	115	13	188	140	115
5	218	152	148	14	200	140	126
6	200	135	134	15	178	135	140
7	210	150	128	16	180	130	130
8	225	165	140	17	150	122	132
9	190	155	138				

(a) Take the differences between the systolic blood pressures on treatments A and C. Use the sign test to test for a treatment A effect (two-sided test; give the p-value).

(b) Take the differences between treatments B and C. Use the sign test to test for a treatment B effect (one-sided test; give the p-value).

(c) Take the differences between treatments B and A. Test for a treatment difference using the sign test (two-sided test; give the p-value).

8.2 Several population studies have demonstrated an inverse correlation of sudden infant death syndrome (SIDS) rate with birthweight. The occurrence of SIDS in one of a pair of twins provides an opportunity to test the hypothesis that birthweight is a major determinant of SIDS. The set of data in Table 8.7 was collected by D. R. Peterson of the

Table 8.7 Birthweight Data for Problem 8.2

Dizygous Twins		Monozygous Twins		Dizygous Twins		Monozygous Twins	
SIDS	Non-SIDS	SIDS	Non-SIDS	SIDS	Non-SIDS	SIDS	Non-SIDS
1474	2098	1701	1956	2381	2608	1956	1588
3657	3119	2580	2438	2892	2693	2296	2183
3005	3515	2750	2807	2920	3232	3232	2778
2041	2126	1956	1843	3005	3005	1446	2268
2325	2211	1871	2041	2268	2325	1559	1304
2296	2750	2296	2183	3260	3686	2835	2892
3430	3402	2268	2495	3260	2778	2495	2353
3515	3232	2070	1673	2155	2552	1559	2466
1956	1701	1786	1843	2835	2693		
2098	2410	3175	3572	2466	1899		
3204	2892	2495	2778	3232	3714		

Department of Epidemiology, University of Washington, consists of the birthweights of each of 22 dizygous twins and each of 19 monozygous twins.

(a) For the dizygous twins test the alternative hypothesis that the SIDS child of each pair has the lower birthweight by taking differences and using the sign test. Find the one-sided p-value.

(b) As in part (a), but do the test for the monozygous twins.

(c) As in part (a), but do the test for the combined data set.

8.3 The following data are from Dobson et al. [1976]. Thirty-six patients with a confirmed diagnosis of phenylketonuria (PKU) were identified and placed on dietary therapy before reaching 121 days of age. The children were tested for IQ (Stanford–Binet test) between the ages of 4 and 6; subsequently, their normal siblings of closest age were also tested with the Stanford–Binet. The 15 pairs shown in Table 8.8 are the first 15 listed in the paper. The null hypothesis is that the PKU children, on average, have the same IQ as their siblings. Using the sign test, find the two-sided p-value for testing against the alternative hypothesis that the IQ levels differ.

Table 8.8 PKU/IQ Data for Problem 8.3

Pair	IQ of PKU Case	IQ of Sibling	Pair	IQ of PKU Case	IQ of Sibling
1	89	77	9	110	88
2	98	110	10	90	91
3	116	94	11	76	99
4	67	91	12	71	93
5	128	122	13	100	104
6	81	94	14	108	102
7	96	121	15	74	82
8	116	114			

8.4 Repeat Problem 8.1 using the signed rank test rather than the sign test. Test at the 0.05 significance level.

8.5 Repeat Problem 8.2, parts (a) and (b), using the signed rank test rather than the sign test. Test at the 0.05 significance level.

8.6 Repeat Problem 8.3 using the signed rank test rather than the sign test. Test at the 0.05 significance level.

8.7 Bednarek and Roloff [1976] deal with the treatment of apnea (a transient cessation of breathing) in premature infants using a drug called aminophylline. The variable of interest, "average number of apneic episodes per hour," was measured before and after treatment with the drug. An episode was defined as the absence of spontaneous breathing for more than 20 seconds, or less if associated with bradycardia or cyanosis. Table 8.9 details the response of 13 patients to aminophylline treatment at 16 hours compared with 24 hours before treatment (in apneic episodes per hour).

