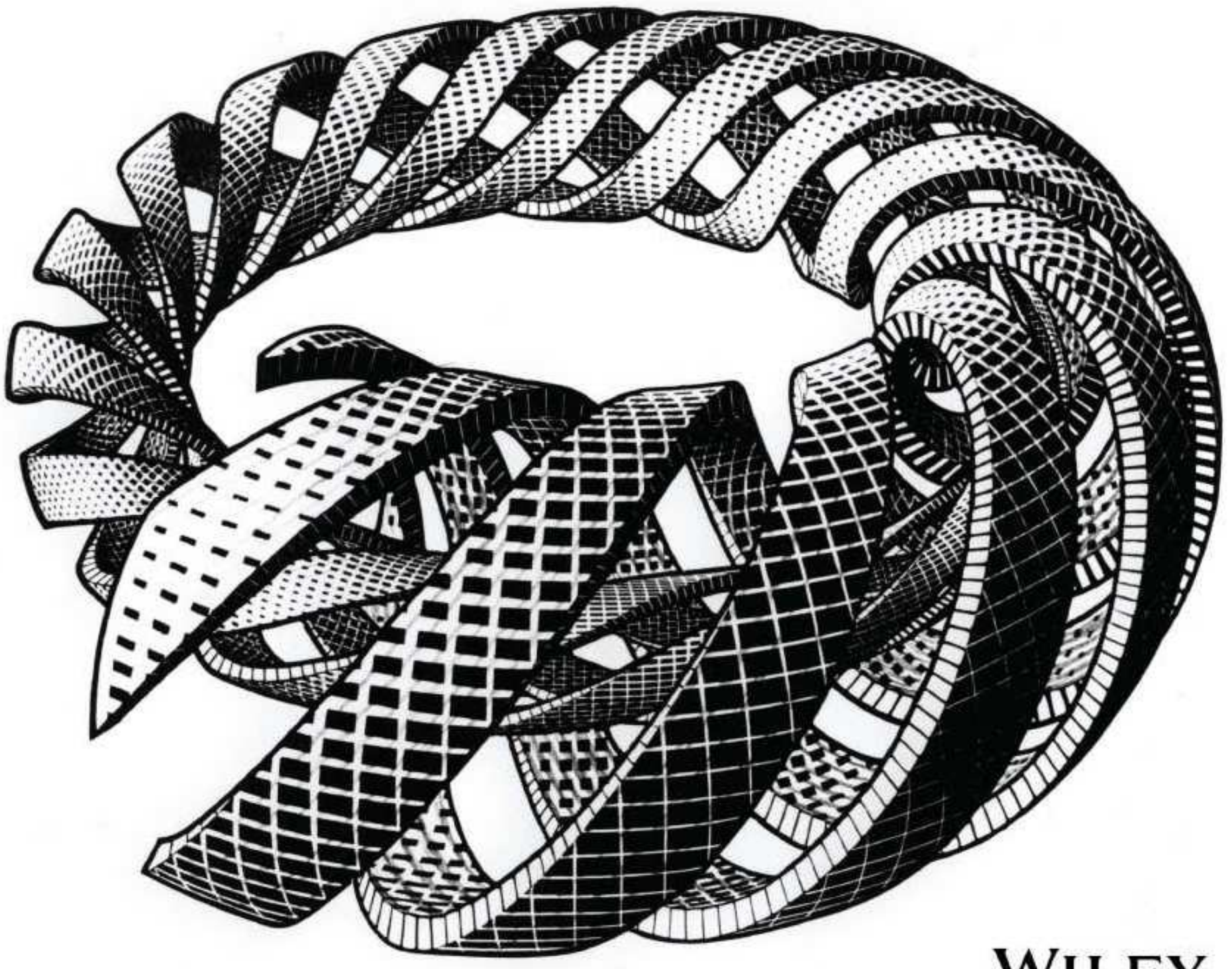
ELEVENTH EDITION

STATISTICS

ROBERT S. **WITTE**

JOHN S. **WITTE**



WILEY

STATISTICS

Eleventh Edition

Robert S. Witte
Emeritus, San Jose State University

John S. Witte
University of California, San Francisco

WILEY

VP AND EDITORIAL DIRECTOR	George Hoffman
EDITORIAL DIRECTOR	Veronica Visentin
EDITORIAL ASSISTANT	Ethan Lipson
EDITORIAL MANAGER	Gladys Soto
CONTENT MANAGEMENT DIRECTOR	Lisa Wojcik
CONTENT MANAGER	Nichole Urban
SENIOR CONTENT SPECIALIST	Nicole Repasky
PRODUCTION EDITOR	Abidha Sulaiman
COVER PHOTO CREDIT	M.C. Escher's Spirals © The M.C. Escher Company - The Netherlands

This book was set in 10/11 Times LT Std by SPi Global and printed and bound by Lightning Source Inc. The cover was printed by Lightning Source Inc.

Founded in 1807, John Wiley & Sons, Inc. has been a valued source of knowledge and understanding for more than 200 years, helping people around the world meet their needs and fulfill their aspirations. Our company is built on a foundation of principles that include responsibility to the communities we serve and where we live and work. In 2008, we launched a Corporate Citizenship Initiative, a global effort to address the environmental, social, economic, and ethical challenges we face in our business. Among the issues we are addressing are carbon impact, paper specifications and procurement, ethical conduct within our business and among our vendors, and community and charitable support. For more information, please visit our website: www.wiley.com/go/citizenship.

ISBN: 978-1-119-25451-5(PBK)
ISBN: 978-1-119-25445-4(EVALC)

Library of Congress Cataloging-in-Publication Data

Names: Witte, Robert S. | Witte, John S.
Title: Statistics / Robert S. Witte, Emeritus, San Jose State University, John S. Witte, University of California, San Francisco.
Description: Eleventh edition. | Hoboken, NJ: John Wiley & Sons, Inc., [2017] | Includes index.
Identifiers: LCCN 2016036766 (print) | LCCN 2016038418 (ebook) | ISBN 9781119254515 (pbk.) | ISBN 9781119299165 (epub)
Subjects: LCSH: Statistics.
Classification: LCC QA276.12 .W57 2017 (print) | LCC QA276.12 (ebook) | DDC 519.5—dc23
LC record available at https://lccn.loc.gov/2016036766

The inside back cover will contain printing identification and country of origin if omitted from this page. In addition, if the ISBN on the back cover differs from the ISBN on this page, the one on the back cover is correct.

To Doris

Preface

TO THE READER

Students often approach statistics with great apprehension. For many, it is a required course to be taken only under the most favorable circumstances, such as during a quarter or semester when carrying a light course load; for others, it is as distasteful as a visit to a credit counselor—to be postponed as long as possible, with the vague hope that mounting debts might miraculously disappear. Much of this apprehension doubtless rests on the widespread fear of mathematics and mathematically related areas.

This book is written to help you overcome any fear about statistics. Unnecessary quantitative considerations have been eliminated. When not obscured by mathematical treatments better reserved for more advanced books, some of the beauty of statistics, as well as its everyday usefulness, becomes more apparent.

You could go through life quite successfully without ever learning statistics. Having learned some statistics, however, you will be less likely to flinch and change the topic when numbers enter a discussion; you will be more skeptical of conclusions based on loose or erroneous interpretations of sets of numbers; you might even be more inclined to initiate a statistical analysis of some problem within your special area of interest.

TO THE INSTRUCTOR

Largely because they panic at the prospect of any math beyond long division, many students view the introductory statistics class as cruel and unjust punishment. A half-dozen years of experimentation, first with assorted handouts and then with an extensive set of lecture notes distributed as a second text, convinced us that a book could be written for these students. Representing the culmination of this effort, the present book provides a simple overview of descriptive and inferential statistics for mathematically unsophisticated students in the behavioral sciences, social sciences, health sciences, and education.

PEDAGOGICAL FEATURES

- Basic concepts and procedures are explained in plain English, and a special effort has been made to clarify such perennially mystifying topics as the standard deviation, normal curve applications, hypothesis tests, degrees of freedom, and analysis of variance. For example, the standard deviation is more than a formula; it roughly reflects the average amount by which individual observations deviate from their mean.
- Unnecessary math, computational busy work, and subtle technical distinctions are avoided without sacrificing either accuracy or realism. Small batches of data define most computational tasks. Single examples permeate entire chapters or even several related chapters, serving as handy frames of reference for new concepts and procedures.

- Each chapter begins with a preview and ends with a summary, lists of important terms and key equations, and review questions.
- Key statements appear in bold type, and step-by-step summaries of important procedures, such as solving normal curve problems, appear in boxes.
- Important definitions and reminders about key points appear in page margins.
- Scattered throughout the book are examples of computer outputs for three of the most prevalent programs: Minitab, SPSS, and SAS. These outputs can be either ignored or expanded without disrupting the continuity of the text.
- Questions are introduced within chapters, often section by section, as Progress Checks. They are designed to minimize the cumulative confusion reported by many students for some chapters and by some students for most chapters. Each chapter ends with Review Questions.
- Questions have been selected to appeal to student interests: for example, probability calculations, based on design flaws, that re-create the chillingly high likelihood of the *Challenger* shuttle catastrophe (8.18, page 165); a *t* test analysis of global temperatures to evaluate a possible greenhouse effect (13.7, page 244); and a chi-square test of the survival rates of cabin and steerage passengers aboard the *Titanic* (19.14, page 384).
- Appendix B supplies answers to questions marked with asterisks. Other appendices provide a practical math review complete with self-diagnostic tests, a glossary of important terms, and tables for important statistical distributions.

INSTRUCTIONAL AIDS

An electronic version of an instructor's manual accompanies the text. The instructor's manual supplies answers omitted in the text (for about one-third of all questions), as well as sets of multiple-choice test items for each chapter, and a chapter-by-chapter commentary that reflects the authors' teaching experiences with this material. Instructors can access this material in the Instructor Companion Site at http://www.wiley.com/college/witte.

An electronic version of a student workbook, prepared by Beverly Dretzke of the University of Minnesota, also accompanies the text. Self-paced and self-correcting, the workbook contains problems, discussions, exercises, and tests that supplement the text. Students can access this material in the Student Companion Site at http://www.wiley.com/college/witte.

CHANGES IN THIS EDITION

- Update discussion of polling and random digit dialing in Section 8.4
- A new Section 14.11 on the "file drawer effect," whereby nonsignificant statistical findings are never published and the importance of replication.
- Updated numerical examples.
- New examples and questions throughout the book.
- Computer outputs and website have been updated.

USING THE BOOK

The book contains more material than is covered in most one-quarter or one-semester courses. Various chapters can be omitted without interrupting the main development. Typically, during a one-semester course we cover the entire book except for analysis of variance (Chapters 16, 17, and 18) and tests of ranked data (Chapter 20). An instructor who wishes to emphasize inferential statistics could skim some of the earlier chapters, particularly *Normal Distributions and Standard Scores (z)* (Chapter 5), and *Regression* (Chapter 7), while an instructor who desires a more applied emphasis could omit *Populations, Samples, and Probability* (Chapter 8) and *More about Hypothesis Testing* (Chapter 11).

ACKNOWLEDGMENTS

The authors wish to acknowledge their immediate family: Doris, Steve, Faith, Mike, Sharon, Andrea, Phil, Katie, Keegan, Camy, Brittany, Brent, Kristen, Scott, Joe, John, Jack, Carson, Sam, Margaret, Gretchen, Carrigan, Kedrick, and Alika. The first author also wishes to acknowledge his brothers and sisters: Henry, the late Lila, J. Stuart, A. Gerhart, and Etz; deceased parents: Henry and Emma; and all friends and relatives, past and present, including Arthur, Betty, Bob, Cal, David, Dick, Ellen, George, Grace, Harold, Helen, John, Joyce, Kayo, Kit, Mary, Paul, Ralph, Ruth, Shirley, and Suzanne.

Numerous helpful comments were made by those who reviewed the current and previous editions of this book: John W. Collins, Jr., Seton Hall University; Jelani Mandara, Northwestern University; L. E. Banderet, Northeastern University; S. Natasha Beretvas, University of Texas at Austin; Patricia M. Berretty, Fordham University; David Coursey, Florida State University; Shelia Kennison, Oklahoma State University; Melanie Kercher, Sam Houston State University; Jennifer H. Nolan, Loyola Marymount University; and Jonathan C. Pettibone, University of Alabama in Huntsville; Kevin Sumrall, Montgomery College; Sky Chafin, Grossmont College; Christine Ferri, Richard Stockton College of NJ; Ann Barich, Lewis University.

Special thanks to Carson Witte who proofread the entire manuscript twice.

Excellent editorial support was supplied by the people at John Wiley & Sons, Inc., most notably Abidha Sulaiman and Gladys Soto.

Contents

Descriptive Statistics

Statistics exists because of the prevalence of variability in the real world. In its simplest form, known as **descriptive statistics,** *statistics provides us with tools—tables, graphs, averages, ranges, correlations—for organizing and summarizing the inevitable variability in collections of actual observations or scores.* Examples are:

1. A tabular listing, ranked from most to least, of the total number of romantic affairs during college reported anonymously by each member of your stat class
2. A graph showing the annual change in global temperature during the last 30 years
3. A report that describes the average difference in grade point average (GPA) between college students who regularly drink alcoholic beverages and those who don't

Inferential Statistics

Statistics also provides tools—a variety of tests and estimates—for generalizing beyond collections of actual observations. This more advanced area is known as **inferential statistics.** Tools from inferential statistics permit us to use a relatively small collection of actual observations to evaluate, for example:

1. A pollster's claim that a majority of all U.S. voters favor stronger gun control laws
2. A researcher's hypothesis that, on average, meditators report fewer headaches than do nonmeditators
3. An assertion about the relationship between job satisfaction and overall happiness

In this book, you will encounter the most essential tools of descriptive statistics (Part 1), beginning with Chapter 2, and those of inferential statistics (Part 2), beginning with Chapter 8.

Progress Check *1.1 Indicate whether each of the following statements typifies descriptive statistics (because it describes sets of actual observations) or inferential statistics (because it generalizes beyond sets of actual observations).

(a) Students in my statistics class are, on average, 23 years old.

(b) The population of the world exceeds 7 billion (that is, 7,000,000,000 or 1 million multiplied by 7000).

(c) Either four or eight years have been the most frequent terms of office actually served by U.S. presidents.

(d) Sixty-four percent of all college students favor right-to-abortion laws.

Answers on page 420.

1.3 MORE ABOUT INFERENTIAL STATISTICS

Population

Any complete collection of observations or potential observations.

Populations and Samples

Inferential statistics is concerned with generalizing beyond sets of actual observations, that is, with generalizing from a sample to a population. In statistics, a **population**

Sample
Any smaller collection of actual observations from a population.

refers to *any complete collection of observations or potential observations,* whereas a **sample** refers to *any smaller collection of actual observations drawn from a population.* In everyday life, populations often are viewed as collections of real objects (e.g., people, whales, automobiles), whereas in statistics, populations may be viewed more abstractly as collections of properties or measurements (e.g., the ethnic backgrounds of people, life spans of whales, gas mileage of automobiles).

Depending on your perspective, a given set of observations can be either a population or a sample. For instance, the weights reported by 53 male statistics students in **Table 1.1** can be viewed either as a population, because you are concerned about exceeding the load-bearing capacity of an excursion boat (chartered by the 53 students to celebrate successfully completing their stat class!), or as a sample from a population because you wish to generalize to the weights of *all* male statistics students or *all* male college students.

Table 1.1
QUANTITATIVE DATA: WEIGHTS (IN POUNDS) OF MALE STATISTICS STUDENTS

160	168	133	170	150	165	158	165
193	169	245	160	152	190	179	157
226	160	170	180	150	156	190	156
157	163	152	158	225	135	165	135
180	172	160	170	145	185	152	
205	151	220	166	152	159	156	
165	157	190	206	172	175	154	

Ordinarily, populations are quite large and exist only as potential observations (e.g., the *potential* scores of all U.S. college students on a test that measures anxiety). On the other hand, samples are relatively small and exist as actual observations (the *actual* scores of 100 college students on the test for anxiety). When using a sample (100 actual scores) to generalize to a population (millions of potential scores), it is important that the sample represent the population; otherwise, any generalization might be erroneous. Although conveniently accessible, the anxiety test scores for the 100 students in stat classes at your college probably would not be representative of the scores for all students. If you think about it, these 100 stat students might tend to have either higher or lower anxiety scores than those in the target population for numerous reasons including, for instance, the fact that the 100 students are mostly psychology majors enrolled in a required stat class at your particular college.

Random Sampling (Surveys)

Random Sampling
A procedure designed to ensure that each potential observation in the population has an equal chance of being selected in a survey.

Whenever possible, a sample should be randomly selected from a population in order to increase the likelihood that the sample accurately represents the population. **Random sampling** *is a procedure designed to ensure that each potential observation in the population has an equal chance of being selected in a survey.* Classic examples of random samples are a state lottery where each number from 1 to 99 in the population has an equal chance of being selected as one of the five winning numbers or a nationwide opinion survey in which each telephone number has an equal chance of being selected as a result of a series of random selections, beginning with a three-digit area code and ending with a specific seven-digit telephone number.

Random sampling can be very difficult when a population lacks structure (e.g., all persons currently in psychotherapy) or specific boundaries (e.g., all volunteers who could conceivably participate in an experiment). In this case, a random sample

becomes an ideal that can only be approximated—always with an effort to remove obvious biases that might cause the sample to misrepresent the population. For example, lacking the resources to sample randomly the target population of all U.S. college students, you might obtain scores by randomly selecting the 100 students, not just from stat classes at your college but also from one or more college directories, possibly using some of the more elaborate techniques described in Chapter 8. Insofar as your sample only approximates a true random sample, any resulting generalizations should be qualified. For example, if the 100 students were randomly selected only from several public colleges in northern California, this fact should be noted, and any generalizations to all college students in the United States would be both provisional and open to criticism.

Random Assignment (Experiments)

Estimating the average anxiety score for all college students probably would not generate much interest. Instead, we might be interested in determining whether relaxation training causes, on average, a reduction in anxiety scores between two groups of otherwise similar college students. Even if relaxation training has no effect on anxiety scores, we would expect average scores for the two groups to differ because of the inevitable variability between groups. The question becomes: How should we interpret the apparent difference between the treatment group and the control group? Once variability has been taken into account, should the difference be viewed as real (and attributable to relaxation training) or as transitory (and merely attributable to variability or chance)?

College students in the relaxation experiment probably are not a random sample from any intact population of interest, but rather a *convenience sample* consisting of volunteers from a limited pool of students fulfilling a course requirement. Accordingly, our focus shifts from random sampling to the random assignment of volunteers to the two groups. **Random assignment** *signifies that each person has an equal chance of being assigned to any group in an experiment.* Using procedures described in Chapter 8, random assignment should be employed whenever possible. Because chance dictates the membership of both groups, not only does random assignment minimize any biases that might favor one group or another, it also serves as a basis for estimating the role of variability in any observed result. Random assignment allows us to evaluate any finding, such as the actual average difference between two groups, to determine whether this difference is larger than expected just by chance, once variability is taken into account. In other words, it permits us to generalize beyond mere appearances and determine whether the average difference merits further attention because it *probably is real* or whether it should be ignored because it *can be attributed to variability or chance.*

Random Assignment

A procedure designed to ensure that each person has an equal chance of being assigned to any group in an experiment.

Overview: Surveys and Experiments

Figure 1.1 compares surveys and experiments. Based on random samples from populations, surveys permit generalizations from samples back to populations. Based on the random assignment of volunteers to groups, experiments permit decisions about whether differences between groups are real or merely transitory.

PROGRESS CHECK *1.2 Indicate whether each of the following terms is associated primarily with a survey (S) or an experiment (E).

(a) random assignment

(b) representative

(c) generalization to the population

(d) control group

Answers on page 420.

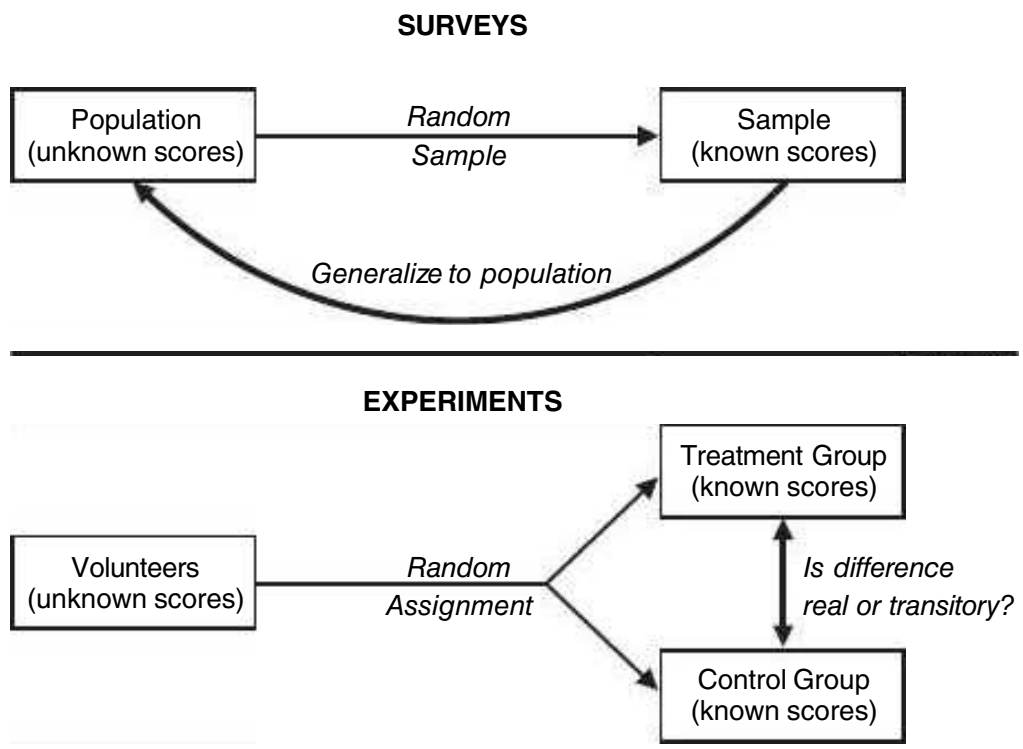


FIGURE 1.1
Overview: surveys and experiments.

(e) real difference

(f) random selection

(g) convenience sample

(h) volunteers

Answers on page 420.

Data
A collection of actual observations or scores in a survey or an experiment

Qualitative Data
A set of observations where any single observation is a word, letter, or numerical code that represents a class or category.

Ranked Data
A set of observations where any single observation is a number that indicates relative standing.

Quantitative Data
A set of observations where any single observation is a number that represents an amount or a count.

1.4 THREE TYPES OF DATA

Any statistical analysis is performed on **data,** a *collection of actual observations or scores in a survey or an experiment.*

The precise form of a statistical analysis often depends on whether data are qualitative, ranked, or quantitative.

Generally, **qualitative data** consist of words (Yes or No), letters (Y or N), or numerical codes (0 or 1) that represent a class or category. **Ranked data** consist of numbers (1st, 2nd, . . . 40th place) that represent relative standing within a group. **Quantitative data** consist of numbers (weights of 238, 170, . . . 185 lbs) that represent an amount or a count. To determine the type of data, focus on a single observation in any collection of observations. For example, the weights reported by 53 male students in Table 1.1 are quantitative data, since any single observation, such as 160 lbs, represents an amount of weight. If the weights in Table 1.1 had been replaced with ranks, beginning with a rank of 1 for the lightest weight of 133 lbs and ending with a rank of 53 for the heaviest weight of 245 lbs, these numbers would have been ranked data, since any single observation represents not an amount, but only relative standing within the group of 53 students. Finally, the Y and N replies of students in **Table 1.2** are qualitative data, since any single observation is a letter that represents a class of replies.

Table 1.2
QUALITATIVE DATA: "DO YOU HAVE A FACEBOOK PROFILE?" YES (Y) OR NO (N) REPLIES OF STATISTICS STUDENTS

Y	Y	Y	N	N	Y	Y	Y
Y	Y	Y	N	N	Y	Y	Y
N	Y	N	Y	Y	Y	Y	Y
Y	Y	N	Y	N	Y	N	Y
Y	N	Y	N	N	Y	Y	Y
Y	Y	N	Y	Y	Y	Y	Y
N	N	N	N	Y	N	N	Y
Y	Y	Y	Y	Y	N	Y	N
Y	Y	Y	Y	N	N	Y	Y
N	Y	N	N	Y	Y	Y	Y
	Y	Y	N				

Progress Check *1.3 Indicate whether each of the following terms is *qualitative* (because it's a word, letter, or numerical code representing a class or category); *ranked* (because it's a number representing relative standing); or *quantitative* (because it's a number representing an amount or a count).

(a) ethnic group

(b) age

(c) family size

(d) academic major

(e) sexual preference

(f) IQ score

(g) net worth (dollars)

(h) third-place finish

(i) gender

(j) temperature

Answers on page 420.

Level of Measurement
Specifies the extent to which a number (or word or letter) actually represents some attribute and, therefore, has implications for the appropriateness of various arithmetic operations and statistical procedures.

1.5 LEVELS OF MEASUREMENT

Learned years ago in grade school, the abstract statement that $2 + 2 = 4$ qualifies as one of life's everyday certainties, along with taxes and death. However, not all numbers have the same interpretation. For instance, it wouldn't make sense to find the sum of two Social Security numbers or to claim that, when viewed as indicators of academic achievement, two GPAs of 2.0 equal a GPA of 4.0. To clarify further the differences among the three types of data, let's introduce the notion of level of measurement. Looming behind any data, the **level of measurement** *specifies the extent to which a number (or word or letter) actually represents some attribute and, therefore, has implications for the appropriateness of various arithmetic operations and statistical procedures.*

Table 2.1
FREQUENCY DISTRIBUTION (UNGROUPED DATA)

WEIGHT	*f*
245	1
244	0
243	0
242	0
*	
*	
*	
161	0
160	4
159	1
158	2
157	3
*	
*	
*	
136	0
135	2
134	0
133	1
Total	53

(Or, more interestingly, is there a difference between two or more sets of data—for instance, between the GRE scores of students who do or do not attend a test-taking workshop; or between the survival rates of coronary bypass patients who do or do not own a dog; or between the starting salaries of male and female executives?) At this point, especially if you are facing a fresh set of data in which you have a special interest, statistics can be exciting as well as challenging. Your initial responsibility is to describe the data as clearly, completely, and concisely as possible. Statistics supplies some tools, including tables and graphs, and some guidelines. Beyond that, it is just the data and you. There is no single right way to describe data. Equally valid descriptions of the same data might appear in tables or graphs with different formats. By following just a few guidelines, your reward will be a well-summarized set of data.

TABLES (FREQUENCY DISTRIBUTIONS)

2.1 FREQUENCY DISTRIBUTIONS FOR QUANTITATIVE DATA

Table 2.1 shows one way to organize the weights of the male statistics students listed in Table 1.1. First, arrange a column of consecutive numbers, beginning with the lightest weight (133) at the bottom and ending with the heaviest weight (245) at the top. (Because of the extreme length of this column, many intermediate numbers have been omitted in Table 2.1, a procedure *never* followed in practice.) Then place a short vertical stroke or tally next to a number each time its value appears in the original set of data; once this process has been completed, substitute for each tally count (not shown in Table 2.1) a number indicating the frequency (*f*) of occurrence of each weight.

A *frequency distribution* is a collection of observations produced by sorting observations into classes and showing their frequency (*f*) of occurrence in each class.

Frequency Distribution
A collection of observations produced by sorting observations into classes and showing their frequency (f) of occurrence in each class.

Frequency Distribution for Ungrouped Data
A frequency distribution produced whenever observations are sorted into classes of single *values.*

When observations are sorted into classes of *single* values, as in Table 2.1, the result is referred to as a **frequency distribution for ungrouped data.**

Not Always Appropriate

The frequency distribution shown in Table 2.1 is only partially displayed because there are more than 100 possible values between the largest and smallest observations. Frequency distributions for ungrouped data are much more informative when the number of possible values is less than about 20. Under these circumstances, they are a straightforward method for organizing data. Otherwise, if there are 20 or more possible values, consider using a frequency distribution for grouped data.

Progress Check *2.1 Students in a theater arts appreciation class rated the classic film *The Wizard of Oz* on a 10-point scale, ranging from 1 (poor) to 10 (excellent), as follows:

3	7	2	7	8
3	1	4	10	3
2	5	3	5	8
9	7	6	3	7
8	9	7	3	6

Since the number of possible values is relatively small—only 10—it's appropriate to construct a frequency distribution for ungrouped data. Do this.

Answer on page 420.

Grouped Data

Table 2.2 shows another way to organize the weights in Table 1.1 according to their frequency of occurrence. When observations are sorted into classes of *more than one value,* as in Table 2.2, the result is referred to as a **frequency distribution for grouped data.** Let's look at the general structure of this frequency distribution. Data are grouped into class intervals with 10 possible values each. The bottom class includes the smallest observation (133), and the top class includes the largest observation (245). The distance between bottom and top is occupied by an orderly series of classes. The frequency (f) column shows the frequency of observations in each class and, at the bottom, the total number of observations in all classes.

Let's summarize the more important properties of the distribution of weights in Table 2.2. Although ranging from the 130s to the 240s, the weights peak in the 150s, with a progressively decreasing but relatively heavy concentration in the 160s and 170s. Furthermore, the distribution of weights is not balanced about its peak, but tilted in the direction of the heavier weights.

Frequency Distribution for Grouped Data

A frequency distribution produced whenever observations are sorted into classes of more than one value.

Table 2.2 FREQUENCY DISTRIBUTION (GROUPED DATA)

WEIGHT	*f*
240–249	1
230–239	0
220–229	3
210–219	0
200–209	2
190–199	4
180–189	3
170–179	7
160–169	12
150–159	17
140–149	1
130–139	3
Total	53

2.2 GUIDELINES

The "Guidelines for Frequency Distributions" box lists seven rules for producing a well-constructed frequency distribution. The first three rules are essential and should not be violated. The last four rules are optional and can be modified or ignored as circumstances warrant. Satisfy yourself that the frequency distribution in Table 2.2 actually complies with these seven rules.

How Many Classes?

The seventh guideline requires a few more comments. The use of too many classes—as in **Table 2.3,** in which the weights are grouped into 24 classes, each with an interval of 5—tends to defeat the purpose of a frequency distribution, namely, to provide a reasonably concise description of data. On the other hand, the use of too few classes—as in **Table 2.4,** in which the weights are grouped into three classes, each with an interval of 50—can mask important data patterns such as the high density of weights in the 150s and 160s.

When There Are Either Many or Few Observations

There is nothing sacred about 10, the recommended number of classes. When describing large sets of data, you might aim for considerably more than 10 classes in order to portray some of the more fine-grained data patterns that otherwise could vanish. On the other hand, when describing small batches of data, you might aim for fewer than 10 classes in order to spotlight data regularities that otherwise could be blurred. It is best, therefore, to think of 10, the recommended number of classes, as a rough rule of thumb to be applied with discretion.

Gaps between Classes

In well-constructed frequency tables, the gaps between classes, such as between 149 and 150 in Table 2.2, show clearly that each observation or score has been assigned to one, and only one, class. The size of the gap should always equal one **unit of measurement;** that is, it should always equal *the smallest possible difference between scores* within a particular set of data. Since the gap is never bigger than one unit of measurement, no score can fall into the gap. In the present case, in which the weights are reported to the nearest pound, one pound is the unit of measurement, and therefore, the gap between classes equals one pound. These gaps would not be appropriate if the weights had been reported to the nearest tenth of a pound. In this case, one-tenth of a pound is the unit of measurement, and therefore, the gap should equal one-tenth of a pound. The smallest class interval would be 130.0–139.9 (not 130–139), and the next class interval would be

Unit of Measurement

The smallest possible difference between scores.

Table 2.3
FREQUENCY DISTRIBUTION WITH TOO MANY INTERVALS

WEIGHT	*f*
245–249	1
240–244	0
235–239	0
230–234	0
225–229	2
220–224	1
215–219	0
210–214	0
205–209	2
200–204	0
195–199	0
190–194	4
185–189	1
180–184	2
175–179	2
170–174	5
165–169	7
160–164	5
155–159	9
150–154	8
145–149	1
140–144	0
135–139	2
130–134	1
Total	53

Table 2.4
FREQUENCY DISTRIBUTION WITH TOO FEW INTERVALS

WEIGHT	*f*
200–249	6
150–199	43
100–149	4
Total	53

GUIDELINES FOR FREQUENCY DISTRIBUTIONS

Essential

1. ***Each observation should be included in one, and only one, class.***
   Example: 130–139, 140–149, 150–159, etc. It would be incorrect to use 130–140, 140–150, 150–160, etc., in which, because the boundaries of classes overlap, an observation of 140 (or 150) could be assigned to either of two classes.
2. ***List all classes, even those with zero frequencies.***
   Example: Listed in Table 2.2 is the class 210–219 and its frequency of zero. It would be incorrect to skip this class because of its zero frequency.
3. ***All classes should have equal intervals.***
   Example: 130–139, 140–149, 150–159, etc. It would be incorrect to use 130–139, 140–159, etc., in which the second class interval (140–159) is twice as wide as the first class interval (130–139).

Optional

4. ***All classes should have both an upper boundary and a lower boundary.***
   Example: 240–249. Less preferred would be 240–above, in which no maximum value can be assigned to observations in this class. (Nevertheless, this type of open-ended class is employed as a space-saving device when many different tables must be listed, as in the *Statistical Abstract of the United States.* An open-ended class appears in the table "Two Age Distributions" in Review Question 2.17 at the end of this chapter.)
5. ***Select the class interval from convenient numbers, such as 1, 2, 3, . . . 10, particularly 5 and 10 or multiples of 5 and 10.***
   Example: 130–139, 140–149, in which the class interval of 10 is a convenient number. Less preferred would be 130–142, 143–155, etc., in which the class interval of 13 is not a convenient number.
6. ***The lower boundary of each class interval should be a multiple of the class interval.***
   Example: 130–139, 140–149, in which the lower boundaries of 130, 140, are multiples of 10, the class interval. Less preferred would be 135–144, 145–154, etc., in which the lower boundaries of 135 and 145 are not multiples of 10, the class interval.
7. ***Aim for a total of approximately 10 classes.***
   Example: The distribution in Table 2.2 uses 12 classes. Less preferred would be the distributions in Tables 2.3 and 2.4. The distribution in Table 2.3 has too many classes (24), whereas the distribution in Table 2.4 has too few classes (3).

140.0–149.9 (not 140–149), and so on. These new boundaries would guarantee that any observation, such as 139.6, would be assigned to one, and only one, class.

Gaps between classes do not signify any disruption in the essentially continuous nature of the data. It would be erroneous to conclude that, because of the gap between 149 and 150 for the frequency distribution in Table 2.2, nobody can weigh between 149 and 150 lbs. As noted in Section 1.6, a man who reports his weight as 150 lbs actually could weigh anywhere between 149.5 and 150.5 lbs, just as a man who reports his weight as 149 lbs actually could weigh anywhere between 148.5 and 149.5 lbs.

Real Limits of Class Intervals

Real Limits
Located at the midpoint of the gap between adjacent tabled boundaries.

Gaps cannot be ignored when you are determining the actual width of any class interval. The **real limits** *are located at the midpoint of the gap between adjacent tabled boundaries;* that is, one-half of one unit of measurement below the lower tabled boundary and one-half of one unit of measurement above the upper tabled boundary.

For example, the real limits for 140–149 in Table 2.2 are 139.5 (140 minus one-half of the unit of measurement of 1) and 149.5 (149 plus one-half of the unit of measurement of 1), and the actual width of the class interval would be 10 (from 149.5 − 139.5 = 10).

If weights had been reported to the nearest tenth of a pound, the real limits for 140.0–149.9 would be 139.95 (140.0 minus one-half of the unit of measurement of .1) and 149.95 (149.9 plus one-half of one unit of measurement of .1), and the actual width of the class interval still would be 10 (from 149.95 139.95 = 10).

Constructing Frequency Distributions

Now that you know the properties of well-constructed frequency distributions, study the step-by-step procedure listed in the "Constructing Frequency Distributions" box, which shows precisely how the distribution in Table 2.2 was constructed from the weight data in Table 1.1. You might want to refer back to this box when you need to construct a frequency distribution for grouped data.

Progress Check *2.2 The IQ scores for a group of 35 high school dropouts are as follows:

91	85	84	79	80
87	96	75	86	104
95	71	105	90	77
123	80	100	93	108
98	69	99	95	90
110	109	94	100	103
112	90	90	98	89

(a) Construct a frequency distribution for grouped data.

(b) Specify the *real* limits for the lowest class interval in this frequency distribution.

Answers on pages 420 and 421.

Progress Check *2.3 What are some possible poor features of the following frequency distribution?

ESTIMATED WEEKLY TV VIEWING TIME (HRS) FOR 250 SIXTH GRADERS

VIEWING TIME	*f*
35–above	2
30–34	5
25–30	29
20–22	60
15–19	60
10–14	34
5–9	31
0–4	29
Total	250

Answers on page 421.

CONSTRUCTING FREQUENCY DISTRIBUTIONS

1. ***Find the range, that is,*** the difference between the largest and smallest observations. The range of weights in Table 1.1 is $245 - 133 = 112$.
2. ***Find the class interval required to span the range*** by dividing the range by the desired number of classes (ordinarily 10). In the present example,

$$\text{Class Interval} = \frac{\text{range}}{\text{desired number of classes}} = \frac{112}{10} = 11.2$$

3. ***Round off to the nearest convenient interval*** (such as 1, 2, 3, . . . 10, particularly 5 or 10 or multiples of 5 or 10). In the present example, the nearest convenient interval is 10.
4. ***Determine where the lowest class should begin.*** (Ordinarily, this number should be a multiple of the class interval.) In the present example, the smallest score is 133, and therefore the lowest class should begin at 130, since 130 is a multiple of 10 (the class interval).
5. ***Determine where the lowest class should end*** by adding the class interval to the lower boundary and then subtracting one unit of measurement. In the present example, add 10 to 130 and then subtract 1, the unit of measurement, to obtain 139—the number at which the lowest class should end.
6. ***Working upward, list as many equivalent classes as are required to include the largest observation.*** In the present example, list 130–139, 140–149, . . . , 240–249, so that the last class includes 245, the largest score.
7. ***Indicate with a tally the class in which each observation falls.*** For example, the first score in Table 1.1, 160, produces a tally next to 160–169; the next score, 193, produces a tally next to 190–199; and so on.
8. ***Replace the tally count for each class with a number—the frequency (f)—and show the total of all frequencies.*** (Tally marks are not usually shown in the final frequency distribution.)
9. ***Supply headings for both columns and a title for the table.***

2.3 OUTLIERS

Outlier

A very extreme score.

Be prepared to deal occasionally with the appearance of one or more *very extreme* scores, or **outliers.** A GPA of 0.06, an IQ of 170, summer wages of $62,000—each requires special attention because of its potential impact on a summary of the data.

Check for Accuracy

Whenever you encounter an outrageously extreme value, such as a GPA of 0.06, attempt to verify its accuracy. For instance, was a respectable GPA of 3.06 recorded erroneously as 0.06? If the outlier survives an accuracy check, it should be treated as a legitimate score.

Stem and Leaf Display
A device for sorting quantitative data on the basis of leading and trailing digits.

Stem and Leaf Displays

Still another technique for summarizing quantitative data is a **stem and leaf display.** Stem and leaf displays are ideal for summarizing distributions, such as that for weight data, without destroying the identities of individual observations.

Constructing a Display

The leftmost panel of **Table 2.9** re-creates the weights of the 53 male statistics students listed in Table 1.1. To construct the stem and leaf display for these data, first note that, when counting by tens, the weights range from the 130s to the 240s. Arrange a column of numbers, the stems, beginning with 13 (representing the 130s) and ending with 24 (representing the 240s). Draw a vertical line to separate the stems, which represent multiples of 10, from the space to be occupied by the leaves, which represent multiples of 1.

Table 2.9
CONSTRUCTING STEM AND LEAF DISPLAY FROM WEIGHTS OF MALE STATISTICS STUDENTS

RAW SCORES					STEM AND LEAF DISPLAY
160	165	135	175		
193	168	245	165	13	3 5 5
226	169	170	185	14	5
152	160	156	154	15	2 7 1 7 8 0 2 0 2 6 9 8 2 6 4 7 6
180	170	160	179	16	0 3 5 8 9 0 0 0 6 5 5 5
205	150	225	165	17	2 0 0 0 2 5 9
163	152	190	206	18	0 0 5
157	160	159	165	19	3 0 0 0
151	190	172	157	20	5 6
157	150	190	156	21	
220	133	166	135	22	6 0 5
145	180	158		23	
158	152	152		24	5
172	170	156			

Next, enter each raw score into the stem and leaf display. As suggested by the shaded coding in Table 2.9, the first raw score of 160 reappears as a leaf of 0 on a stem of 16. The next raw score of 193 reappears as a leaf of 3 on a stem of 19, and the third raw score of 226 reappears as a leaf of 6 on a stem of 22, and so on, until each raw score reappears as a leaf on its appropriate stem.