(a) Use the sign test to examine a treatment effect (give the two-sided p-value).

(b) Use the signed rank test to examine a treatment effect (two-sided test at the 0.05 significance level).

Table 8.9 Before/After Treatment Data for Problem 8.7

Patient	24 Hours Before	16 Hours After	Before–After (Difference)
1	1.71	0.13	1.58
2	1.25	0.88	0.37
3	2.13	1.38	0.75
4	1.29	0.13	1.16
5	1.58	0.25	1.33
6	4.00	2.63	1.37
7	1.42	1.38	0.04
8	1.08	0.50	0.58
9	1.83	1.25	0.58
10	0.67	0.75	−0.08
11	1.13	0.00	1.13
12	2.71	2.38	0.33
13	1.96	1.13	0.83

8.8 The following data from Schechter et al. [1973] deal with sodium chloride preference as related to hypertension. Two groups, 12 normal and 10 hypertensive subjects, were isolated for a week and compared with respect to Na^+ intake. The average daily Na^+ intakes are listed in Table 8.10. Compare the average daily Na^+ intake of the hypertensive subjects with that of the normal volunteers by means of the Wilcoxon two-sample test at the 5% significance level.

Table 8.10 Sodium Data for Problem 8.8

Normal	Hypertensive	Normal	Hypertensive
10.2	92.8	45.8	34.7
2.2	54.8	63.6	62.2
0.0	51.6	1.8	11.0
2.6	61.7	0.0	39.1
0.0	250.8	3.7	
43.1	84.5	0.0	

8.9 During July and August 1976, a large number of Legionnaires attending a convention died of a mysterious and unknown cause. Epidemiologists have talked of "an outbreak of Legionnaires' disease." Chen et al. [1977] examined the hypothesis of nickel contamination as a toxin. They examined the nickel levels in the lungs of nine cases and nine controls. The authors point out that contamination at autopsy is a possibility. The data are as follows (μg per 100 g dry weight):

Legionnaire Cases	65	24	52	86	120	82	399	87	139
Control Cases	12	10	31	6	5	5	29	9	12

Note that there was no attempt to match cases and controls. Use the Wilcoxon test at the one-sided 5% level to test the null hypothesis that the numbers are samples from similar populations.

Table 8.11 Plasma iPGE Data for Problem 8.10

Patient Number	Mean Plasma iPGE (pg/mL)	Mean Serum Calcium (ml/dL)
	Patients with Hypercalcemia	
1	500	13.3
2	500	11.2
3	301	13.4
4	272	11.5
5	226	11.4
6	183	11.6
7	183	11.7
8	177	12.1
9	136	12.5
10	118	12.2
11	60	18.0
	Patients without Hypercalcemia	
12	254	10.1
13	172	9.4
14	168	9.3
15	150	8.6
16	148	10.5
17	144	10.3
18	130	10.5
19	121	10.2
20	100	9.7
21	88	9.2

8.10 Robertson et al. [1976] discuss the level of plasma prostaglandin E (iPGE in pg/mL) in patients with cancer with and without hypercalcemia. The data are given in Table 8.11. Note that the variables are "mean plasma iPGE" and "mean serum Ca" levels; presumably more than one assay was carried out for each patient's level. The number of such tests for each patient is not indicated, nor is the criterion for the number. Using the Wilcoxon two-sample test, test for differences between the two groups in:

(a) Mean plasma iPGE.

(b) Mean serum Ca.

8.11 Sherwin and Layfield [1976] present data about protein leakage in the lungs of male mice exposed to 0.5 part per million of nitrogen dioxide (NO_2). Serum fluorescence data were obtained by sacrificing animals at various intervals. Use the two-sided Wilcoxon test, 0.05 significance level, to look for differences between controls and exposed mice.

(a) At 10 days:

Controls	143	169	95	111	132	150	141
Exposed	152	83	91	86	150	108	78

(b) At 14 days:

Controls	76	40	119	72	163	78
Exposed	119	104	125	147	200	173

8.12 Using the data of Problem 8.8:

(a) Find the value of the Kolmogorov–Smirnov statistic.

(b) Plot the two empirical distribution functions.

(c) Do the curves differ at the 5% significance level? For sample sizes 10 and 12, the 10%, 5%, and 1% critical values for *mnD* are 60, 66, and 80, respectively.

8.13 Using the data of Problem 8.9:

(a) Find the value of the Kolmogorov–Smirnov statistic.

(b) Do you reject the null hypothesis at the 5% level? For $m = 9$ and $n = 9$, the 10%, 5%, and 1% critical values of *mnD* are 54, 54, and 63, respectively.

8.14 Using the data of Problem 8.10:

(a) Find the value of the Kolmogorov–Smirnov statistic for both variables.

(b) What can you say about the p-value? For $m = 10$ and $n = 11$, the 10%, 5%, and 1% critical values of *mnD* are 57, 60, and 77, respectively.

8.15 Using the data of Problem 8.11:

(a) Find the value of the Kolmogorov–Smirnov statistic.

(b) Do you reject at 10%, 5%, and 1%, respectively? Do this for parts (a) and (b) of Problem 8.11. For $m = 7$ and $n = 7$, the 10%, 5%, and 1% critical values of *mnD* are 35, 42, and 42, respectively. The corresponding critical values for $m = 6$ and $n = 6$ are 30, 30, and 36.

8.16 Test at the 0.05 significance level for a significant improvement with the cream treatment of Example 8.2.

(a) Use the sign test.

(b) Use the signed rank test.

(c) Use the t-test.

8.17 Use the expression of colostrum data of Example 8.2, and test at the 0.10 significance level the null hypothesis of no treatment effect.

(a) Use the sign test.

(b) Use the signed rank test.

(c) Use the usual t-test.

8.18 Test the null hypothesis of no treatment difference from Example 8.2 using each of the tests in parts (a), (b), and (c).

(a) The Wilcoxon two-sample test.

(b) The Kolmogorov–Smirnov two-sample test. For $m = n = 19$, the 20%, 10%, 5%, 1%, and 0.1% critical values for *mnD* are 133, 152, 171, 190, and 228, respectively.

(c) The two-sample t-test.

Compare the two-sided p-values to the extent possible. Using the data of Example 8.2, examine each treatment.

(d) Nipple-rolling vs. masse cream.

(e) Nipple-rolling vs. expression of colostrum.

(f) Masse cream vs. expression of colostrum.

8.19 As discussed in Chapter 3, Winkelstein et al. [1975] studied systolic blood pressures of three groups of Japanese men: native Japanese, first-generation immigrants to the United States (Issei), and second-generation Japanese in the United States (Nisei). The data are listed in Table 8.12. Use the asymptotic Wilcoxon two-sample statistic to test:

(a) Native Japanese vs. California Issei.

(b) Native Japanese vs. California Nisei.

(c) California Issei vs. California Nisei.

Table 8.12 Blood Pressure Data for Problem 8.19

Blood Pressure (mmHg)	Native Japanese	Issei	Nisei
<106	218	4	23
106–114	272	23	132
116–124	337	49	290
126–134	362	33	347
136–144	302	41	346
146–154	261	38	202
156–164	166	23	109
>166	314	52	112

***8.20** Rascati et al. [2001] report a study of medical costs for children with asthma in which children prescribed steroids had a higher mean cost than other children, but lower costs according to a Wilcoxon rank-sum test. How can this happen, and what conclusions should be drawn?

***8.21** An outlier is an observation far from the rest of the data. This may represent valid data or a mistake in experimentation, data collection, or data entry. At any rate, a few outlying observations may have an extremely large effect. Consider a one-sample t-test of mean zero based on 10 observations with

$$\overline{x} = 10 \quad \text{and} \quad s^2 = 1$$

Suppose now that one observation of value x is added to the sample.