Interpretation

Notice that the weight data have been sorted by the stems. All weights in the 130s are listed together; all of those in the 140s are listed together, and so on. A glance at the stem and leaf display in Table 2.9 shows essentially the same pattern of weights depicted by the frequency distribution in Table 2.2 and the histogram in Figure 2.1. (If you rotate the book counterclockwise one-quarter of a full turn, the silhouette of the stem and leaf display is the same as the histogram for the weight data. This simple maneuver only works if, as in the present display, stem values are listed from smallest at the top to largest at the bottom—one reason why the customary ranking for most tables in this book has been reversed for stem and leaf displays.)

Selection of Stems

Stem values are not limited to units of 10. Depending on the data, you might identify the stem with one or more leading digits that culminates in some variation on a stem

value of 10, such as 1, 100, 1000, or even .1, .01, .001, and so on. For instance, an annual income of $23,784 could be displayed as a stem of 23 (thousands) and a leaf of 784. (Leaves consisting of two or more digits, such as 784, are separated by commas.) An SAT test score of 689 could be displayed as a stem of 6 (hundreds) and a leaf of 89. A GPA of 3.25 could be displayed as a stem of 3 (ones) and a leaf of 25, or if you wanted more than a few stems, 3.25 could be displayed as a stem of 3.2 (one-tenths) and a leaf of 5.

Stem and leaf displays represent statistical bargains. Just a few minutes of work produces a description of data that is both clear and complete. Even though rarely appearing in published reports, stem and leaf displays often serve as the first step toward organizing data.

Progress Check *2.10 Construct a stem and leaf display for the following IQ scores obtained from a group of four-year-old children.

120	98	118	117	99	111
126	85	88	124	104	113
108	141	123	137	78	96
102	132	109	106	143»	

Answers on page 422.

2.9 TYPICAL SHAPES

Whether expressed as a histogram, a frequency polygon, or a stem and leaf display, an important characteristic of a frequency distribution is its shape. **Figure 2.3** shows some of the more typical shapes for smoothed frequency polygons (which ignore the inevitable irregularities of real data).

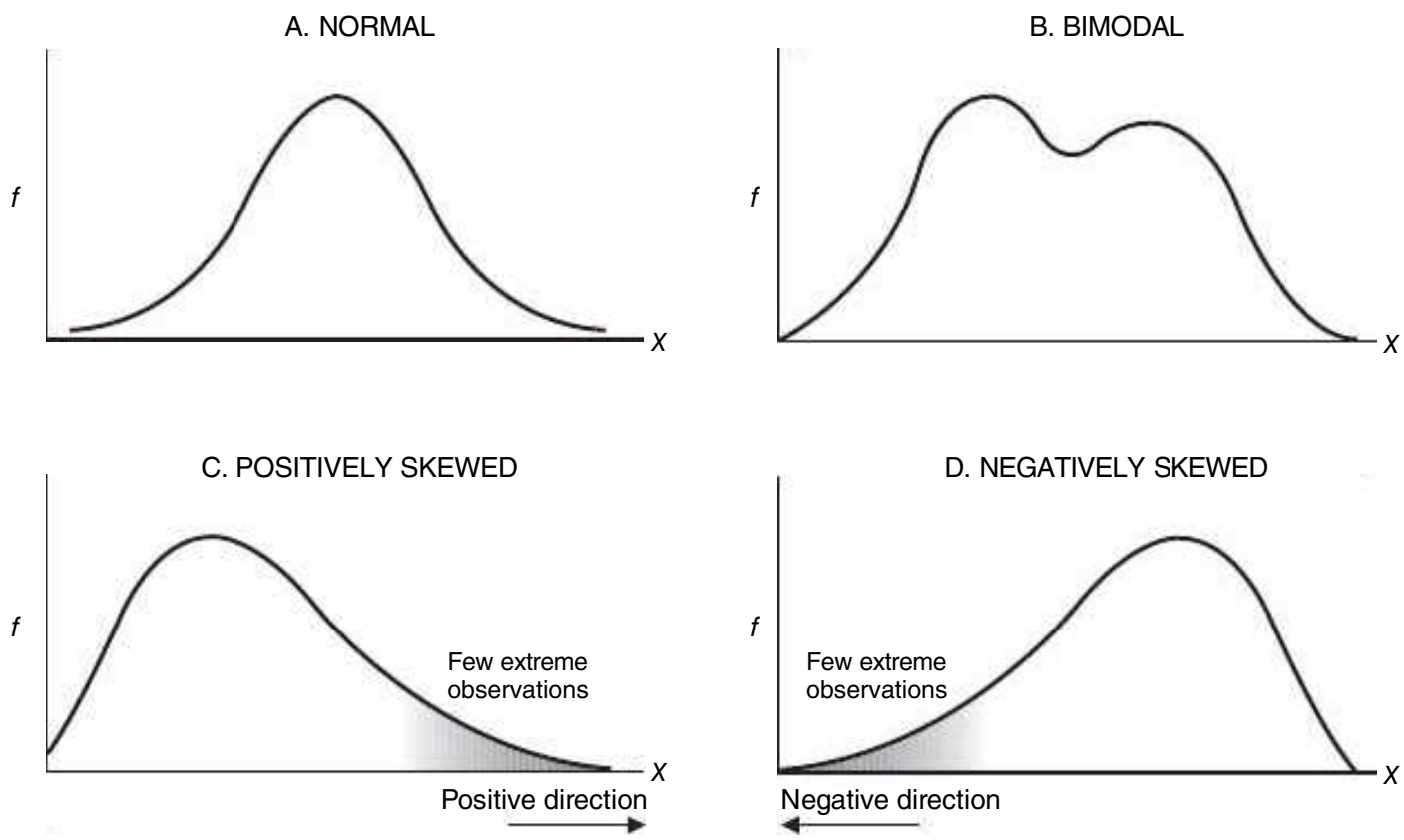


FIGURE 2.3
Typical shapes.

Normal

Any distribution that approximates the normal shape in panel A of Figure 2.3 can be analyzed, as we will see in Chapter 5, with the aid of the well-documented normal curve. The familiar bell-shaped silhouette of the normal curve can be superimposed on many frequency distributions, including those for uninterrupted gestation periods of human fetuses, scores on standardized tests, and even the popping times of individual kernels in a batch of popcorn.

Bimodal

Any distribution that approximates the bimodal shape in panel B of Figure 2.3 might, as suggested previously, reflect the coexistence of two different types of observations in the same distribution. For instance, the distribution of the ages of residents in a neighborhood consisting largely of either new parents or their infants has a bimodal shape.

Positively Skewed

Positively Skewed Distribution
A distribution that includes a few extreme observations in the positive direction (to the right of the majority of observations).

The two remaining shapes in Figure 2.3 are lopsided. A *lopsided distribution caused by a few extreme observations in the positive direction (to the right of the majority of observations), as in panel C of* Figure 2.3*, is* a **positively skewed distribution.** The distribution of incomes among U.S. families has a pronounced positive skew, with most family incomes under $200,000 and relatively few family incomes spanning a wide range of values above $200,000. The distribution of weights in Figure 2.1 also is positively skewed.

Negatively Skewed

Negatively Skewed Distribution
A distribution that includes a few extreme observations in the negative direction (to the left of the majority of observations).

A lopsided distribution caused by a few extreme observations in the negative direction (to the left of the majority of observations), as in panel D of Figure 2.3*, is* a **negatively skewed distribution.** The distribution of ages at retirement among U.S. job holders has a pronounced negative skew, with most retirement ages at 60 years or older and relatively few retirement ages spanning the wide range of ages younger than 60.

Positively or Negatively Skewed?

Some people have difficulty with this terminology, probably because an entire distribution is labeled on the basis of the relative location, in the positive or negative direction, of a few extreme observations, rather than on the basis of the location of the majority of observations. To make this distinction, always force yourself to focus on the relative locations of the few extreme observations. If you get confused, use panels C and D of Figure 2.3 as guides, noting which silhouette in these two panels best approximates the shape of the distribution in question.

Progress Check *2.11 Describe the probable shape—normal, bimodal, positively skewed, or negatively skewed—for each of the following distributions:

(a) female beauty contestants' scores on a masculinity test, with a higher score indicating a greater degree of masculinity

(b) scores on a standardized IQ test for a group of people selected from the general population

(c) test scores for a group of high school students on a very difficult college-level math exam

(d) reading achievement scores for a third-grade class consisting of about equal numbers of regular students and learning-challenged students

(e) scores of students at the Eastman School of Music on a test of music aptitude (designed for use with the general population)

Answers on page 422.

2.10 A GRAPH FOR QUALITATIVE (NOMINAL) DATA

Bar Graph

A bar-type graph for qualitative data. Gaps between adjacent bars emphasize the discontinuous nature of the data.

The distribution in Table 2.7, based on replies to the question "Do you have a Facebook profile?" appears as a **bar graph** in **Figure 2.4.** A glance at this graph confirms that Yes replies occur approximately twice as often as No replies.

As with histograms, equal segments along the horizontal axis are allocated to the different words or classes that appear in the frequency distribution for qualitative data. Likewise, equal segments along the vertical axis reflect increases in frequency. The body of the bar graph consists of a series of bars whose heights reflect the frequencies for the various words or classes.

A person's answer to the question "Do you have a Facebook profile?" is either Yes or No, not some impossible intermediate value, such as 40 percent Yes and 60 percent No. Gaps are placed between adjacent bars of bar graphs to emphasize the discontinuous nature of qualitative data. A bar graph also can be used with quantitative data to emphasize the discontinuous nature of a discrete variable, such as the number of children in a family.

Progress Check *2.12 Referring to the box "Constructing Graphs" on page 47 for step-by-step instructions, construct a bar graph for the data shown in the following table:

RACE/ETHNICITY OF U.S. POPULATION, 2010 (IN MILLIONS)

Race/Ethnicity	*f*
African American	37.7
Asian American*	17.2
Hispanic	50.5
White	196.8
Total**	302.2

**Mostly Asians, but also other races, such as Native Americans and Eskimos.*
*** Total does not include 6.6 million non-Hispanics reporting two or more races.*
Source: www.uscensus.gov/prod/census2010/

Answer on page 423.

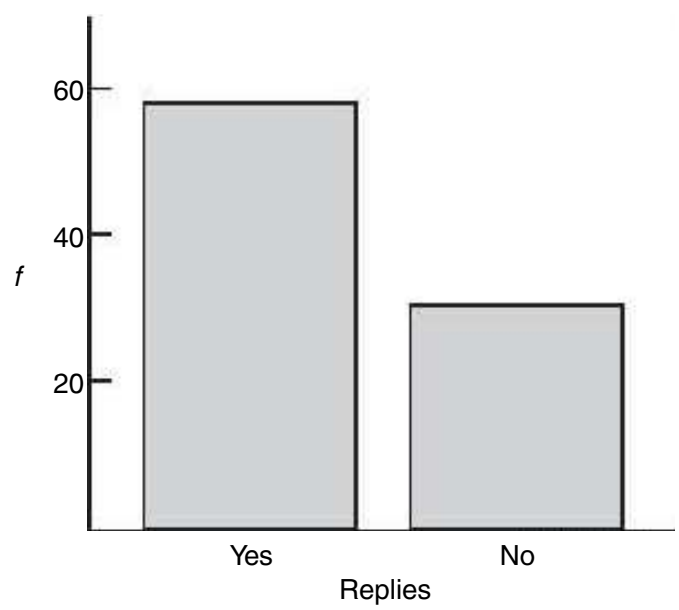


FIGURE 2.4
Bar graph.

2.11 MISLEADING GRAPHS

Graphs can be constructed in an unscrupulous manner to support a particular point of view. Indeed, this type of statistical fraud gives credibility to popular sayings, including "Numbers don't lie, but statisticians do" and "There are three kinds of lies—lies, damned lies, and statistics."

For example, to imply that comparatively many students responded Yes to the Facebook profile question, an unscrupulous person might resort to the various tricks shown in **Figure 2.5:**

- The width of the Yes bar is more than three times that of the No bar, thus violating the custom that bars be equal in width.
- The lower end of the frequency scale is omitted, thus violating the custom that the entire scale be reproduced, beginning with zero. (Otherwise, a broken scale should be highlighted by crossover lines, as in Figures 2.1 and 2.2.)
- The height of the vertical axis is several times the width of the horizontal axis, thus violating the custom, heretofore unmentioned, that the vertical axis be *approximately* as tall as the horizontal axis is wide. Beware of graphs in which, because the vertical axis is many times larger than the horizontal axis (as in Figure 2.5), frequency differences are exaggerated, or in which, because the vertical axis is many times smaller than the horizontal axis, frequency differences are suppressed.

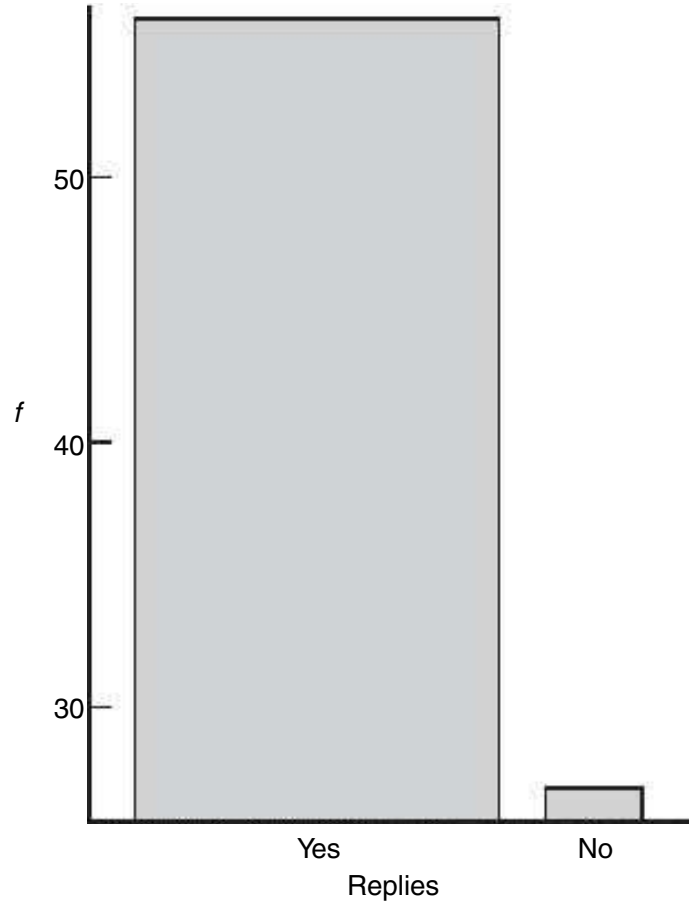


FIGURE 2.5
Distorted bar graph.

CHAPTER 3

Describing Data with Averages

Preview

Tables and graphs of frequency distributions are important points of departure when attempting to describe data. More precise summaries, such as averages, provide additional valuable information. Long-term investors in the stock market are able to ignore, with only an occasional sleepless night, daily fluctuations in their stocks by remembering that, ***on average,*** *the annual growth rate of stocks during the past 50 years has exceeded by several percentage points that of more conservative investments in bonds. You might stop smoking because,* ***on average,*** *nonsmokers can expect to live longer than heavy smokers (by as much as 10 years, according to some researchers). You might strengthen your resolve to graduate from college upon hearing that,* ***on average,*** *the lifetime earnings of college graduates are almost double those of high school graduates.*

Averages consist of numbers (or words) about which the data are, in some sense, centered. They are often referred to as ***measures of central tendency,*** *the several types of average yield numbers or words that attempt to describe, most generally, the middle or typical value for a distribution. This chapter focuses on three different measures of central tendency—the mode, median, and mean. Each of these has its special uses, but the mean is the most important average in both descriptive and inferential statistics.*

Measures of Central Tendency
Numbers or words that attempt to describe, most generally, the middle or typical value for a distribution.

3.1 MODE

Mode
The value of the most frequent score.

The mode reflects the value of the most frequently occurring score.

Table 3.1 shows the number of years served by 20 recent U.S. presidents, beginning with Benjamin Harrison (4 years) and ending with Bill Clinton (8 years). Four years is the modal term, since the greatest number of presidents, 7, served this term. Note that the mode equals 4 years, the *value* of the most frequently occurring term, not 7, the *frequency* with which that term occurred.

Bimodal
Describes any distribution with two obvious peaks.

It is easy to assign a value to the mode. If the data are organized, as in **Figure 3.1,** a glance will often be enough. However, if the data are not organized, as in Table 3.1, some counting may be required. The mode is readily understood as the most prevalent or typical value.

More Than One Mode

Distributions can have more than one mode (or no mode at all). *Distributions with two obvious peaks, even though they are not exactly the same height, are referred to as* **bimodal.** Distributions with more than two peaks are referred to as **multimodal.** The presence of more than one mode might reflect important differences among subsets of data. For instance, the distribution of weights for both male and female statistics students would most likely be bimodal, reflecting the combination of two separate weight distributions—a heavier one for males and a lighter one for females. Notice that even the distribution of presidential terms in Figure 3.1 tends to be bimodal, with a major peak at 4 years and a minor peak at 8 years, reflecting the two most typical terms of office.

Progress Check *3.1 Determine the mode for the following retirement ages: 60, 63, 45, 63, 65, 70, 55, 63, 60, 65, 63.

Progress Check *3.2 The owner of a new car conducts six gas mileage tests and obtains the following results, expressed in miles per gallon: 26.3, 28.7, 27.4, 26.6, 27.4, 26.9. Find the mode for these data.

Answers on page 424.

Table 3.1
TERMS IN YEARS OF 20 RECENT U.S. PRESIDENTS, LISTED CHRONOLOGICALLY

4	(Harrison)
4	
4	
8	
4	
8	
2	
6	
4	
12	
8	
8	
2	
6	
5	
3	
4	
8	
4	
8	(Clinton)

Source: The New York Times Almanac (2012).

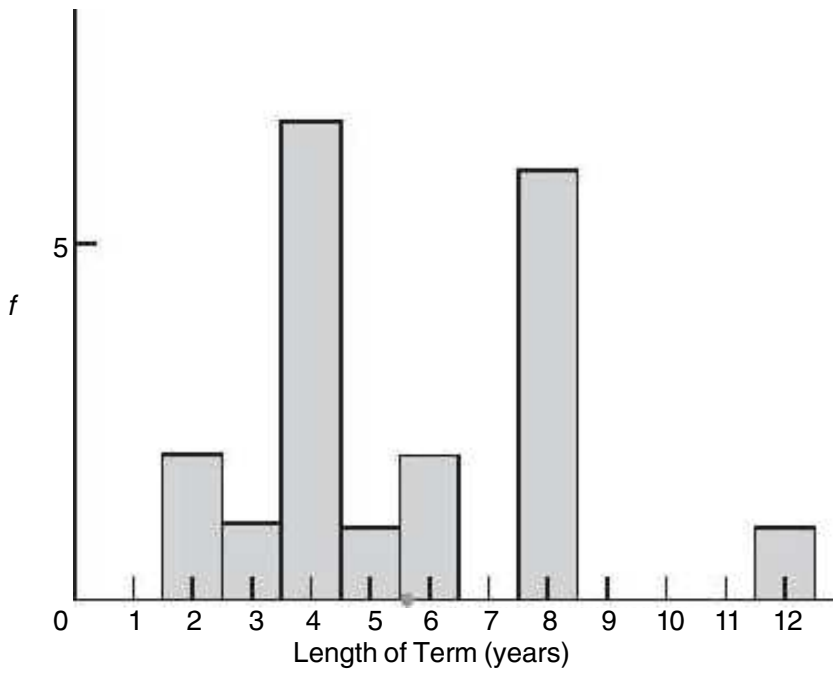


FIGURE 3.1
Distribution of presidential terms.

3.2 MEDIAN

Median

The middle value when observations are ordered from least to most.

The median reflects the middle value when observations are ordered from least to most.

The median splits a set of ordered observations into two equal parts, the upper and lower halves. In other words, the median has a percentile rank of 50, since observations with equal or smaller values constitute 50 percent of the entire distribution.*

Finding the Median

Table 3.2 shows how to find the median for two different sets of scores. The numbers in shaded squares cross-reference instructions in the top panel with examples in the bottom panel. Study Table 3.2 before reading on.

Table 3.2
FINDING THE MEDIAN

A. INSTRUCTIONS

1. Order scores from least to most.
2. Find the middle position by adding one to the total number of scores and dividing by 2.
3. *If the middle position is a whole number*, as in the left-hand panel below, use this number to *count* into the set of ordered scores.
4. The value of the median equals the value of the score located at the middle position.
5. *If the middle position is not a whole number,* as in the right-hand panel below, use the two nearest whole numbers to *count* into the set of ordered scores.
6. The value of the median equals the value midway between those of the two middlemost scores; to find the midway value, add the two given values and divide by 2.

B. EXAMPLES

Set of five scores:	Set of six scores:
2, 8, 2, 7, 6	3, 8, 9, 3, 1, 8
1 2, 2, 6, 7, 8	1 1, 3, 3, 8, 8, 9
2 $\frac{5+1}{2} = 3$	2 $\frac{6+1}{2} = 3.5$
2, 2, 6, 7, 8 ↑	
3 1, 2, 3	
4 median = 6	1, 3, 3, 8, 8, 9 ↑ ↑
	5 1, 2, 3, 4
	6 median $= \frac{3+8}{2} = 5.5$

*Strictly speaking, the median always has a percentile rank of *exactly* 50 only insofar as interpolation procedures, not discussed in this book, identify the value of the median with a single point along the numerical scale for the data.

To find the median, *scores always must be ordered from least to most (or vice versa)*. This task is straightforward with small sets of data but becomes increasingly cumbersome with larger sets of data that must be ordered manually.

When the total number of scores is odd, as in the lower left-hand panel of Table 3.2, there is a single middle-ranked score, and the value of the median equals the value of this score. When the total number of scores is even, as in the lower right-hand panel of Table 3.2, the value of the median equals a value midway between the values of the two middlemost scores. In either case, the value of the median always reflects the *value* of middle-ranked scores, not the *position* of these scores among the set of ordered scores.

The median term can be found for the 20 presidents. First, rank the terms from longest (12 for Franklin Roosevelt) to shortest (2 for Harding and Kennedy), as shown in the left-hand column of **Table 3.3.** Then, following the instructions in Table 3.2, verify that the median term for the 20 presidents equals 4.5 years, since 4.5 is the value midway between the values (4 and 5) of the two middlemost (10th- and 11th-ranked) terms in Table 3.3.

Notice that although the values for median and modal presidential terms are quite similar, they have different interpretations. The median term (4.5 years) describes the *middle-ranked* term; the modal term (4 years) describes the *most frequent* term in the distribution.

Progress Check *3.3 Find the median for the following retirement ages: 60, 63, 45, 63, 65, 70, 55, 63, 60, 65, 63.

Progress Check *3.4 Find the median for the following gas mileage tests: 26.3, 28.7, 27.4, 26.6, 27.4, 26.9.

Answers on page 424.

Table 3.3
TERMS IN YEARS OF 20 RECENT U.S. PRESIDENTS

ARRANGED BY LENGTH	DEVIATION FROM MEAN	SUM OF DEVIATIONS
12	6.40	
8	2.40	
8	2.40	
8	2.40	21.6
8	2.40	
8	2.40	
8	2.40	
8	2.40	
6	0.40	
6	0.40	
	(mean = 5.60)	0
5	−0.60	
4	−1.60	
4	−1.60	
4	−1.60	
4	−1.60	
4	−1.60	−21.6
4	−1.60	
4	−1.60	
3	−2.60	
2	−3.60	
2	−3.60	

3.3 MEAN

The mean is the most common average, one you have doubtless calculated many times.

The mean is found by adding all scores and then dividing by the number of scores.

That is,

$$Mean = \frac{sum\ of\ all\ scores}{number\ of\ scores}$$

To find the mean term for the 20 presidents, add all 20 terms in Table 3.1 (4 + . . . + 4 + 8) to obtain a sum of 112 years, and then divide this sum by 20, the number of presidents, to obtain a mean of 5.60 years.

There is no requirement that presidential terms be ranked before calculating the mean. Even when large sets of unorganized data are involved, the calculation of the mean is usually straightforward, particularly with the aid of a calculator or computer.

Sample or Population?

Population

A complete set of scores.

Sample

A subset of scores.

Statisticians distinguish between two types of means—the population mean and the sample mean—depending on whether the data are viewed as a **population** (*a complete set of scores*) or as a **sample** (*a subset of scores*). For example, if the terms of the 20 U.S. presidents are viewed as a population, then 5.60 years qualifies as a population mean. On the other hand, if the terms of the 20 U.S. presidents are viewed as a sample from the terms of *all* U.S. presidents, then 5.60 years qualifies as a sample mean. Not only is the present distinction entirely a matter of perspective, but it also produces exactly the same numerical value of 5.60 for both means. This distinction is introduced here because of its importance in later chapters, where the population mean usually is unknown but fixed as a constant, while the sample mean is known but varies from sample to sample. *Until then, unless noted otherwise, you can assume that we are dealing with the sample mean*.

Formula for Sample Mean

Sample Mean ($\bar{X}$)

The balance point for a sample, found by dividing the sum for the values of all scores in the sample by the number of scores in the sample.

Sample Size (n)

The total number of scores in the sample.

It's usually more efficient to substitute symbols for words in statistical formulas, including the word formula given above for the mean. When symbols are used, $\bar{X}$ designates the **sample mean,** and the formula becomes

SAMPLE MEAN

$$\bar{X} = \frac{\Sigma X}{n} \qquad (3.1)$$

and reads: "*X*-bar equals the sum of the variable *X* divided by the **sample size *n*.**" [Note that the uppercase Greek letter sigma (Σ) is read as *the sum of,* not as *sigma*. To avoid confusion, read only the lowercase Greek letter sigma (σ) as *sigma* since it has an entirely different meaning in statistics, as described in Chapter 4.]

In Formula 3.1, the variable *X* can be replaced, in turn, by each of the 20 presidential terms in Table 3.1, beginning with 4 and ending with 8. The symbol Σ, the uppercase Greek letter sigma, specifies that all scores represented by the variable *X* be added (4 +. . . + 4 + 8) to find the sum of 112. (Notice that this sum contains the values of *all*

Table 4.4
CALCULATION OF SAMPLE STANDARD DEVIATION (*S*)
(COMPUTATION FORMULA)

A. COMPUTATIONAL SEQUENCE

Assign a value to *n* representing the number of *X* scores **1**
Sum all *X* scores **2**
Square the sum of all *X* scores **3**
Square each *X* score **4**
Sum all squared *X* scores **5**
Substitute numbers into the formula to obtain the sum of squares, *SS* **6**
Substitute numbers into the formula to obtain the sample variance, s^2 **7**
Take the square root of s^2 to obtain the sample standard deviation, *s* **8**

B. DATA AND COMPUTATIONS

	4
X	X^2
7	49
3	9
1	1
0	0
4	16

1 $n = 5$ **2** $\Sigma X = 15$ **5** $\Sigma X^2 = 75$
3 $(\Sigma X)^2 = 225$

6 $$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = 75 - \frac{225}{5} = 75 - 45 = 30$$

7 $$s^2 = \frac{SS}{n-1} = \frac{30}{4} = 7.50 \quad \textbf{8}\ s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{30}{4}} = \sqrt{7.50} = 2.74$$

Progress Check * 4.5 Using the computation formula for the sum of squares, calculate the population standard deviation for the scores in (a) and the sample standard deviation for the scores in (b).

(a) 1, 3, 7, 2, 0, 4, 7, 3 **(b)** 10, 8, 5, 0, 1, 1, 7, 9, 2

Progress Check *4.6 Days absent from school for a *sample* of 10 first-grade children are: 8, 5, 7, 1, 4, 0, 5, 7, 2, 9.

a) Before calculating the standard deviation, decide whether the definitional or computational formula would be more efficient. Why?

b) Use the more efficient formula to calculate the sample standard deviation.

Answers on page 425.

Why *n* – 1?

Using $n - 1$ in the denominator of Formulas 4.7 and 4.8 solves a problem in inferential statistics associated with generalizations from samples to populations. The adequacy of these generalizations usually depends on *accurately* estimating unknown variability in the population with known variability in the sample. But if we were to use

n rather than $n - 1$ in the denominator of our estimates, they would tend to underestimate variability in the population because n is too large. This tendency would compromise any subsequent generalizations, such as whether observed mean differences are real or merely transitory. On the other hand, when the denominator is made smaller by using $n - 1$, variability in the population is estimated more accurately, and subsequent generalizations are more likely to be valid.

Assume that the five scores (7, 3, 1, 0, 4) in Table 4.3 are a random sample from some population whose unknown variability is to be estimated with the sample variability. To understand why $n - 1$ works, let's look more closely at deviation scores. Formula 4.3, the definition formula for the sample sum of squares, specifies that each of the five original scores, X, be expressed as positive or negative deviations from their sample mean, $\bar{X}$, of 3. At this point, a subtle mathematical restriction causes a complication. It's always true, as demonstrated on the left-hand side of **Table 4.5,** that *the sum of all scores, when expressed as deviations about their own mean, equals zero.* (If you're skeptical, recall the discussion on page 52 about the mean as a balance point that equalizes the sums of all positive and negative deviations.) Given values for *any* four of the five deviations on the left-hand side of Table 4.5, the value of the remaining deviation is not free to vary. Instead, its value is completely fixed because it must comply with the mathematical restriction that the sum of all deviations about *their own mean* equals zero. For instance, given the sum for the four top deviations on the left-hand side of Table 4.5, that is, [4 + 0 + (–2) + (–3) = –1], the value of the bottom deviation must equal 1, as it does, because of the zero-sum restriction, that is, [–1 + 1 = 0]. Or since this mathematical restriction applies to *any* four of the five deviations, given the sum for the four bottom deviations in Table 4.5, that is, [0 + (–2) + (–3) + 1 = – 4], the value of the top deviation must equal 4 because [– 4 + 4 = 0].

If μ Is Known

For the sake of the present discussion, now assume that we know the value of the population mean, μ —let's say it equals 2. (Any value assigned to μ other than 3, the value of $\bar{X}$, would satisfy the current argument. It's reasonable to assume that the values of μ and $\bar{X}$ will differ because a random sample exactly replicates its population rarely, if at all.) Furthermore, assume that we take a random sample of $n = 5$

Table 4.5
TWO ESTIMATES OF POPULATION VARIABILITY

WHEN μ IS UNKNOWN ($\bar{X} = 3$)			WHEN μ IS KNOWN ($\mu = 2$)		
X	$X - \bar{X}$	$(X - \bar{X})^2$	X	$X - \mu$	$(X - \mu)^2$
7	$7 - 3 = 4$	16	7	$7 - 2 = 5$	25
3	$3 - 3 = 0$	0	3	$3 - 2 = 1$	1
1	$1 - 3 = -2$	4	1	$1 - 2 = -1$	1
0	$0 - 3 = -3$	9	0	$0 - 2 = -2$	4
4	$4 - 3 = 1$	1	4	$4 - 2 = 2$	4
	$\Sigma(X - \bar{X}) = 0$	$\Sigma(X - \bar{X})^2 = 30$		$\Sigma(X - \mu) = 5$	$\Sigma(X - \mu)^2 = 35$
	$df = n - 1 = 5 - 1 = 4$			$df = n = 5$	
	$s^2(df = n-1) = \frac{\Sigma(X-\bar{X})^2}{n-1} = \frac{30}{4} = 7.50$			$s^2(df = n) = \frac{\Sigma(X-\mu)^2}{n} = \frac{35}{5} = 7.00$	

population deviation scores, $X - \mu$. Then these five known deviation scores will serve as the initial basis for estimating the unknown variability in the population. As demonstrated on the right-hand side of Table 4.5, the sum of the five deviations about μ, that is, $[5 + 1 + (-1) + (-2) + 2]$, equals not 0 but 5. The zero-sum restriction applies only if the five deviations are expressed around *their own* mean—that is, the sample mean, $\bar{X}$, of 3. It does not apply when the five deviations are expressed around *some other* mean, such as the population mean, μ, of 2 for the entire population. In this case, since all five deviations are free to vary, *each* provides valid information about the variability in the population. Therefore, when calculating the sample variance based on a random sample of five population deviation scores, $X - \mu$, it would be appropriate to divide this sample sum of squares by the n of 5, as shown on the right-hand side of Table 4.5.

If μ Is Unknown

It would be most efficient if, as above, we could use a random sample of n deviations expressed around the population mean, $X - \mu$, to estimate variability in the population. But this is usually impossible because, in fact, the population mean is unknown. Therefore, we must substitute the known sample mean, $\bar{X}$, for the unknown population mean, μ, and we must use a random sample of n deviations expressed around their own sample mean, $X - \bar{X}$, to estimate variability in the population. Although there are $n = 5$ deviations in the sample, only $n - 1 = 4$ of these deviations are free to vary because the sum of the $n = 5$ deviations from *their own sample mean* always equals zero.

Only $n - 1$ of the sample deviations supply valid information for estimating variability. One bit of valid information has been lost because of the zero-sum restriction when the sample mean replaces the population mean. And that's why we divide the sum of squares for $X - \bar{X}$ by $n - 1$, as on the left-hand side of Table 4.5.

4.6 DEGREES OF FREEDOM (*df*)

Technically, we have been discussing a very important notion in inferential statistics known as degrees of freedom.

Degrees of Freedom (df)
The number of values free to vary, given one or more mathematical restrictions.

***Degrees of freedom (df)* refers to the number of values that are free to vary, given one or more mathematical restrictions, in a sample being used to *estimate* a population characteristic.**

The concept of degrees of freedom is introduced only because we are using scores in a sample to *estimate* some unknown characteristic of the population. Typically, when used as an estimate, not all observed values in the sample are free to vary because of one or more mathematical restrictions. As has been noted, when n deviations about the sample mean are used to estimate variability in the population, only $n - 1$ are free to vary. As a result, there are only $n - 1$ degrees of freedom, that is, $df = n - 1$. One df is lost because of the zero-sum restriction.

If the sample sum of squares were divided by n, it would tend to underestimate variability in the population. (In Table 4.5, when μ is unknown, division by n instead of $n - 1$ would produce a smaller estimate of 6.00 instead of 7.50.) This would occur because there are only $n - 1$ independent deviations (estimates of variability) in the sample sum of squares. A more accurate estimate is obtained when the denominator term reflects the number of independent deviations—that is, the number of degrees of

freedom—in the numerator, as in the formulas for s^2 and s. In fact, we can use degrees of freedom to rewrite the formulas for the sample variance and standard deviation:

VARIANCE FOR SAMPLE

$$s^2 = \frac{SS}{n-1} = \frac{SS}{df} \tag{4.9}$$

STANDARD DEVIATION FOR SAMPLE

$$s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{SS}{df}} \tag{4.10}$$

where s^2 and s represent the sample variance and standard deviation, SS is the sum of squares as defined in either Formula 4.3 or 4.4, and df is the degrees of freedom and equals $n - 1$.

Other Mathematical Restrictions

The notion of degrees of freedom is used extensively in inferential statistics. We'll encounter other mathematical restrictions, and sometimes more than one degree of freedom will be lost. In any event, however, degrees of freedom (df) always indicate the number of values that are free to vary, given one or more mathematical restrictions, in a set of values used to estimate some unknown population characteristic.

Progress Check *4.7 As a first step toward modifying his study habits, Phil keeps daily records of his study time.

(a) During the first two weeks, Phil's mean study time equals 20 hours per week. If he studied 22 hours during the first week, how many hours did he study during the second week?

(b) During the first four weeks, Phil's mean study time equals 21 hours. If he studied 22, 18, and 21 hours during the first, second, and third weeks, respectively, how many hours did he study during the fourth week?

(c) If the information in (a) and (b) is to be used to estimate some unknown population characteristic, the notion of degrees of freedom can be introduced. How many degrees of freedom are associated with (a) and (b)?

(d) Describe the mathematical restriction that causes a loss of degrees of freedom in (a) and (b).

Answers on page 425.

4.7 INTERQUARTILE RANGE (IQR)

Interquartile Range (IQR)
The range for the middle 50 percent of the scores.

The most important spinoff of the range, the **interquartile range (IQR),** *is simply the range for the middle 50 percent of the scores.* More specifically, the IQR equals the distance between the third quartile (or 75th percentile) and the first quartile (or 25th percentile), that is, after the highest quarter (or top 25 percent) and the lowest quarter (or bottom 25 percent) have been trimmed from the original set of scores. Since most distributions are spread more widely in their extremities than their middle, the IQR tends to be less than half the size of the range.

Table 4.6
CALCULATION OF THE IQR

A. INSTRUCTIONS

1. Order scores from least to most.
2. To determine how far to penetrate the set of ordered scores, begin at either end, then add 1 to the total number of scores and divide by 4. If necessary, round the result to the nearest whole number.
3. Beginning with the largest score, count the requisite number of steps (calculated in step 2) into the ordered scores to find the location of the third quartile.
4. The third quartile equals the value of the score at this location.
5. Beginning with the smallest score, again count the requisite number of steps into the ordered scores to find the location of the first quartile.
6. The first quartile equals the value of the score at this location.
7. The IQR equals the third quartile minus the first quartile.

B. EXAMPLE

1. 7, 9, 9, 10, 11, 11, 13
2. $(7 + 1)/4 = 2$
3. 7, 9, 9, 10, 11, 11, 13
   2 1
4. third quartile = 11
5. 7, 9, 9, 10, 11, 11, 13
   1 2
6. first quartile = 9
7. IQR = 11 − 9 = 2

The calculation of the IQR is relatively straightforward, as you can see by studying **Table 4.6.** This table shows that the IQR equals 2 for distribution C (7, 9, 9, 10, 11, 11, 13) shown in Figure 4.1.

Not Sensitive to Extreme Scores

A key property of the IQR is its resistance to the distorting effect of extreme scores, or outliers. For example, if the smallest score (7) in distribution C of Figure 4.1 were replaced by a much smaller score (for instance, 1), the value of the IQR would remain the same (2), although the value of the original range (6) would be larger (12). Thus, if you are concerned about possible distortions caused by extreme scores, or outliers, use the IQR as the measure of variability, along with the median (or second quartile) as the measure of central tendency.

Progress Check *4.8 Determine the values of the range and the IQR for the following sets of data.

(a) Retirement ages: 60, 63, 45, 63, 65, 70, 55, 63, 60, 65, 63

(b) Residence changes: 1, 3, 4, 1, 0, 2, 5, 8, 0, 2, 3, 4, 7, 11, 0, 2, 3, 4

Answers on page 425.

4.8 MEASURES OF VARIABILITY FOR QUALITATIVE AND RANKED DATA

Qualitative Data

Measures of variability are virtually nonexistent for qualitative or nominal data. It is probably adequate to note merely whether scores are evenly divided among the various classes (maximum variability), unevenly divided among the various classes (intermediate variability), or concentrated mostly in one class (minimum variability). For example, if the ethnic composition of the residents of a city is about evenly divided among several groups, the variability with respect to ethnic groups is maximum; there is considerable heterogeneity. (An inspection of county population data from the 2010 census, available on the Internet at http://factfinder.census.gov, reveals that the greatest ethnic variability occurs in large urban counties, such as Bronx County in New York and San Francisco County in California.) At the other extreme, if almost all the residents are concentrated in a single ethnic group, the variability will be minimum; there is little heterogeneity. (According to the previous source, virtually no ethnic variability occurs in sparsely populated rural counties, such as Hooker County in Nebraska and King County in Texas, with an almost exclusively white population.) If the ethnic composition falls between these two extremes—because of an uneven division among several large ethnic groups—the variability will be intermediate, as is true of many U.S. cities and counties.

Ordered Qualitative and Ranked Data

If qualitative data can be ordered because measurement is ordinal (or if the data are ranked), then it's appropriate to describe variability by identifying extreme scores (or ranks). For instance, the active membership of an officers' club might include no one with a rank below first lieutenant or above brigadier general.

Summary

Measures of variability reflect the amount by which observations are dispersed or scattered in a distribution. These measures assume a key role in the analysis of research results.

The simplest measure of variability, the range, is readily calculated and understood, but it has two shortcomings.

Among measures of variability, the variance and particularly the standard deviation occupy the same exalted position as does the mean among measures of central tendency.

The variance is a type of mean, that is, the mean of all squared deviations about their mean. To avoid mind-boggling squared units of measurement, we take the square root of the variance to obtain the standard deviation.

The standard deviation is a rough measure of the average or typical amount by which scores deviate on either side of their mean.

For most frequency distributions, a majority of all scores are within one standard deviation of their mean, and a small minority of all scores deviate more than two standard deviations on either side of their mean.

Unlike the mean, which is a measure of position, the standard deviation is a measure of distance.

Calculation of either the population standard deviation (σ) or the sample standard deviation (s) requires three steps:

1. Calculate the sum of all squared deviation scores (*SS)* using either the definition or computation formula.

2. Divide the *SS* by *N*, the population size, to obtain the population variance (σ^2) or divide the *SS* by $n - 1$, the sample size minus 1, to obtain the sample variance (s^2).
3. Take the square root of the variance to obtain the population standard deviation (σ) or the sample standard deviation (s).

The denominator of the formulas for sample variance and standard deviation reflects the fact that, because of the zero-sum restriction, only $n - 1$ of the sample deviation scores provide valid estimates of population variability.

Whenever we estimate unknown population characteristics, we must be concerned about the number of degrees of freedom (*df*) associated with our estimate. Degrees of freedom specify the number of values that are free to vary, given one or more mathematical restrictions. When estimating the population variance and standard deviation, degrees of freedom equal $n - 1$.