(a) Show that the value of the new sample mean, variance, and t-statistic are

$$\overline{x} = \frac{100 + x}{11}$$

$$s^2 = \frac{10x^2 - 200x + 1099}{11 \times 10}$$

$$t = \frac{100 + x}{\sqrt{x^2 - 20x + 109.9}}$$

***(b)** Graph t as a function of x.

(c) For which values of x would one reject the null hypothesis of mean zero? What does the effect of an outlier (large absolute value) do in this case?

(d) Would you reject the null hypothesis without the outlier?

(e) What would the graph look like for the Wilcoxon signed rank test? For the sign test?

***8.22** Using the ideas of Note 8.4 about the signed rank test, verify the values shown in Table 8.13 when $n = 4$.

Table 8.13 Signed-Rank Test Data for Problem 8.23

s	$P[S \leq s]$	s	$P[S \leq s]$
0	0.062	6	0.688
1	0.125	7	0.812
2	0.188	8	0.875
3	0.312	9	0.938
4	0.438	10	1.000
5	0.562		

Source: Owen [1962]; by permission of Addison-Wesley Publishing Company.

***8.23** The Wilcoxon two-sample test depends on the fact that under the null hypothesis, if two samples are drawn without ties, all $\binom{n+m}{n}$ arrangements of the n ranks from the first sample, and the m ranks from the second sample, are equally likely. That is, if $n = 1$ and $m = 2$, the three arrangements

$$\begin{array}{llll} \mathbf{1} & 2 & 3; & W = 1 \\ 1 & \mathbf{2} & 3; & W = 2 \\ 1 & 2 & \mathbf{3}; & W = 3 \end{array}$$

are equally likely. Here, the rank from population 1 appears in bold type.

(a) If $n = 2$ and $m = 4$, graph the distribution function of the Wilcoxon two-sample statistic when the null hypothesis holds.

(b) Find $E(W)$. Does it agree with equation (5)?

(c) Find $\text{var}(W)$. Does it agree with equation (6)?

***8.24** (Permutation Two-Sample t-Test) To use the permutation two-sample t-test, the text (in Section *8.9) used the fact that for $n + m$ fixed values, the t-test was a monotone function of $\overline{x} - \overline{y}$. To show this, prove the following equality:

$$t = \frac{1}{\sqrt{\dfrac{(n+m)\left(\sum_i x_i^2 + \sum_i y_i^2\right) - \left(\sum_i x_i + \sum_i y_i\right)^2 - nm(\overline{x} - \overline{y})^2}{nm(n+m-2)(\overline{x} - \overline{y})^2}}}$$

Note that the first two terms in the numerator of the square root are constant for all permutations, so t is a function of $\overline{x} - \overline{y}$.

***8.25** (One-Sample Randomization t-Test) For the randomization one-sample t-test, the paired x_i and y_i values give $\overline{x}-\overline{y}$ values. Assume that the $|x_i - y_i|$ are known but the signs are random, independently $+$ or $-$ with probability 1/2. The $2^n (i = 1, 2, \dots, n)$ patterns of pluses and minuses are equally likely.

(a) Show that the one-sample t-statistic is a monotone function of $\overline{x-y}$ when the $|x_i - y_i|$ are known. Do this by showing that

$$t = \frac{\overline{x-y}}{\sqrt{\left[-n(\overline{x-y})^2 + \sum_i (x_i - y_i)^2\right]/n(n-1)}}$$

(b) For the data

i	X_i	Y_i
1	1	2
2	3	1
3	1	5

compute the eight possible randomization values of t. What is the two-sided randomization p-value for the t observed?

***8.26** (Robust Estimation of the Mean) Show that the α-trimmed mean and the α-Winsorized mean are weighted means by explicitly showing the weights W_i that are given the two means.

***8.27** (Robust Estimation of the Mean)

(a) For the combined data for SIDS in Problem 8.2, compute (**i**) the 0.05 trimmed mean; (**ii**) the 0.05 Winsorized mean; (**iii**) the weighted mean with weights $W_i = i(n + 1 - i)$, where n is the number of observations.

(b) The same as in Problem 8.27(a), but do this for the non-SIDS twins.

REFERENCES

Alderman, E., Fisher, L. D., Maynard, C., Mock, M. B., Ringqvist, I., Bourassa, M. G., Kaiser, G. C., and Gillespie, M. J. [1982]. Determinants of coronary surgery in a consecutive patient series from geographically dispersed medical centers: the Coronary Artery Surgery Study. *Circulation*, **66**: 562–568.

Bednarek, E., and Roloff, D. W. [1976]. Treatment of apnea of prematurity with aminophylline. *Pediatrics*, **58**: 335–339.

Beyer, W. H. (ed.) [1990]. *CRC Handbook of Tables for Probability and Statistics*. 2nd ed. CRC Press, Boca Raton, FL.

Bradley, J. V. [1968]. *Distribution-Free Statistical Tests*. Prentice Hall, Englewood Cliffs, NJ.

Brown, M. S., and Hurlock, J. T. [1975]. Preparation of the breast for breast-feeding. *Nursing Research*, **24**: 448–451.

CASS [1981]. (Principal investigators of CASS and their associates; Killip, T. (ed.); Fisher, L. D., and Mock, M. (assoc. eds.) National Heart, Lung and Blood Institute Coronary Artery Surgery Study. *Circulation*, **63**: part II, I–1 to I–81. Used with permission from the American Heart Association.

Chaitin, G. J. [1975]. Randomness and mathematical proof, *Scientific American*, **232**(5): 47–52.

Chen, J. R., Francisco, R. B., and Miller, T. E. [1977]. Legionnaires' disease: nickel levels. *Science*, **196**: 906–908.

Church, J. D., and Harris, B. [1970]. The estimation of reliability from stress–strength relationships. *Technometrics*, **12**: 49–54.

Davison, A. C., and Hinckley, D. V. [1997]. *Bootstrap Methods and Their Application*. Cambridge University Press, New York.

Dennett, D. C. [1984]. *Elbow Room: The Varieties of Free Will Worth Wanting*. MIT Press, Cambridge, MA.

Dobson, J. C., Kushida, E., Williamson, M., and Friedman, E. [1976]. Intellectual performance of 36 phenylketonuria patients and their nonaffected siblings. *Pediatrics*, **58**: 53–58.

Edgington, E. S. [1995]. *Randomization Tests*, 3rd ed. Marcel Dekker, New York.

Efron, B. [1979]. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, **7**: 1–26.

Efron, B. [1982]. *The Jackknife, Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia.

Efron, B., and Tibshirani, R. [1986]. The bootstrap (with discussion). *Statistical Science*, **1**: 54–77.

Efron, B., and Tibshirani, R. [1993]. *An Introduction to the Bootstrap*. Chapman & Hall, London.

Hajek, J. [1969]. *A Course in Nonparametric Statistics*. Holden-Day, San Francisco.

Hajek, J., and Sidak, Z. [1999]. *Theory of Rank Tests*. 2nd ed. Academic Press, New York.

Hoffman, D. T. [1979]. *Monte Carlo: The Use of Random Digits to Simulate Experiments*. Models and monographs in undergraduate mathematics and its Applications, Unit 269, EDC/UMAP, Newton, MA.

Hollander, M., and Wolfe, D. A. [1999]. *Nonparametric Statistical Methods*, 2nd ed. Wiley, New York.

Huber, P. J. [2003]. *Robust Statistics*. Wiley, New York.

Johnson, R. A., Verill, S., and Moore D. H. [1987]. Two-sample rank tests for detecting changes that occur in a small proportion of the treated population. *Biometrics*, **43**: 641–655

Kraft, C. H., and van Eeden, C. [1968]. *A Nonparametric Introduction to Statistics*. Macmillan, New York.

Lehmann, E. L., and D'Abrera, H. J. M. [1998]. *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco.