The interquartile range (IQR) is resistant to the distorting effects of extreme scores.

Measures of variability are virtually nonexistent for qualitative and ranked data.

Important Terms

Measures of variability
Range
Standard deviation
Population standard deviation (σ)
Degrees of freedom (*df*)
Variance
Sum of squares (*SS*)
Sample standard deviation (s)
Interquartile range (IQR)

Key Equations

STANDARD DEVIATION FOR SAMPLE

$$s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{SS}{df}}$$

where $$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

REVIEW QUESTIONS

***4. 9** For each of the following pairs of distributions, first decide whether their standard deviations are about the same or different. If their standard deviations are different, indicate which distribution should have the larger standard deviation. **Hint:** The distribution with the more dissimilar set of scores or individuals should produce the larger standard deviation regardless of whether, *on average*, scores or individuals in one distribution differ from those in the other distribution.

(a) SAT scores for all graduating high school seniors (a_1) or all college freshmen (a_2)

(b) Ages of patients in a community hospital (b_1) or a children's hospital (b_2)

(c) Motor skill reaction times of professional baseball players (c_1) or college students (c_2)

(d) GPAs of students at some university as revealed by a random sample (d_1) or a census of the entire student body (d_2)

(e) Anxiety scores (on a scale from 0 to 50) of a random sample of college students taken from the senior class (e_1) or those who plan to attend an anxiety-reduction clinic (e_2)

(f) Annual incomes of recent college graduates (f_1) or of 20-year alumni (f_2)

Answers on page 425.

4.10 When not interrupted artificially, the duration of human pregnancies can be described, we'll assume, by a mean of 9 months (270 days) and a standard deviation of one-half month (15 days).

(a) Between what two times, in days, will a majority of babies arrive?

(b) A small minority of all babies will arrive sooner than _____.

(c) A small minority of all babies will arrive later than _____.

(d) In a paternity suit, the suspected father claims that, since he was overseas during the entire 10 months prior to the baby's birth, he could not possibly be the father. Any comment?

4.11 Add 10 to each of the scores in Question 4.4 (1, 3, 4, 4) to produce a new distribution (11, 13, 14, 14). Would you expect the value of the sample standard deviation to be the same for both the original and new distributions? Explain your answer, and then calculate *s* for the new distribution.

4.12 Add 10 to only the smallest score in Question 4.4 (1, 3, 4, 4) to produce another new distribution (11, 3, 4, 4). Would you expect the value of *s* to be the same for both the original and new distributions? Explain your answer, and then calculate *s* for the new distribution.

***4.13** **(a)** While in office, a former governor of California proposed that all state employees receive the same pay raise of $70 per month. What effect, if any, would this raise have had on the mean and the standard deviation for the distribution of monthly wages in existence before the proposed raise? **Hint:** Imagine the effect of adding $70 to the monthly wages of each state employee on the mean and on the standard deviation (or on a more easily visualized measure of variability, such as the range).

(b) Other California officials suggested that all state employees receive a pay raise of 5 percent. What effect, if any, would this raise have had on the mean and the standard deviation for the distribution of monthly wages in existence before the proposed raise? **Hint:** Imagine the effect of multiplying the monthly wages of each state employee by 5 percent on the mean and on the standard deviation or on the range.

Answers on page 426.

4.14 **(a)** Using the computation formula for the sample sum of squares, verify that the sample standard deviation, *s*, equals 23.33 lbs for the distribution of 53 weights in Table 1.1.

(b) Verify that a majority of all weights fall within one standard deviation of the mean (169.51) and that a small minority of all weights deviate more than two standard deviations from the mean.

4.15 In what sense is the variance

(a) a type of mean?

(b) not a readily understood measure of variability?

(c) a stepping stone to the standard deviation?

4.16 Specify an important difference between the standard deviation and the mean.

4. 17 Why can't the value of the standard deviation ever be negative?

***4. 18** Indicate whether each of the following statements about degrees of freedom is true or false.

(a) Degrees of freedom refer to the number of values free to vary in the population.

(b) One degree of freedom is lost because, when expressed as a deviation from the sample mean, the final deviation in the sample fails to supply information about population variability.

(c) Degrees of freedom makes sense only if we wish to estimate some unknown characteristic of a population.

(d) Degrees of freedom reflect the poor quality of one or more observations.

Answers on page 426.

4. 19 Referring to Review Question 2.18 on page 46, would you describe the distribution of majors for all male graduates as having maximum, intermediate, or minimum variability?

CHAPTER

5

Normal Distributions and Standard (z) Scores

Preview

The familiar bell-shaped normal curve describes many observed frequency distributions, including scores on IQ tests, slight measurement errors made by a succession of people who attempt to measure precisely the same thing, the useful lives of 100-watt electric light bulbs, and even the heights of stalks in a field of corn. As will become apparent in later chapters, the normal curve also describes some important theoretical distributions in inferential statistics.

Thanks to the standard normal table, we can answer questions about any normal distribution whose mean and standard deviation are known. In the long run, this proves to be both more accurate and more efficient than dealing directly with each observed frequency distribution. Use of the standard normal table requires a familiarity with z scores. Regardless of the original measurements—whether IQ points, measurement errors in millimeters, or reaction times in milliseconds—z scores are "pure" or unit-free numbers that indicate how many standard deviation units an observation is above or below the mean.

In the classic movie *The President's Analyst*, the director of the Federal Bureau of Investigation, rather short himself, encourages the recruitment of similarly short FBI agents. If, in fact, FBI agents are to be selected only from among applicants who are no taller than exactly 66 inches, what proportion of all of the original applicants will be eligible? This question can't be answered without additional information.

One source of additional information is the relative frequency distribution of heights for the 3091 men shown in **Figure 5.1**. To find the proportion of men who are a particular height, merely note the value of the vertical scale that corresponds to the top of any bar in the histogram. For example, .10 of these men, that is, one-tenth of 3091, or about 309 men, are 70 inches tall.

When expressed as a proportion, any conclusion based on the 3091 men can be generalized to other comparable sets of men, even sets containing an unspecified number. For instance, if the distribution in Figure 5.1 is viewed as representative of all men who apply for FBI jobs, we can estimate that .10 of all applicants will be 70 inches tall. Or, given the director's preference for shorter agents, we can use the same distribution to estimate the proportion of applicants who will be eligible. To obtain the estimated proportion of eligible applicants (.165) from Figure 5.1, add the values associated with the shaded bars. (Only half of the bar at 66 inches is shaded to adjust for the fact that any height between 65.5 and 66.5 inches is reported as 66 inches, whereas eligible applicants must be shorter than *exactly* 66 inches, that is, 66.0 inches.)

The distribution in Figure 5.1 has an obvious limitation: It is based on a group of just 3091 men that, at most, only resembles the distributions for other groups of men, including the group of FBI applicants. Therefore, any generalization will contain inaccuracies due to chance irregularities in the original distribution.

5.1 THE NORMAL CURVE

More accurate generalizations usually can be obtained from distributions based on larger numbers of men. A distribution based on 30,910 men usually is more accurate than one based on 3091, and a distribution based on 3,091,000 usually is even more accurate. But it is prohibitively expensive in both time and money to even survey 30,910 people. Fortunately, it is a fact that the distribution of heights for all

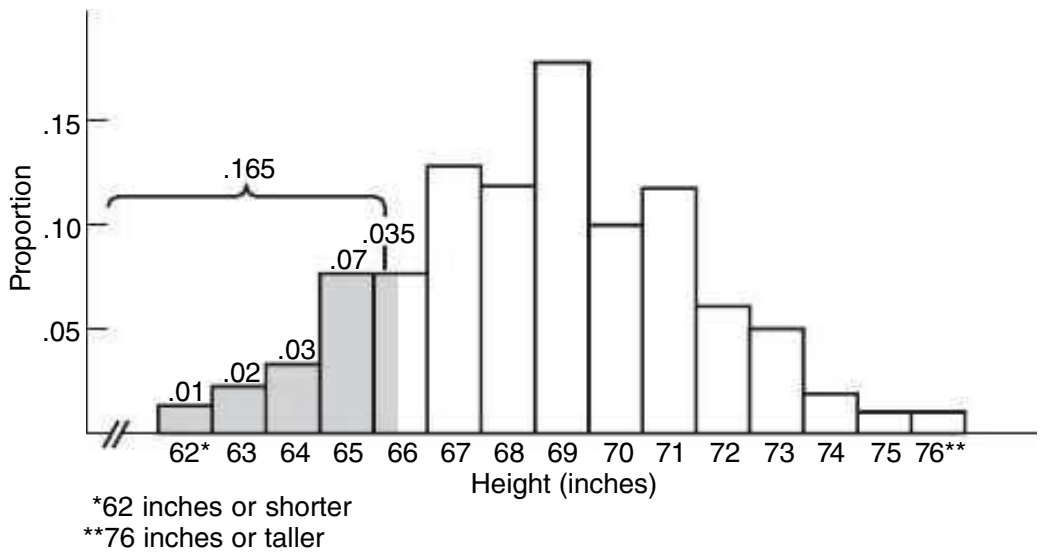


FIGURE 5.1
Relative frequency distribution for heights of 3091 men.
Source: National Center for Health Statistics, 1960–62, Series 11, No.14. Mean updated by authors.

American men—not just 3091 or even 3,091,000—approximates the normal curve, a well-documented theoretical curve.

In **Figure 5.2**, the idealized normal curve has been superimposed on the original distribution for 3091 men. Irregularities in the original distribution, most likely due to chance, are ignored by the smooth normal curve. Accordingly, any generalizations based on the smooth normal curve will tend to be more accurate than those based on the original distribution.

Interpreting the Shaded Area

The total area under the normal curve in Figure 5.2 can be identified with all FBI applicants. Viewed relative to the total area, the shaded area represents the proportion of applicants who will be eligible because they are shorter than exactly 66 inches. This new, more accurate proportion will differ from that obtained from the original histogram (.165) because of discrepancies between the two distributions.

Finding a Proportion for the Shaded Area

To find this new proportion, we cannot rely on the vertical scale in Figure 5.2, because it describes as proportions the areas in the rectangular bars of histograms, not the areas in the various curved sectors of the normal curve. Instead, in Section 5.3 we will learn how to use a special table to find the proportion represented by any area under the normal curve, including that represented by the shaded area in Figure 5.2.

Properties of the Normal Curve

Let's note several important properties of the normal curve:

Normal Curve

A theoretical curve noted for its symmetrical bell-shaped form.

- Obtained from a mathematical equation, the **normal curve** is a theoretical curve defined for a continuous variable, as described in Section 1.6, and noted for its symmetrical bell-shaped form, as revealed in Figure 5.2.
- Because the normal curve is symmetrical, its lower half is the mirror image of its upper half.
- Being bell shaped, the normal curve peaks above a point midway along the horizontal spread and then tapers off gradually in either direction from the peak (without actually touching the horizontal axis, since, in theory, the tails of a normal curve extend infinitely far).
- The values of the mean, median (or 50th percentile), and mode, located at a point midway along the horizontal spread, are the same for the normal curve.

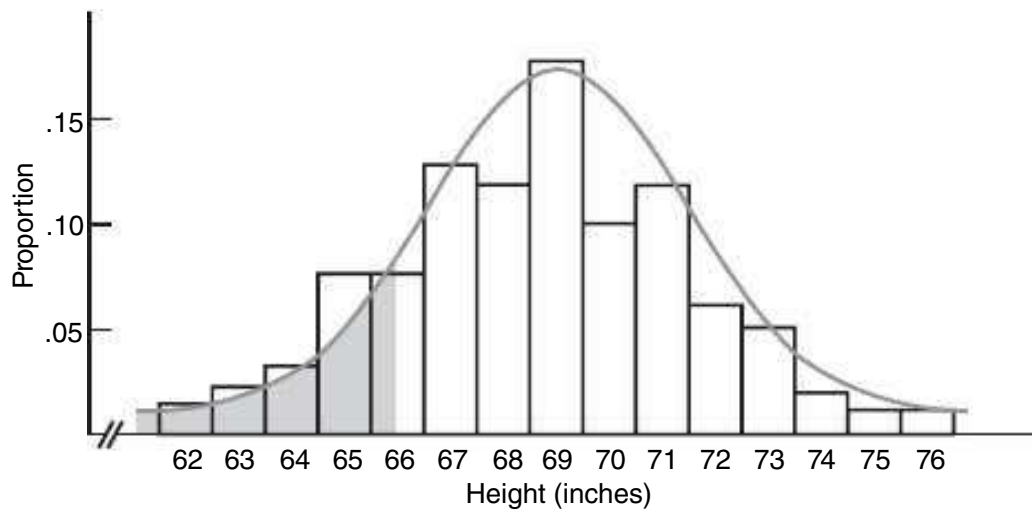


FIGURE 5.2
Normal curve superimposed on the distribution of heights.

Importance of Mean and Standard Deviation

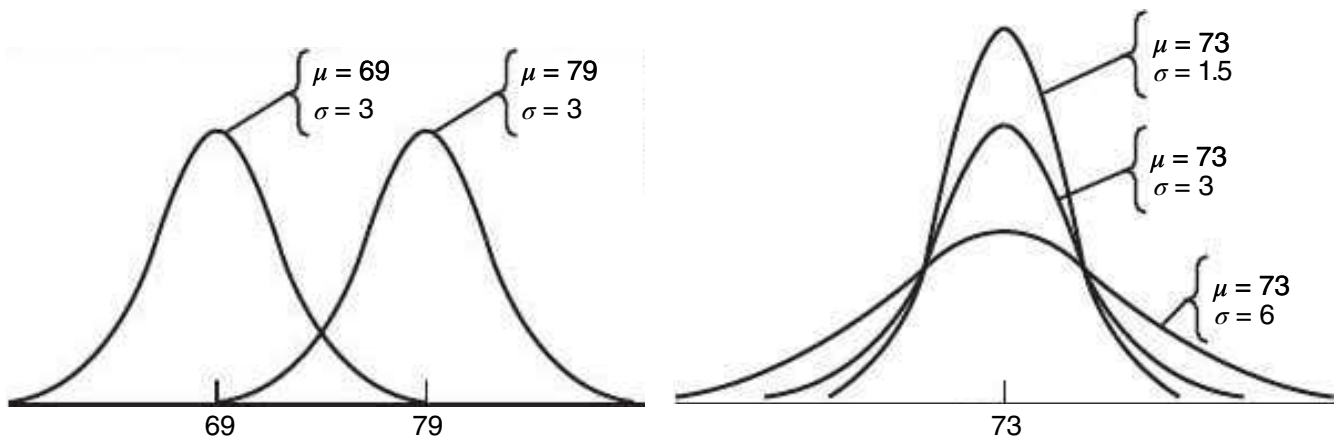
When you're using the normal curve, two bits of information are indispensable: values for the mean and the standard deviation. For example, before the normal curve can be used to answer the question about eligible FBI applicants, it must be established that, for the original distribution of 3091 men, the mean height equals 69 inches and the standard deviation equals 3 inches.

Different Normal Curves

Having established that a particular normal curve has a mean of 69 inches and a standard deviation of 3 inches, we can't arbitrarily change these values, as any change in the value of either the mean or the standard deviation (or both) would create a new normal curve that no longer describes the original distribution of heights. Nevertheless, as a theoretical exercise, it is instructive to note the various types of normal curves that are produced by an arbitrary change in the value of either the mean (μ) or the standard deviation (σ).*

For example, changing the mean height from 69 to 79 inches produces a new normal curve that, as shown in panel A of **Figure 5.3**, is displaced 10 inches to the right of the original curve. Dramatically new normal curves are produced by changing the value of the standard deviation. As shown in panel B of Figure 5.3, changing the standard deviation from 3 to 1.5 inches produces a more peaked normal curve with smaller variability, whereas changing the standard deviation from 3 to 6 inches produces a shallower normal curve with greater variability.

Obvious differences in appearance among normal curves are less important than you might suspect. Because of their common mathematical origin, every normal curve can be interpreted in exactly the same way *once any distance from the mean is expressed in standard deviation units*. For example, .68, or 68 percent of the total area under a normal curve—any normal curve—is within one standard deviation above and below the mean, and only .05, or 5 percent, of the total area is more than two standard deviations above and below the mean. And this is only the tip of the iceberg. Once any distance from the mean has been expressed in standard deviation units, we will be able to consult the standard normal table, described in Section 5.3, to determine the corresponding proportion of the area under the normal curve.



A. Different Means, Same Standard Deviation

B. Same Mean, Different Standard Deviations

FIGURE 5.3
Different normal curves.

*Since the normal curve is an idealized curve that is presumed to describe a complete set of observations or a population, the symbols μ and σ, representing the mean and standard deviation of the population, respectively, will be used in this chapter.

5.2 z SCORES

***z* Score**
A unit-free, standardized score that indicates how many standard deviations a score is above or below the mean of its distribution.

A *z* score is a unit-free, standardized score that, regardless of the original units of measurement, indicates how many standard deviations a score is above or below the mean of its distribution.

To obtain a z score, express any original score, whether measured in inches, milliseconds, dollars, IQ points, etc., as a deviation from its mean (by subtracting its mean) and then split this deviation into standard deviation units (by dividing by its standard deviation), that is,

z SCORE

$$z = \frac{X - \mu}{\sigma} \tag{5.1}$$

where X is the original score and μ and σ are the mean and the standard deviation, respectively, for the normal distribution of the original scores. Since identical units of measurement appear in both the numerator and denominator of the ratio for z, the original units of measurement cancel each other and the z score emerges as a unit-free or standardized number, often referred to as a standard score.

A z score consists of two parts:

1. *a positive or negative* sign indicating whether it's above or below the mean; and
2. *a number* indicating the size of its deviation from the mean in standard deviation units.

A z score of 2.00 always signifies that the original score is exactly two standard deviations above its mean. Similarly, a z score of –1.27 signifies that the original score is exactly 1.27 standard deviations below its mean. A z score of 0 signifies that the original score coincides with the mean.

Converting to *z* Scores

To answer the question about eligible FBI applicants, replace X with 66 (the maximum permissible height), μ with 69 (the mean height), and σ with 3 (the standard deviation of heights) and solve for z as follows:

$$\frac{66 - 69}{3} = \frac{-3}{3} = -1$$

This informs us that the cutoff height is exactly one standard deviation below the mean. Knowing the value of z, we can use the table for the standard normal curve to find the proportion of eligible FBI applicants. First, however, we'll make a few comments about the standard normal curve.

Progress Check *5.1 Express each of the following scores as a z score:

(a) Margaret's IQ of 135, given a mean of 100 and a standard deviation of 15

(b) a score of 470 on the SAT math test, given a mean of 500 and a standard deviation of 100

(c) a daily production of 2100 loaves of bread by a bakery, given a mean of 2180 and a standard deviation of 50

(d) Sam's height of 69 inches, given a mean of 69 and a standard deviation of 3

(e) a thermometer-reading error of –3 degrees, given a mean of 0 degrees and a standard deviation of 2 degrees

Answers on page 426.

5.3 STANDARD NORMAL CURVE

Standard Normal Curve

The tabled normal curve for Z scores, with a mean of 0 and a standard deviation of 1.

If the original distribution approximates a normal curve, then the shift to standard or z scores will always produce a new distribution that approximates the **standard normal curve**. This is the one normal curve for which a table is actually available. It is a mathematical fact—not proven in this book—that the standard normal curve always has a mean of 0 and a standard deviation of 1. However, to verify (rather than prove) that the mean of a standard normal distribution equals 0, replace X in the z score formula with μ, the mean of any (nonstandard) normal distribution, and then solve for z:

$$\text{Mean of } z = \frac{X-\mu}{\sigma} = \frac{\mu-\mu}{\sigma} = \frac{0}{\sigma} = 0$$

Likewise, to verify that the standard deviation of the standard normal distribution equals 1, replace X in the z score formula with $\mu + 1\sigma$, the value corresponding to one standard deviation above the mean for any (nonstandard) normal distribution, and then solve for z:

$$\text{Standard deviation of } z = \frac{X-\mu}{\sigma} = \frac{\mu+1\sigma-\mu}{\sigma} = \frac{1\sigma}{\sigma} = 1$$

Although there is an infinite number of different normal curves, each with its own mean and standard deviation, there is only one standard normal curve, with a mean of 0 and a standard deviation of 1.

Figure 5.4 illustrates the emergence of the standard normal curve from three different normal curves: that for the men's heights, with a mean of 69 inches and a standard deviation of 3 inches; that for the useful lives of 100-watt electric light bulbs, with a mean of 1200 hours and a standard deviation of 120 hours; and that for the IQ scores of fourth graders, with a mean of 105 points and a standard deviation of 15 points.

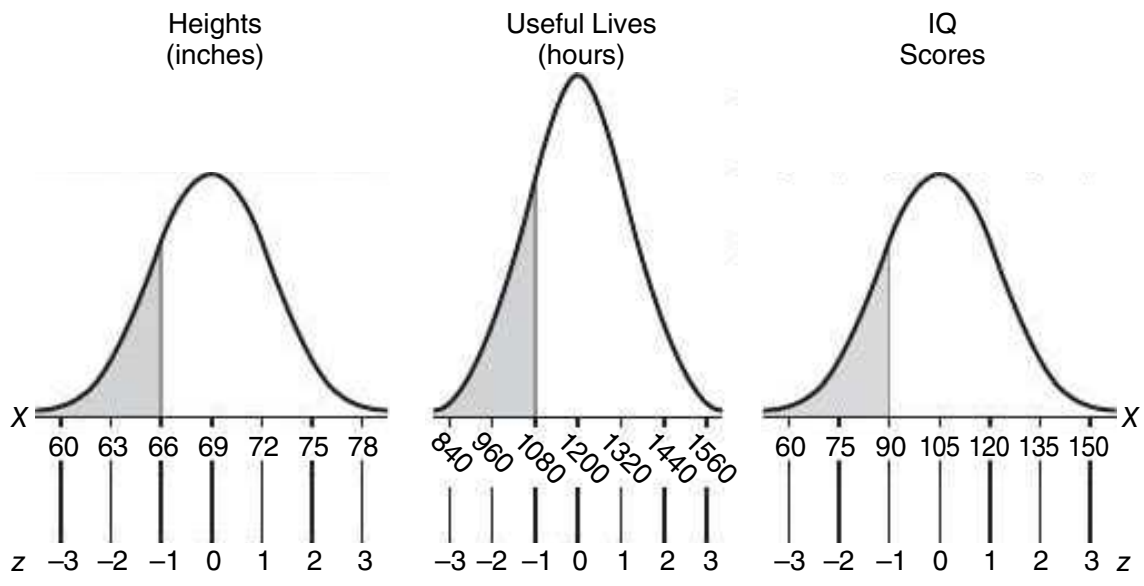


FIGURE 5.4

Converting three normal curves to the standard normal curve.

Converting all original observations into z scores leaves the normal shape intact but not the units of measurement. Shaded observations of 66 inches, 1080 hours, and 90 IQ points all reappear as a z score of –1.00. Verify this by using the z score formula. Showing no traces of the original units of measurement, this z score contains the one crucial bit of information common to the three original observations: All are located one standard deviation below the mean. Accordingly, to find the proportion for the shaded areas in Figure 5.4 (that is, the proportion of applicants who are less than exactly 66 inches tall, or light bulbs that burn for fewer than 1080 hours, or fourth graders whose IQ scores are less than 90), we can use the same z score of –1.00 when referring to the table for the standard normal curve, the one table for all normal curves.

Standard Normal Table

Essentially, the standard normal table consists of columns of z scores coordinated with columns of proportions. In a typical problem, access to the table is gained through a z score, such as –1.00, and the answer is read as a proportion, such as the proportion of eligible FBI applicants.

Using the Top Legend of the Table

Table 5.1 shows an abbreviated version of the standard normal curve, while Table A in Appendix C on page 458 shows a more complete version of the same curve. Notice that columns are arranged in sets of three, designated as A, B, and C in the legend at the top of the table. When using the top legend, all entries refer to the upper half of the standard normal curve. The entries in column A are z scores, beginning with 0.00 and ending (in the full-length table of Appendix C) with 4.00. Given a z score of zero or more, columns B and C indicate how the z score splits the area in the upper half of the normal curve. As suggested by the shading in the top legend, column B indicates the proportion of area between the mean and the z score, and column C indicates the proportion of area beyond the z score, in the upper tail of the standard normal curve.

Using the Bottom Legend of the Table

Because of the symmetry of the normal curve, the entries in Table 5.1 and Table A of Appendix C also can refer to the lower half of the normal curve. Now the columns are designated as A′, B′, and C′ in the legend at the bottom of the table. When using the bottom legend, all entries refer to the lower half of the standard normal curve.

Imagine that the nonzero entries in column A′ are negative z scores, beginning with –0.01 and ending (in the full-length table of Appendix C) with –4.00. Given a negative z score, columns B′ and C′ indicate how that z score splits the lower half of the normal curve. As suggested by the shading in the bottom legend of the table, column B′ indicates the proportion of area between the mean and the negative z score, and column C′ indicates the proportion of area beyond the negative z score, in the lower tail of the standard normal curve.

Progress Check *5.2 Using Table A in Appendix C, find the proportion of the total area identified with the following statements:

(a) above a *z* score of 1.80

(b) between the mean and a *z* score of –0.43

(c) below a *z* score of –3.00

(d) between the mean and a *z* score of 1.65

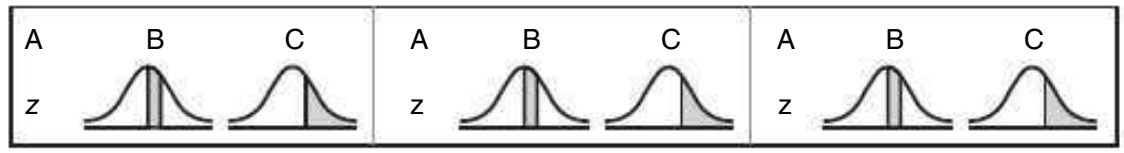
(e) between *z* scores of 0 and –1.96

Answers on page 426.

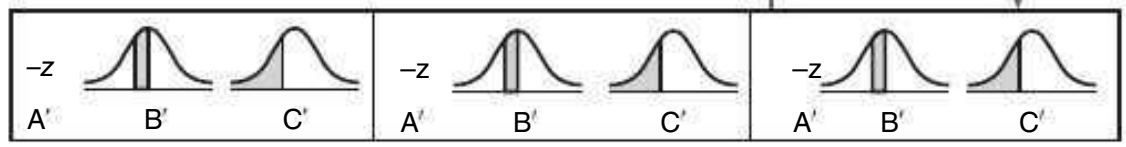
Reminder:

Use of a standard normal table always involves Z scores.

TABLE 5.1
PROPORTIONS (OF AREAS) UNDER THE STANDARD NORMAL CURVE FOR VALUES OF *z* (FROM TABLE A OF APPENDIX C)



A z	B	C	A z	B	C	A z	B	C
0.00	.0000	.5000	0.40	.1554	.3446	0.80	.2881	.2119
0.01	.0040	.4960	0.41	.1591	.3409	0.81	.2910	.2090
•	•	•	•	•	•	•	•	•
						•	•	•
						•	•	•
						•	•	•
•	•	•	•	•	•	•	•	•
						0.99	.3389	.1611
						1.00	.3413	.1587
•	•	•	•	•	•	1.01	.3438	.1562
						•	•	•
						•	•	•
•	•	•	•	•	•	•	•	•
0.38	.1480	.3520	0.78	.2823	.2711	1.18	.3810	.1190
0.39	.1517	.3483	0.79	.2852	.2148	1.19	.3830	.1170



5.4 SOLVING NORMAL CURVE PROBLEMS

Sections 5.5 and 5.6 give examples of two main types of normal curve problems. In the first type of problem, we use a known score (or scores) to find an unknown *proportion*. For instance, we use the known score of 66 inches to find the unknown proportion of eligible FBI applicants. In the second type of problem, the procedure is reversed. Now we use a known proportion to find an unknown *score* (or *scores*). For instance, if the FBI director had specified that applicants' heights must not exceed the 25th percentile (the shortest .25) of the population, we would use the known proportion of .25 to find the unknown cutoff height in inches.

Solve Problems Logically

Do not rush through these examples, memorizing solutions to particular problems or looking for some magic formula. Concentrate on the logic of the solution, *using rough graphs of normal curves as an aid to visualizing the solution.* Only after thinking through to a solution should you do any calculations and consult the normal tables. Then, with just a little practice, you will view the wide variety of normal curve problems not as a bewildering assortment but as many slight variations on two distinctive types.

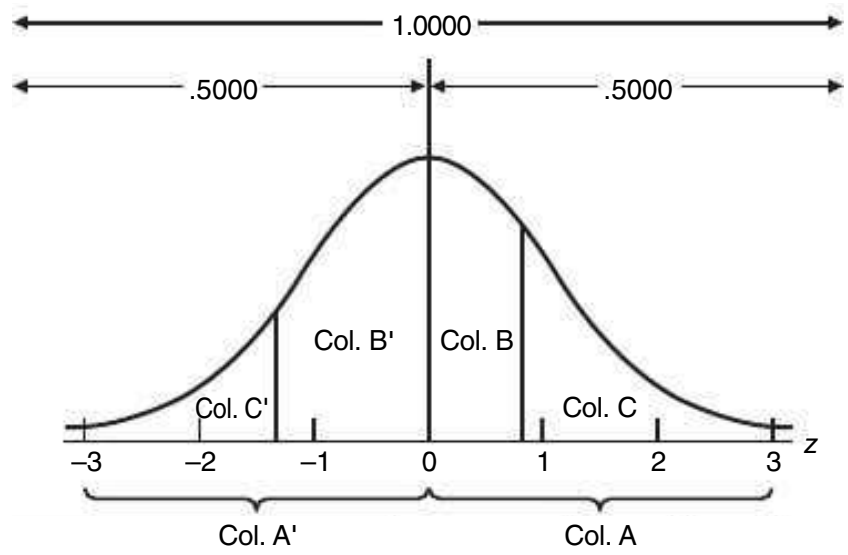


FIGURE 5.5
Interpretation of Table A, Appendix C.

Key Facts to Remember

Reminder:

Z scores can be negative, but areas under the normal curve cannot.

When using the standard normal table, it is important to remember that for any z score, the corresponding proportions in columns B and C (or columns B′ and C′) always sum to .5000. Similarly, the total area under the normal curve always equals 1.0000, the sum of the proportions in the lower and upper halves, that is, .5000 + .5000. Finally, although a z score can be either positive or negative, the proportions of area under the curve are always positive or zero but *never* negative (because an area cannot be negative). **Figure 5.5** summarizes how to interpret the normal curve table in this book.

5.5 FINDING PROPORTIONS

Example: Finding Proportions for *One* Score

Now we'll use a step-by-step procedure, adopted throughout this chapter, to find the proportion of all FBI applicants who are shorter than exactly 66 inches, given that the distribution of heights approximates a normal curve with a mean of 69 inches and a standard deviation of 3 inches.

1. **Sketch a normal curve and shade in the target area**, as in the left part of **Figure 5.6**. Being less than the mean of 69, 66 is located to the left of the mean. Furthermore, since the unknown proportion represents those applicants who are shorter than 66 inches, the shaded target sector is located to the left of 66.
2. **Plan your solution according to the normal table**. Decide precisely how you will find the value of the target area. In the present case, the answer will be obtained from column C′ of the standard normal table, since the target area coincides with the type of area identified with column C′, that is, the area in the lower tail beyond a negative *z*.
3. **Convert *X* to *z***. Express 66 as a z score:

$$z = \frac{X - \mu}{\sigma} = \frac{66 - 69}{3} = \frac{-3}{3} = -1$$

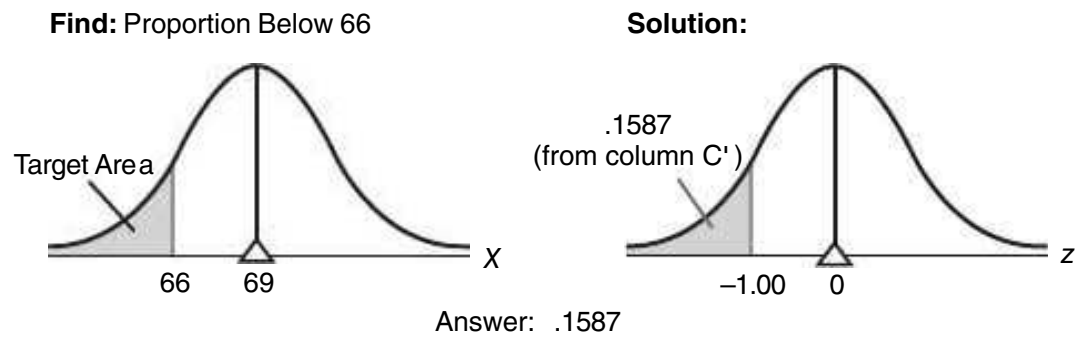


FIGURE 5.6
Finding proportions.

4. **Find the target area**. Refer to the standard normal table, using the bottom legend, as the z score is negative. The arrows in Table 5.1 show how to read the table. Look up column A' to 1.00 (representing a z score of –1.00), and note the corresponding proportion of .1587 in column C': This is the answer, as suggested in the right part of Figure 5.6. It can be concluded that only .1587 (or .16) of all of the FBI applicants will be shorter than 66 inches.

A Clarification

Because the normal curve is defined for continuous variables, such as height, the same proportion of .1587 would describe not only FBI applicants who are shorter than 66 inches, but also FBI applicants who are shorter than *or equal* to 66 inches. If you think about it, equal to 66 inches translates into a height of *exactly* 66 inches—that is, 66.0000 with a string of zeros out to infinity! No measured height can coincide with exactly 66 inches since, in theory, however long the string of zeros for someone's height, measurement always can be carried additional steps until a nonzero appears.

Exactly 66 inches translates into a point along the horizontal base of the normal curve. The vertical line through this point defines one side of the desired area—the portion below 66 inches—but the line itself has no area. Therefore, when doing normal curve problems, you need not agonize over, for example, whether the desired proportion is below exactly 66 inches or below *and equal to* exactly 66 inches. The answer is the same.

Read Carefully

Carefully read normal curve problems. A single word can change the entire problem as, for example, if you had been asked to find the proportion of applicants who are *taller* than 66 inches. Now we must find the total area to the right, not to the left, of 66 inches (or a z score of –1.00) in Figure 5.6. This requires that we add the proportions for two sectors: the unshaded sector between 66 inches and the mean of 69 inches and the unshaded sector above the mean of 69 inches. To find the proportion between 66 and 69 inches, refer to the standard normal table. Use the bottom legend, as the z score is negative; look up column A′ to 1.00 (representing a z score of –1.00); and note the proportion of .3438 in column B′ (which corresponds to the sector between 66 and 69 inches). Recalling that .5000 always equals the proportion in the upper half of the curve (above the mean of 69 inches), add these two proportions (.3438 + .5000 = .8438) to determine that .8438 of all FBI applicants will be taller than 66 inches.

Reminder about Interpreting Areas

When read from left to right, the X and z scales along the base of the normal curve, as in Figure 5.6, always increase in value. Accordingly, the area under the normal curve to the left of any given score represents the proportion of shorter applicants (or, more generally, smaller or lower scores), and the area to the right of any given score represents the proportion of taller applicants (or larger or higher scores).

Progress Check *5.3 Assume that GRE scores approximate a normal curve with a mean of 500 and a standard deviation of 100.

(a) Sketch a normal curve and shade in the target area described by each of the following statements:

(i) less than 400

(ii) more than 650

(iii) less than 700

(b) Plan solutions (in terms of columns B, C, B′, or C′ of the standard normal table, as well as the fact that the proportion for either the entire upper half or lower half always equals .5000) for the target areas in part (a).

(c) Convert to z scores and find the proportions that correspond to the target areas in part (a).

Answers on page 426.

Example: Finding Proportions *between* Two Scores

Assume that, when not interrupted artificially, the gestation periods for human fetuses approximate a normal curve with a mean of 270 days (9 months) and a standard deviation of 15 days. What proportion of gestation periods will be between 245 and 255 days?

1. **Sketch a normal curve and shade in the target area**, as in the top panel of **Figure 5.7**. Satisfy yourself that, in fact, the shaded area represents just those gestation periods between 245 and 255 days.
2. **Plan your solution according to the normal table**. This type of problem requires more effort to solve because the value of the target area cannot be read directly from Table A. As suggested in the bottom two panels of Figure 5.7, the basic idea is to identify the target area with the difference between two overlapping areas whose values can be read from column C′ of Table A. The larger area (less than 255 days) contains two sectors: the target area (between 245 and 255 days) and a remainder (less than 245 days). The smaller area contains only the remainder (less than 245 days). Subtracting the smaller area (less than 245 days) from the larger area (less than 255 days), therefore, eliminates the common remainder (less than 245 days), leaving only the target area (between 245 and 255 days).
3. **Convert X to z** by expressing 255 as

$$z = \frac{255 - 270}{15} = \frac{-15}{15} = -1.00$$

and by expressing 245 as

$$z = \frac{245 - 270}{15} = \frac{-25}{15} = -1.67$$

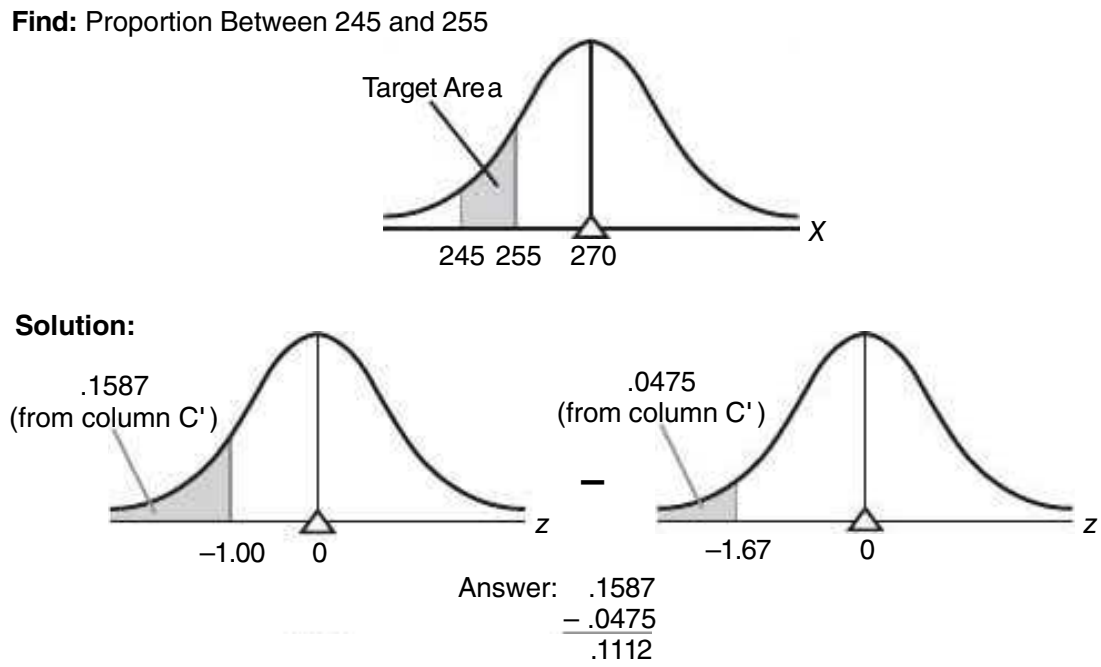


FIGURE 5.7
Finding proportions.

4. **Find the target area**. Look up column A′ to a negative z score of –1.00 (remember, you must imagine the negative sign), and note the corresponding proportion of .1587 in column C′. Likewise, look up column A′ to a z score of –1.67, and note the corresponding proportion of .0475 in column C′. Subtract the smaller proportion from the larger proportion to obtain the answer, .1112. Thus, only .11, or 11 percent, of all gestation periods will be between 245 and 255 days.

Warning: Enter Table Only with Single *z* Score

When solving problems with two z scores, as above, resist the temptation to subtract one z score directly from the other and to enter Table A with this difference. Table A is designed only for individual z scores, not for differences between z scores.

Progress Check 5.4 The previous problem can be solved in another way, using entries from column B′ rather than column C′. Visualize this alternative solution as a graph of the normal curve, and verify that, even though column B′ is used, the answer still equals .1112.

Example: Finding Proportions *beyond* Two Scores

Assume that high school students' IQ scores approximate a normal distribution with a mean of 105 and a standard deviation of 15. What proportion of IQs are more than 30 points either above or below the mean?

1. **Sketch a normal curve and shade in the two target areas**, as in the top panel of **Figure 5.8**.
2. **Plan your solution according to the normal table**. The solution to this type of problem is straightforward because each of the target areas can be read directly from Table A. The target area in the tail to the right can be obtained from column C, and that in the tail to the left can be obtained from column C′, as shown in the bottom two panels of Figure 5.8.

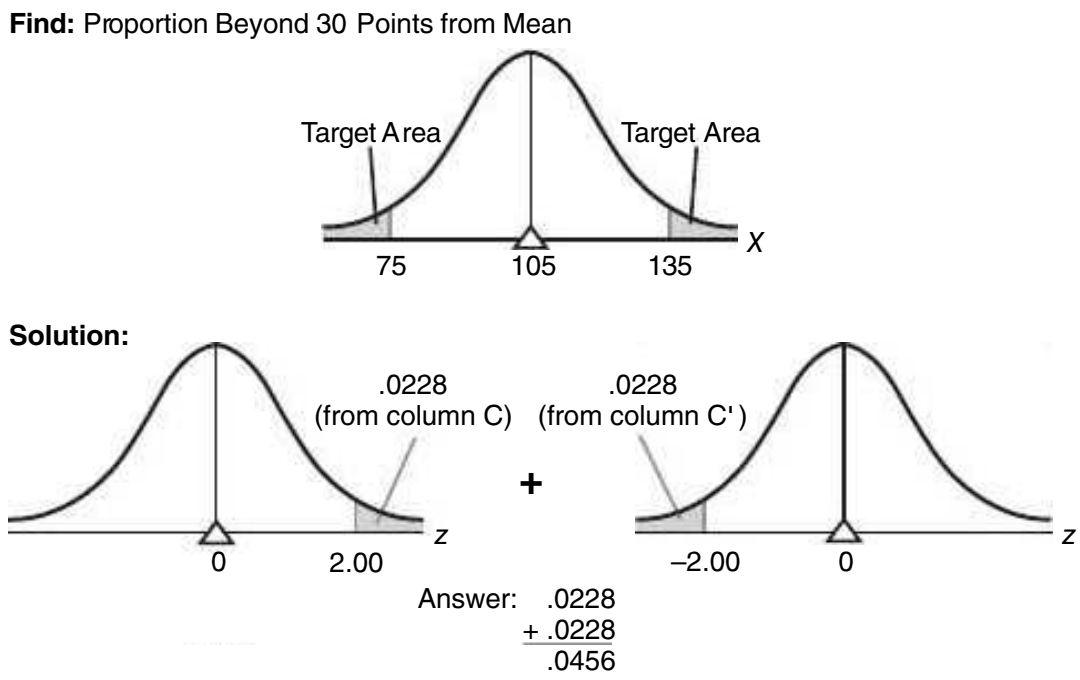


FIGURE 5.8
Finding proportions.

3. Convert *X* to *z* by expressing IQ scores of 135 and 75 as

$$z = \frac{135 - 105}{15} = \frac{30}{15} = 2.00$$

$$z = \frac{75 - 105}{15} = \frac{-30}{15} = -2.00$$

4. Find the target area. In Table A, locate a z score of 2.00 in column A, and note the corresponding proportion of .0228 in column C. Because of the symmetry of the normal curve, you need not enter the table again to find the proportion below a z score of –2.00. Instead, merely double the above proportion of .0228 to obtain .0456, which represents the proportion of students with IQs more than 30 points either above or below the mean.

Semantic Alert

"*More* than 30 points either above or below the mean" translates into two target areas, one in each tail of the normal curve. " *Within* 30 points either above or below the mean" translates into two entirely new target areas corresponding to the two unshaded sectors in Figure 5.8. These "within" sectors share a common boundary at the mean, but one sector extends 30 points above the mean and the other sector extends 30 points below the mean.

Progress Check *5.5 Assume that SAT math scores approximate a normal curve with a mean of 500 and a standard deviation of 100.

(a) Sketch a normal curve and shade in the target area(s) described by each of the following statements:

(i) *more* than 570

(ii) *less* than 515

(iii) *between* 520 and 540

(iv) between 470 and 520

(v) more than 50 points above the mean

(vi) more than 100 points either above or below the mean

(vii) within 50 points either above or below the mean </RLLNL>

(b) Plan solutions (in terms of columns B, C, B′, and C′) for the target areas in part (a).

(c) Convert to *z* scores and find the target areas in part (a).

Answers on page 426.

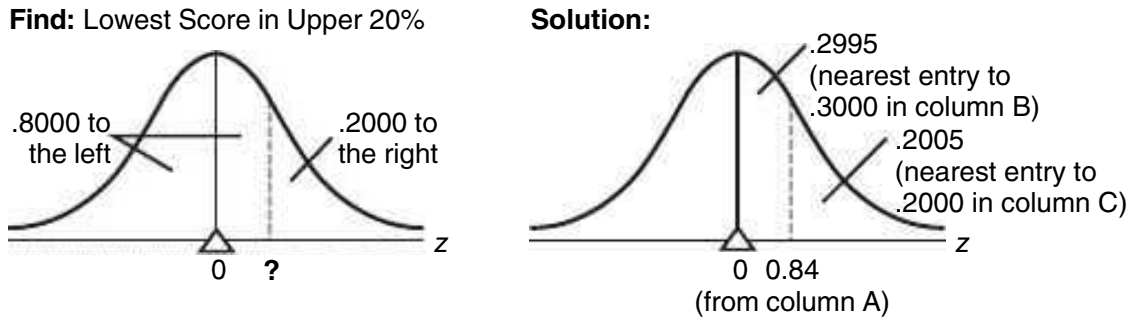
5.6 FINDING SCORES

So far, we have concentrated on normal curve problems for which Table A must be consulted to find the unknown proportion (of area) associated with some known score or pair of known scores. For instance, given a GRE score of 650, we found that the unknown proportion of scores larger than 650 equals .07. *Now we will concentrate on the opposite type of normal curve problem for which Table A must be consulted to find the unknown score or scores associated with some known proportion.* For instance, given that a GRE score must be in the upper 25 percent of the distribution (in order for an applicant to be considered for admission to graduate school), we must find the unknown minimum GRE score. Essentially, this type of problem requires that we reverse our use of Table A by entering proportions in columns B, C, B′, or C′ and finding *z* scores listed in columns A or A′.

Example: Finding *One* Score

Exam scores for a large psychology class approximate a normal curve with a mean of 230 and a standard deviation of 50. Furthermore, students are graded "on a curve," with only the upper 20 percent being awarded grades of A. What is the lowest score on the exam that receives an A?

1. **Sketch a normal curve and, on the correct side of the mean, draw a line representing the target score,** as in **Figure 5.9.** This is often the most difficult step, and it involves semantics rather than statistics. It's often helpful to visualize the target



Answer: $X = \mu + (z)(\sigma)$
$= 230 + (0.84)(50)$
$= 230 + 42$
$= 272$

FIGURE 5.9
Finding scores.

score as splitting the total area into two sectors—one to the left of (below) the target score and one to the right of (above) the target score. For example, in the present case, the target score is the point along the base of the curve that splits the total area into 80 percent, or .8000 to the left, and 20 percent, or .2000 to the right. The mean of a normal curve serves as a valuable frame of reference since it always splits the total area into two equal halves—.5000 to the left of the mean and .5000 to the right of the mean. Since more than .5000—that is, .8000—of the total area is to the left of the target score, this score must be on the upper or right side of the mean. On the other hand, if less than .5000 of the total area had been to the left of the target score, this score would have been placed on the lower or left side of the mean.

2. **Plan your solution according to the normal table.** In problems of this type, you must plan how to find the z score for the target score. Because the target score is on the right side of the mean, concentrate on the area in the upper half of the normal curve, as described in columns B and C. The right panel of Figure 5.9 indicates that either column B or C can be used to locate a z score in column A. It is crucial, however, to search for the single value (.3000) that is valid for column B or the single value (.2000) that is valid for column C. Note that we look in column B for .3000, not for .8000. Table A is not designed for sectors, such as the lower .8000, that span the mean of the normal curve.

3. **Find z.** Refer to Table A. Scan column C to find .2000. If this value does not appear in column C, as typically will be the case, approximate the desired value (and the correct score) by locating the entry in column C nearest to .2000. If adjacent entries are equally close to the target value, use either entry—it is your choice. As shown in the right panel of Figure 5.9, the entry in column C closest to .2000 is .2005, and the corresponding z score in column A equals 0.84. Verify this by checking Table A. Also note that exactly the same z score of 0.84 would have been identified if column B had been searched to find the entry (.2995) nearest to .3000. The z score of 0.84 represents the point that separates the upper 20 percent of the area from the rest of the area under the normal curve.

4. **Convert z to the target score.** Finally, convert the z score of 0.84 into an exam score, given a distribution with a mean of 230 and a standard deviation of 50. You'll recall that a z score indicates how many standard deviations the original score is above or below its mean. In the present case, the target score must be located .84 of a standard deviation above its mean. The distance of the target score above its mean equals 42 (from $.84 \times 50$), which, when added to the mean of 230, yields a value of 272. Therefore, 272 is the lowest score on the exam that receives an A.

When converting z scores to original scores, you will probably find it more efficient to use the following equation (derived from the z score equation on page 86):

CONVERTING z SCORE TO ORIGINAL SCORE

$$X = \mu + (z)(\sigma) \tag{5.2}$$

where X is the target score, expressed in original units of measurement; μ and σ are the mean and the standard deviation, respectively, for the original normal curve; and z is the standard score read from column A or A′ of Table A. When appropriate numerical substitutions are made, as shown in the bottom of Figure 5.9, 272 is found to be the answer, in agreement with our earlier conclusion.

Comment: Place Target Score on Correct Side of Mean

When finding scores, it is crucial that the target score be placed on the correct side of the mean. This placement dictates how the normal table will be read—whether down from the top legend, with entries in column A interpreted as positive z scores, or up from the bottom legend, with entries in column A′ interpreted as negative z scores. In the previous problem, the incorrect placement of the target score on the left side of the mean would have led to a z score of –0.84, rather than 0.84, and an erroneous answer of 188 (230 − 42), rather than the correct answer of 272 (230 + 42).

To make correct placements, you must properly interpret the specifications for the target score. Expand potentially confusing one-sided specifications, such as the "upper 20 percent, or upper .2000," into "left .8000 and right .2000." Having identified the left and right areas of the target score, which sum to 1.0000, you can compare the specifications of the target score with those of the mean. Remember that the mean of a normal curve always splits the total area into .5000 to the left of the mean and .5000 to the right of the mean. Accordingly, if the area to the left of the target score is more than .5000, the target score should be placed on the upper or right side of the mean. Otherwise, if the area to the left of the target score is less than .5000, the target score should be placed on the lower or left side of the mean.

Example: Finding *Two* Scores

Assume that the annual rainfall in the San Francisco area approximates a normal curve with a mean of 22 inches and a standard deviation of 4 inches. What are the rainfalls for the more atypical years, defined as the driest 2.5 percent of all years and the wettest 2.5 percent of all years?

1. **Sketch a normal curve. On either side of the mean, draw two lines representing the two target scores,** as in **Figure 5.10.** The smaller (driest) target score splits the total area into .0250 to the left and .9750 to the right, and the larger (wettest) target score does the exact opposite.
2. **Plan your solution according to the normal table.** Because the smaller target score is located on the lower or left side of the mean, we will concentrate on the area in the lower half of the normal curve, as described in columns B′ and C′. The target z score can be found by scanning either column B′ for .4750 or column C′

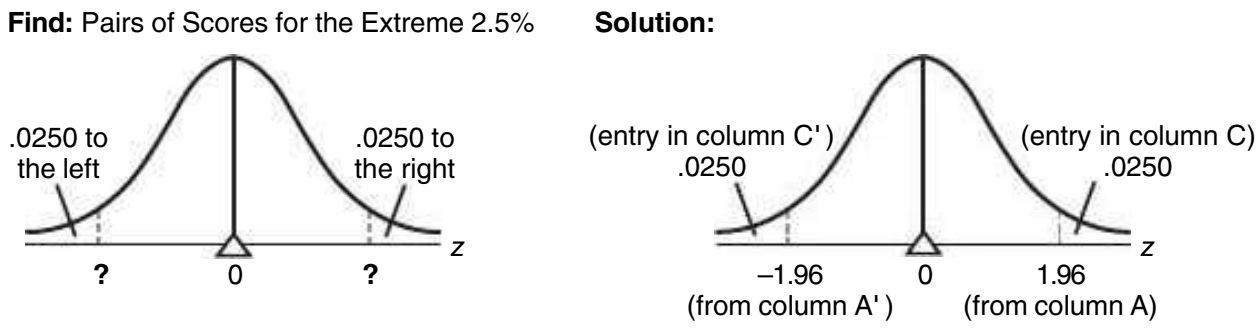


FIGURE 5.10
Finding scores.

Table 6.3
CALCULATION OF *r*: COMPUTATION FORMULA

A. COMPUTATIONAL SEQUENCE

Assign a value to *n* (1), representing the number of pairs of scores.
Sum all scores for *X* (2) and for *Y* (3).
Find the product of each pair of *X* and *Y* scores (4), one at a time, then add all of these products (5).
Square each *X* score (6), one at a time, then add all squared *X* scores (7).
Square each *Y* score (8), one at a time, then add all squared *Y* scores (9).
Substitute numbers into formulas (10) and solve for SP_{xy}, SS_x, and SS_y.
Substitute into formula (11) and solve for *r*.

B. DATA AND COMPUTATIONS

	CARDS		4	6	8
FRIEND	**SENT, *X***	**RECEIVED, *Y***	***XY***	***X*²**	***Y*²**
Doris	13	14	182	169	196
Steve	9	18	162	81	324
Mike	7	12	84	49	144
Andrea	5	10	50	25	100
John	1	6	6	1	36

1 $n = 5$ 2 $\Sigma X = 35$ 3 $\Sigma Y = 60$ 5 $\Sigma XY = 484$ 7 $\Sigma X^2 = 325$ 9 $\Sigma Y^2 = 800$

10 $$SP_{xy} = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n} = 484 - \frac{(35)(60)}{5} = 484 - 420 = 64$$

$$SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = 325 - \frac{(35)^2}{5} = 325 - 245 = 80$$

$$SS_y = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n} = 800 - \frac{(60)^2}{5} = 800 - 720 = 80$$

11 $$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}} = \frac{64}{\sqrt{(80)(80)}} = \frac{64}{80} = .80$$

6.5 OUTLIERS AGAIN

In Section 2.3, *outliers* were defined as very extreme scores that require special attention because of their potential impact on a summary of data. This is also true when outliers appear among sets of paired scores. Although quantitative techniques can be used to detect these outliers, we simply focus on dots in scatterplots that deviate conspicuously from the main dot cluster.

Greeting Card Study Revisited

Figure 6.6 shows the effect of each of two possible outliers, substituted one at a time for Doris's dot (13, 14), on the original value of *r* (.80) for the greeting card data.

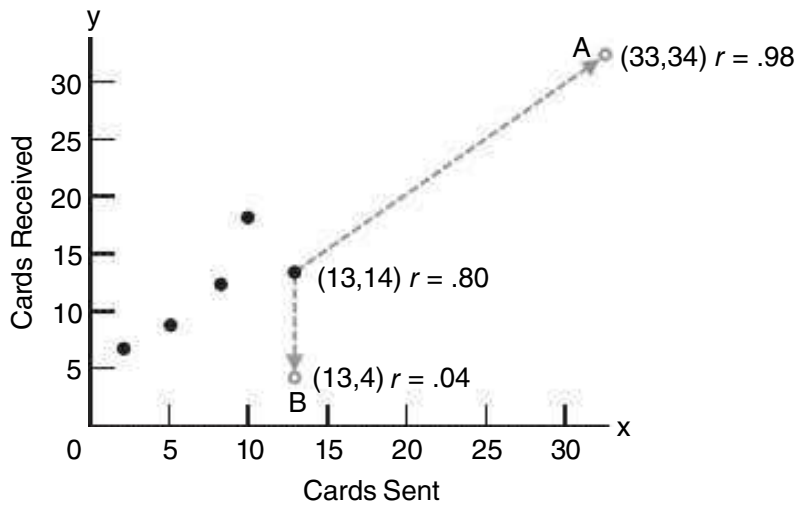


FIGURE 6.6
Effect of each of two outliers on the value of r.

Although both outliers A and B deviate conspicuously from the dot cluster, they have radically different effects on the value of *r*. Outlier A (33, 34) contributes to a new value of .98 for *r* that merely reaffirms the original positive relationship between cards sent and received. On the other hand, outlier B (13, 4) causes a dramatically new value of .04 for *r* that entirely neutralizes the original positive relationship. Neither of the values for outlier B, taken singularly, is extreme. Rather, it is their unusual combination—13 cards sent and only 4 received—that yields the radically different value of .04 for *r*, indicating that the new dot cluster is not remotely approximated by a straight line.

Dealing with Outliers

Of course, serious investigators would use many more than five pairs of scores, and therefore the effect of outliers on the value of *r* would tend not to be as dramatic as the one above. Nevertheless, outliers can have a considerable impact on the value of *r* and, therefore, pose problems of interpretation. Unless there is some reason for discarding an outlier—because of a failed accuracy check or because, for example, you establish that the friend who received only 4 cards had sent 13 cards that failed to include an expected monetary gift—the most defensible strategy is to report the values of *r* both with and without any outliers.

6.6 OTHER TYPES OF CORRELATION COEFFICIENTS

There are many other types of correlation coefficients, but we will discuss only several that are direct descendants of the Pearson correlation coefficient. Although designed originally for use with quantitative data, the Pearson *r* has been extended, sometimes under the guise of new names and customized versions of Formula 6.1, to other kinds of situations. For example, to describe the correlation between *ranks* assigned independently by two judges to a set of science projects, simply substitute the numerical ranks into Formula 6.1, then solve for a value of the Pearson *r* (also referred to as *Spearman's rho* coefficient for ranked or ordinal data). To describe the correlation between quantitative data (for example, annual income) and *qualitative or nominal data with only two categories* (for example, male and female), assign arbitrary numerical codes, such as 1 and 2, to the two qualitative categories, then solve Formula 6.1 for a value of the Pearson *r* (also referred to as a *point biserial* correlation coefficient). Or to describe the relationship between *two ordered qualitative variables,* such as the attitude toward legal abortion (favorable, neutral, or opposed) and educational level (high school only,

some college, college graduate), assign any *ordered* numerical codes, such as 1, 2, and 3, to the categories for both qualitative variables, then solve Formula 6.1 for a value of the Pearson *r* (also referred to as *Cramer's phi* coefficient).

Most computer outputs would simply report each of these correlations as a Pearson *r*. Given the widespread use of computers, the more specialized names for the Pearson *r* will probably survive, if at all, as artifacts of an earlier age, when calculations were manual and some computational relief was obtained by customizing Formula 6.1 for situations involving ranks and qualitative data.

6.7 COMPUTER OUTPUT

Most analyses in this book are performed by hand on small batches of data. When analyses are based on large batches of data, as often happens in practice, it is much more efficient to use a computer. Although we will not show how to enter commands and data into a computer, we will describe the most relevant portions of some computer outputs. Once you have learned to ignore irrelevant details and references to more advanced statistical procedures, you'll find that statistical results produced by computers are as easy to interpret as those produced by hand.

Three of the most widely used statistical programs—Minitab, SPSS (Statistical Package for the Social Sciences), and SAS (Statistical Analysis System)—generate the computer outputs in this book. As interpretive aids, some outputs are cross-referenced with explanatory comments at the bottom of the printout. Since these outputs are based on data already analyzed by hand, computer-produced results can be compared with familiar results. For example, the computer-produced scatterplot, as well as the correlation of .800 in **Table 6.4** can be compared with the manually produced scatterplot in Figure 6.1 and the correlation of .80 in Table 6.3.

WWW

INTERNET SITES

Go to the website for this book (http://www.wiley.com/college/witte). Click on the *Student Companion Site,* then *Internet Sites,* and finally ***Minitab, SPSS, or SAS*** to obtain more information about these statistical packages, as well as demonstration software.

Correlation Matrix

Table showing correlations for all possible pairs of variables.

When every possible pairing of variables is reported, as in lower half of the output in Table 6.4, a **correlation matrix** is produced. The value of .800 occurs twice in the matrix, since the correlation is the same whether the relationship is described as that between cards sent and cards received or vice versa. The value of 1.000, which also occurs twice, reflects the trivial fact that any variable correlates perfectly with itself.

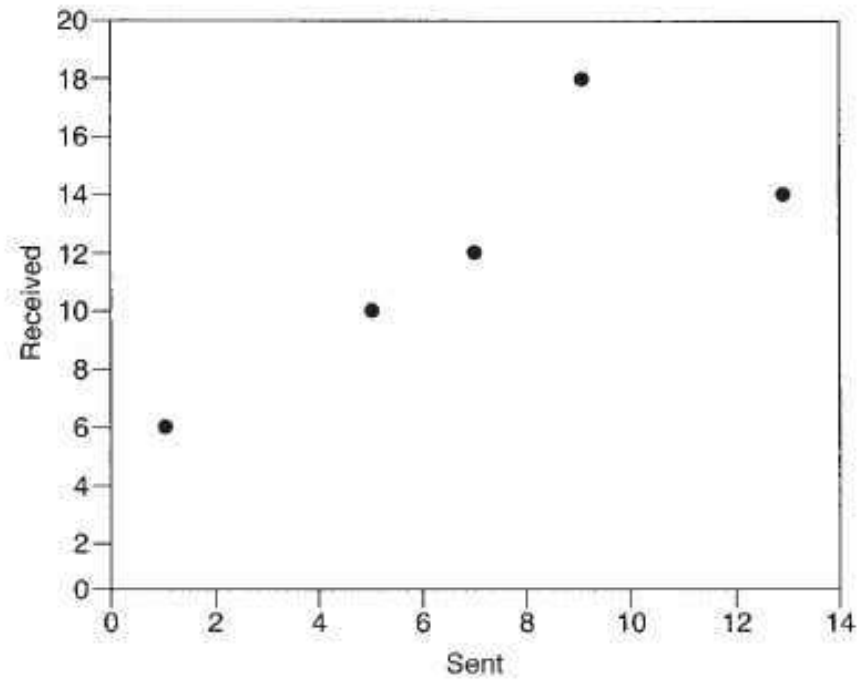
Reading a Larger Correlation Matrix

Since correlation matrices can be expanded to incorporate any number of variables, they are useful devices for showing correlations between all possible pairs of variables when, in fact, many variables are being studied. For example, in **Table 6.5,** four variables generate a correlation matrix with 4×4, or 16, correlation coefficients. The four perfect (but trivial) correlations of 1.000, produced by pairing each variable with itself,

Table 6.4
SPSS OUTPUT: SCATTERPLOT AND CORRELATION FOR GREETING CARD DATA

GRAPH

1



CORRELATIONS

		SENT	RECEIVED
Sent	Pearson Correlation	1.000	.800
	Sig. (2-tailed)	–	.104
	N	5	5
Received	Pearson Correlation	2 .800	1.000
	Sig. (2-tailed)	3 .104	–
	N	4 5	5

Comments:

1. Scatterplot for greeting card data (using slightly different scales than in Figure 6.1).

2. The correlation for cards sent and cards received equals .800, in agreement with the calculations in Table 6.3.

3. The value of Sig. *helps us interpret the statistical significance of a correlation by evaluating the observed value of* r *relative to the actual number of pairs of scores used to calculate r. Discussed later in Section 14.6,* Sig.*-values are referred to as p-values in this book. At this point, perhaps the easiest way to view a* Sig.*-value is as follows: The smaller the value of* Sig. *(on a scale from 0 to 1), the more likely that you would observe a correlation with the same sign, either positive or negative, if the study were repeated with new observations. Investigators often focus only on those correlations with* Sig.*-values smaller than .05.*

4. Number of cases or paired scores.

6.8 Calculate the value of *r* using the computational formula (6.1) for the following data.

X	*Y*
2	8
4	6
5	2
3	3
1	4
7	1
2	4

6.9 Indicate whether the following generalizations suggest a positive or negative relationship. Also speculate about whether or not these generalizations reflect simple cause-effect relationships.

(a) Preschool children who delay gratification (postponing eating one marshmallow to win two) subsequently receive higher teacher evaluations of adolescent competencies.

(b) College students who take longer to finish a test perform more poorly on that test.

(c) Heavy smokers have shorter life expectancies.

(d) Infants who experience longer durations of breastfeeding score higher on IQ tests in later childhood.

***6.10** On the basis of an extensive survey, the California Department of Education reported an *r* of –.32 for the relationship between the amount of time spent watching TV and the achievement test scores of schoolchildren. Each of the following statements represents a possible interpretation of this finding. Indicate whether each is True or False.

(a) *Every* child who watches a lot of TV will perform poorly on the achievement tests.

(b) Extensive TV viewing causes a decline in test scores.

(c) Children who watch little TV will tend to perform well on the tests.

(d) Children who perform well on the tests will tend to watch little TV.

(e) If Gretchen's TV-viewing time is reduced by one-half, we can expect a substantial improvement in her test scores.

(f) TV viewing could not possibly cause a decline in test scores.

Answers on page 428.

6.11 Assume that an *r* of .80 describes the relationship between daily food intake, measured in ounces, and body weight, measured in pounds, for a group of adults. Would a shift in the units of measurement from ounces to grams and from pounds to kilograms change the value of *r*? Justify your answer.

6.12 An extensive correlation study indicates that a longer life is experienced by people who follow the seven "golden rules" of behavior, including moderate drinking, no smoking, regular meals, some exercise, and eight hours of sleep each night. Can we conclude, therefore, that this type of behavior *causes* a longer life?

CHAPTER 7

Regression

Preview

If two variables are correlated, description can lead to prediction. For example, if computer skills and GPAs are related, level of computer skills can be used to predict GPAs. Predictive accuracy increases with the strength of the underlying correlation.

Also discussed is a prevalent phenomenon known as "regression toward the mean." It often occurs over time to subsets of extreme observations, such as after the superior performance of professional athletes or after the poor performance of learning-challenged children. If misinterpreted as a real effect, regression toward the mean can lead to erroneous conclusions.

A correlation analysis of the exchange of greeting cards by five friends for the most recent holiday season suggests a strong positive relationship between cards sent and cards received. When informed of these results, another friend, Emma, who enjoys receiving greeting cards, asks you to predict how many cards she will receive during the next holiday season, assuming that she plans to send 11 cards.

7.1 TWO ROUGH PREDICTIONS

Predict "Relatively Large Number"

You could offer Emma a very rough prediction by recalling that cards sent and received tend to occupy *similar* relative locations in their respective distributions. Therefore, Emma can expect to receive a *relatively large* number of cards, since she plans to send a *relatively large* number of cards.

Predict "between 14 and 18 Cards"

To obtain a slightly more precise prediction for Emma, refer to the scatter plot for the original five friends shown in **Figure 7.1.** Notice that Emma's plan to send 11 cards locates her along the X axis between the 9 cards sent by Steve and the 13 sent by Doris. Using the dots for Steve and Doris as guides, construct two strings of arrows, one beginning at 9 and ending at 18 for Steve and the other beginning at 13 and ending at 14 for Doris. [The direction of the arrows reflects our attempt to predict cards received (Y) from cards sent (X). Although not required, it is customary to predict from X to Y.] Focusing on the interval along the Y axis between the two strings of arrows, you could predict that Emma's return should be between 14 and 18 cards, the numbers received by Doris and Steve.

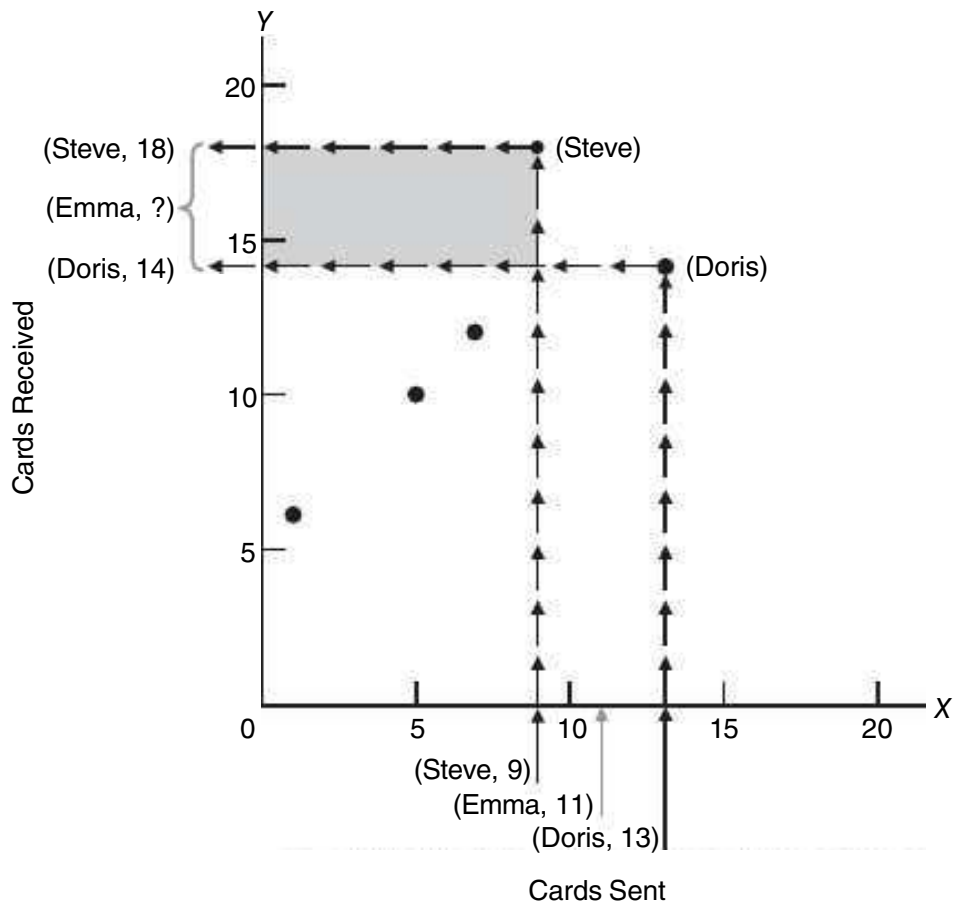


FIGURE 7.1
A rough prediction for Emma (using dots for Steve and Doris).

The latter prediction might satisfy Emma, but it would not win any statistical awards. Although each of the five dots in Figure 7.1 supplies valuable information about the exchange of greeting cards, our prediction for Emma is based only on the two dots for Steve and Doris.

7.2 A REGRESSION LINE

All five dots contribute to the more precise prediction, illustrated in **Figure 7.2,** that Emma will receive 15.20 cards. Look more closely at the solid line designated as the regression line in Figure 7.2, which guides the string of arrows, beginning at 11, toward the predicted value of 15.20. The regression line is a straight line rather than a curved line because of the linear relationship between cards sent and cards received. As will become apparent, it can be used repeatedly to predict cards received. Regardless of whether Emma decides to send 5, 15, or 25 cards, it will guide a new string of arrows, beginning at 5 or 15 or 25, toward a new predicted value along the *Y* axis.

Placement of Line

For the time being, forget about any prediction for Emma and concentrate on how the five dots dictate the placement of the regression line. If all five dots had defined a single straight line, placement of the regression line would have been simple; merely let it pass through all dots. When the dots fail to define a single straight line, as in the scatterplot for the five friends, placement of the regression line represents a compromise. It passes through the main cluster, possibly touching some dots but missing others.

Predictive Errors

Figure 7.3 illustrates the predictive errors that would have occurred if the regression line had been used to predict the number of cards received by the five friends. Solid dots reflect the *actual* number of cards received, and open dots, always located along

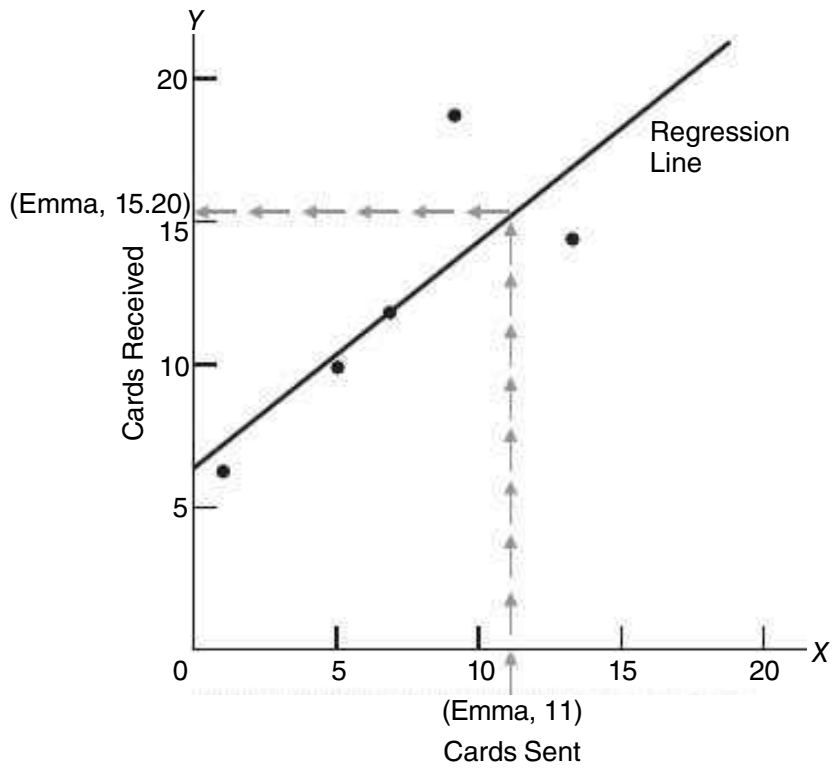


FIGURE 7.2
Prediction of 15.20 for Emma (using the regression line).

scores of these five students continue to reflect an above-average permanent component, some of their scores will suffer because of less good luck or even bad luck. The net effect is that the scores of at least some of the original five top students will drop below the top five scores—that is, regress *back* toward the mean—on the second exam. (When significant regression toward the mean occurs after a spectacular performance by, for example, a rookie athlete or a first-time author, the term *sophomore jinx* often is invoked.)

There is good news for those students who made the five lowest scores on the first exam. Although all five students might score below the mean on the second exam, some of their scores probably will regress *up* toward the mean. On the second exam, some of them will not be as unlucky. The net effect is that the scores of at least some of the original five lowest scoring students will move above the bottom five scores—that is, regress up toward the mean—on the second exam.

Appears in Many Distributions

Regression toward the mean appears among subsets of extreme observations for a wide variety of distributions. For example, it appears for the subset of best (or worst) performing stocks on the New York Stock Exchange across any period, such as a week, month, or year. It also appears for the top (or bottom) major league baseball hitters during consecutive seasons. **Table 7.4** lists the top 10 hitters in the major leagues during 2014 and shows how they fared during 2015. Notice that 7 of the top 10 batting averages regressed downward, toward 260s, the approximate mean for all hitters during 2015. Incidentally, it is not true that, viewed as a group, all major league hitters are headed toward mediocrity. Hitters among the top 10 in 2014, who were not among the top 10 in 2015, were replaced by other mostly above-average hitters, who also were very lucky during 2015. Observed regression toward the mean occurs for individuals or subsets of individuals, not for entire groups.

The Regression Fallacy

Regression Fallacy
Occurs whenever regression toward the mean is interpreted as a real, rather than a chance, effect.

The **regression fallacy** *is committed whenever regression toward the mean is interpreted as a real, rather than a chance, effect.* A classic example of the regression fallacy occurred in an Israeli Air Force study of pilot training [as described by Tversky,

Table 7.4
REGRESSION TOWARD THE MEAN: BATTING AVERAGES OF TOP 10 HITTERS IN MAJOR LEAGUE BASEBALL DURING 2014 AND HOW THEY FARED DURING 2015

	BATTING AVERAGES*		
TOP 10 HITTERS (2014)	2014	2015	REGRESS TOWARD MEAN?
1. J. Alture	.341	.313	Yes
2. V. Martinez	.335	.282	Yes
3. M. Brantley	.327	.310	Yes
4. A. Beltre	.324	.287	Yes
5. J. Abreu	.317	.290	Yes
6. R. Cano	.314	.287	Yes
7. A. McCutchen	.314	.292	Yes
8. M. Cabrera	.313	.338	No
9. B. Posey	.311	.318	No
10. B. Revere	.306	.306	No

** Proportion of hits per official number of times at bat.*
Source: http://sports.espn.go.com/mlb/stats/batting.

A., & Kahnemann, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.] Some trainees were praised after very good landings, while others were reprimanded after very bad landings. On their next landings, praised trainees did more poorly and reprimanded trainees did better. It was concluded, therefore, that praise hinders but a reprimand helps performance!

A valid conclusion considers regression toward the mean. It's reasonable to assume that, in addition to skill, chance plays a role in landings. Some trainees who made very good landings were lucky, while some who made very bad landings were unlucky. Therefore, there would be a tendency, attributable to chance, that good landings would be followed by less good landings and poor landings would be followed by less poor landings—even if trainees had not been praised after very good landings or reprimanded after very bad landings.

Avoiding the Regression Fallacy

The regression fallacy can be avoided by splitting the subset of extreme observations into two groups. In the previous example, one group of trainees would continue to be praised after very good landings and reprimanded after very poor landings. A second group of trainees would receive no feedback whatsoever after very good and very bad landings. In effect, the second group would serve as a control for regression toward the mean, since any shift toward the mean on their second landings would be due to chance. Most important, any observed difference between the two groups (that survives a statistical analysis described in Part 2) would be viewed as a real difference not attributable to the regression effect.

Watch out for the regression fallacy in educational research involving groups of underachievers. For example, a group of fourth graders, selected to attend a special program for underachieving readers, might show an improvement. Whether this improvement can be attributed to the special program or to a regression effect requires information from a control group of similarly underachieving fourth graders who did not attend the special program. It is crucial, therefore, that research with underachievers always includes a control group for regression toward the mean.

Progress Check *7.6 After a group of college students attended a stress-reduction clinic, declines were observed in the anxiety scores of those who, prior to attending the clinic, had scored high on a test for anxiety.

(a) Can this decline be attributed to the stress-reduction clinic? Explain your answer.

(b) What type of study, if any, would permit valid conclusions about the effect of the stress-reduction clinic?

Answers on page 429.

Summary

If a linear relationship exists between two variables, then one variable can be predicted from the other by using the least squares regression equation, as described in Formulas 7.1, 7.2, and 7.3.

The least squares equation minimizes the total of all squared predictive errors that would have occurred if the equation had been used to predict known Y scores from the original correlation analysis.

An estimate of predictive error can be obtained from Formula 7.5. Known as the *standard error of estimate,* this estimate is a special kind of standard deviation that roughly reflects the average amount of predictive error. The value of the standard error of estimate depends mainly on the size of the correlation coefficient. The larger the

correlation coefficient, in either the positive or negative direction, the smaller the standard error of estimate and the more favorable the prognosis for predictions.

The regression equation assumes a linear relationship between variables, and the standard error of estimate assumes homoscedasticity—approximately equal dispersion of data points about all segments of the regression line.

The square of the correlation coefficient, r^2, indicates the proportion of total variability in one variable that is predictable from its relationship with the other variable.

Serious predictive efforts usually involve multiple regression equations composed of more than one predictor, or X, variable. These multiple regression equations share many common features with the simple regression equations discussed in this chapter.

Regression toward the mean refers to a tendency for scores, particularly extreme scores, to shrink toward the mean. The regression fallacy is committed whenever regression toward the mean is interpreted as a real rather than a chance effect. To guard against the regression fallacy, control groups should be used to estimate the regression effect.

Important Terms

Least squares regression equation
Standard error of estimate ($s_{y|x}$)
Regression fallacy
Multiple regression equation
Squared correlation (r^2)
Regression toward the mean

Key Equations

PREDICTION EQUATION

$$Y' = bx + a$$

where $b = r\sqrt{\frac{SS_Y}{SS_X}}$

and $a = \overline{Y} - b\overline{X}$

REVIEW QUESTIONS

7.7 Assume that an r of –.80 describes the strong negative relationship between years of heavy smoking (X) and life expectancy (Y). Assume, furthermore, that the distributions of heavy smoking and life expectancy each have the following means and sums of squares:

$$\overline{X} = 5 \qquad \overline{Y} = 60$$
$$SS_x = 35 \qquad SS_y = 70$$

(a) Determine the least squares regression equation for predicting life expectancy from years of heavy smoking.

(b) Determine the standard error of estimate, $s_{y|x}$, assuming that the correlation of –.80 was based on $n = 50$ pairs of observations.

(c) Supply a rough interpretation of $s_{y|x}$.

(d) Predict the life expectancy for John , who has smoked heavily for 8 years.

(e) Predict the life expectancy for Katie, who has never smoked heavily.

7.8 Each of the following pairs represents the number of licensed drivers (X) and the number of cars (Y) for seven houses in my neighborhood:

DRIVERS (X)	CARS (Y)
5	4
5	3
2	2
2	2
3	2
1	1
2	2

(a) Construct a scatterplot to verify a lack of pronounced curvilinearity.

(b) Determine the least squares equation for these data. (Remember, you will first have to calculate r, SS_y and SS_x)

(c) Determine the standard error of estimate, $s_{y|x}$, given that $n = 7$.

(d) Predict the number of cars for each of two new families with two and five drivers.

7.9 At a large bank, length of service is the best single predictor of employees' salaries. Can we conclude, therefore, that there is a cause-effect relationship between length of service and salary?

7.10 Assume that r^2 equals .50 for the relationship between height and weight for adults. Indicate whether the following statements are true or false.

(a) Fifty percent of the variability in heights can be explained by variability in weights.

(b) There is a cause-effect relationship between height and weight.

(c) The heights of 50 percent of adults can be predicted exactly from their weights.

(d) Fifty percent of the variability in weights is predictable from heights.

***7.11** In studies dating back over 100 years, it's well established that regression toward the mean occurs between the heights of fathers and the heights of their *adult* sons. Indicate whether the following statements are true or false.

(a) Sons of tall fathers will tend to be shorter than their fathers.

(b) Sons of short fathers will tend to be taller than the mean for all sons.

(c) Every son of a tall father will be shorter than his father.