Lumley, T., Diehr, P., Emerson, S., and Chen, L. [2002]. The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, **23**: 151–169.

Marascuilo, L. A., and McSweeney, M. [1977]. *Nonparametric and Distribution-Free Methods for the Social Sciences*. Brooks/Cole, Scituate, MA.

Massart, P. [1990]. The tight constant in the Dvoretsky-Kiefer-Wolfowitz inequality. *Annals of Probability*, **18**: 897–919.

Odeh, R. E., Owen, D. B., Birnbaum, Z. W., and Fisher, L. D. [1977]. *Pocket Book of Statistical Tables*. Marcel Dekker, New York.

Owen, D. B. [1962]. *Handbook of Statistical Tables*. Addison-Wesley, Reading, MA.

Peterson, A. P., and Fisher, L. D. [1980]. Teaching the principles of clinical trials design. *Biometrics*, **36**: 687–697.

Rascati, K. L., Smith, M. J., and Neilands, T. [2001]. Dealing with skewed data: an example using asthma-related costs of Medicaid clients. *Clinical Therapeutics*, **23**: 481–498.

Ripley B. D. [1987]. *Stochastic Simulation*. Wiley, New York.

Robertson, R. P., Baylink, D. J., Metz, S. A., and Cummings, K. B. [1976]. Plasma prostaglandin in patients with cancer with and without hypercalcemia. *Journal of Clinical Endocrinology and Metabolism*, **43**: 1330–1335.

Schechter, P. J., Horwitz, D., and Henkin, R. I. [1973]. Sodium chloride preference in essential hypertension. *Journal of the American Medical Association*, **225**: 1311–1315.

Sherwin, R. P., and Layfield, L. J. [1976]. Protein leakage in the lungs of mice exposed to 0.5 ppm nitrogen dioxide: a fluorescence assay for protein. *Archives of Environmental Health*, **31**: 116–118.

Siegel, S., and Castellan, N. J., Jr. [1990]. *Nonparametric Statistics for the Behavioral Sciences*, 2nd ed. McGraw-Hill, New York.

Ulam, S. M. [1976]. *Adventures of a Mathematician*. Charles Scribner's Sons, New York.

U.S. EPA [1994]. *Statistical Methods for Evaluating the Attainment of Cleanup Standards*, Vol. 3, *Reference-Based Standards for Soils and Solid Media*. EPA/600/R-96/005. Office of Research and Development, U.S. EPA, Washington, DC.

Vlachakis, N. D., and Mendlowitz, M. [1976]. Alpha- and beta-adrenergic receptor blocking agents combined with a diuretic in the treatment of essential hypertension. *Journal of Clinical Pharmacology*, **16**: 352–360.

Winkelstein, W., Jr., Kazan, A., Kato, H., and Sachs, S. T. [1975]. Epidemiologic studies of coronary heart disease and stroke in Japanese men living in Japan, Hawaii, and California: blood pressure distributions. *American Journal of Epidemiology*, **102**: 502–513.

CHAPTER 9

Association and Prediction: Linear Models with One Predictor Variable

9.1 INTRODUCTION

Motivation for the methods of this chapter is aided by the use of examples. For this reason, we first consider three data sets. These data are used to motivate the methods to follow. The data are also used to illustrate the methods used in Chapter 11. After the three examples are presented, we return to this introduction.

Example 9.1. Table 9.1 and Figure 9.1 contain data on mortality due to malignant melanoma of the skin of white males during the period 1950–1969 for each state in the United States as well as the District of Columbia. No mortality data are available for Alaska and Hawaii for this period. It is well known that the incidence of melanoma can be related to the amount of sunshine and, somewhat equivalently, the latitude of the area. The table contains the latitude as well as the longitude for each state. These numbers were obtained simply by estimating the center of the state and reading off the latitude as given in a standard atlas. Finally, the 1965 population and contiguity to an ocean are noted, where "1" indicates contiguity: the state borders one of the oceans.

In the next section we shall be particularly interested in the relationship between the melanoma mortality and the latitude of the states. These data are presented in Figure 9.1.

Definition 9.1. When two variables are collected for each data point, a plot is very useful. Such plots of the two values for each of the data points are called *scatter diagrams* or *scattergrams*.

Note several things about the scattergram of malignant melanoma rates vs. latitude. There appears to be a rough relationship. As the latitude increases, the melanoma rate decreases. Nevertheless, there is no one-to-one relationship between the values. There is considerable scatter in the picture. One problem is to decide whether or not the scatter could be due to chance or whether there is some relationship. It might be of interest to estimate the melanoma rate for various latitudes. In this case, how would we estimate the relationship? To convey the relationship to others, it would also be useful to have some simple way of summarizing the relationship. There are two aspects of the relationship that might be summarized. One is how the melanoma rate changes with latitude; it would also be useful to summarize the variability of the scattergram.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2 Copyright © 2004 John Wiley & Sons, Inc.

Table 9.1 Mortality Rate [per 10 Million (10^7)] of White Males Due to Malignant Melanoma of the Skin for the Period 1950–1959 by State and Some Related Variables

State	Mortality per 10,000,000	Latitude (deg)	Longitude (deg)	Population (millions, 1965)	Ocean State[a]
Alabama	219	33.0	87.0	3.46	1
Arizona	160	34.5	112.0	1.61	0
Arkansas	170	35.0	92.5	1.96	0
California	182	37.5	119.5	18.60	1
Colorado	149	39.0	105.5	1.97	0
Connecticut	159	41.8	72.8	2.83	1
Delaware	200	39.0	75.5	0.50	1
Washington, DC	177	39.0	77.0	0.76	0
Florida	197	28.0	82.0	5.80	1
Georgia	214	33.0	83.5	4.36	1
Idaho	116	44.5	114.0	0.69	0
Illinois	124	40.0	89.5	10.64	0
Indiana	128	40.2	86.2	4.88	0
Iowa	128	42.2	93.8	2.76	0
Kansas	166	38.5	98.5	2.23	0
Kentucky	147	37.8	85.0	3.18	0
Louisiana	190	31.2	91.8	3.53	1
Maine	117	45.2	69.0	0.99	1
Maryland	162	39.0	76.5	3.52	1
Massachusetts	143	42.2	71.8	5.35	1
Michigan	117	43.5	84.5	8.22	0
Minnesota	116	46.0	94.5	3.55	0
Mississippi	207	32.8	90.0	2.32	1
Missouri	131	38.5	92.0	4.50	0
Montana	109	47.0	110.5	0.71	0
Nebraska	122	41.5	99.5	1.48	0
Nevada	191	39.0	117.0	0.44	0
New Hampshire	129	43.8	71.5	0.67	1
New Jersey	159	40.2	74.5	6.77	1
New Mexico	141	35.0	106.0	1.03	0
New York	152	43.0	75.5	18.07	1
North Carolina	199	35.5	79.5	4.91	1
North Dakota	115	47.5	100.5	0.65	0
Ohio	131	40.2	82.8	10.24	0
Oklahoma	182	35.5	97.2	2.48	0
Oregon	136	44.0	120.5	1.90	1
Pennsylvania	132	40.8	77.8	11.52	0
Rhode Island	137	41.8	71.5	0.92	1
South Carolina	178	33.8	81.0	2.54	1
South Dakota	86	44.8	100.0	0.70	0
Tennessee	186	36.0	86.2	3.84	0
Texas	229	31.5	98.0	10.55	1
Utah	142	39.5	111.5	0.99	0
Vermont	153	44.0	72.5	0.40	1
Virginia	166	37.5	78.5	4.46	1
Washington	117	47.5	121.0	2.99	1
West Virginia	136	38.8	80.8	1.81	0
Wisconsin	110	44.5	90.2	4.14	0
Wyoming	134	43.0	107.5	0.34	0

Source: U.S. Department of Health, Education, and Welfare [1974].

[a] 1 = state borders on ocean.