(d) Taken as a group, adult sons are shorter than their fathers.

(e) Fathers of tall sons will tend to be taller than their sons.

(f) Fathers of short sons will tend to be taller than their sons but shorter than the mean for all fathers.

Answers on page 429.

7.12 Someone suggests that it would be a good investment strategy to buy the five poorest-performing stocks on the New York Stock Exchange and capitalize on regression toward the mean. Comments?

7.13 In the original study of regression toward the mean, Sir Francis Galton noted a tendency for offspring of both tall and short parents to drift toward the mean height for offspring and referred to this tendency as "regression toward mediocrity." What is wrong with the conclusion that eventually all heights will be close to their mean?

PART 2

Inferential Statistics

Generalizing Beyond Data

Preview

The remaining chapters deal with the problem of generalizing beyond sets of actual observations. The next two chapters develop essential concepts and tools for inferential statistics, while subsequent chapters introduce a series of statistical tests or procedures, all of which permit us to generalize beyond an observed result, whether from a survey or an experiment, by considering the effects of chance.

Many pollsters use *random digit dialing* in an effort to give each telephone number—whether landline or wireless—in the United States an equal chance of being called for an interview. Essentially, the first six digits of a 10-digit phone number, including the area code, are randomly selected from tens of thousands of telephone exchanges, while the final four digits are taken directly from random numbers. Although random digit dialing ensures that all unlisted telephone numbers will be sampled, it has lost some of its appeal recently because of a federal prohibition against its use to contact wireless numbers and also because of the excessively high nonresponse rates, often as high as 91 percent. In an effort to approximate a more representative sample, pollsters have been exploring other techniques, such as online polling.*
(http://www.stat.columbia.edu/~gelman/research/published/forecasting-with-nonrepresentative-polls.pdf).

Hypothetical Populations

As has been noted, the researcher, unlike the pollster, usually deals with hypothetical populations. Unfortunately, it is impossible to take random samples from hypothetical populations. All potential observations cannot have an equal chance of being included in the sample if, in fact, some observations are not accessible at the time of sampling. It is a common practice, nonetheless, for researchers to treat samples from hypothetical populations *as if* they were random samples and to analyze sample results with techniques from inferential statistics. Our adoption of this practice—to provide a common basis for discussing both surveys and experiments—is less troublesome than you might think inasmuch as random assignment replaces random sampling in well-designed experiments.

8.5 RANDOM ASSIGNMENT OF SUBJECTS

Random Assignment
A procedure designed to ensure that each subject has an equal chance of being assigned to any group in an experiment.

Typically, experiments evaluate an independent variable by focusing on a treatment group and a control group. Although subjects in experiments can't be selected randomly from any real population, they can be **assigned randomly**, that is, with equal likelihood, to these two groups. This procedure has a number of desirable consequences:

- Since random assignment or chance determines the membership for each group, all possible configurations of subjects are equally likely. This provides a basis for calculating the chances of observing any specific difference between groups and ultimately deciding whether, for instance, the one observed mean difference between groups is real or merely transitory.
- Random assignment generates groups of subjects that, except for random differences, are similar with respect to any uncontrolled variables at the outset of the experiment.

For instance, to determine whether a study-skill workshop improves academic performance, volunteer subjects should be assigned randomly either to the treatment group (attendance at the workshop) or to the control group. This ensures that, except for random differences, both groups are similar initially with respect to any uncontrolled variables, such as academic preparation, motivation, IQ, etc. At the conclusion of such an experiment, therefore, any observed differences in academic performance between these two groups, *not attributable to random differences,* would provide the most clear-cut evidence of a cause-effect relationship between the independent variable (attendance at the workshop) and the dependent variable (academic performance).

*See introductory comments in http://dx.doi.org/10.1016/j.ijforcast.2014.06.001.

How to Assign Subjects

The random assignment of subjects can be accomplished in a number of ways. For instance, as each new subject arrives to participate in the experiment, a flip of a coin can decide whether that subject should be assigned to the treatment group (if heads turns up) or the control group (if tails turn up). An even better procedure, because it eliminates any biases of a live coin tosser, relies on tables of random numbers. Once the tables have been entered at some arbitrary point, they can be consulted, much like a string of coin tosses, to determine whether each new subject should be assigned to the treatment group (if, for instance, the random number is odd) or to the control group (if the random number is even).

Creating Equal Groups

Equal numbers of subjects should be assigned to the treatment and control groups for a variety of reasons, including the increased likelihood of detecting any difference between the two groups. To achieve this goal, the random assignment should involve pairs of subjects. If the table of random numbers assigns the first volunteer to the treatment group, the second volunteer should be assigned *automatically* to the control group. If the random numbers assign the third volunteer to the control group, the fourth volunteer should be assigned *automatically* to the treatment group, and so forth. This procedure guarantees that at any stage of the random assignment, equal numbers of subjects will be assigned to the two groups.

More Extensive Sets of Random Numbers

Incidentally, the page of random numbers in Table H, Appendix C, serves only as a specimen. For serious applications, refer to a more extensive collection of random numbers, such as that in the book by the Rand Corporation cited on page 470 of Appendix C. If you have access to a computer, you might refer to the list of random numbers that can be generated, almost effortlessly, by computers.

Progress Check * 8.5 Assume that 12 subjects arrive, one at a time, to participate in an experiment. Use random numbers to assign these subjects in equal numbers to group A and group B. Specifically, random numbers should be used to identify the first subject as either A or B, the second subject as either A or B, and so forth, until all subjects have been identified. There should be six subjects identified with A and six with B.

(a) Formulate an acceptable rule for single-digit random numbers. Incorporate into this rule a procedure that will ensure equal numbers of subjects in the two groups. *Check your answer in Appendix B before proceeding.*

(b) Reading from left to right in the top row of the random number page (Table H, Appendix C), use the random digits of each random number in conjunction with your assignment rule to determine whether the first subject is A or B, and so forth. List the assignment for each subject.

Answers on pages 429 and 430.

Reminder:

Random sampling occurs in well-designed surveys*, and random assignment occurs in well-designed* experiments.

8.6 SURVEYS OR EXPERIMENTS?

When using random numbers, it's important to have a general perspective. Are you engaged in a *survey* (because subjects have been sampled from a real population) or in an *experiment* (because subjects have been assigned to various groups)? In the case of surveys, the object is to obtain a random sample from some real population.

Short-circuit unnecessary clerical work as much as possible, but use random numbers in a fashion that complies with the notion of *random sampling—that all subjects in the population have an equal opportunity of being sampled*. In the case of experiments, the objective is to obtain, at the outset of the experiment, equivalent groups whose membership has been determined by chance. Introduce any restrictions required to generate equal group sizes (for example, the restriction that every other subject be assigned to the smaller group), but use random numbers in a fashion that complies with the notion of *random assignment—that all subjects have an equal opportunity of being assigned to each of the various groups.*

PROBABILITY

Probability
The proportion or fraction of times that a particular event is likely to occur.

Probability considerations are prevalent in everyday life: the *probability* that it will rain this weekend (20 percent, or one in five, according to our morning weather report), that a projected family of two children will consist of a boy and a girl (one-half, on the assumption that boys and girls are equally likely), or that you'll win a state lottery (one in many millions, unfortunately). Probability considerations also are prevalent in inferential statistics and, therefore, in the remainder of this book.

8.7 DEFINITION

***Probability* refers to the proportion or fraction of times that a particular event is likely to occur.**

The probability of an event can be determined in several ways. We can *speculate* that if a coin is truly fair, heads and tails should be equally likely to occur whenever the coin is tossed, and therefore, the probability of heads should equal .50, or $^1/_2$. By the same token, ignoring the slight differences in the lengths of the months of the year, we can *speculate* that if a couple's wedding is equally likely to occur in each of the months, then the probability of a June wedding should be .08 or $^1/_{12}$.

On the other hand, we might actually *observe* a long string of coin tosses and conclude, on the basis of these observations, that the probability of heads should equal .50, or $^1/_2$. Or we might collect extensive data on wedding months and *observe* that the probability of a June wedding actually is not only much higher than the speculated .08 or $^1/_{12}$, but higher than that for any other month. In this case, assuming that the observed probability is well substantiated, we would use it rather than the erroneous speculative probability.

Table 8.1
PROBABILITY DISTRIBUTION FOR HEIGHTS OF AMERICAN MEN

HEIGHT (INCHES)	RELATIVE FREQUENCY
76 or taller	.02
75	.02
74	.03
73	.05
72	.08
71	.11
70	.12
69	.14
68	.12
67	.11
66	.08
65	.05
64	.03
63	.02
62 or shorter	.02
Total	1.00

Source: See Figure 5.2.

Probability Distribution of Heights

Sometimes we'll use probabilities that are based on a mixture of observation and speculation, as in **Table 8.1.** This table shows a probability distribution of heights for all American men (derived from the *observed* distribution of heights for 3091 men by superimposing—and this is the *speculative* component—the idealized normal curve, originally shown in Figure 5.2). These probabilities indicate the proportion of men in the population who attain a particular height. They also indicate the likelihood of observing a particular height when a single man is randomly selected from the population. For example, the probability is .14 that a randomly selected man will stand 69 inches tall. Each of the probabilities in Table 8.1 can vary in value between 0 (impossible) and 1 (certain). Furthermore, an entire set of probabilities always sums to 1.

Probabilities of Complex Events

Often you can find the probabilities of more complex events by using two rules—the addition and multiplication rules—for combining the probabilities of various simple events. Each rule will be introduced, in turn, to solve a problem based on the probabilities from Table 8.1.

8.8 ADDITION RULE

What's the probability that a randomly selected man will be at least 73 inches tall? That's the same as asking, "What's the probability that a man will stand 73 inches tall *or* taller?" To answer this type of question, which involves a cluster of simple events connected by the word *or,* merely add the respective probabilities. The probability that a man, X, will stand 73 or more inches tall, symbolized as $\Pr(X \geq 73)$, equals the sum of the probabilities in Table 8.1 that a man will stand 73 or 74 or 75 or 76 inches or taller, that is,

$$\Pr(X \geq 73) = \Pr(73) + \Pr(74) + \Pr(75) + \Pr(76 \textit{ or taller})$$
$$= .05 + .03 + .02 + .02 = .12$$

Mutually Exclusive Events

Mutually Exclusive Events

Events that cannot occur together.

Addition Rule

Add together the separate probabilities of several mutually exclusive events to find the probability that any one of these events will occur.

The addition of probabilities, as just stated, works only when none of the events can occur together. This is true in the present case because, for instance, a single man can't stand both 73 and 74 inches tall. By the same token, a single person's blood type can't be both O and B (or any other type); nor can a single person's birth month be both January and February (or any other month). Whenever events can't occur together—that is, more technically, when there are **mutually exclusive events**—the probability that any one of these several events will occur is given by the addition rule. Therefore, whenever you must find the probability for two or more sets of mutually exclusive events connected by the word *or,* use the addition rule.

The **addition rule** *tells us to add together the separate probabilities of several mutually exclusive events in order to find the probability that any one of these events will occur*. Stated generally, the addition rule reads:

ADDITION RULE FOR MUTUALLY EXCLUSIVE EVENTS

$$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) \qquad (8.1)$$

where Pr( ) refers to the probability of the event in parentheses and A and B are mutually exclusive events.

When Events Aren't Mutually Exclusive

When events aren't mutually exclusive because they can occur together, the addition rule must be adjusted for the overlap between outcomes. For example, assume that students in your class are seniors with probability .20, psychology majors with probability .70, and both seniors and psychology majors with probability .10. To determine the probability that a student is *either* a senior *or* a psychology major, add the first two probabilities (.20 + .70 = .90), but then subtract the third probability (.90 − .10 = .80), because students who are both seniors and psychology majors are counted twice—once because they are seniors and once because they are psychology majors.

Ordinarily, in this book you will be able to use the addition rule for mutually exclusive outcomes. Before doing so, however, always satisfy yourself that the various events are, in

fact, mutually exclusive. Otherwise, the addition rule yields an inflated answer that must be reduced by subtracting the overlap between events that are not mutually exclusive.

Progress Check *8.6 Assuming that people are equally likely to be born during any one of the months, what is the probability of Jack being born during

(a) June?

(b) any month other than June?

(c) either May or June?

Answers on page 430.

8.9 MULTIPLICATION RULE

Given a probability of .12 that a randomly selected man will be at least 73 inches tall, what is the probability that two randomly selected men will be at least 73 inches tall? That is the same as asking, "What is the probability that the first man will stand at least 73 inches tall *and* that the second man will stand at least 73 inches tall?"

To answer this type of question, which involves clusters of simple events connected by the word *and,* merely multiply the respective probabilities. The probability that both men will stand at least 73 inches tall equals the product of the probabilities in Table 8.1 that the first man, X_1, will stand at least 73 inches tall *and* that the second man, X_2, will stand at least 73 inches tall, that is,

$$\Pr(X_1 \geq 73 \text{ and } X_2 \geq 73) = [\Pr(X_1) \geq 73][\Pr(X_2) \geq 73] = (.12)(.12) = .0144$$

Notice that the probability of two events occurring together (.0144) is smaller than the probability of either event occurring alone (.12). If you think about it, this should make sense. The combined occurrence of two events is less likely than the solitary occurrence of just one of the two events.

Independent Events

Independent Events

The occurrence of one event has no effect on the probability that the other event will occur.

Multiplication Rule

Multiply together the separate probabilities of several independent events to find the probability that these events will occur together.

The multiplication of probabilities, as discussed, works only because the occurrence of one event has no effect on the probability of the other event. This is true in the present case because, when randomly selecting from the population of American men, the initial appearance of a man at least 73 inches tall has no effect, practically speaking, on the probability that the next man also will be at least 73 inches tall. By the same token, the birth of a girl in a family has no effect on the probability of .50 that the next family addition also will be a girl, and the winning lottery number for this week has no effect on the probability that a particular lottery number will be a winner for the next week. Whenever one event has no effect on the other—that is, more technically, when there are **independent events**—the probability of the combined or joint occurrence of both events is given by the multiplication rule.

Whenever you must find the probability for two or more sets of independent events connected by the word *and,* use the multiplication rule. The **multiplication rule** *tells us to multiply together the separate probabilities of several independent events in order to find the probability that these events will occur together* . Stated generally, for the independent events A and B, the multiplication rule reads:

MULTIPLICATION RULE FOR INDEPENDENT EVENT

$$\Pr(A \text{ and } B) = [\Pr(A)][\Pr(B)] \qquad (8.2)$$

where A and B are independent events.

Progress Check * 8.7 Assuming that people are equally likely to be born during any of the months, and also assuming (possibly over the objections of astrology fans) that the birthdays of married couples are independent, what's the probability of

(a) the husband being born during January and the wife being born during February?

(b) both husband and wife being born during December?

(c) both husband and wife being born during the spring (April or May)? (**Hint:** First, find the probability of just one person being born during April or May.)

Answers on page 430.

Dependent Events

When the occurrence of one event affects the probability of the other event, these events are dependent. Although the heights of randomly selected pairs of men are independent, the heights of brothers are dependent. Knowing, for instance, that one person is relatively tall increases the probability that his brother also will be relatively tall. Among students in your class, being a senior and a psychology major are dependent if knowing that a student is a senior automatically changes (either increases or decreases) the probability of being a psychology major.

Conditional Probabilities

Conditional Probability

The probability of one event, given the occurrence of another event.

Before multiplying to obtain the probability that two dependent events occur together, the probability of the second event must be adjusted to reflect its dependency on the prior occurrence of the first event. This new probability is the **conditional probability** of the second event, given the first event. Examples of conditional probabilities are the probability that you will earn a grade of A in a course, given that you have already gotten an A on the midterm, or the probability that you'll graduate from college, given that you've already completed the first two years. Notice that, in both examples, these conditional probabilities are different—they happen to be larger—than the regular or unconditional probability of your earning a grade of A in a course (without knowing your grade on the midterm) or of graduating from college (without knowing whether or not you've completed the first two years). Incidentally, a conditional probability also can be smaller than its corresponding unconditional probability. Such is the case for the conditional probability that you'll earn a grade of A in a course, given that you have already gotten (alas) a C on the midterm.

If, as already assumed, being a senior and a psychology major are dependent events among students in your class, then it would be incorrect to use the multiplication rule for two independent outcomes. More specifically, it would be incorrect to simply multiply the observed unconditional probability of being a senior (.20) and the observed unconditional probability of being a psych major (.70), and to find

$$\Pr(\textit{senior and psych}) \neq (.20)(.70) = .14$$

Instead, you must go beyond knowing merely the proportion of students who are psych majors (the unconditional probability of being a psych major) to find the proportion of psych majors *among the subset of seniors* (the conditional probability of being a psych major, given that a student is a senior). For example, a survey of your class might reveal that, although .70 of *all* students are psych majors (for an observed unconditional probability of .70), only *.50 of all seniors are also psych majors* (for an observed conditional probability of .50 for being a psych major given that a student is a senior). Therefore, it would be correct to multiply the observed unconditional probability of being a senior (.20) and the observed conditional probability of being a

psych major, given that a student is a senior (.50), and to find the correct probability, that is,

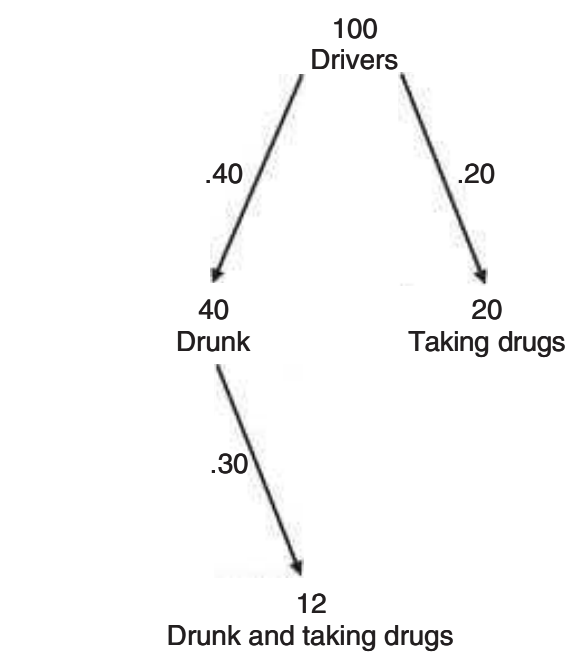
$$\Pr(\textit{senior and psych}) = \Pr(\textit{senior})\Pr(\textit{psych}, \textit{given senior}) = (.20)(.50) = .10$$

rather than the (previous and) erroneous .14, when the dependency between events was ignored.

Alternative Approach to Conditional Probabilities

Conditional probabilities can be easily misinterpreted. Sometimes it is helpful to convert probabilities to frequencies (which, for example, total 100); solve the problem with frequencies; and then convert the answer back to a probability.* **Figure 8.1** shows a frequency analysis for the 100 drivers involved in fatal accidents. Working from the top down, notice that among the 100 drivers, 40 are drunk (from .40 × 100 = 40) and 20 take drugs (from .20 × 100 = 20). Also notice that 12 of the 40 drunk drivers also take drugs (from .30 × 40 = 12). Now, it is fairly straightforward to establish that the probability of drivers both being drunk *and* taking drugs. It is simply the number of drivers who are drunk and take drugs, 12, divided by the total number of drivers, 100, that is, 12/100 =.12, which, of course, is the same as the previous answer.

Once a frequency analysis has been done, it often is easy to answer other questions. For example, you might ask "What is the conditional probability of being drunk, given that the driver takes illegal drugs?" Referring to Figure 8.1, divide the number of drivers who are drunk and take drugs, 12, by the number of drivers who take drugs, 20, that is, 12/20 = .60. (This conditional probability of .60, given drivers who take drugs,



Pr (drivers who are drunk and taking drugs) = 12/100 = .12

FIGURE 8.1
A frequency analysis of 100 drivers who caused fatal accidents.

*See Gigerenzer, G. (2002). *Calculated Risks*. New York, NY; Simon & Schuster, for a more extensive discussion of using frequencies to simplify work with conditional probabilities.

is twice that of .30, given drunk drivers, because the fixed number of drivers who are drunk and take drugs, 12, represents proportionately more (.60) among the relatively small number of drivers who take drugs, 20, and proportionately less (.30) among the relatively large number of drunk drivers, 40.)

Whenever appropriate, as in Progress Check 8.8 and Review Exercise 8.21, construct a line diagram, similar to the one in Figure 8.1, and use frequencies to answer questions involving conditional probabilities. Incidentally, when doing a frequency analysis, there is nothing sacred about a convenient number of 100. As events become rarer and their probabilities become smaller, more convenient numbers might equal 1,000, 10,000, or even 100,000. The choice of a convenient number does not affect the accuracy of the answer since frequencies are converted back to probabilities.

Ordinarily, in this book, you'll be able to use the multiplication rule for independent outcomes (including when it appears, slightly disguised, in Chapter 19 as a key ingredient in the important statistical test known as chi-square). Before using this rule to calculate the probabilities of complex events, however, satisfy yourself—mustering any information at your disposal, whether speculative or observational—that the various outcomes lack any obvious dependency. That is, satisfy yourself that, just as the outcome of the last coin toss has no obvious effect on the outcome of the next toss, the prior occurrence of one event has no obvious effect on the occurrence of the other event. If this is not the case, you should only proceed if one outcome can be expressed (most likely on the basis of some data collection) as a conditional probability of the other.*

Progress Check *8.8 Among 100 couples who had undergone marital counseling, 60 couples described their relationships as improved, and among this latter group, 45 couples had children. The remaining couples described their relationships as unimproved, and among this group, 5 couples had children. (Hint: Using a frequency analysis, begin with the 100 couples, first branch into the number of couples with improved and unimproved relationships, then under each of these numbers, branch into the number of couples with children and without children. Enter a number at each point of the diagram before proceeding.)

(a) What is the probability of randomly selecting a couple who described their relationship as improved?

(b) What is the probability of randomly selecting a couple with children?

(c) What is the conditional probability of randomly selecting a couple with children, given that their relationship was described as improved?

(d) What is the conditional probability of randomly selecting a couple without children, given that their relationship was described as not improved?

(e) What is the conditional probability of an improved relationship, given that a couple has children?

Answers on page 430.

*Don't confuse independent and dependent *outcomes* with independent and dependent *variables*. Independent and dependent *outcomes* refer to whether or not the occurrence of one outcome affects the probability that the other outcome will occur and dictates the precise form of the multiplication rule. On the other hand, as described in Chapter 1, independent and dependent *variables* refer to the manipulated and measured variables in experiments. Usually, the context—whether calculating the probabilities of complex outcomes or describing the essential features of an experiment—will make the meanings of these terms clear.

8.10 PROBABILITY AND STATISTICS

Probability assumes a key role in inferential statistics including, for instance, the important area known as *hypothesis testing*. Because of the inevitable variability that accompanies any observed result, such as a mean difference between two groups, its value must be viewed within the context of the many possible results that could have occurred just by chance. With the aid of some theoretical curve, such as the normal curve, and a provisional assumption, known as the *null hypothesis,* that chance can reasonably account for the result, probabilities are assigned to the one observed mean difference. If this probability is very small, the result is viewed as a rare outcome, and we conclude that something real—that is, something that can't reasonably be attributed to chance—has occurred. On the other hand, if this probability isn't very small, the result is viewed as a common outcome, and we conclude that something transitory—that is, something that can reasonably be attributed to chance—has occurred.

Common Outcomes

Common outcomes signify, most generally, a lack of evidence that something special has occurred. For instance, they suggest that the observed mean difference—whatever its value—might signify that the true mean difference could equal zero and, therefore, that any comparable study would just as likely produce either a positive or negative mean difference. Therefore, the observed mean difference should not be taken seriously because, in the language of statistics, it lacks *statistical significance.*

Rare Outcomes

On the other hand, rare outcomes signify that something special has occurred. For instance, they suggest that the observed mean difference probably signifies a true mean difference equal to some nonzero value and, therefore, that any comparable study would most likely produce a mean difference with the same sign and a value in the neighborhood of the one originally observed. Therefore, the observed mean difference should be taken seriously because it has statistical significance.

Common or Rare?

As an aid to determining whether observed results should be viewed as common or rare, statisticians interpret different *proportions of area under theoretical curves,* such as the normal curve shown in **Figure 8.2**, *as probabilities of random outcomes.* For instance, the standard normal table indicates that .9500 is the proportion of total area between z scores of –1.96 and +1.96. (Verify this proportion by referring to Table A in Appendix C and, if necessary, to the latter part of Section 5.6.) Accordingly, the probability of a randomly selected z score anywhere between ±1.96 equals .95. Because it should happen about 95 times out of 100, this is often designated as a *common* event signifying that, once variability is considered, nothing special is happening. On the other hand, since the standard normal curve indicates that .025 is the proportion of total area above a z score of +1.96, and also that .025 is the proportion of total area below a z score of –1.96, then the probability of a randomly selected z score anywhere beyond either +1.96 or –1.96 equals .05 (from .025 + .025, thanks to the addition rule). Because it should happen only about 5 times in 100, this is often designated as a *rare* outcome signifying that something special is happening.

At this point, you're not expected to understand the rationale behind this perspective, but merely that, once identified with a particular result, a specified sector of area under a curve will be interpreted as the probability of that outcome. Furthermore, since the probability of an outcome has important implications for generalizing beyond actual results, probabilities play a key role in inferential statistics.

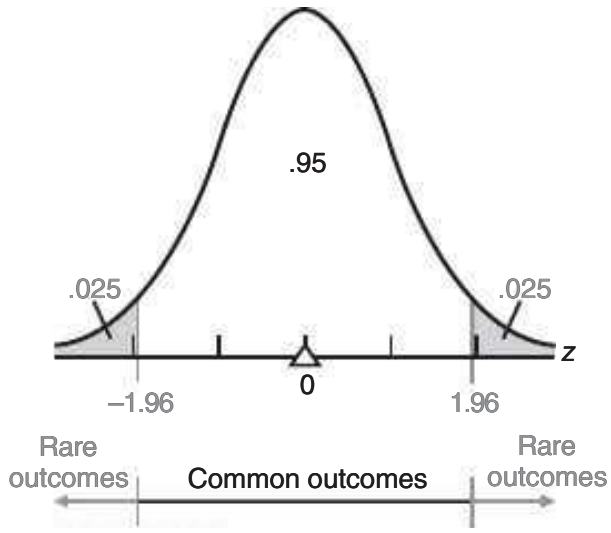


FIGURE 8.2
One possible model for determining common and rare outcomes.

Progress Check * 8.9 Referring to the standard normal table (Table A, Appendix C), find the probability that a randomly selected *z* score will be

(a) above 1.96

(b) either above 1.96 or below –1.96

(c) between –1.96 and 1.96

(d) either above 2.58 or below –2.58

Answers on page 431.

Summary

Any set of potential observations may be characterized as a population. Any subset of observations constitutes a sample.

Populations are either real or hypothetical, depending on whether or not all observations are accessible at the time of sampling.

The valid application of techniques from inferential statistics requires that the samples be random or that subjects be randomly assigned. A sample is random if at each stage of sampling the selection process guarantees that all remaining observations in the population have an equal chance of being included in the sample. Random assignment occurs whenever all subjects have an equal opportunity of being assigned to each of the various groups.

Tables of random numbers provide one method both for taking random samples in surveys and for randomly assigning subjects to various groups in experiments. Some type of randomization always should occur during the early stages of any investigation, whether a survey or an experiment.

The probability of an event specifies the proportion of times that this event is likely to occur.

Whenever you must find the probability of sets of mutually exclusive events connected with the word *or,* use the addition rule: Add together the separate probabilities of each of the mutually exclusive events to find the probability that any one of these events will occur. Whenever events aren't mutually exclusive, the addition rule must be adjusted for the overlap between outcomes.

Whenever you must find the probability of sets of independent events connected with the word *and,* use the multiplication rule: Multiply together the separate probabilities of each of the independent events to find the probability that these events will occur together. Whenever events are dependent, the multiplication rule must be adjusted by using the conditional probability of the second outcome, given the occurrence of the first outcome.

In inferential statistics, sectors of area under various theoretical curves are interpreted as probabilities, and these probabilities play a key role in inferential statistics.

Important terms

Population
Random sampling
Probability
Addition rule
Multiplication rule
Sample
Random assignment
Mutually exclusive events
Independent events
Conditional probability

Key equations

ADDITION RULE

$$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$$

MULTIPLICATION RULE

$$\Pr(A \text{ and } B) = [\Pr(A)][\Pr(B)]$$

REVIEW QUESTIONS

8.10 Television stations sometimes solicit feedback volunteered by viewers about a televised event. Following a televised debate between Barack Obama and Mitt Romney in the 2012 presidential election campaign, a TV station conducted a telephone poll to determine the "winner." Callers were given two phone numbers, one for Obama and the other for Romney, to register their opinions automatically.

(a) Comment on whether or not this was a random sample.

(b) How might this poll have been improved?

8.11 You want to take a random sample of 30 from a population described by telephone directory with a single telephone area code. Indicate whether or not each of the following selection techniques would be a random sample and, if not, why. Using the telephone directory,

(a) make 30 blind pencil stabs.

(b) refer to tables of random numbers to determine the page and then the position of the selected person on that page. Repeat 30 times.

(c) refer to tables of random numbers to find six-digit numbers that identify the page number and line on that page for each of 30 people.

(d) select 30 people haphazardly.

8.12 Indicate whether the following terms are associated with surveys (S) or experiments (E).

(a) random sample

(b) two groups

(c) real population

(d) real difference

(e) population directory

(f) digit dialing

(g) similar groups

(h) random assignment

(i) independent variable

8.13 As subjects arrive to participate in an experiment, tables of random numbers are used to make random assignments to either group A or group B. (To ensure equal numbers of subjects in the two groups, alternate subjects are automatically assigned to the other, smaller group.) Indicate with a Yes or No whether each of the following rules would work:

(a) Assign the subject to group A if the random number is **even** and to group B if the random number is **odd**.

(b) Assign the subject to group A if the first digit of the random number is between **0 and 4** and to group B if the first digit is between **5 and 9**.

(c) Assign the subject to group A if the first two digits of the random number are between **00 and 40** and to group B if the first two digits are between **41 and 99**.

(d) Assign the subject to group A if the first three digits of the random number are between **000 and 499** and to group B if the first three digits are between **500 and 999**.

***8.14** The probability of a boy being born equals .50, or $^1/_2$, as does the probability of a girl being born. For a randomly selected family with two children, what's the probability of

(a) two boys, that is, a boy and a boy? (Reminder: Before using either the addition or multiplication rule, satisfy yourself that the various events are either mutually exclusive or independent, respectively.)

(b) two girls?

(c) either two boys or two girls?

Answers on page 431.

8.15 Assume the same probabilities as in the previous question. For a randomly selected family with three children, what's the probability of

(a) three boys?

(b) three girls?

(c) either three boys or three girls?

(d) neither three boys nor three girls? (**Hint:** This question can be answered indirectly by first finding the opposite of the specified outcome, then subtracting from 1.)

8.16 A traditional test for extrasensory perception (ESP) involves a set of playing cards, each of which shows a different symbol (circle, square, cross, star, or wavy lines). If C represents a correct guess and I an incorrect guess, what is the probability of

(a) C?

(b) CI (in that order) for two guesses?

(c) CCC for three guesses?

(d) III for three guesses?

8.17 In a school population, assume that the probability of being white equals .40; black equals .30; Hispanic equals .20; and Asian-American equals .10. What is the probability of

(a) a student being either white or black.

(b) a student being neither white nor black.

(c) pairs of black and white students being selected together, assuming ethnic background has no role.

(d) given that a black student has been selected, that his/her companion is white, assuming ethnic background has no role.

(e) given that a black student has been selected, that his/her companion is white, assuming students tend to congregate with companions with similar ethnic backgrounds. In this case, would the probability of the companion being white be less than .40, equal to .40, or more than .40?

***8.18** In *Against All Odds*, the TV series on statistics (available at http://www.learner.org/resources/series65.html), statistician Bruce Hoadley discusses the catastrophic failure of the *Challenger* space shuttle in 1986. Hoadley estimates that there was a *failure* probability of .02 for each of the six O-rings (designed to prevent the escape of potentially explosive burning gases from the joints of the segmented rocket boosters).

(a) What was the *success* probability of *each* O-ring?

(b) Given that the six O-rings function independently of each other, what was the probability that *all* six O-rings would succeed, that is, perform as designed? In other words, what was the success probability of the first O-ring and the second O-ring and the third O-ring, and so forth?

(c) Given that you know the probability that all six O-rings would succeed (from the previous question), what was the probability that at least one O-ring would fail? (Hint: Use your answer to the previous question to solve this problem.)

(d) Given the abysmal failure rate revealed by your answer to the previous question, why, you might wonder, was this space mission even attempted? According to Hoadley, missile engineers thought that a secondary set of O-rings would function independently of the primary set of O-rings. If true and if the failure probability of each of the secondary O-rings was the same as that for each primary O-ring (.02), what would be the probability that both the primary and secondary O-rings would fail at any one joint? (Hint: Concentrate on the present question, ignoring your answers to previous questions.)

(e) In fact, under conditions of low temperature, as on the morning of the **Challenger** catastrophe, both primary and secondary O-rings lost their flexibility, and whenever the primary O-ring failed, its associated secondary O-ring also failed. Under these conditions, what would be the *conditional* probability of a secondary O-ring failure, *given* the failure of its associated primary O-ring? (**Note:** Any probability, including a conditional probability, can vary between 0 and 1.)

Answers on page 431.

8.19 A sensor is used to monitor the performance of a nuclear reactor. The sensor accurately reflects the state of the reactor with a probability of .97. But with a probability of .02, it gives a false alarm (by reporting excessive radiation even though the reactor is performing normally), and with a probability of .01, it misses excessive radiation (by failing to report excessive radiation even though the reactor is performing abnormally).

(a) What is the probability that a sensor will give an incorrect report, that is, either a false alarm or a miss?

(b) To reduce costly shutdowns caused by false alarms, management introduces a second completely independent sensor, and the reactor is shut down only when both sensors report excessive radiation. (According to this perspective, solitary reports of excessive radiation should be viewed as false alarms and ignored, since both sensors provide accurate information much of the time.) What is the new probability that the reactor will be shut down because of simultaneous false alarms by both the first and second sensors?

(c) Being more concerned about failures to detect excessive radiation, someone who lives near the nuclear reactor proposes an entirely different strategy: Shut down the reactor whenever either sensor reports excessive radiation. (According to this point of view, even a solitary report of excessive radiation should trigger a shutdown, since a failure to detect excessive radiation is potentially catastrophic.) If this policy were adopted, what is the new probability that excessive radiation will be missed simultaneously by both the first and second sensors?

***8.20** Continue to assume that people are equally likely to be born during any of the months. However, just for the sake of this exercise, assume that there is a tendency for married couples to have been born during the same month. Furthermore, we wish to calculate the probability of a husband and wife both being born during December.

(a) It would be appropriate to use the multiplication rule for independent outcomes. True or False?

(b) The probability of a married couple both being born during December is smaller than, equal to, or larger than $(1/12)(1/12)=1/144$.

(c) With only the previous information, it would be possible to calculate the actual probability of a married couple both being born during December. True or False?

Answers on page 431.

***8.21** Assume that the probability of breast cancer equals .01 for women in the 50–59 age group. Furthermore, if a woman does have breast cancer, the probability of a true positive mammogram (correct detection of breast cancer) equals .80 and the probability of a false negative mammogram (a miss) equals .20. On the other hand, if a woman does not have breast cancer, the probability of a true negative mammogram

(correct nondetection) equals .90 and the probability of a false positive mammogram (a false alarm) equals .10. (Hint: Use a frequency analysis to answer questions. To facilitate checking your answers with those in the book, begin with a total of 1,000 women, then branch into the number of women who do or do not have breast cancer, and finally, under each of these numbers, branch into the number of women with positive and negative mammograms.)

(a) What is the probability that a randomly selected woman will have a positive mammogram?

(b) What is the probability of having breast cancer, given a positive mammogram?

(c) What is the probability of not having breast cancer, given a negative mammogram?

Answers on page 431.

CHAPTER 9

Sampling Distribution of the Mean

Preview

This chapter focuses on the single most important concept in inferential statistics—the concept of a sampling distribution. A sampling distribution serves as a frame of reference for every outcome, among all possible outcomes, that could occur just by chance. It reappears in every subsequent chapter as the key to understanding how, once variability has been estimated, we can generalize beyond a limited set of actual observations. In order to use a sampling distribution, we must identify its mean, its standard deviation, and its shape—a seemingly difficult task that, thanks to the theory of statistics, can be performed by invoking the population mean, the population standard deviation, and the normal curve, respectively.

There's a good chance that you've taken the SAT test, and you probably remember your scores. Assume that the SAT math scores for all college-bound students during a recent year were distributed around a mean of 500 with a standard deviation of 110. An investigator at a university wishes to test the claim that, on the average, the SAT math scores for local freshmen equals the national average of 500. His task would be straightforward if, in fact, the math scores for all local freshmen were readily available. Then, after calculating the mean score for all local freshmen, a direct comparison would indicate whether, on the average, local freshmen score below, at, or above the national average.

Assume that it is not possible to obtain scores for the entire freshman class. Instead, SAT math scores are obtained for a random sample of 100 students from the local population of freshmen, and the mean score for this sample equals 533. If each sample were an exact replica of the population, generalizations from the sample to the population would be most straightforward. Having observed a mean score of 533 for a sample of 100 freshmen, we could have concluded, without even a pause, that the mean math score for the entire freshman class also equals 533 and, therefore, exceeds the national average.

9.1 WHAT IS A SAMPLING DISTRIBUTION?

Random samples rarely represent the underlying population exactly. Even a mean math score of 533 could originate, just by chance, from a population of freshmen whose mean equals the national average of 500. Accordingly, generalizations from a single sample to a population are much more tentative. Indeed, generalizations are based not merely on the single sample mean of 533 but also on its distribution—a distribution of sample means for all possible random samples. Representing the statistician's model of random outcomes,

Sampling Distribution of the Mean

Probability distribution of means for all possible random samples of a given size from some population.

the *sampling distribution of the mean* refers to the probability distribution of means for all possible random samples of a given size from some population.

In effect, this distribution describes the variability among sample means that could occur just by chance and thereby serves as a frame of reference for generalizing from a single sample mean to a population mean.

The sampling distribution of the mean allows us to determine whether, given the variability among all possible sample means, the one observed sample mean can be viewed as a *common* outcome or as a *rare* outcome (from a distribution centered, in this case, about a value of 500). If the sample mean of 533 qualifies as a *common* outcome in this sampling distribution, then the difference between 533 and 500 isn't large enough, relative to the variability of all possible sample means, to signify that anything special is happening in the underlying population. Therefore, we can conclude that the mean math score for the entire freshman class could be the same as the national average of 500. On the other hand, if the sample mean of 533 qualifies as a *rare* outcome in this sampling distribution, then the difference between 533 and 500 is large enough, relative to the variability of all possible sample means, to signify that something special probably is happening in the underlying population. Therefore, we can conclude that the mean math score for the entire freshman class probably exceeds the national average of 500.

All Possible Random Samples

When attempting to generalize from a single sample mean to a population mean, we must consult the sampling distribution of the mean. In the present case, this distribution is based on *all possible* random samples, each of size 100 that can be taken from the

local population of freshmen. *All possible random samples* refers not to the number of samples of size 100 required to *survey completely* the local population of freshmen but to the number of different ways in which a *single* sample of size 100 can be selected from this population.

"All possible random samples" tends to be a huge number. For instance, if the local population contained at least 1,000 freshmen, the total number of possible random samples, each of size 100, would be astronomical in size. The 301 digits in this number would dwarf even the national debt. Even with the aid of a computer, it would be a horrendous task to construct this sampling distribution from scratch, itemizing each mean for all possible random samples.

Fortunately, statistical theory supplies us with considerable information about the sampling distribution of the mean, as will be discussed in the remainder of this chapter. Armed with this information about sampling distributions, we'll return to the current example in the next chapter and test the claim that the mean math score for the local population of freshmen equals the national average of 500. Only at that point—and not at the end of this chapter—should you expect to understand completely the role of sampling distributions in practical applications.

9.2 CREATING A SAMPLING DISTRIBUTION FROM SCRATCH

Let's establish precisely what constitutes a sampling distribution by creating one from scratch under highly simplified conditions. Imagine some ridiculously small population of four observations with values of 2, 3, 4, and 5, as shown in **Figure 9.1.** Next, itemize all possible random samples, each of size two, that could be taken from this population. There are four possibilities on the first draw from the population and also four possibilities on the second draw from the population, as indicated in **Table 9.1.***

The two sets of possibilities combine to yield a total of 16 possible samples. At this point, remember, we're clarifying the notion of a sampling distribution of the mean. In practice, only a single random sample, not 16 possible samples, would be taken from the population; the sample size would be very small relative to a much larger population size, and, of course, not all observations in the population would be known.

For each of the 16 possible samples, Table 9.1 also lists a sample mean (found by adding the two observations and dividing by 2) and its probability of occurrence (expressed as $\frac{1}{16}$, since each of the 16 possible samples is equally likely). When cast into a relative frequency or probability distribution, as in **Table 9.2,** the 16 sample means constitute the sampling distribution of the mean, previously defined as the probability distribution of means for all possible random samples of a given size from some population. Not all values of the sample mean occur with equal probabilities in Table 9.2 since some values occur more than once among the 16 possible samples. For instance, a sample mean value of 3.5 appears among 4 of 16 possibilities and has a probability of $\frac{4}{16}$.

Probability of a Particular Sample Mean

The distribution in Table 9.2 can be consulted to determine the probability of obtaining a particular sample mean or set of sample means. For example, the probability of a randomly selected sample mean of 5.0 equals $\frac{1}{16}$ or .0625. According to the addition

*Ordinarily, a single observation is sampled only once, that is, sampling is *without replacement*. If employed with the present, highly simplified example, however, sampling without replacement would magnify an unimportant technical adjustment.

than the full spectrum of scores from the parent population. Sometimes, because of the relatively large number of extreme scores in a particular direction, the calculation of a mean only slightly dilutes their effect, and the sample mean emerges with some more extreme value. The likelihood of obtaining extreme sample mean values declines with the extremity of the value, producing the smoothly tapered, slender tails that characterize a normal curve.

Progress Check *9.5 Indicate whether the following statements are True or False. The central limit theorem

(a) states that, with sufficiently large sample sizes, the shape of the population is normal.

(b) states that, regardless of sample size, the shape of the sampling distribution of the mean is normal.

(c) ensures that the shape of the sampling distribution of the mean equals the shape of the population.

(d) applies to the shape of the sampling distribution—not to the shape of the population and not to the shape of the sample.

Answers on page 432.

9.7 OTHER SAMPLING DISTRIBUTIONS

For the Mean

There are many different sampling distributions of means. A new sampling distribution is created by a switch to another population. Furthermore, for any single population, there are as many different sampling distributions as there are possible sample sizes. Although each of these sampling distributions has the same mean, the value of the standard error always differs and depends upon the size of the sample.

For Other Measures

There are sampling distributions for measures other than a single mean. For instance, there are sampling distributions for medians, proportions, standard deviations, variances, and correlations, as well as for differences between pairs of means, pairs of proportions, and so forth. We'll have occasion to work with some of these distributions in later chapters.

Summary

The notion of a sampling distribution is the most important concept in inferential statistics. The sampling distribution of the mean is defined as the probability distribution of means for all possible random samples of a given size from some population.

Statistical theory pinpoints three important characteristics of the sampling distribution of the mean:

- The mean of the sampling distribution equals the mean of the population.
- The standard deviation of the sampling distribution, that is, the standard error of the mean, equals the standard deviation of the population divided by the square root of the sample size. An important implication of this formula is that a larger sample size translates into a sampling distribution with a smaller variability,

allowing more precise generalizations from samples to populations. The standard error of the mean serves as a rough measure of the average amount by which sample means deviate from the mean of the sampling distribution or from the population mean.

- According to the central limit theorem, regardless of the shape of the population, the shape of the sampling distribution approximates a normal curve if the sample size is sufficiently large. Depending on the degree of non-normality in the parent population, a sample size of between 25 and 100 is sufficiently large.

Any single sample mean can be viewed as originating from a sampling distribution whose (1) mean equals the population mean (whatever its value); whose (2) standard error equals the population standard deviation divided by the square root of the sample size; and whose (3) shape approximates a normal curve (if the sample size satisfies the requirements of the central limit theorem).

Important Terms

Mean of the sampling distribution of the mean ($\mu_{\overline{X}}$)
Sampling distribution of the mean
Standard error of the mean ($\sigma_{\overline{X}}$)
Central limit theorem

Key Equations

SAMPLING DISTRIBUTION MEAN

$$\mu_{\overline{X}} = \mu$$

STANDARD ERROR

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

REVIEW QUESTIONS

9.6 A random sample tends not to be an exact replica of its parent population. This fact has a number of implications. Indicate which are true and which are false.

(a) All possible random samples can include a few samples that are exact replicas of the population, but most samples aren't exact replicas.

(b) A more representative sample can be obtained by handpicking (rather than randomly selecting) observations.

(c) Insofar as it misrepresents the parent population, a random sample can cause an erroneous generalization.

(d) In practice, the mean of a single random sample is evaluated relative to the variability of means for all possible random samples.

9.7 Define the sampling distribution of the mean.

9.8 Specify three important properties of the sampling distribution of the mean.

9.9 Indicate whether the following statements are true or false. If we took a random sample of 35 subjects from some population, the associated sampling distribution of the mean would have the following properties:

(a) Shape would approximate a normal curve.

(b) Mean would equal the one sample mean.

(c) Shape would approximate the shape of the population.

(d) Compared to the population variability, the variability would be reduced by a factor equal to the square root of 35.

(e) Mean would equal the population mean.

(f) Variability would equal the population variability.

9.10 Indicate whether the following statements are true or false. The sampling distribution of the mean

(a) is always constructed from scratch, even when the population is large.

(b) serves as a bridge to aid generalizations from a sample to a population.

(c) is the same as the sample mean.

(d) always reflects the shape of the underlying population.

(e) has a mean that always coincides with the population mean.

(f) is a device used to determine the effect of variability (that is, what can happen, just by chance, when samples are random).

(g) remains unchanged even with shifts to a new population or sample size.

(h) supplies a spectrum of possibilities against which to evaluate the one observed sample mean.

(i) tends to cluster more closely about the population mean with increases in sample size.

9.11 Someone claims that, since the mean of the sampling distribution equals the population mean, any single sample mean must also equal the population mean. Any comment?

9.12 Given that population standard deviation equals 24, how large must the sample size, *n*, be in order for the standard error to equal

(a) 8 ?

(b) 6 ?

(c) 3 ?

(d) 2 ?

9.13 Given a sample size of 36, how large does the population standard deviation have to be in order for the standard error to be

(a) 1 ?

(b) 2 ?

(c) 5 ?

(d) 100 ?

9.14 **(a)** A random sample of size 144 is taken from the local population of grade-school children. Each child estimates the number of hours per week spent watching TV. At this point, what can be said about the sampling distribution?

(b) Assume that a standard deviation, σ, of 8 hours describes the TV estimates for the local population of schoolchildren. At this point, what can be said about the sampling distribution?

(c) Assume that a mean, μ, of 21 hours describes the TV estimates for the local population of schoolchildren. Now what can be said about the sampling distribution?

(d) Roughly speaking, the sample means in the sampling distribution should deviate, on average, about ___ hours from the mean of the sampling distribution and from the mean of the population.

(e) About 95 percent of the sample means in this sampling distribution should be between ___ hours and ___ hours.

CHAPTER

10

Introduction to Hypothesis Testing: The *z* Test

Preview

This chapter describes the first in a series of hypothesis tests. Learning the vocabulary of special terms for hypothesis tests will be most helpful throughout the remainder of the book. However, do not become so concerned about either terminology or computational mechanics that you lose sight of the essential role of the sampling distribution—the model of everything that could happen just by chance—in any hypothesis test.

Using the sampling distribution as our frame of reference, the one observed outcome is characterized as either a common outcome or a rare outcome. A common outcome is readily attributable to chance, and therefore, the hypothesis that nothing special is happening—the null hypothesis—is retained. On the other hand, a rare outcome isn't readily attributable to chance, and therefore, the null hypothesis is rejected (usually to the delight of the researcher).

10.1 TESTING A HYPOTHESIS ABOUT SAT SCORES

In the previous chapter, we postponed a test of the hypothesis that the mean SAT math score for all local freshmen equals the national average of 500. Now, given a mean math score of 533 for a random sample of 100 freshmen, let's test the hypothesis that, with respect to the national average, nothing special is happening in the local population. Insofar as an investigator usually suspects just the opposite—namely, that something special is happening in the local population—he or she hopes to reject the hypothesis that nothing special is happening, henceforth referred to as the *null* hypothesis and defined more formally in a later section.

Hypothesized Sampling Distribution

If the null hypothesis is true, then the distribution of sample means—that is, the sampling distribution of the mean for all possible random samples, each of size 100, from the local population of freshmen—will be centered about the national average of 500. (Remember, the mean of the sampling distribution always equals the population mean.) In **Figure 10.1**, this sampling distribution is referred to as the *hypothesized* sampling distribution, since its mean equals 500, the hypothesized mean reading score for the local population of freshmen.

Anticipating the key role of the hypothesized sampling distribution in our hypothesis test, let's focus on two more properties of this distribution:

1. In Figure 10.1, vertical lines appear, at intervals of size 11, on either side of the hypothesized population mean of 500. These intervals reflect the size of the standard error of the mean, $\sigma_{\bar{X}}$. To verify this fact, originally demonstrated in Chapter 9, substitute 110 for the population standard deviation, σ, and 100 for the sample size, n, in Formula 9.2 to obtain

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{110}{\sqrt{100}} = \frac{110}{10} = 11$$

2. Notice that the shape of the hypothesized sampling distribution in Figure 10.1 approximates a normal curve, since the sample size of 100 is large enough to satisfy the requirements of the central limit theorem. Eventually, with the aid of normal curve tables, we will be able to construct boundaries for common and rare outcomes under the null hypothesis.

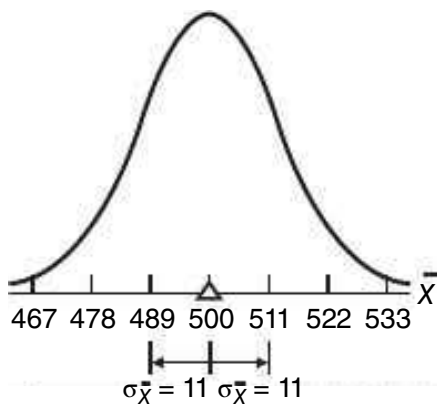


FIGURE 10.1
Hypothesized sampling distribution of the mean centered about a hypothesized population mean of 500.

The null hypothesis that the population mean for the freshman class equals 500 is *tentatively* assumed to be true. It is tested by determining whether the one observed sample mean qualifies as a common outcome or a rare outcome in the hypothesized sampling distribution of Figure 10.1.

Common Outcomes

An observed sample mean qualifies as a *common* outcome if the difference between its value and that of the hypothesized population mean is small enough to be viewed as a probable outcome under the null hypothesis.

That is, a sample mean qualifies as a common outcome if it doesn't deviate too far from the hypothesized population mean but appears to emerge from the dense concentration of possible sample means in the middle of the sampling distribution. *A common outcome signifies a lack of evidence that, with respect to the null hypothesis, something special is happening in the underlying population*. Because now there is no compelling reason for rejecting the null hypothesis, it is retained.

Key Point:

Does the one observed sample mean qualify as a common or a rare outcome?

Rare Outcomes

An observed sample mean qualifies as a *rare* outcome if the difference between its value and the hypothesized population mean is too large to be reasonably viewed as a probable outcome under the null hypothesis.

That is, a sample mean qualifies as a rare outcome if it deviates too far from the hypothesized mean and appears to emerge from the sparse concentration of possible sample means in either tail of the sampling distribution. *A rare outcome signifies that, with respect to the null hypothesis, something special probably is happening in the underlying population.* Because now there are grounds for suspecting the null hypothesis, it is rejected.

Boundaries for Common and Rare Outcomes

Superimposed on the hypothesized sampling distribution in **Figure 10.2** is one possible set of boundaries for common and rare outcomes, expressed in values of $\overline{X}$. (Techniques for constructing these boundaries are described in Section 10.7.) If the one observed sample mean is located between 478 and 522, it will qualify as a common outcome (readily attributed to variability) under the null hypothesis, and the null

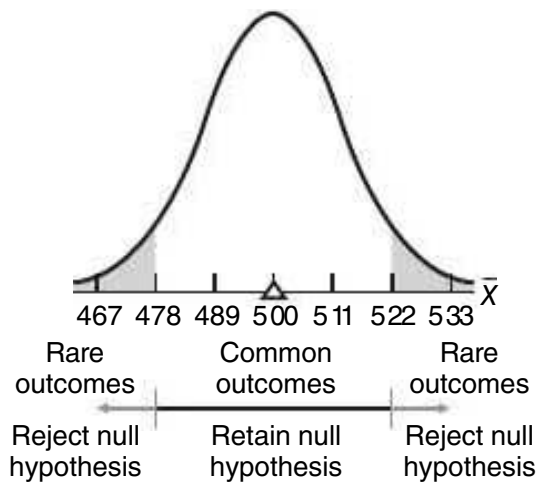


FIGURE 10.2

One possible set of common and rare outcomes (values of $\overline{X}$).

hypothesis will be retained. If, however, the one observed sample mean is greater than 522 or less than 478, it will qualify as a rare outcome (not readily attributed to variability) under the null hypothesis, and the null hypothesis will be rejected. Because the observed sample mean of 533 does exceed 522, the null hypothesis is rejected. On the basis of the present test, it is unlikely that the sample of 100 freshmen, with a mean math score of 533, originates from a population whose mean equals the national average of 500, and, therefore, the investigator can conclude that the mean math score for the local population of freshmen probably differs from (exceeds) the national average.

10.2 *z* TEST FOR A POPULATION MEAN

For the hypothesis test with SAT math scores, it is customary to base the test not on the hypothesized sampling distribution of $\bar{X}$ shown in Figure 10.2, but on its standardized counterpart, the hypothesized sampling distribution of z shown in **Figure 10.3.** Now z represents a variation on the familiar standard score, and it displays all of the properties of standard scores described in Chapter 5. Furthermore, like the sampling distribution of $\bar{X}$, the **sampling distribution of *z*** *represents the distribution of z values that would be obtained if a value of z were calculated for each sample mean for all possible random samples of a given size from some population* .

Sampling Distribution of z
The distribution of z values that would be obtained if a value of z were calculated for each sample mean for all possible random samples of a given size from some population.

The conversion from $\bar{X}$ to z yields a distribution that approximates the standard normal curve in Table A of Appendix C, since, as indicated in Figure 10.3, the original hypothesized population mean (500) emerges as a z score of 0 and the original standard error of the mean (11) emerges as a z score of 1. The shift from $\bar{X}$ to z eliminates the original units of measurement and standardizes the hypothesis test across all situations without, however, affecting the test results.

Reminder: Converting a Raw Score to *z*

To convert a raw score into a standard score (also described in Chapter 5), express the raw score as a distance from its mean (by subtracting the mean from the raw score), and then split this distance into standard deviation units (by dividing with the standard deviation). Expressing this definition as a word formula, we have

$$Standard\ score = \frac{raw\ score - mean}{standard\ deviation}$$

in which, of course, the standard score indicates the deviation of the raw score in standard deviation units, above or below the mean.

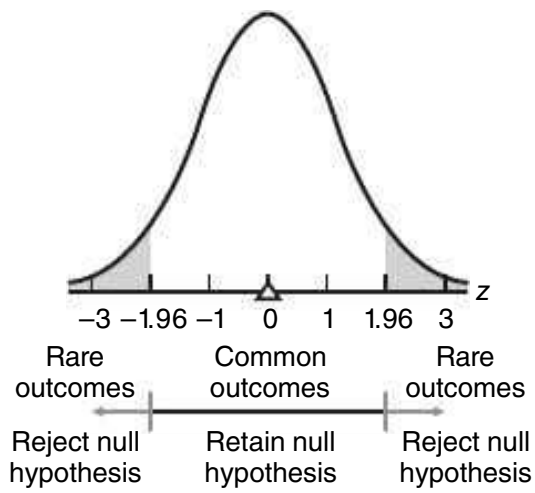


FIGURE 10.3
Common and rare outcomes (values of z).

Converting a Sample Mean to *z*

The z for the present situation emerges as a slight variation of this word formula: Replace the *raw score* with the one observed sample mean $\bar{X}$; replace the *mean* with the mean of the sampling distribution, that is, the hypothesized population mean μ_{hyp}; and replace the *standard deviation* with the standard error of the mean $\sigma_{\bar{x}}$. Now

z RATIO FOR A SINGLE POPULATION MEAN

$$z = \frac{\bar{X} - \mu_{hyp}}{\sigma_{\bar{x}}} \qquad (10.1)$$

where z indicates the deviation of the observed sample mean in standard error units, above or below the hypothesized population mean.

To test the hypothesis for SAT scores, we must determine the value of z from Formula 10.1. Given a sample mean of 533, a hypothesized population mean of 500, and a standard error of 11, we find

$$z = \frac{533 - 500}{11} = \frac{33}{11} = 3$$

The observed z of 3 exceeds the value of 1.96 specified in the hypothesized sampling distribution in Figure 10.3. Thus, the observed z qualifies as a rare outcome under the null hypothesis, and the null hypothesis is rejected. The results of this test with z are the same as those for the original hypothesis test with $\bar{X}$.

Assumptions of *z* Test

***z* Test for a Population Mean**
A hypothesis test that evaluates how far the observed sample mean deviates, in standard error units, from the hypothesized population mean.

When a hypothesis test evaluates how far the observed sample mean deviates, in standard error units, from the hypothesized population mean, as in the present example, *it is referred to as a z test* or, more accurately, as a ***z* test for a population mean.** This z test is accurate only when (1) the population is normally distributed or the sample size is large enough to satisfy the requirements of the central limit theorem and (2) the population standard deviation is known. In the present example, the z test is appropriate because the sample size of 100 is large enough to satisfy the central limit theorem and the population standard deviation is known to be 110.

Progress Check *10.1 Calculate the value of the z test for each of the following situations:

(a) $\bar{X} = 566; \sigma = 30; n = 36; \mu_{hyp} = 560$

(b) $\bar{X} = 24; \sigma = 4; n = 64; \mu_{hyp} = 25$

(c) $\bar{X} = 82; \sigma = 14; n = 49; \mu_{hyp} = 75$

(d) $\bar{X} = 136; \sigma = 15; n = 25; \mu_{hyp} = 146$

Answers on page 432.

10.3 STEP-BY-STEP PROCEDURE

Having been exposed to some of the more important features of hypothesis testing, let's take a detailed look at the test for SAT scores. The test procedure lends itself to a step-by-step description, beginning with a brief statement of the problem that inspired

the test and ending with an interpretation of the test results. The following box summarizes the step-by-step procedure for the current hypothesis test. Whenever appropriate, this format will be used in the remainder of the book. Refer to it while reading the remainder of the chapter.

10.4 STATEMENT OF THE RESEARCH PROBLEM

The formulation of a research problem often represents the most crucial and exciting phase of an investigation. Indeed, the mark of a skillful investigator is to focus on an important research problem that can be answered. Do children from broken families score lower on tests of personal adjustment? Do aggressive TV cartoons incite more disruptive behavior in preschool children? Does profit sharing increase the productivity of employees? Because of our emphasis on hypothesis testing, research problems appear in this book as finished products, usually in the first one or two sentences of a new example.

HYPOTHESIS TEST SUMMARY: *z* TEST FOR A POPULATION MEAN (SAT SCORES)

Research Problem

Does the mean SAT math score for all local freshmen differ from the national average of 500?

Statistical Hypotheses

$$H_0 : \mu = 500$$
$$H_1 : \mu \neq 500$$

Decision Rule

Reject H_0 at the .05 level of significance if $z \geq 1.96$ or if $z \leq -1.96$.

Calculations

Given

$$\bar{X} = 533; \mu_{\text{hyp}} = 500; \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{110}{\sqrt{100}} = 11$$

$$z = \frac{533 - 500}{11} = 3$$

Decision

Reject H_0 at the .05 level of significance because $z = 3$ exceeds 1.96.

Interpretation

The mean SAT math score for all local freshmen does not equal—it exceeds—the national average of 500.

10.5 NULL HYPOTHESIS (H_0)

Once the problem has been described, it must be translated into a statistical hypothesis regarding some population characteristic. Abbreviated as H_0, the null hypothesis becomes the focal point for the entire test procedure (even though we usually hope to reject it). In the test with SAT scores, the null hypothesis asserts that, with respect to the national average of 500, nothing special is happening to the mean score for the local population of freshmen. An equivalent statement, in symbols, reads:

$$H_0 : \mu = 500$$

where H_0 represents the null hypothesis and μ is the population mean for the local freshman class.

Null Hypothesis (H_0)
A statistical hypothesis that usually asserts that nothing special is happening with respect to some characteristic of the underlying population.

Generally speaking, the **null hypothesis (H_0)** *is a statistical hypothesis that usually asserts that nothing special is happening with respect to some characteristic of the underlying population*. Because the hypothesis testing procedure requires that the hypothesized sampling distribution of the mean be centered about a single number (500), the null hypothesis equals a single number (H_0: $\mu = 500$). Furthermore, the null hypothesis always makes a precise statement about a characteristic of the population, never about a sample. Remember, the purpose of a hypothesis test is to determine whether a particular outcome, such as an observed sample mean, could have reasonably originated from a population with the hypothesized characteristic.

Finding the Single Number for H_0

The single number actually used in H_0 varies from problem to problem. Even for a given problem, this number could originate from any of several sources. For instance, it could be based on available information about some relevant population other than the target population, as in the present example in which 500 reflects the mean SAT math scores for all college-bound students during a recent year. It also could be based on some existing standard or theory—for example, that the mean math score for the current population of local freshmen should equal 540 because that happens to be the mean score achieved by all local freshmen during recent years.

If, as sometimes happens, it's impossible to identify a meaningful null hypothesis, don't try to salvage the situation with arbitrary numbers. Instead, use another entirely different technique, known as *estimation,* which is described in Chapter 12.

10.6 ALTERNATIVE HYPOTHESIS (H_1)

Alternative Hypothesis (H_1)
The opposite of the null hypothesis.

In the present example, the alternative hypothesis asserts that, with respect to the national average of 500, something special is happening to the mean math score for the local population of freshmen (because the mean for the local population doesn't equal the national average of 500). An equivalent statement, in symbols, reads:

$$H_1 : \mu \neq 500$$

where H_1 represents the alternative hypothesis, μ is the population mean for the local freshman class, and $\neq$ signifies, "is not equal to."

The **alternative hypothesis (H_1)** *asserts the opposite of the null hypothesis.* A decision to retain the null hypothesis implies a lack of support for the alternative hypothesis, and a decision to reject the null hypothesis implies support for the alternative hypothesis.

As will be described in the next chapter, the alternative hypothesis may assume any one of three different forms, depending on the perspective of the investigator. In its present form, H_1 specifies a *range* of possible values about the *single* number (500) that appears in H_0.

Research Hypothesis
Usually identified with the alternative hypothesis, this is the informal hypothesis or hunch that inspires the entire investigation.

Regardless of its form, H_1 usually is identified with the **research hypothesis,** *the informal hypothesis or hunch that, by implying the presence of something special in the underlying population, serves as inspiration for the entire investigation* . "Something special" might be, as in the current example, a deviation from a national average, or it could be, as in later chapters, a deviation from some control condition produced by a new teaching method, a weight-reduction diet, or a self-improvement workshop. In any event, it is this research hypothesis—and certainly not the null hypothesis—that supplies the motive behind an investigation.

Progress Check *10.2 Indicate what's wrong with each of the following statistical hypotheses:

(a) $H_0: \mu = 155$
$H_1: \mu \neq 160$

(b) $H_0: \overline{X} = 241$
$H_1: \overline{X} \neq 241$

Progress Check *10.3 First using words, then symbols, identify the null hypothesis for each of the following situations. (Don't concern yourself about the precise form of the alternative hypothesis at this point.)

(a) A school administrator wishes to determine whether sixth-grade boys in her school district differ, on average, from the national norms of 10.2 pushups for sixth-grade boys.

(b) A consumer group investigates whether, on average, the true weights of packages of ground beef sold by a large supermarket chain differ from the specified 16 ounces.

(c) A marriage counselor wishes to determine whether, during a standard conflict-resolution session, his clients differ, on average, from the 11 verbal interruptions reported for "well-adjusted couples."

Answers on page 432.

10.7 DECISION RULE

Decision Rule
Specifies precisely when H_0 should be rejected (because the observed z qualifies as a rare outcome).

A **decision rule** *specifies precisely when H_0 should be rejected* (because the observed z qualifies as a rare outcome). There are many possible decision rules, as will be seen in Section 11.3. A very common one, already introduced in Figure 10.3, specifies that H_0 should be *rejected* if the observed z equals or is more positive than 1.96 or if the observed z equals or is more negative than –1.96. Conversely, H_0 should be *retained* if the observed z falls between ± 1.96.

Critical *z* Scores

Figure 10.4 indicates that z scores of ± 1.96 define the boundaries for the middle .95 of the total area (1.00) under the hypothesized sampling distribution for z. Derived from the normal curve table, as you can verify by checking Table A in Appendix C, these two z scores *separate common from rare outcomes and hence dictate whether H_0 should be retained or rejected*. Because of their vital role in the decision about H_0, these scores are referred to as **critical *z* scores.**

Critical z Score
A z score that separates common from rare outcomes and hence dictates whether H_0 should be retained or rejected.

Level of Significance (α)

Figure 10.4 also indicates the proportion (.025 + .025 = .05) of the total area that is identified with rare outcomes. Often referred to as the level of significance of the statistical test, this proportion is symbolized by the Greek letter α (alpha) and discussed more thoroughly in Section 11.4. In the present example, the level of significance, α, equals .05.

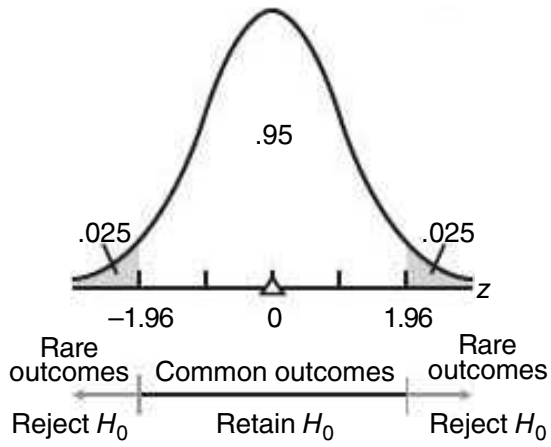


FIGURE 10.4
Proportions of area associated with common and rare outcomes ($\alpha = 05$).

Level of Significance (α)
The degree of rarity required of an observed outcome in order to reject the null hypothesis (H_0).

The **level of significance (α)** *indicates the degree of rarity required of an observed outcome in order to reject the null hypothesis* (H_0). For instance, the .05 level of significance indicates that H_0 should be rejected if the observed z could have occurred just by chance with a probability of only .05 (one chance out of twenty) *or less*.

10.8 CALCULATIONS

We can use information from the sample to calculate a value for z. As has been noted previously, use Formula 10.1 to convert the observed sample mean of 533 into a z of 3.

10.9 DECISION

Either retain or reject H_0, depending on the location of the observed z value relative to the critical z values specified in the decision rule. According to the present rule, H_0 should be rejected at the .05 level of significance because the observed z of 3 exceeds the critical z of 1.96 and, therefore, qualifies as a rare outcome, that is, an unlikely outcome from a population centered about the null hypothesis.

Retain or Reject H_0?

If you are ever confused about whether to retain or reject H_0, recall the logic behind the hypothesis test. You want to reject H_0 only if the observed value of z qualifies as a rare outcome because it deviates too far into the tails of the sampling distribution. Therefore, you want to reject H_0 only if the observed value of z equals or is more positive than the upper critical z (1.96) or if it equals or is more negative than the lower critical z (–1.96). Before deciding, you might find it helpful to sketch the hypothesized sampling distribution, along with its critical z values and shaded rejection regions, and then use some mark, such as an arrow (↑), to designate the location of the observed value of z (3) along the z scale. If this mark is located in the shaded rejection region—or farther out than this region, as in Figure 10.4—then H_0 should be rejected.

Progress Check *10.4 For each of the following situations, indicate whether H_0 should be retained or rejected and justify your answer by specifying the precise relationship between observed and critical z scores. Should H_0 be retained or rejected, given a hypothesis test with critical z scores of $\pm$ 1.96 and

(a) $z = 1.74$ **(b)** $z = 0.13$ **(c)** $z = -2.51$

Answers on page 432.

10.10 INTERPRETATION

Finally, interpret the decision in terms of the original research problem. In the present example, it can be concluded that, since the null hypothesis was rejected, the mean SAT math score for the local freshman class probably differs from the national average of 500.

Although not a strict consequence of the present test, a more specific conclusion is possible. Since the sample mean of 533 (or its equivalent z of 3) falls in the *upper* rejection region of the hypothesized sampling distribution, it can be concluded that the population mean SAT math score for all local freshmen probably *exceeds* the national average of 500. By the same token, if the observed sample mean or its equivalent z had fallen in the *lower* rejection region of the hypothesized sampling distribution, it could have been concluded that the population mean for all local freshmen probably is *below* the national average.

If the observed sample mean or its equivalent z had fallen in the retention region of the hypothesized sampling distribution, it would have been concluded (somewhat weakly, as discussed in Section 11.2) that there is no evidence that the population mean for all local freshmen differs from the national average of 500.

Progress Check *10.5 According to the American Psychological Association, members with a doctorate and a full-time teaching appointment earn, on the average, $82,500 per year, with a standard deviation of $6,000. An investigator wishes to determine whether $82,500 is also the mean salary for all female members with a doctorate and a full-time teaching appointment. Salaries are obtained for a random sample of 100 women from this population, and the mean salary equals $80,100.

(a) Someone claims that the observed difference between $80,100 and $82,500 is large enough by itself to support the conclusion that female members earn less than male members. Explain why it is important to conduct a hypothesis test.

(b) The investigator wishes to conduct a hypothesis test for what population?

(c) What is the null hypothesis, H_0?

(d) What is the alternative hypothesis, H_1?

(e) Specify the decision rule, using the .05 level of significance.

(f) Calculate the value of z. (Remember to convert the standard deviation to a standard error.)

(g) What is your decision about H_0?

(h) Using words, interpret this decision in terms of the original problem.

Answers on page 433.

Summary

To test a hypothesis about the population mean, a single observed sample mean is viewed within the context of a hypothesized sampling distribution, itself centered about the null-hypothesized population mean. If the sample mean appears to emerge from the dense concentration of possible sample means in the middle of the sampling distribution, it qualifies as a common outcome, and the null hypothesis is retained. On the other hand, if the sample mean appears to emerge from the sparse concentration of possible *sample* means at the extremities of the *sampling* distribution, it qualifies as a rare outcome, and the null hypothesis is rejected.

Hypothesis tests are based not on the sampling distribution of $\overline{X}$ expressed in original units of measurement, but on its standardized counterpart, the sampling distribution of z. Referred to as the z test for a single population mean, this test is appropriate only when (1) the population is normally distributed or the sample size is large enough to satisfy the central limit theorem, and (2) the population standard deviation is known.

When testing a hypothesis, adopt the following step-by-step procedure:

- **State the research problem.** Using words, state the problem to be resolved by the investigation.
- **Identify the statistical hypotheses.** The statistical hypotheses consist of a null hypothesis (H_0) and an alternative (or research) hypothesis (H_1). The null hypothesis supplies the value about which the hypothesized sampling distribution is centered. Depending on the outcome of the hypothesis test, H_0 will either be retained or rejected. Insofar as H_0 implies that nothing special is happening in the underlying population, the investigator usually hopes to reject it in favor of H_1, the research hypothesis. In the present chapter, the statistical hypotheses take the form

$$H_0 : \mu = \text{some number}$$
$$H_1 : \mu \neq \text{some number}$$

(Two other possible forms for statistical hypotheses will be described in Chapter 11.)

- **Specify a decision rule.** This rule indicates precisely when H_0 should be rejected. The exact form of the decision rule depends on a number of factors, to be discussed in Chapter 11. In any event, H_0 is rejected whenever the observed z deviates from 0 as far as, or farther than, the critical z does.

  The level of significance indicates how rare an observed z must be (assuming that H_0 is true) before H_0 can be rejected.

- **Calculate the value of the observed *z*.** Express the one observed sample mean as an observed z, using Formula 10.1.
- **Make a decision.** Either retain or reject H_0 at the specified level of significance, justifying this decision by noting the relationship between observed and critical z scores.
- **Interpret the decision**. Using words, interpret the decision in terms of the original research problem. Rejection of the null hypothesis supports the research hypothesis, while retention of the null hypothesis fails to support the research hypothesis.

Important Terms

Sampling distribution of *z*
Null hypothesis (H_0)
Research hypothesis
Critical *z* score
***z* Test for a population mean**
Alternative hypothesis (H_1)
Decision rule
Level of significance (α)

Key Equations

***z* RATIO**

$$z = \frac{\overline{X} - \mu_{\text{hyp}}}{\sigma_{\overline{x}}}$$

where $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$

REVIEW QUESTIONS

10.6 Calculate the value of the *z* test for each of the following situations.

(a) $\overline{X} = 12; \sigma = 9; n = 25, \mu_{hyp} = 15$

(b) $\overline{X} = 3600; \sigma = 4000; n = 100; \mu_{hyp} = 3500$

(c) $\overline{X} = 0.25; \sigma = 010; n = 36; \mu_{hyp} = 0.22$

10.7 Given critical *z* scores of ±1.96, should H_0 be accepted or rejected for each of the *z* scores calculated in Exercise 10.6?

***10.8** For the population at large, the Wechsler Adult Intelligence Scale is designed to yield a normal distribution of test scores with a mean of 100 and a standard deviation of 15. School district officials wonder whether, on the average, an IQ score different from 100 describes the intellectual aptitudes of all students in their district. Wechsler IQ scores are obtained for a random sample of 25 of their students, and the mean IQ is found to equal 105. Using the step-by-step procedure described in this chapter, test the null hypothesis at the .05 level of significance.

Answers on page 433.

10.9 The normal range for a widely accepted measure of body size, the body mass index (BMI), ranges from 18.5 to 25. Using the midrange BMI score of 21.75 as the null hypothesized value for the population mean, test this hypothesis at the .01 level of significance given a random sample of 30 weight-watcher participants who show a mean BMI = 22.2 and a standard deviation of 3.1.

10.10 Let's assume that, over the years, a paper and pencil test of anxiety yields a mean score of 35 for all incoming college freshmen. We wish to determine whether the scores of a random sample of 20 new freshmen, with a mean of 30 and a standard deviation of 10, can be viewed as coming from this population. Test at the .05 level of significance.

10.11 According to the California Educational Code (http://www.cde.ca.gov/ls/fa/sf/peguidemidhi.asp), students in grades 7 through 12 should receive 400 minutes of physical education every 10 school days. A random sample of 48 students has a mean of 385 minutes and a standard deviation of 53 minutes. Test the hypothesis at the .05 level of significance that the sampled population satisfies the requirement.

10.12 According to a 2009 survey based on the United States census (http://www.census.gov/prod/2011pubs/acs-15.pdf), the daily one-way commute time of U.S. workers averages 25 minutes with, we'll assume, a standard deviation of 13 minutes. An investigator wishes to determine whether the national average describes the mean commute time for all workers in the Chicago area. Commute times are obtained for a random sample of 169 workers from this area, and the mean time is found to be 22.5 minutes. Test the null hypothesis at the .05 level of significance.

10.13 Supply the missing word(s) in the following statements:

If the one observed sample mean can be viewed as a __(a)__ outcome under the hypothesis, H_0 will be __(b)__. Otherwise, if the one observed sample mean can be viewed as a __(c)__ outcome under the hypothesis, H_0 will be __(d)__.

The pair of *z* scores that separates common and rare outcomes is referred to as **(e)** *z* scores. Within the hypothesized sampling distribution, the proportion of area allocated to rare outcomes is referred to as the **(f)** and is symbolized by the Greek letter **(g)** .

When based on the sampling distribution of *z*, the hypothesis test is referred to as a **(h)** test. This test is appropriate if the sample size is sufficiently large to satisfy the **(i)** and if the **(j)** is known.

CHAPTER

11

More about Hypothesis Testing

Preview

Based on the notion of everything that could possibly happen just by chance—in other words, based on the concept of a sampling distribution—hypothesis tests permit us to draw conclusions that go beyond a limited set of actual observations. This chapter describes why rejecting the null hypothesis is stronger than retaining the null hypothesis and why a one-tailed test is more likely than a two-tailed test to detect a false null hypothesis.

We speculate about how the hypothesis test fares if we assume, in turn, that the null hypothesis is true and then that it is false. The two types of incorrect decisions—rejecting a true null hypothesis (a false alarm) or retaining a false null hypothesis (a miss)—can be controlled by our selection of the level of significance and of the sample size.

11.1 WHY HYPOTHESIS TESTS?

There is a crucial link between hypothesis tests and the need of investigators, whether pollsters or researchers, to generalize beyond existing data. If the 100 freshmen in the SAT example of the previous chapter had been not a sample but a *census* of the entire freshman class, there wouldn't have been any need to generalize beyond existing data, and it would have been inappropriate to conduct a hypothesis test. Now, the observed difference between the newly observed population mean of 533 and the national average of 500, by itself, would have been sufficient grounds for concluding that the mean SAT math score for all local freshmen exceeds the national average. Indeed, *any* observed difference in favor of the local freshmen, regardless of the size of the difference, would have supported this conclusion.

If we must generalize beyond the 100 freshmen to a larger local population, as was actually the case, the observed difference between 533 and 500 cannot be interpreted at face value. The basic problem is that the sample mean for a second random sample of 100 freshmen probably would differ, just by chance, from the sample mean of 533 for the first sample. Accordingly, the variability among sample means must be considered when we attempt to decide whether the observed difference between 533 and 500 is real or merely transitory.

Importance of the Standard Error

To evaluate the effect of chance, we use the concept of a sampling distribution, that is, the concept of the sample means for all possible random outcomes. A key element in this concept is the standard error of the mean, a measure of the average amount by which sample means differ, just by chance, from the population mean. Dividing the observed difference (533−500) by the standard error (11) to obtain a value of z (3) locates the original observed difference along a z scale of either common outcomes (reasonably attributable to chance) or rare outcomes (not reasonably attributable to chance). If, when expressed as z, the ratio of the observed difference to the standard error is small enough to be reasonably attributed to chance, we retain H_0. Otherwise, if the ratio of the observed difference to the standard error is too large to be reasonably attributed to chance, as in the SAT example, we reject H_0.

Before generalizing beyond the existing data, we must always measure the effect of chance; that is, we must obtain a value for the standard error . To appreciate the vital role of the standard error in the SAT example, increase its value from 11 to 33 and note that even though the observed difference remains the same (533 −500), we would retain, not reject, H_0 because now z would equal 1 (rather than 3) and be less than the critical z of 1.96.

Possibility of Incorrect Decisions

Having made a decision about the null hypothesis, we never know absolutely whether that decision is correct or incorrect, unless, of course, we survey the entire population. Even if H_0 is true (and, therefore, the hypothesized distribution of z about H_0 also is true), there is a *slight* possibility that, just by chance, the one observed z actually originates from one of the shaded rejection regions of the hypothesized distribution of z, thus causing the true H_0 to be rejected. This type of incorrect decision—rejecting a true H_0—is referred to as a *type I error* or a *false alarm*.

On first impulse, it might seem desirable to abolish the shaded rejection regions in the hypothesized sampling distribution to ensure that a true H_0 never is rejected. A most unfortunate consequence of this strategy, however, is that no H_0, not even a radically false H_0, ever would be rejected. This second type of incorrect decision—retaining a false H_0—is referred to as a *type II error* or a *miss*. Both type I and type II errors are described in more detail later in this chapter.

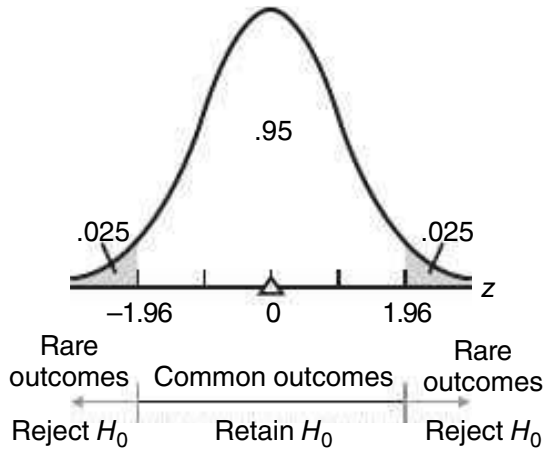


FIGURE 11.1
Proportions of area associated with common and rare outcomes ($\alpha = .05$).

Minimizing Incorrect Decisions

Traditional hypothesis-testing procedures, such as the one illustrated in **Figure 11.1**, tend to minimize both types of incorrect decisions. *If H_0 is true,* there is a high probability that the observed z will qualify as a common outcome under the hypothesized sampling distribution and that the true H_0 will be retained. (In Figure 11.1, this probability equals the proportion of white area (.95) in the hypothesized sampling distribution.) On the other hand, *if H_0 is seriously false,* because the hypothesized population mean differs considerably from the true population mean, there is also a high probability that the observed z will qualify as a rare outcome under the hypothesized distribution and that the false H_0 will be rejected. (In Figure 11.1, this probability can't be determined since, in this case, the hypothesized sampling distribution does not actually reflect the true sampling distribution. More about this later in the chapter.)

Even though we never really know whether a particular decision is correct or incorrect, it is reassuring that in the long run, *most* decisions will be correct—assuming the null hypotheses are *either true or seriously false.*

11.2 STRONG OR WEAK DECISIONS

Retaining H_0 Is a *Weak* Decision

There are subtle but important differences in the interpretation of decisions to retain H_0 and to reject H_0. H_0 is retained whenever the observed z qualifies as a common outcome on the assumption that H_0 is true. Therefore, H_0 *could* be true. However, the same observed result also would qualify as a common outcome when the original value in H_0 (500) is replaced with a slightly different value. Thus, the retention of H_0 must be viewed as a relatively weak decision. Because of this weakness, many statisticians prefer to describe this decision as simply a *failure to reject* H_0 rather than as the retention of H_0. In any event, the retention of H_0 can't be interpreted as proving H_0 to be true. If H_0 had been retained in the present example, it would have been appropriate to conclude not that the mean SAT math score for all local freshmen equals the national average, but that the mean SAT math score *could* equal the national average, as well as many other possible values in the general vicinity of the national average.

Rejecting H_0 Is a *Strong* Decision

On the other hand, H_0 is rejected whenever the observed z qualifies as a rare outcome—one that could have occurred just by chance with a probability of .05 or less—on the assumption that H_0 is true. This suspiciously rare outcome implies that H_0 is probably false (and conversely, that H_1 is probably true). Therefore, the rejection of H_0 can be viewed as a strong decision. When H_0 was rejected in the present example, it was appropriate to report a definitive conclusion that the mean SAT math score for all local freshmen probably exceeds the national average.
To summarize,

the decision to retain H_0 implies not that H_0 is probably true, but only that H_0 *could* be true, whereas the decision to reject H_0 implies that H_0 is *probably* false (and that H_1 is *probably* true).

Since most investigators hope to reject H_0 in favor of H_1, the relative weakness of the decision to retain H_0 usually does not pose a serious problem.

Why the Research Hypothesis Isn't Tested Directly

Even though H_0, the null hypothesis, is the focus of a statistical test, it is usually of secondary concern to the investigator. Nevertheless, there are several reasons why, although of primary concern, the research hypothesis is identified with H_1 and tested indirectly.

Lacks Necessary Precision

The research hypothesis, but not the null hypothesis, lacks the necessary precision to be tested directly.

To be tested, a hypothesis must specify a single number about which the hypothesized sampling distribution can be constructed. *Because it specifies a single number, the null hypothesis, rather than the research hypothesis, is tested directly.* In the SAT example, the null hypothesis specifies that a precise value (the national average of 500) describes the mean for the current population of interest (all local freshmen). Typically, the research hypothesis lacks the required precision. It merely specifies that some inequality exists between the hypothesized value (500) and the mean for the current population of interest (all local freshmen).

Supported by a Strong Decision to Reject

Logical considerations also argue for the indirect testing of the research hypothesis and the direct testing of the null hypothesis.

Because the research hypothesis is identified with the alternative hypothesis, the decision to reject the null hypothesis, should it be made, will provide *strong* support for the research hypothesis, while the decision to retain the null hypothesis, should it be made, will provide, at most, *weak* support for the null hypothesis.

As mentioned, the decision to reject the null hypothesis is stronger than the decision to retain it. Logically, a statement such as "All cows have four legs" can never be proven in spite of a steady stream of positive instances. It only takes one negative instance—one cow with three legs—to disprove the statement. By the same token, one positive instance (common outcome) doesn't prove the null hypothesis, but one

negative instance (rare outcome) disproves the null hypothesis. (Strictly speaking, however, since a rare outcome implies that the null hypothesis is probably *but not definitely* false, remember that there always is a very small possibility that the rare outcome reflects a true null hypothesis.)

Reminder:
Rejecting H_0 implies that it probably is false, while retaining H_0 implies only that it might be true.

Logically, therefore, it makes sense to identify the research hypothesis with the alternative hypothesis. If, as hoped, the data favor the research hypothesis, the test will generate strong support for your hunch: It's *probably* true. If the data do not favor the research hypothesis, the hypothesis test will generate, at most, weak support for the null hypothesis: It *could* be true. *Weak support for the null hypothesis is of little consequence, as this hypothesis—that nothing special is happening in the population—usually serves only as a convenient testing device.*

11.3 ONE-TAILED AND TWO-TAILED TESTS

Let's consider some techniques that make the hypothesis test more responsive to special conditions.

Two-Tailed Test

Generally, the alternative hypothesis, H_1, is the complement of the null hypothesis, H_0. Under typical conditions, the form of H_1 resembles that shown for the SAT example, namely,

$$H_1 : \mu \neq 500$$

Two-Tailed or Nondirectional Test
Rejection regions are located in both tails of the sampling distribution.

This alternative hypothesis says that the null hypothesis should be rejected if the mean reading score for the population of local freshmen differs in either direction from the national average of 500. An observed z will qualify as a rare outcome if it deviates too far either below or above the national average. Panel A of **Figure 11.2** shows rejection regions that are associated with both tails of the hypothesized sampling distribution. The corresponding decision rule, with its pair of critical z scores of ±1.96, is referred to as a **two-tailed** or **nondirectional test**.

One-Tailed Test (Lower Tail Critical)

Now let's assume that the research hypothesis for the investigation of SAT math scores was based on complaints from instructors about the poor preparation of local freshmen. Assume also that if the investigation supports these complaints, a remedial program will be instituted. Under these circumstances, the investigator might prefer a hypothesis test that is specially designed to detect only whether the population mean math score for all local freshmen is *less* than the national average.

This alternative hypothesis reads:

$$H_1 : \mu \leq 500$$

It reflects a concern that the null hypothesis should be rejected only if the population mean math score for all local freshmen is less than the national average of 500. Accordingly, an observed z triggers the decision to reject H_0 only if z deviates too far below the national average. Panel B of Figure 11.2 illustrates a rejection region that is associated with only the lower tail of the hypothesized sampling distribution. The corresponding decision rule, with its critical z of –1.65, is referred to as a **one-tailed** or **directional test** *with the lower tail critical.* Use Table A in Appendix C to verify that if the critical z equals –1.65; then .05 of the total area under the distribution of z has been allocated to the lower rejection region. Notice that the level of significance, α, equals .05 for this one-tailed test and also for the original two-tailed test.

One-Tailed or Directional Test
Rejection region is located in just one tail of the sampling distribution.

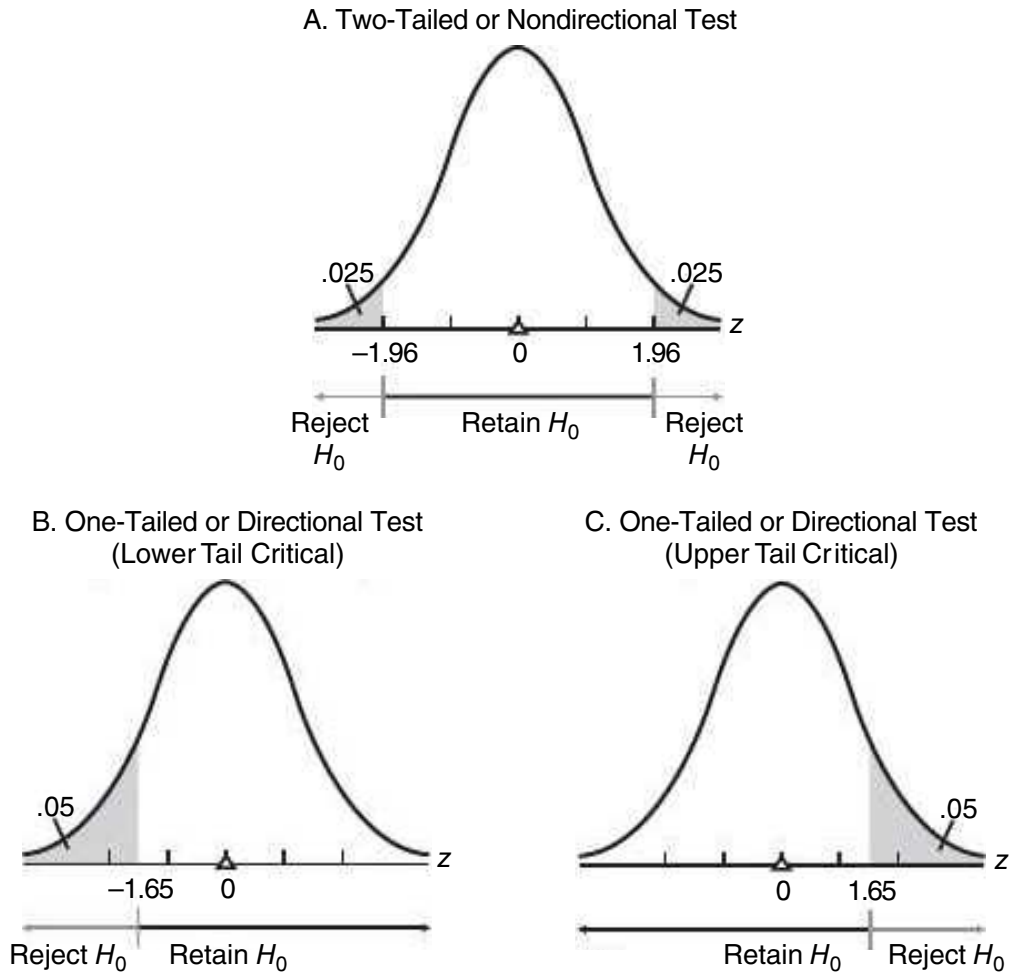


FIGURE 11.2
Three different types of tests ($\alpha = .05$).

Extra Sensitivity of One-Tailed Tests

This new one-tailed test is extra sensitive to any drop in the population mean for the local freshmen below the national average. If H_0 is false because a drop has occurred, then the observed z will be more likely to deviate below the national average. As can be seen in panels A and B of Figure 11.2, an observed deviation in the direction of concern—below the national average—is more likely to penetrate the broader rejection region for the one-tailed test than that for the two-tailed test. Therefore, the decision to reject a *false* H_0 (in favor of the research hypothesis) is more likely to occur in the one-tailed test than in the two-tailed test.

One-Tailed Test (Upper Tail Critical)

Panel C of Figure 11.2 illustrates a **one-tailed** or **directional test** *with the upper tail critical.* This one-tailed test is the mirror image of the previous test. Now the alternative hypothesis reads:

$$H_1 : \mu > 500$$

and its critical z equals 1.65. This test is specially designed to detect only whether the population mean math score for all local freshmen *exceeds* the national average. For example, the research hypothesis for this investigation might have been inspired by the possibility of eliminating an existing remedial math program if it can be demonstrated that, on the average, the SAT math scores of all local freshmen exceed the national average.

One or Two Tails?

Before a hypothesis test, if there is a concern that the true population mean differs from the hypothesized population mean *only* in a particular direction, use the appropriate one-tailed or directional test for extra sensitivity. Otherwise, use the more customary two-tailed or nondirectional test.

Having committed yourself to a one-tailed test with its single rejection region, you must retain H_0, regardless of how far the observed z deviates from the hypothesized population mean in the direction of "no concern." For instance, if a one-tailed test with the lower tail critical had been used with the data for 100 freshmen from the SAT example, H_0 would have been retained because, even though the observed z equals an impressive value of 3, it deviates in the direction of no concern—in this case, above the national average. Clearly, a one-tailed test should be adopted only when there is absolutely no concern about deviations, even very large deviations, in one direction. If there is the slightest concern about these deviations, use a two-tailed test.

The selection of a one- or two-tailed test should be made before the data are collected. Never "peek" at the value of the observed z to determine whether to locate the rejection region for a one-tailed test in the upper or the lower tail of the distribution of z. To qualify as a one-tailed test, the location of the rejection region must reflect the investigator's concern only about deviations in a particular direction *before any inspection of the data*. Indeed, the investigator should be able to muster a compelling reason, based on an understanding of the research hypothesis, to support the direction of the one-tailed test.

New Null Hypothesis for One-Tailed Tests

When tests are one-tailed, a complete statement of the null hypothesis also should include all possible values of the population mean in the direction of no concern. For example, given a one-tailed test with the lower tail critical, such as H_1: $\mu < 500$, the complete null hypothesis should be stated as H_0: $\mu \geq 500$ instead of H_0: $\mu = 500$. By the same token, given a one-tailed test with the upper tail critical, such as H_1: $\mu > 500$, the complete null hypothesis should be stated as H_0: $\mu \leq 500$.

If you think about it, the complete H_0 describes all of the population means that could be true if a one-tailed test results in the retention of the null hypothesis. For instance, if a one-tailed test with the lower tail critical results in the retention of H_0: $\mu \geq 500$, the complete H_0 accurately reflects the fact that not only $\mu = 500$ could be true, but also that any other value of the population mean in the direction of no concern, that is, $\mu > 500$, could be true. (Remember, when the test is one-tailed, even a very deviant result in the direction of no concern—possibly reflecting a mean much larger than 500—still would trigger the decision to retain H_0.) Henceforth, whenever a one-tailed test is employed, write H_0 to include values of the population mean in the direction of no concern—*even though the single number in the complete H_0 identified by the equality sign is the one value about which the hypothesized sampling distribution is centered and, therefore, the one value actually used in the hypothesis test.*

Reminder:

In the absence of compelling reasons for a one-tailed test, use a two-tailed test.

Progress Check *11.1 Each of the following statements could represent the point of departure for a hypothesis test. Given only the information in each statement, would you use a two-tailed (or nondirectional) test, a one-tailed (or directional) test with the lower tail critical, or a one-tailed (or directional) test with the upper tail critical? Indicate your decision by specifying the appropriate H_0 and H_1. Furthermore, whenever you conclude that the test is one-tailed, indicate the precise word (or words) in the statement that justifies the one-tailed test.

(a) An investigator wishes to determine whether, for a sample of drug addicts, the mean score on the depression scale of a personality test differs from a score of 60, which, according to the test documentation, represents the mean score for the general population.

(b) To increase rainfall, extensive cloud-seeding experiments are to be conducted, and the results are to be compared with a baseline figure of 0.54 inch of rainfall (for comparable periods when cloud seeding was not done).

(c) Public health statistics indicate, we will assume, that American males gain an average of 23 lbs during the 20-year period after age 40. An ambitious weight-reduction program, spanning 20 years, is being tested with a sample of 40-year-old men.

(d) When untreated during their lifetimes, cancer-susceptible mice have an average life span of 134 days. To determine the effects of a potentially life-prolonging (and cancer-retarding) drug, the average life span is determined for a group of mice that receives this drug.

Progress Check *11.2 For each of the following situations, indicate whether H_0 should be retained or rejected.

Given a one-tailed test, lower tail critical with $\alpha = .01$, and

(a) $z = -2.34$ **(b)** $z = -5.13$ **(c)** $z = 4.04$

Given a one-tailed test, upper tail critical with $\alpha = .05$, and

(d) $z = 2.00$ **(e)** $z = -1.80$ **(f)** $z = 1.61$

Answers on pages 433 and 434.

11.4 CHOOSING A LEVEL OF SIGNIFICANCE (α)

The level of significance indicates how rare an observed z must be before H_0 can be rejected. To reject H_0 at the .05 level of significance implies that the observed z would have occurred, just by chance, with a probability of only .05 (one chance out of twenty) *or less*.

The level of significance also spotlights an inherent risk in hypothesis testing, that is, the risk of rejecting a true H_0. When the level of significance equals .05, there is a probability of .05 that, even though H_0 is true, the observed z will stray into the rejection region and cause the true H_0 to be rejected.

Which Level of Significance?

When the rejection of a true H_0 is particularly serious, a smaller level of significance can be selected. For example, the .01 level of significance implies that before H_0 can be rejected, the observed z must achieve a degree of rarity equal to .01 (one chance out of one hundred) *or less;* it also limits, to a probability of .01, the risk of rejecting a true H_0. The .01 level might be used in a hypothesis test in which the rejection of a true H_0 would cause the introduction of a costly new remedial education program, even though the population mean math score for all local freshmen really equals the national average. An even smaller level of significance, such as the .001 level, might be used when the rejection of a true H_0 would have horrendous consequences—for instance, the treatment of serious illnesses, such as AIDS, exclusively with a new, very expensive drug that not only is worthless but also has severe side effects.

Although many different levels of significance are possible, most tables for hypothesis tests are geared to the .05 and .01 levels. In this book, the level of significance will be specified for you. However, in real-life applications, you, as an investigator, might

Table 11.1
CRITICAL *z* VALUES

TYPE OF TEST	LEVEL OF SIGNIFICANCE (α) .05	.01
Two-tailed or nondirectional test (H_0: $\mu =$ some number) (H_1: $\mu \neq$ some number)	±1.96	±2.58
One-tailed or directional test, lower tail critical (H_0: $\mu \geq$ some number) (H_1: $\mu <$ some number)	–1.65	–2.33
One-tailed or directional test, upper tail critical (H_0: $\mu \leq$ some number) (H_1: $\mu >$ some number)	+1.65	+2.33

have to select a level of significance. *Unless there are obvious reasons for selecting either a larger or a smaller level of significance, use the customary .05 level* —the largest level of significance reported in most professional journals.

When testing hypotheses with the z test, you may find it helpful to refer to **Table 11.1**, which lists the critical z values for one- and two-tailed tests at the .05 and .01 levels of significance. These z values were obtained from Table A in Appendix C.

Progress Check *11.3 Specify the decision rule for each of the following situations (referring to Table 11.1 to find critical z values):

(a) a two-tailed test with $\alpha = .05$

(b) a one-tailed test, upper tail critical, with $\alpha = .01$

(c) a one-tailed test, lower tail critical, with $\alpha = .05$

(d) a two-tailed test with $\alpha = .01$

Answers on page 434.

11.5 TESTING A HYPOTHESIS ABOUT VITAMIN C

Let's look more closely at the four possible outcomes of a hypothesis test by focusing on a study to determine whether vitamin C increases the intellectual aptitude of high school students. After being randomly selected from some large school district, each of 36 students takes a daily dose of 90 milligrams of vitamin C for a period of two months before being tested for IQ.

Ordinarily, IQ scores for all students in this school district approximate a normal distribution with a mean of 100 and a standard deviation of 15. According to the null hypothesis, a mean of 100 still would describe the distribution of IQ scores even if all of the students in the district were to receive the vitamin C treatment. Furthermore, given our exclusive concern about detecting only any deviation of the population mean *above* 100, the null hypothesis takes the form appropriate for a one-tailed test with the upper tail critical, namely:

$$H_0 : \mu \leq 100$$

The rejection of H_0 would support H_1, the research hypothesis that something special is happening in the underlying population (because vitamin C increases intellectual aptitude), namely:

$$H_0 : \mu > 100$$

z Test Is Appropriate

To determine whether the sample mean IQ for the 36 students qualifies as a common or a rare outcome under the null hypothesis, a z test will be used. The z test for a population mean is appropriate since, for IQ scores, the population standard deviation is known to be 15 and the shape of the population is known to be normal.

Two Groups Would Have Been Better

Although poorly designed, the present experiment supplies a perspective that will be most useful in later chapters. A better-designed experiment would contrast the IQ scores for the group of subjects who receive vitamin C with the IQ scores for a *placebo control group* of subjects who receive fake vitamin C—thereby controlling for the "*placebo effect*," *a self-induced improvement in performance caused by the subject's awareness of being treated in a special way.* Hypothesis tests for experiments with two groups are described in Chapters 14 and 15.

The box on page 205 summarizes those features of the hypothesis test that can be identified before the collection of any data.

11.6 FOUR POSSIBLE OUTCOMES

Table 11.2 summarizes the four possible outcomes of any hypothesis test. Before testing a hypothesis, we must be concerned about all four possible outcomes because we don't know whether H_0 is true or false—that's why we're testing the hypothesis. If, unknown to us, H_0 really is true, a well-designed hypothesis test will tend to confirm this fact; that is, it will cause us to retain H_0 and conclude that H_0 could be true. To conclude otherwise, as is always a slight possibility, reflects a type I error. On the other hand, if, unknown to us, H_0 really is *seriously* false, a well-designed hypothesis test also will tend to confirm this fact; that is, it will cause us to reject H_0 and conclude that H_0 is false. To conclude otherwise, as is always a slight possibility, reflects a type II error.

Four Possible Outcomes of the Vitamin C Experiment

It's instructive to describe the four possible outcomes in Table 11.2 in terms of the vitamin C experiment.

Table 11.2
POSSIBLE OUTCOMES OF A HYPOTHESIS TEST

	STATUS OF H_0	
DECISION	**TRUE H_0**	**FALSE H_0**
Retain H_0	(1) Correct decision	(3) Type II error (miss)
Reject H_0	(2) Type I error (false alarm)	(4) Correct decision

HYPOTHESIS TEST SUMMARY: *z* TEST FOR A POPULATION MEAN (*PRIOR* TO THE VITAMIN C EXPERIMENT)

Research Problem

Does the daily ingestion of vitamin C cause an increase, on average, in IQ scores among all students in the school district?

Statistical Hypotheses

$$H_0 : \mu \leq 100$$
$$H_1 : \mu > 100$$

Decision Rule

Reject H_0 at the .05 level of significance if $z \geq 1.65$.

Calculations

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{36}} = \frac{15}{6} = 2.5$$

1. If H_0 really is true (because vitamin C does not cause an increase in the population mean IQ), then *it is a correct decision to retain the true* H_0. In this case, we would conclude correctly that there is no evidence that vitamin C increases IQ.

Type I Error

Rejecting a true null hypothesis.

2. If H_0 really is true, then *it is a* **type I error** *to reject the true* H_0 and conclude that vitamin C increases IQ when, in fact, it doesn't. Type I errors are sometimes called *false alarms* because, as with their firehouse counterparts, they trigger wild goose chases after something that does not exist. For instance, a type I error might encourage a batch of worthless experimental efforts to discover precisely what dosage of vitamin C maximizes the nonexistent "increase" in IQ.

Type II Error

Retaining a false null hypothesis.

3. If H_0 really is false (because vitamin C really causes an increase in the population mean IQ), then *it is a* **type II error** *to retain the false* H_0 and conclude that there is no evidence that vitamin C increases IQ when, in fact, it does. Type II errors are sometimes called *misses* because they fail to detect a potentially important relationship, such as that between vitamin C and IQ.

4. If H_0 really is false, then *it is a correct decision to reject the false* H_0 and conclude that vitamin C increases IQ.

Importance of Null Hypothesis

Refer to Table 11.2 when, as in the following exercise, you must describe the four possible outcomes for a particular hypothesis test. To avoid confusing the type I and II errors, first identify the null hypothesis, H_0. Typically, *the null hypothesis asserts that there is no effect, thereby contradicting the research hypothesis* . In the present case, contrary to the research hypothesis, the null hypothesis (H_0: $\mu \leq 100$) assumes that vitamin C has no positive effect on IQ.

Decisions Usually Are Correct

When generalizing beyond existing observations, there is always the possibility of a type I or type II error, and we never can be absolutely certain of having made the correct decision. At best, we can use a test procedure that *usually* produces a correct decision when H_0 is either true or seriously false. This claim will be examined in the context of the vitamin C experiment, assuming first that H_0 really is true and then that H_0 really is false. Although you might view this approach as hopelessly theoretical, *since we never know whether H_0 really is true or false,* read the next few sections carefully, for they have important implications for any hypothesis test.

Progress Check *11.4

(a) List the four possible outcomes for any hypothesis test.

(b) Under the U.S. Criminal Code, a defendant is presumed innocent until proven guilty. Viewing a criminal trial as a hypothesis test (with H_0 specifying that the defendant is innocent), describe each of the four possible outcomes.

Answers on page 434.

11.7 IF H_0 REALLY IS TRUE

Assume that H_0 really is true because vitamin C doesn't increase the population mean IQ. In this case, we need be concerned only about either retaining or rejecting a true H_0 (the two leftmost outcomes in Table 11.2). It's instructive to view these two possible outcomes in terms of the sampling distribution in **Figure 11.3**. Centered about a value of 100, the hypothesized sampling distribution in Figure 11.3 reflects the properties of the projected one-tailed test for vitamin C. If H_0 really is true—and this is a crucial point—the hypothesized sampling distribution also can be viewed as the *true* sampling distribution (from which the one observed sample mean actually originates). Therefore, the one observed sample mean (or z) in the experiment can be viewed as being randomly selected from the hypothesized distribution.

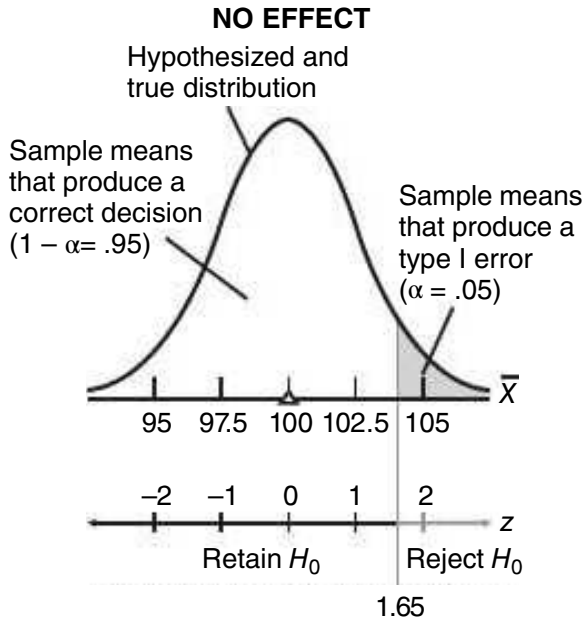


FIGURE 11.3
Hypothesized and true sampling distribution when H_0 is true (because vitamin C causes no increase in IQ).

Probability of a Type I Error

Alpha (α)

The probability of a type I error, that is, the probability of rejecting a true null hypothesis.

When, just by chance, a randomly selected sample mean originates from the small, shaded portion of the sampling distribution in Figure 11.3, its z value equals or exceeds 1.65, and hence H_0 is rejected. Because H_0 really is true, this is an incorrect decision or type I error—a false alarm, announced as evidence that vitamin C increases IQ, even though it really does not. *The probability of a type I error equals* **alpha** (α), the level of significance. (The level of significance, remember, indicates the proportion of the total area of the sampling distribution in the rejection region for H_0.) In the present case, the probability of a type I error equals .05, as indicated in Figure 11.3.

Probability of a Correct Decision

When, just by chance, a randomly selected sample mean originates from the large white portion of the sampling distribution in Figure 11.3, its z value is less than 1.65 and H_0 is retained. Because H_0 really is true, this is a correct decision—announced as a lack of evidence that vitamin C increases IQ. The probability of a correct decision equals $1 - \alpha$, that is, .95.

Reducing the Probability of a Type I Error

If H_0 really is true, the present test will produce a correct decision with a probability of .95 and a type I error with a probability of .05.* If a false alarm has serious consequences, the probability of a type I error can be reduced to .01 or even to .001 simply by using the .01 or .001 level of significance, respectively. One of these levels of significance might be preferred for the vitamin C test if, for instance, a false alarm could cause the adoption of an expensive program to supply worthless vitamin C to all students in the district and, perhaps, the creation of an accelerated curriculum to accommodate the fictitious increase in intellectual aptitude.

True H_0 Usually Retained

Reminder:

If H_0 is true and an error is committed, it must be a type I error.

If H_0 really is true, the probability of a type I error, α, equals the level of significance, and the probability of a correct decision equals $1 - \alpha$.

Because values of .05 or less are usually selected for α, we can conclude that if H_0 really is true, correct decisions will occur much more frequently than will type I errors.

Progress Check *11.5 In order to eliminate the type I error, someone decides to use the .00 level of significance. What's wrong with this procedure?

Answer on page 434.

11.8 IF H_0 REALLY IS FALSE BECAUSE OF A *LARGE* EFFECT

Next, assume that H_0 really is false because vitamin C increases the population mean by not just a few points, but *by many points*—for example, by ten points. Using the vocabulary of most investigators, we also could describe this increase as a "ten-point

*Strictly speaking, if H_0: $\mu \leq 100$ really is true, the true sampling distribution also could be centered about some value less than 100, in the direction of no concern. In this case, the consequences of the hypothesis test would be even more favorable than suggested. Essentially, because the true sampling distribution would be shifted to the left of the one shown in Figure 11.3, while everything else remains the same, the type I error would have a smaller probability than .05, and a correct decision would have a larger probability than .95.

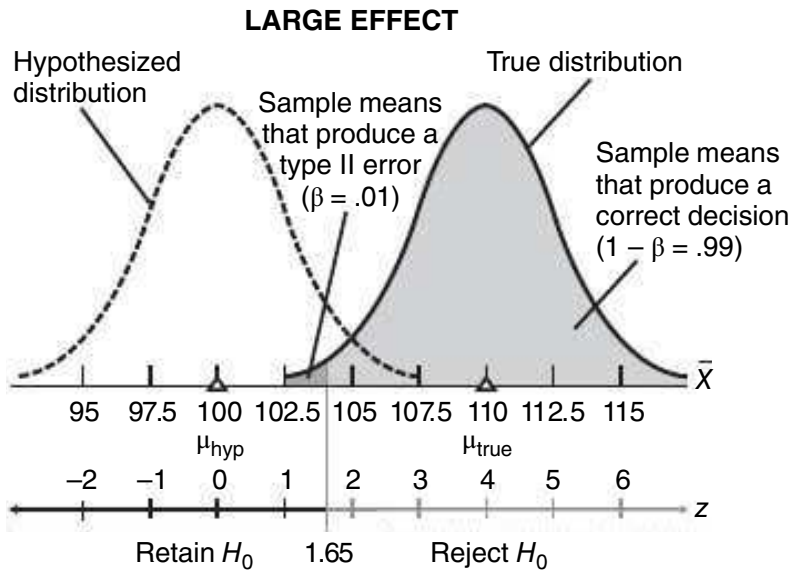


FIGURE 11.4
Hypothesized and true sampling distribution when H_0 is false because of a large effect.

effect," since *any difference between a true and a hypothesized population mean* is referred to as an **effect.** If H_0 really is false, because of the relatively large ten-point effect of vitamin C on IQ, we need be concerned only about either retaining or rejecting a false H_0 (the two rightmost outcomes in Table 11.2). Let's view each of these two possible outcomes in terms of the sampling distributions in **Figure 11.4**.

Effect
Any difference between a true and a hypothesized population mean.

Hypothesized Sampling Distribution

Hypothesized Sampling Distribution
Centered about the hypothesized population mean, this distribution is used to generate the decision rule.

It is essential to distinguish between the *hypothesized* sampling distribution and the *true* sampling distribution shown in **Figure 11.4**. Centered about the hypothesized population mean of 100, the **hypothesized sampling distribution** serves as the parent distribution for the familiar decision rule with a critical z of 1.65 for the projected one-tailed test. Once the decision rule has been identified, attention shifts from the hypothesized sampling distribution to the true sampling distribution.

True Sampling Distribution
Centered about the true population mean, this distribution produces the one observed mean (or z).

True Sampling Distribution

Centered about the true population mean of 110 (which reflects the ten-point effect, that is, 100 + 10 = 110), the **true sampling distribution** serves as the parent distribution for the one randomly selected sample mean (or z) that will be observed in the experiment. Viewed relative to the decision rule (based on the hypothesized sampling distribution), the one randomly selected sample mean (originating from the true sampling distribution) dictates whether we retain or reject the false H_0.

Low Probability of a Type II Error for a *Large* Effect

Beta (β)
The probability of a type II error, that is, the probability of retaining a false null hypothesis.

When, just by chance, a randomly selected sample mean originates from the very small black portion of the true sampling distribution of the mean, its z value is less than 1.65, and therefore, in compliance with the decision rule, H_0 is retained. Because H_0 really is false, this is an incorrect decision or type II error—a miss, announced as a lack of evidence that vitamin C increases IQ, even though, in fact, it does. With the aid of tables for the normal curve, it can be demonstrated that in the present case, *the probability of a type II error,* symbolized by the Greek letter **beta (*β*),** equals .01.

of significance. By the same token, any p-value greater than .05, such as $p > .05$, $p < .10$, $p < .20$, or $p = .18$ implies that, with the same data, H_0 would have been retained at the .05 level of significance.

Progress Check *14.5 Indicate which member of each of the following pairs of p-values describes the *more rare* test result:

(a_1) $p > .05$ **(a_2)** $p < .05$

(b_1) $p < .001$ **(b_2)** $p < .01$

(c_1) $p < .05$ **(c_2)** $p < .01$

(d_1) $p < .10$ **(d_2)** $p < .20$

(e_1) $p = .04$ **(e_2)** $p = .02$

Progress Check *14.6 Treating each of the p-values in the previous exercise separately, indicate those that would cause you to reject the null hypothesis at the .05 level of significance.

Answers on page 438.

14.7 STATISTICALLY SIGNIFICANT RESULTS

It's important that you accurately interpret the findings of others—often reported as "having statistical significance." Tests of hypotheses often are referred to as *tests of significance,* and test results are described as being *statistically significant* (if the null hypothesis has been rejected) or as not being statistically significant (if the null hypothesis has been retained). *Rejecting the null hypothesis* and *statistically significant* both signify that the test result can't be attributed to chance. However, correct usage dictates that *rejecting the null hypothesis* always refers to the population, such as rejecting the hypothesized zero difference between two population means, while *statistically significant* always refers to the sample, such as assigning statistical significance to the observed difference between two sample means. Either phrase can be used. However, assigning *statistical significance* to a population mean difference would be misleading, since a population mean difference equals a fixed value controlled by "nature," not something controlled by the results of a statistical test. *Rejecting* a sample mean difference also would be misleading, since a sample mean difference is an observed result that serves as the basis for statistical tests, not something to be rejected.

Statistical Significance

Implies only that the null hypothesis is probably false, and not whether it's false because of a large or small difference between population means.

Statistical significance doesn't imply that the underlying effect is important. **Statistical significance** between pairs of sample means *implies only that the null hypothesis is probably false, and not whether it's false because of a large or small difference between population means.*

Beware of Excessively Large Sample Sizes

Using excessively large sample sizes can produce statistically significant results that lack importance. For instance, assume a new EPO experiment with the same amount of variability among endurance scores as in the original experiment, that is, with a pooled variance, s_p^2, equal to 16.2 (from Table 14.1). But assume that the new experiment has a *much smaller* mean difference, $\bar{X}_1 - \bar{X}_2$, equal to only 0.50 minutes (instead of 5 minutes in the original experiment) and much *larger* sample sizes each equal to 500 patients (instead of 6). Because of these much larger sample sizes, the new standard error would equal only 0.25 (instead of 2.32) and the new t would equal 2.00. Now we would have rejected the null hypothesis at the .05 level, even though the new

difference between sample means is only one-tenth the size of the original difference. With large sample sizes and, therefore, with a small standard error, even a very small and unimportant *effect* (*difference between population means*) will be detected, and the test will be reported as statistically significant.

Statistical significance merely indicates that an observed effect, such as an observed difference between the sample means, is sufficiently large, relative to the standard error, to be viewed as a rare outcome. (Statistical significance also implies that the observed outcome is *reliable,* that is, it would reappear as a similarly rare outcome in a repeat experiment.) It's very desirable, therefore, that we go beyond reports of statistical significance by estimating the size of the effect and, if possible, judging its importance.

Avoid an Erroneous Conditional Probability

Rejecting H_0 at, for instance, the .05 level of significance, signifies that the probability of the observed, or a more extreme, result is less than or equal to .05 *assuming H_0 is true*. This is a conditional probability that takes the form:

$$\Pr(\text{the observed result, given } H_0 \text{ is true}) \leq .05.$$

The probability of .05 depends entirely on the *assumption* that H_0 is true since that probability of .05 originates from the hypothesized sampling distribution centered about H_0.

This statement often is confused with another enticing but erroneous statement, namely H_0 itself is true with probability .05 or less, that reverses the order of events in the conditional probability. The new, erroneous conditional probability takes the form:

$$\Pr(H_0 \text{ is true, given the observed result}) \leq .05.$$

At issue is the question of what the probability of .05 refers to. Our hypothesis testing procedure only supports the first, not the second conditional probability. Having rejected H_0 at the .05 level of significance, we can conclude, without indicating a specific probability, that H_0 is *probably false,* but we can't reverse the original conditional probability and conclude that it's true with only probability .05 or less. We have not tested the truth of H_0 on the basis of the observed result. To do so goes beyond the scope of our statistical test and makes an unwarranted claim regarding the probability that the null hypothesis actually is true.

14.8 ESTIMATING EFFECT SIZE: POINT ESTIMATES AND CONFIDENCE INTERVALS

It would make sense to estimate the effect for the EPO experiment featured in this chapter since the results are statistically significant. (But, strictly speaking, *only* if the results are statistically significant. Otherwise, we would be estimating an "effect" that could be merely transitory and attributed to chance.)

Point Estimate ($\bar{X}_1 - \bar{X}_2$)

As you probably recall from Chapter 12, a point estimate is the most straightforward type of estimate. It identifies the observed difference for $\bar{X}_1 - \bar{X}_2$, in this case, 5 minutes, as an estimate of the unknown effect, that is, the unknown difference between population means, $\mu_1 - \mu_2$. On average, the treatment patients stay on the treadmill for 11 minutes, which is almost twice as long as the 6 minutes for the control patients. If you think about it, this impressive estimate of effect size isn't surprising. With the very small groups of only 6 patients, we had to create a large, fictitious mean

difference of 5 minutes in order to claim a statistically significant result. If this result had occurred in a real experiment, it would have signified a powerful effect of EPO on endurance that could be detected even with very small samples.

Confidence Interval

Although simple, straightforward, and precise, point estimates tend to be inaccurate because they ignore sampling variability. Confidence intervals do not because, as noted in Chapter 12, they are based on the variability in the sampling distribution of $\overline{X}_1 - \overline{X}_2$. To estimate the range of possible effects of EPO on endurance, a confidence interval can be constructed for the difference between population means, $\mu_1 - \mu_2$.

Confidence intervals for $\mu_1 - \mu_2$ specify ranges of values that, in the long run, include the unknown effect (difference between population means) a certain percent of the time.

Confidence Intervals for $\mu_1 - \mu_2$
Ranges of values that, in the long run, include the unknown effect a certain percent of the time.

Given two independent samples, a confidence interval for $\mu_1 - \mu_2$ can be constructed from the following expression:

CONFIDENCE INTERVAL (CI) FOR $\mu_1 - \mu_2$ (TWO INDEPENDENT SAMPLES)

$$\overline{X}_1 - \overline{X}_2 \pm \left(t_{conf}\right)\left(s_{\overline{X}_1 - \overline{X}_2}\right) \tag{14.4}$$

where $\overline{X}_1 - \overline{X}_2$ represents the difference between sample means; t_{conf} represents a number, distributed with $n_1 + n_2 - 2$ degrees of freedom, from the t tables, which satisfies the confidence specifications; and $s_{\overline{X}_1 - \overline{X}_2}$ represents the estimated standard error defined in Formula 14.3.

To find the appropriate value of t_{conf} in Formula 14.4, refer to Table B in Appendix C and follow essentially the same procedure described earlier. For example, if a 95 percent confidence interval is desired for the EPO experiment, first locate the row corresponding to 10 degrees of freedom (from $df = n_1 + n_2 - 2 = 6 + 6 - 2 = 10$) and then locate the column for the 95 percent level of confidence, that is, the column heading identified with a single asterisk. The intersected cell specifies a value of 2.228 to be entered for t_{conf} in Formula 14.4. Given this value for t_{conf}, and values of 5 for the difference between sample means, $\overline{X}_1 - \overline{X}_2$, and of 2.32 for the estimated standard error, $s_{\overline{X}_1 - \overline{X}_2}$ (from Table 14.1), Formula 14.4 becomes

$$5 \pm (2.228)(2.32) = 5 \pm 5.17 = \begin{cases} 10.17 \\ -0.17 \end{cases}$$

Now it can be claimed, with 95 percent confidence, that the interval between –0.17 minutes and 10.17 minutes includes the true effect size, that is, the true difference between population means for endurance scores.

Interpreting Confidence Intervals for $\mu_1 - \mu_2$

The numbers in this confidence interval refer to *differences* between population means, and the signs are particularly important since they indicate the *direction* of these differences. Otherwise, the interpretation of a confidence interval for $\mu_1 - \mu_2$ is the same as that for μ. In the long run, 95 percent of all confidence intervals, similar to the one just stated, will include the unknown difference between population means.

Although we never really know whether this particular confidence interval is true or false, we can be *reasonably confident* that the true effect (or true difference between population means) is neither less than –0.17 minutes nor more than 10.17 minutes If only positive differences had appeared in this confidence interval, a single interpretation would have been possible. However, the appearance of a negative difference in the lower limit indicates that EPO might hinder endurance, and therefore, no single interpretation is possible. Furthermore, the automatic inclusion of a zero difference in an interval with dissimilar signs indicates that EPO may have had no effect whatsoever on endurance.*

Key Point

A single interpretation is possible only if the two limits of the confidence interval for $\mu_1 - \mu_2$ share the same signs, either both positive or both negative.

The range of possible differences (from a low of –0.17 minute to a high of 10.17 minutes) is very large and imprecise—as you would expect, given the very small sample sizes and, therefore, the relatively large standard error. A repeat experiment should use larger sample sizes in order to produce a narrower, more precise confidence interval that would reduce the range of possible population mean differences and effect sizes.

Progress Check *14.7 Imagine that one of the following 95 percent confidence intervals is based on an EPO experiment. (Because of the appearance of pairs of limits with dissimilar signs, a statistically significant result wasn't required as a preliminary screen for constructing the confidence interval—possibly because, in the early stages of research, the investigator simply wanted to know the range of estimates, whether positive or negative, for any possible effect of EPO.)

95% CONFIDENCE INTERVAL	LOWER LIMIT	UPPER LIMIT
1	–3.45	4.25
2	1.89	2.21
3	–1.54	–0.32
4	0.21	1.53
5	–2.53	1.78

(a) Which confidence interval is most precise?

(b) Which confidence interval most strongly supports the conclusion that EPO *facilitates* endurance?

(c) Which confidence interval most strongly supports the conclusion that EPO *hinders* endurance?

(d) Which confidence interval would most likely stimulate the investigator to conduct an additional experiment using larger sample sizes?

Answers on page 438.

*Because of the common statistical origins of confidence intervals and hypothesis tests, the appearance of only positive limits (and the automatic absence of a zero difference) in the 95 percent confidence interval signifies that the null hypothesis would have been rejected if the *same data* were used to conduct a comparable hypothesis test—that is, in this case, a *two-tailed* test at the .05 level of significance. The seemingly contradictory conclusions between the previous hypothesis test and the current confidence interval for the EPO data indicate that a new hypothesis test would *not* have rejected the null hypothesis if a two-tailed rather than a one-tailed test had been used.

14.9 ESTIMATING EFFECT SIZE: COHEN'S *d*

Using a variation of the z score formula in Chapter 5, **Cohen's *d*** describes effect size by expressing *the observed mean difference in standard deviation units* . To calculate *d*, divide the observed mean difference by the standard deviation, that is,

***Standardized Effect Estimate, Cohen's* d**
Describes effect size by expressing the observed mean difference in standard deviation units.

STANDARDIZED EFFECT SIZE, COHEN'S *d* (TWO INDEPENDENT SAMPLES)

$$d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2}} \qquad (14.5)$$

where, according to current usage, *d* refers to a standardized *estimate* of the effect size; $\bar{X}_1$ and $\bar{X}_2$ are the two sample means; and $\sqrt{s_p^2}$ is the sample standard deviation obtained from the square root of the pooled variance estimate.

Division of the mean difference by the standard deviation has several desirable consequences:

- The standard deviation supplies a *stable* frame of reference not influenced by increases in sample size. Unlike the standard error, whose value decreases as sample size increases, the value of the standard deviation remains the same, except for chance, as sample size increases. Therefore, straightforward comparisons can be made between *d* values based on studies with appreciably different sample sizes.
- The original units of measurement cancel out because of their appearance in both the numerator and denominator. Subsequently, *d* always emerges as an estimate in standard deviation units, regardless of whether the original mean difference is based on, for example, reaction times in *milliseconds* of pilots to two different cockpit alarms or weight losses in *pounds* of overweight subjects to two types of dietary restrictions. Except for chance, comparisons are straightforward between values of *d*—with larger values of *d* reflecting larger effect sizes—even though the original mean differences are based on very different units of measurement, such as milliseconds and pounds.

Cohen's Guidelines for *d*

After surveying the research literature, Jacob Cohen suggested a number of general guidelines (see Table 14.2) for interpreting values of *d*:

Table 14.2 COHEN'S GUIDELINES FOR *d*

d	EFFECT SIZE
.20	Small
.50	Medium
.80	Large

- Effect size is *small if d* is less than or in the vicinity of 0.20, that is, one-fifth of a standard deviation.
- Effect size is *medium* if *d* is in the vicinity of 0.50, that is, one-half of a standard deviation.
- Effect size is *large* if *d* is more than or in the vicinity of 0.80, that is, four-fifths of a standard deviation.*

Although widely adopted, Cohen's abstract guidelines for small, medium, and large effects can be difficult to interpret. You might find these guidelines more comprehensible by referring to **Table 14.3**, where Cohen's guidelines for *d* are converted into more

*Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Table 14.3
COHEN'S GUIDELINES FOR *d* AND MEAN DIFFERENCES FOR GPA, IQ, AND SAT SCORES

		MEAN DIFFERENCE		
		GPA	IQ	SAT
d	EFFECT SIZE	$s_P = 0.50$	$s_P = 15$	$s_P = 100$
.20	**Small**	0.10	3	20
.50	**Medium**	0.25	7.5	50
.80	**Large**	0.40	12	80

concrete mean differences involving GPAs, IQs, and SAT scores. Notice that Cohen's medium effect, a *d* value of .50, translates into mean differences of .25 for GPAs, 7.5 for IQs, and 50 for SAT scores. To qualify as medium effects, the average GPA would have to increase, for example, from 3.00 to 3.25; the average IQ from 100 to 107.5; and the average SAT score from 500 to 550.

Furthermore, for a particular measure, such as SAT scores, a 20-point mean difference corresponds to Cohen's small effect, while an 80-point mean difference corresponds to his large effect. However, *do not interpret Cohen's guidelines without regard to special circumstances.* A "small" 20-point increase in SAT scores might be viewed as virtually worthless if it occurred after a lengthy series of workshops on taking SAT tests, but viewed as worthwhile if it occurred after a brief study session.

You might also find it helpful to visualize the impact of each of Cohen's guidelines on the degree of separation between pairs of normal curves. Although, of course, not every distribution is normal, these curves serve as a convenient frame of reference to render values of *d* more meaningful. As shown in **Figure 14.4**, separation between pairs of normal curves is nonexistent (and overlap is complete) when $d = 0$. Separation becomes progressively more conspicuous as the values of *d*, corresponding to Cohen's

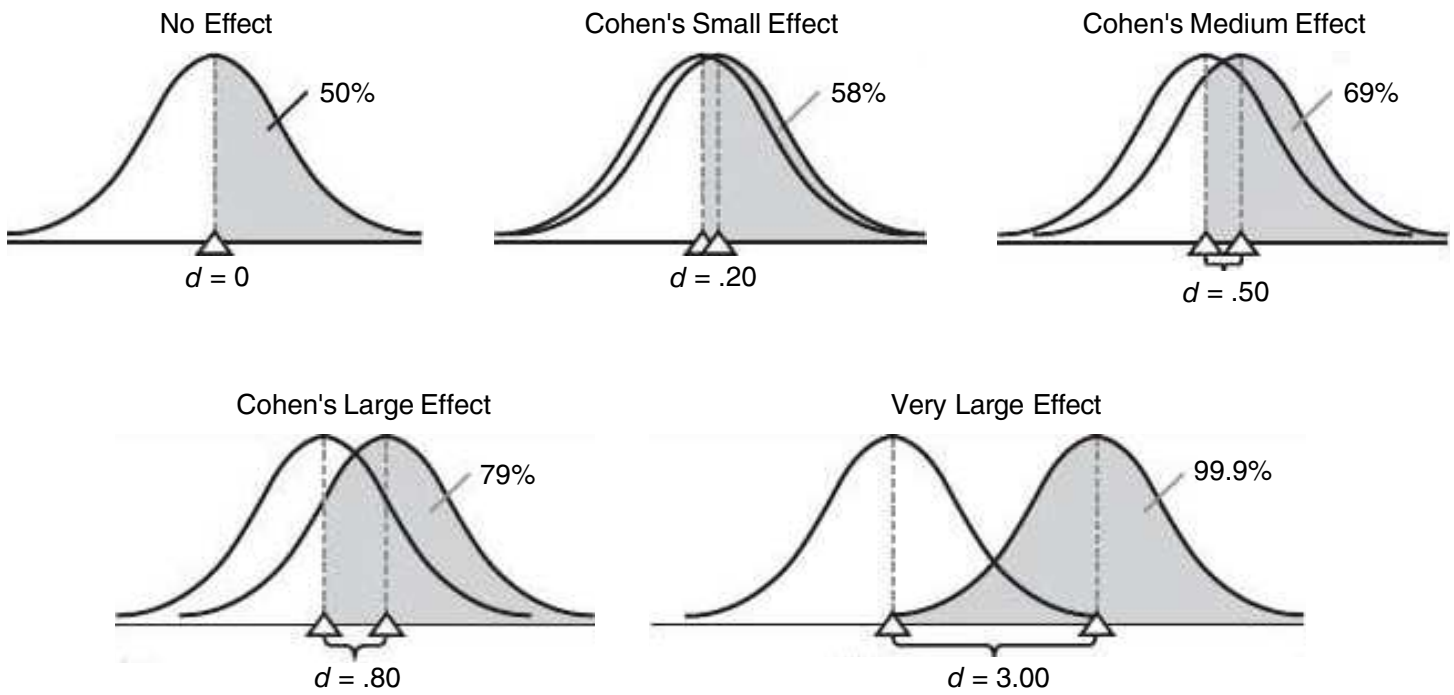


FIGURE 14.4
Separation between pairs of normal curves for selected values of d. *Shaded sectors reflect the percent of scores in one curve that exceed the mean of the other curve.*

small, medium, and large effects, increase from .20 to .50 and then to .80. Separation becomes very conspicuous, with relatively little overlap, given a d value of 3.00, equivalent to three standard deviations, for a very large effect.

To dramatize further the differences between selected d values, the percents (and shaded sectors) in **Figure 14.4** reflect scores in the higher curve that exceed the mean of the lower curve. When $d = 0$, the two curves coincide, and it's a tossup, 50%, whether or not the scores in one curve exceed the mean of the other curve. As values of d increase, the percent of scores in the higher curve that exceed the mean of the lower curve varies from a modest 58% (six out of ten) when $d = .20$ to a more impressive 79% (eight out of ten) when $d = .80$ to a most impressive 99.9% (ten out of ten) when $d = 3.00$.

We can use d to estimate the standardized effect size for the statistically significant results in the EPO experiment described in this chapter. When the mean difference of 5 is divided by the standard deviation of 4.02 (from the square root of the pooled variance estimate of 16.2 in Table 14.1), the value of d equals a large 1.24, that is, a mean difference equivalent to one and one-quarter standard deviations. (Being itself a product of chance sampling variability, this value of d—even if based on real data—would be highly speculative because of the instability of d when sample sizes are small.)

14.10 META-ANALYSIS

The most recent *Publication Manual of the American Psychological Association* recommends that reports of statistical significance tests include some estimate of effect size. Beginning in the next section, we'll adopt this recommendation by including the standardized estimate of effect size, d, in reports of statistically significant mean differences. (A slightly more complicated estimate will be used in later chapters when effect size can't be conceptualized as a simple mean difference.) The routine reporting of effect sizes will greatly facilitate efforts to summarize research findings.

Because of the inevitable variability, attributable to differences in design, subject populations, measurements, etc., as well as chance, the size of effects differs among similar studies. Traditional literature reviews attempt to make sense out of these differences on the basis of expert judgment. Within the last couple of decades, literature reviews have been supplemented by more systematic reviews, referred to as "meta-analysis." A **meta-analysis** begins with an intensive review of all relevant studies. This includes small and even unpublished studies, to try to limit potential "publication bias" arising from only reporting statistically significant results. Typically, extensive details are recorded for each study, such as estimates of effect, design (for example, experimental versus observational), subject population, variability, sample size, etc. Then the collection of previous findings are combined using statistical procedures to obtain either a composite estimate (for example, a standardized mean difference, such as Cohen's d) of the overall effect and its confidence interval, or estimates of subsets of similar effects, if required by the excessive variability among the original effects.*

Meta-analysis
A set of data-collecting and statistical procedures designed to summarize the various effects reported by groups of similar studies.

14.11 IMPORTANCE OF REPLICATION

Over the past few decades there have been a series of widely publicized, seemingly transitory—if not outright contradictory—health-related research findings. For example, initial research suggested that hormonal replacement therapy in women *decreases* the risk of heart attacks and cancer. However, subsequent, more extensive research

*An excellent introduction to meta-analysis can be found in Chapter 1 of Lipsey, M. W., & Wilson, D. B. (2000). *Practical Meta-Analysis*. Thousand Oaks, CA: Sage.

findings suggested that this therapy has no effect or may even *increase* these risks. One precaution you might adopt is to wait for the replication of any new findings, especially for complex, controversial phenomena. Of course, the media most likely would ignore such advice given the competitive climate for breaking news.

There is a well-known bias—often called the **file drawer effect**—that favours the publication only of reports that are statistically significant. Typically, reports of non-significant findings are never published, but are simply put away in file drawers or wastebaskets. A solitary significant finding—much like the tip of an iceberg—could be a false positive result reflecting the high cumulative probability of a type I error when there exist many unpublicized studies with nonsignificant findings. This could contribute to the seemingly transitory nature of some widely publicized reports of research findings. Ideally, a remedy to the file drawer effect would be to have all researchers initially register their research project and then report actual data and results of all statistical analyses, whether significant or nonsignificant, to a repository of research findings.

File Drawer Effect

A false positive caused by an inflated type I error attributable to reports of nonsignificant findings being tossed in a file drawer or waste basket.

More replication of statistically significant findings is needed. A single publicized statistically significant finding may simply reflect a large unknown and unreported type I error, due to many unreported non-statistically significant findings relegated to the file drawer. The wise consumer of research findings withholds a complete acceptance of a single significant finding until it is replicated.

14.12 REPORTS IN THE LITERATURE

It's become common practice to report means and standard deviations, as well as the results of statistical tests. Reports of statistical tests usually are brief, often consisting only of a parenthetical statement that summarizes the statistical analysis and usually includes a p-value and some estimate of effect size, such as Cohen's d *or a confidence interval*. A published report of the EPO experiment might read as follows:

> ***Endurance scores for the EPO group ($\bar{X} = 11$, $s = 4.24$) significantly exceed those for the control group ($\bar{X} = 6$, $s = 3.79$), according to a t test [$t(10) = 2.16$, $p < .05$ and $d = 1.24$].***

Or expressed in a format prevalent in the current psychological literature, where the mean and standard deviation are symbolized as M and SD, respectively:

> ***The endurance scores for the EPO group ($M = 11$, $SD = 4.24$) and control group ($M = 6$, $SD = 3.79$) differed significantly [$t(10) = 2.16$, $p < .05$ and $d = 1.24$].***

In both examples, the parenthetical statement indicates that a t based on 10 degrees of freedom was found to equal 2.16. Since the p-value of less than .05 reflects a rare test result, given that the null hypothesis is true, this result supports the research hypothesis, as implied in the interpretative statements. The d of 1.24 suggests that the observed mean difference of 5 is equivalent to one and one-quarter standard deviations and qualifies as a large effect size. Values for the two standard deviations were obtained by converting $SS_1 = 90$ and $SS_2 = 72$ in Table 14.1 into their respective sample variances and standard deviations, using Formulas 4.9 and 4.10. (For your convenience, values of standard deviations will be supplied in subsequent questions requiring a literature report.)

It's also become common practice to describe results with data graphs. In data graphs, such as that shown in **Figure 14.5**, the dependent variable, mean endurance, is identified with the vertical axis, while values of the independent variable, EPO and control, are located as points along the horizontal axis. Dots identify the mean endur-

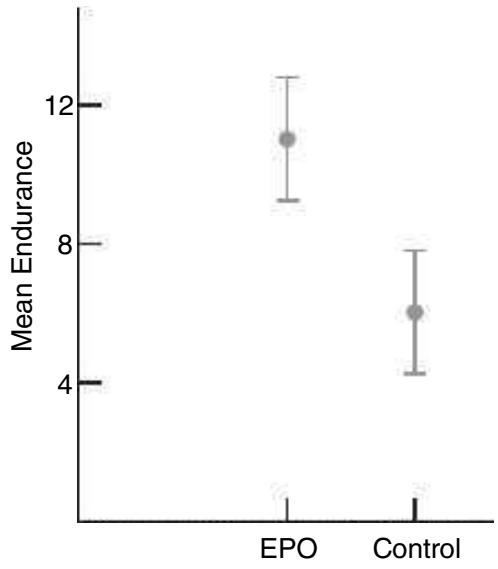


FIGURE 14.5
Data graph where dots identify the mean endurance scores for EPO and control groups. Error bars show deviations equal to one standard error above and below each mean.

ance score for EPO and control groups, while error bars reflect the standard error associated with each dot. (Since error bars could reflect other measures of variability, such as standard deviations or 95 percent confidence intervals, it's important to identify which measure is being used.) Generally speaking, nonoverlapping error bars (for standard errors) imply that differences between means *might* be statistically significant, as, in fact, is the case for the mean differences shown in Figure 14.5. Incidentally, data graphs also can appear as bar charts, where error bars are centered on bar tops and extend vertically above and below bar tops.

Progress Check *14.8 Recall that in Question 14.3, a psychologist determined the effect of instructions on the time required by subjects to solve the same puzzle. For two independent samples of ten subjects per group, mean solution times, in minutes, were longer for subjects given "difficult" instructions ($\overline{X} = 15.8$, $s = 8.64$) than for subjects given "easy" instructions ($\overline{X} = 9.0$, $s = 5.01$). A t ratio of 2.15 led to the rejection of the null hypothesis.

(a) Given a standard deviation, s_p, of 7.06, calculate the value of the standardized effect size, d.

(b) Indicate how these results might be described in the literature.

Answers on page 438.

14.13 ASSUMPTIONS

Whether testing a hypothesis or constructing a confidence interval, t assumes that both underlying populations are normally distributed with equal variances. You need not be too concerned about violations of these assumptions, particularly if both sample sizes are equal and each is fairly large (greater than about 10). Otherwise, in the *unlikely* event that you observe conspicuous departures from normality or equality of variances in the data for the two groups, consider the following possibilities:

1. Increase sample sizes to minimize the effect of any non-normality.
2. Equate sample sizes to minimize the effect of unequal population variances.

3. Use a slightly less sensitive, more complex version of *t* designed for unequal variances, alluded to in the next section and described more fully in Chapter 7 of Howell, D. C. (2013). *Statistical Methods for Psychology* (8th ed.). Belmont, CA: Wadsworth.
4. Use a less sensitive but more assumption-free test, such as the Mann-Whitney U test described in Chapter 20.

14.14 COMPUTER OUTPUT

Table 14.4 shows an SAS output for the *t* test for EPO, as summarized in the box on page 252. Spend a few moments reviewing this material.

Progress Check *14.9 The following questions refer to the SAS output in Table 14.4.

(a) Although, in this case, the results are the same for the *t* test for equal variances and for the *t* test for unequal variances, which test should be reported? Why?

(b) The exact *p*-value equals .0569 for a two-tailed test, the default test for SAS. What is the more appropriate (exact) one-tailed *p*-value?

Table 14.4
SAS OUTPUT: *t* TEST FOR ENDURANCE SCORES

The SAS System
12:17 Thursday, Jan. 14, 2016
t test Procedure

Variable	group	*N*	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
endure	EPO	6	6.5476	11	15.452	2.6483	4.2426	10.406	1.7321
endure	control	6	2.0177	6	9.9823	2.3687	3.7947	9.307	1.5492
endure	Diff (1–2)		–0.178	5	10.1777	2.8123	4.0249	7.0635	2.3238

t Tests

Variable	Method	Variances	*df*	*t* Value	Pr > |*t*|
endure	Pooled	Equal	10	1 2.16	0.0569
endure	Satterthwaite	Unequal	9.88	2.16	0.0572

Equality of Variances

Variable	Method	Num *df*	Den *df*	*F* Value	Pr > *F*
endure	2 Folded *F*	5	5	1.25	0.81250

Comments:
1. Compare the value of t *with that given in Table 14.1. Report the results for the customary* t *test (discussed in this book) that assumes equal variances rather than the more generalized* t *test (not discussed in this book) that accommodates unequal variances unless, as explained in comment 2, the assumption of equal population variances has been rejected.*
2. Given in Table 14.4 are results of the folded F *(or two-tailed* F*) test for equal population variances or, as it is often called, the "*F *test for homogeneity of variance." The folded* F *value of 1.25 is found by dividing the square of the larger standard deviation (4.24)(4.24) by the square of the smaller standard deviation (3.79)(3.79). When the* p*-value for* F*, shown as Pr >* F *in the SAS output, is too small—say, less than .10—there is a possibility that the population variances are not equal. In this case, the more accurate results for the* t *test for unequal variances should be reported. (Because the* F *test responds to any non-normality, as well as to unequal population variances, some practitioners prefer other tests such as Levene's test for equal population variances as screening devices for reporting* t *test results based on unequal variances. For more information about both the* t *test that accommodates unequal population variances and Levene's test, see Chapter 7 in Howell, D. C. (2013). Statistical Methods for Psychology (8th ed.). Belmont, CA: Wadsworth.*

(c) SAS gives the upper and lower confidence limits (CL) for each of six different 95 percent confidence intervals, three for means and three for standard deviations. Is the single set of CLs for the difference between population means, that is, Diff (1-2), consistent with the two-tailed *p*-values for the *t* test?

Answers on page 438.

Summary

Statistical hypotheses for the difference between two population means must be selected from among the following three possibilities:

Nondirectional:
$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

Directional, lower tail critical:
$$H_0 : \mu_1 - \mu_2 \geq 0$$
$$H_1 : \mu_1 - \mu_2 < 0$$

Directional, upper tail critical:
$$H_0 : \mu_1 - \mu_2 \leq 0$$
$$H_1 : \mu_1 - \mu_2 > 0$$

Tests of this null hypothesis are based on the sampling distribution of the difference between sample means, $\bar{X}_1 - \bar{X}_2$. The mean of this sampling distribution equals the difference between population means, and its standard error roughly measures the average amount by which any difference between sample means deviates from the difference between population means.

Because the standard error must be estimated, hypothesis tests use the *t* ratio for two independent samples.

The *p*-value for a test result indicates its degree of rarity, given that the null hypothesis is true. Smaller *p*-values tend to discredit the null hypothesis.

A confidence interval also can be constructed to estimate differences between population means. A single interpretation is possible only if the two limits of the confidence interval share the same sign, either both positive or both negative.

The importance of a statistically significant result can be evaluated with Cohen's *d*, the unit-free, standardized estimate of effect size. Cohen's guidelines identify *d* values in the vicinity of .20, .50, and .80 with small, medium, and large effects, respectively.

Beware of the file drawer effect, that is, a false positive caused by an inflated type I error due to nonsignificant findings never being published.

The *t* test assumes that both underlying populations are normally distributed with equal variances. Except under rare circumstances, you need not be concerned about violations of these assumptions.

Important Terms

Two independent samples
Sampling distribution of $\bar{X}_1 - \bar{X}_2$
Pooled variance estimate $\left(s_p^2\right)$
Effect
Estimated standard error, $s_{\bar{X}_1 - \bar{X}_2}$
Statistical significance

Standard error of the difference between means, $\sigma_{\bar{X}_1-\bar{X}_2}$
Confidence intervals for $\mu_1 - \mu_2$
Meta-analysis
***p*-value**
Standardized effect estimate, Cohen's *d*
File drawer effect

Key Equations

RATIO

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_{hyp}}{s_{\bar{X}_1-\bar{X}_2}}$$

where $s_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$

and $s_p^2 = \frac{SS_1 + SS_2}{n_1 + n_2 - 2}$

STANDARDIZED EFFECT

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2}}$$

REVIEW QUESTIONS

***14.10** Figure 4.2 on page 62 describes the results for two fictitious experiments, each with the same mean difference of 2 but with noticeably different variabilities. Unresolved was the question "Once variability has been considered, should the difference between each pair of means be viewed as real or merely transitory?" A t test for two independent samples permits us to answer this question for each experimental result.

(a) Referring to Figure 4.2, again decide which of the two identical differences between pairs of means—that for Experiment B or for Experiment C—is more likely to be viewed as real.

(b) Given that $s_p^2 = .33$ for Experiment B, test the null hypothesis at the .05 level of significance.

(c) Given that $s_p^2 = 3.67$ for Experiment C, test the null hypothesis at the .05 level of significance. You needn't repeat the usual step-by-step hypothesis test procedure, but specify the observed value of t and the decision about the null hypothesis.

(d) Specify the approximate p-values for both t tests.

(e) Answer the original question about whether the difference between each pair of means is real or merely transitory.

(f) If a difference is real, use Cohen's *d* to estimate the effect size.

Answers on pages 438 and 439.

14.11 To test compliance with authority, a classical experiment in social psychology requires subjects to administer increasingly painful electric shocks to seemingly helpless victims who agonize in an adjacent room.* Each subject earns a score between 0 and 30, depending on the point at which the subject refuses to comply with authority—an investigator, dressed in a white lab coat, who orders the administration of increasingly intense shocks. A score of 0 signifies the subject's unwillingness to comply at the very outset, and a score of 30 signifies the subject's willingness to comply completely with the experimenter's orders.

Ignore the very real ethical issues raised by this type of experiment, and assume that you want to study the effect of a "committee atmosphere" on compliance with authority. In one condition, shocks are administered only after an affirmative decision by the committee, consisting of one real subject and two associates of the investigator, who act as subjects but, in fact, merely go along with the decision of the real subject. In the other condition, shocks are administered only after an affirmative decision by a solitary real subject.

A total of 12 subjects are randomly assigned, in equal numbers, to the committee condition (X_1) and to the solitary condition (X_2). A compliance score is obtained for each subject. Use *t* to test the null hypothesis at the .05 level of significance.

COMPLIANCE SCORES	
COMMITTEE	SOLITARY
2	3
5	8
20	7
15	10
4	14
10	0

14.12 To determine whether training in a series of workshops on creative thinking increases IQ scores, a total of 70 students are randomly divided into treatment and control groups of 35 each. After two months of training, the sample mean IQ $(\bar{X}_1)$ for the treatment group equals 110, and the sample mean IQ $(\bar{X}_2)$ for the control group equals 108. The estimated standard error equals 1.80.

(a) Using *t*, test the null hypothesis at the .01 level of significance.

(b) If appropriate (because the null hypothesis has been rejected), estimate the standardized effect size, construct a 99 percent confidence interval for the true population mean difference, and interpret these estimates.

***14.13** Is the performance of college students affected by the grading policy? In an introductory biology class, a total of 40 student volunteers are randomly assigned,

*See S. Milgram, S. (1975). *Obedience to Authority: An Experimental View*. New York, NY: HarperPerennial.

in equal numbers, to take the course for either letter grades or a simple pass/fail. At the end of the academic term, the mean achievement score for the letter grade students $(\bar{X}_1)$ equals 86.2, and the mean achievement score for pass/fail students $(\bar{X}_2)$ equals 81.6. The estimated standard error is 1.50.

(a) Use t to test the null hypothesis at the .05 level of significance.

(b) How would the above hypothesis test change if the roles of X_1 and X_2 were reversed—that is, if X_1 were identified with pass/fail students and X_2 were identified with letter grade students?

(c) Most students would doubtless prefer to select their favorite grading policy rather than be randomly assigned to a particular grading policy. Therefore, why not replace random assignment with self-selection?

(d) Specify the p-value for this test result.

(e) If the test result is statistically significant, estimate the standardized effect size, given that the standard deviation, s_p, equals 5.

(f) State how the test results might be reported in the literature, given that $s_1 = 5.39$ and $s_2 = 4.58$.

Answers on page 439.

***14.14** An investigator wishes to determine whether alcohol consumption causes a deterioration in the performance of automobile drivers. Before the driving test, subjects drink a glass of orange juice, which, in the case of the treatment group, is laced with two ounces of vodka. Performance is measured by the number of errors made on a driving simulator. A total of 120 volunteer subjects are randomly assigned, in equal numbers, to the two groups. For subjects in the treatment group, the mean number of errors $(\bar{X}_1)$ equals 26.4, and for subjects in the control group, the mean number of errors $(\bar{X}_2)$ equals 18.6. The estimated standard error equals 2.4.

(a) Use t to test the null hypothesis at the .05 level of significance.

(b) Specify the p-value for this test result.

(c) If appropriate, construct a 95 percent confidence interval for the true population mean difference and interpret this interval.

(d) If the test result is statistically significant, use Cohen's d to estimate the effect size, given that the standard deviation, s_p, equals 13.15.

(e) State how these test results might be reported in the literature, given $s_1 = 13.99$ and $s_2 = 12.15$.

Answers on pages 439 and 440.

14.15 Review Question 2.16 on page 44 lists the GPAs for groups of 27 meditators and 27 nonmeditators.

(a) Given that the mean GPA equals 3.19 for the meditators and 2.90 for the nonmeditators, and that s_p^2 equals .20, specify the observed value of t and its approximate p-value.

(b) Answer the original question about whether these two groups tend to differ.

(c) If the p-value is less than .05, use Cohen's d to estimate the effect size.

14.16 After testing several thousand high school seniors, a state department of education reported a statistically significant difference between the mean GPAs for female and male students. Comments?

14.17 Someone claims that, given a p-value less than .01, the corresponding null hypothesis also must be true with probability less than .01. Comments?

14.18 Indicate (Yes or No) whether each of the following statements is a desirable property of Cohen's d.

(a) immune to changes in sample size

(b) reflects the size of the p-value

(c) increases with sample size

(d) reflects the size of the effect

(e) independent of the particular measuring units

(f) facilitates comparisons across studies

(g) bypasses hypothesis test

***14.19** During recent decades, there have been a series of widely publicized, seemingly transitory, often contradictory research findings reported in newspapers and on television. For example, a few initial research findings suggested that vaccination causes autism in children. However, subsequent, more extensive research findings, as well as a more critical look at the original findings, suggested that vaccination doesn't cause autism (https://www.sciencebasedmedicine.org/reference/vaccines-and-autism/#overview).

What might be one explanation for the seemingly erroneous initial research finding?

Answers on page 440.

CHAPTER

15

t Test for Two Related Samples (Repeated Measures)

Preview

Although differences among individuals make life interesting, they also can blunt the precision of a statistical analysis because of their considerable impact on the overall variability among scores. You can control for individual differences by measuring each subject twice and using a t *test for repeated measures. This t test can be extra sensitive to detecting a false null hypothesis. However, several potential problems must be addressed before adopting a repeated-measures design.*

15.1 EPO EXPERIMENT WITH REPEATED MEASURES

In the EPO experiment of Chapter 14, the endurance scores of patients reflect not only the effect of EPO, *if it exists,* but also the random effects of many uncontrolled factors. One very important type of uncontrolled factor, referred to as *individual differences,* reflects the array of characteristics, such as differences in attitude, physical fitness, personality, etc., that distinguishes one person from another. If uncontrolled, individual differences can cause appreciable random variations among endurance scores and, therefore, make it more difficult to detect any treatment effect. When each subject is measured twice, as in the experiment described in this chapter, the t test for repeated measures can be extra sensitive to detecting a treatment effect by eliminating the distorting effect of variability due to individual differences.

Difference (D) Scores

Computations can be simplified by working directly with the difference between pairs of endurance scores, that is, by working directly with

DIFFERENCE SCORE (D)

$$D = X_1 - X_2 \tag{15.1}$$

Difference Score (D)
The arithmetic difference between each pair of scores in repeated measures or, more generally, in two related samples.

where D is the **difference score** and X_1 and X_2 are the paired endurance scores for each patient measured twice, once under the treatment condition and once under the control condition, respectively. Essentially, the use of difference scores converts a two-sample problem with X_1 and X_2 scores into a one-sample problem with D scores.

Mean Difference Score ($\bar{D}$)

To obtain the mean for a set of difference scores, add all difference scores and divide by the number of scores, that is,

MEAN DIFFERENCE SCORE ($\bar{D}$)

$$\bar{D} = \frac{\Sigma D}{n} \tag{15.2}$$

where $\bar{D}$ is the mean difference score, ΣD is the sum of all positive difference scores *minus* the sum of all negative difference scores, and n is the number of difference scores. The sign of $\bar{D}$ is crucial. For example, in the current experiment, a positive value of $\bar{D}$ would signify that EPO facilitates endurance, while a negative value of $\bar{D}$ would signify that EPO hinders endurance.

Table 15.1
SCORES FOR TWO EPO EXPERIMENTS

ORIGINAL X_1	ORIGINAL X_2	NEW D
12	7	5
5	3	2
11	4	7
11	6	5
9	3	6
18	13	5

Comparing the Two Experiments

To simplify comparisons, exactly the same six X_1 scores and six X_2 scores in the original EPO experiment with two independent samples are used to generate the six D scores in the new EPO experiment with repeated measures, as indicated in **Table 15.1**. Therefore, the sample mean difference also is the same, both for the original experiment, where $\bar{X}_1 - \bar{X}_2 = 11 - 6 = 5$, and for the new experiment, where $\bar{D} = 5$. To dramatize the beneficial effects of repeated measures, highly similar pairs of X_1 and X_2 scores appear in the new experiment. For example, high endurance scores of 18 and 13 minutes are paired, presumably for a very physically fit patient, while low scores of only

5 and 3 minutes are paired, presumably for another patient in terrible shape. Since in real applications there is no guarantee that individual differences will be this large, the net effect of a repeated-measures experiment might not be as beneficial as that described in the current analysis.

Figure 15.1 shows the much smaller variability among paired differences in endurance scores, D, for the new experiment. The range of scores in the top histogram for X_1 and X_2 equals 15 (from $18 - 3$), while that in the bottom histogram for D equals only 5 (from $7 - 2$). This suggests that once the new data have been analyzed with a t test for repeated measures, it should be possible not only to reject the null hypothesis again, but also to claim a much smaller p-value than that ($p < .05$) for the t test for the original experiment with two independent samples.

Repeated Measures

Repeated Measures
Whenever the same subject is measured more than once.

A favorite technique for controlling individual differences is referred to as **repeated measures**, *because each subject is measured more than once* . By focusing on the *differences* between pairs of scores for each subject, the investigator effectively eliminates, by the simple act of subtraction, each individual's unique impact on both endurance scores. Accordingly, an analysis of the resulting difference scores reflects only any effects due to EPO, *if it exists,* and random variations of other uncontrolled factors or *experimental errors* not attributable to individual differences. (Experimental errors refer to random variations in endurance scores due to the combined impact of numerous uncontrolled changes, such as slight changes in temperature, treadmill speed, etc., as well as any changes in a particular subject's motivation, health, etc., between the two experimental sessions.) Because of the smaller standard error term, the result is a test with an increased likelihood of detecting any effect due to EPO.

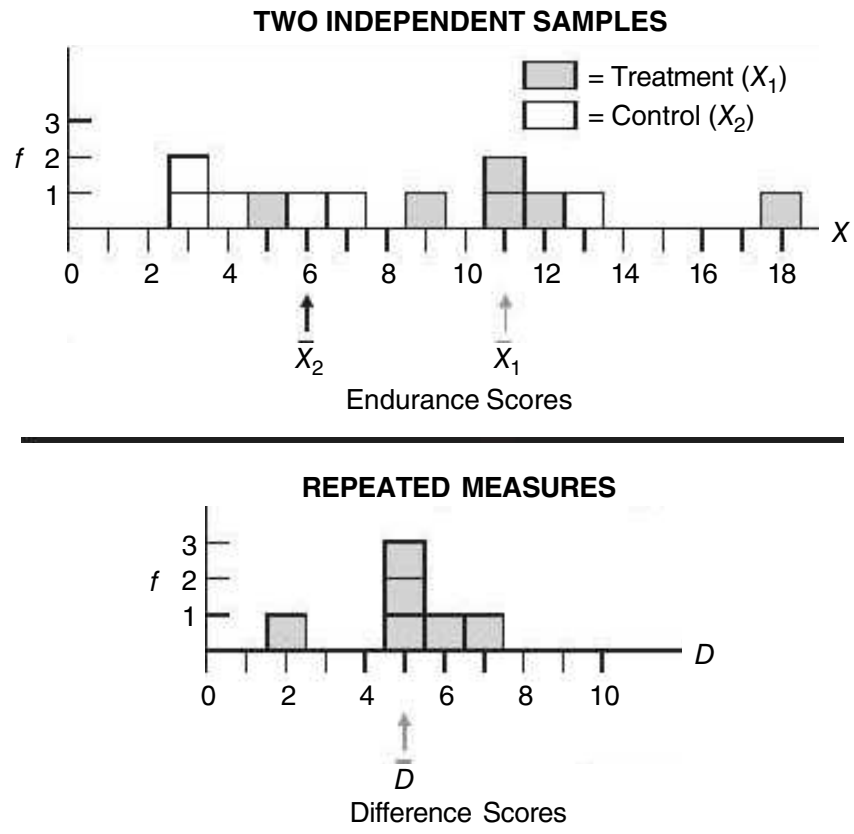


FIGURE 15.1
Different variabilities (with identical mean differences of 5) in EPO experiments with two independent samples and with repeated measures.

Two Related Samples

Two Related Samples
Each observation in one sample is paired, on a one-to-one basis, with a single observation in the other sample.

Favored by investigators who wish to control for individual differences, repeated measures represent the most important special case of two related samples. **Two related samples** occur whenever *each observation in one sample is paired, on a one-to-one basis, with a single observation in the other sample* .

Repeated measures might not always be feasible since, as discussed below, several potential complications must be resolved before measuring subjects twice. An investigator still might choose to use two related samples by matching pairs of different subjects in terms of some uncontrolled variable that appears to have a considerable impact on the dependent variable. For example, patients might be matched for their body weight because preliminary studies revealed that, regardless of whether or not they received EPO, lightweight patients have better endurance scores than heavyweight patients. Before collecting data, patients could be matched, beginning with the two lightest and ending with the two heaviest (and random assignment dictating which member of each pair receives EPO). Now, as with repeated measures, the endurance scores for pairs of matched patients tend to be more similar (than those of unmatched subjects in two independent samples), and so the statistical test must be altered to reflect this new dependency between pairs of matched endurance scores.

Progress Check *15.1 Indicate whether each of the following studies involves two independent samples or two related samples, and in the latter case, indicate whether the study involves repeated measures for the same subjects or matched pairs of different subjects.

(a) Estimates of weekly TV-viewing time of third-grade girls compared with those of third-grade boys

(b) Number of cigarettes smoked by participants before and after an antismoking workshop

(c) Annual incomes of husbands compared with those of their wives

(d) Problem-solving skills of recognized scientists compared with those of recognized artists, given that scientists and artists have been matched for IQ

Answers on page 440.

Some Complications with Repeated Measurements

Unfortunately, the attractiveness of repeated measures sometimes fades upon closer inspection. For instance, since each patient is measured twice, once in the treatment condition and once in the control condition, sufficient time must elapse between these two conditions to eliminate any lingering effects due to the treatment. If there is any concern that these effects cannot be eliminated, use each subject in only one condition.

Counterbalancing

Counterbalancing
Reversing the order of conditions for equal numbers of all subjects.

Otherwise, when subjects do perform double duty in both conditions, *it is customary to randomly assign half of the subjects to experience the two conditions in a particular order*—say, first the treatment and then the control condition— *while the other half of the subjects experience the two conditions in the reverse order* . Known as **counterbalancing**, this adjustment controls for any sequence effect, that is, any potential bias in favor of one condition merely because subjects happen to experience it first (or second).*

*Counterbalancing would be inappropriate for repeated-measure experiments that focus on any changes in the dependent variable *before* and *after* some special event, such as the anti-smoking workshop described in Questions 15.1(b) and 15.8.

Presumably, the investigator considered these potential complications before beginning the EPO experiment with repeated measures. The two sessions should have been separated by a sufficiently long period of time—possibly several weeks—in order to dissipate any lingering effects of EPO. In addition, a randomly selected half of the six patients should have experienced the two conditions in one order, while the remaining patients should have experienced the two conditions in the reverse order.

15.2 STATISTICAL HYPOTHESES

Null Hypothesis

Converting to difference scores generates a single population of difference scores, and the null hypothesis is expressed in terms of this new population. If EPO has either no consistent effect or a negative effect on endurance scores when patients are measured twice, the population mean of all difference scores, μ_D, should equal zero or less. In symbols, an equivalent statement reads:

$$H_0: \mu_D \leq 0$$

Alternative (or Research) Hypothesis

As before, the investigator wants to reject the null hypothesis only if EPO actually increases endurance scores. An equivalent statement in symbols reads:

$$H_1: \mu_D > 0$$

This directional alternative hypothesis translates into a one-tailed test with the upper tail critical.

Two Other Possible Alternative Hypotheses

Although not appropriate for the current experiment, there are two other possible alternative hypotheses. Another directional hypothesis, expressed as

$$H_1: \mu_D < 0$$

translates into a one-tailed test with the lower tail critical, and a nondirectional hypothesis, expressed as

$$H_1: \mu_D \neq 0$$

translates into a two-tailed test.

15.3 SAMPLING DISTRIBUTION OF $\bar{D}$

The sample mean of the difference scores, $\bar{D}$, varies from sample to sample, and it has a sampling distribution with its own mean, $\mu_{\bar{D}}$, and standard error, $\sigma_{\bar{D}}$. When $\bar{D}$ is viewed as the mean for a single sample of difference scores, its sampling distribution can be depicted as a straightforward extension of the sampling distribution of $\bar{X}$, the mean for a single sample of original scores, as described in Chapter 9. Therefore, the mean, $\mu_{\bar{D}}$, and standard error, $\sigma_{\bar{D}}$, of the sampling distribution of $\bar{D}$ have essentially the same properties as the mean, $\mu_{\bar{X}}$, and standard error, $\sigma_{\bar{X}}$, respectively, of the sampling distribution of $\bar{X}$.

Since the mean of the sampling distribution of $\bar{X}$ equals the population mean, that is, since $\mu_{\bar{X}} = \mu$, the mean of the sampling distribution of $\bar{D}$ equals the corresponding population mean (for difference scores), that is,

$$\mu_{\bar{D}} = \mu_D$$

Likewise, since the standard error of $\overline{X}$ equals the population standard deviation divided by the square root of the sample size, that is, since $\sigma_{\overline{X}} = \sigma/\sqrt{n}$, the standard error of $\overline{D}$ equals the corresponding population standard deviation (for difference scores) divided by the square root of the sample size, that is,

$$\sigma_{\overline{D}} = \frac{\sigma_D}{\sqrt{n}}$$

15.4 *t* TEST

The null hypothesis for two related samples can be tested with a *t* ratio. Expressed in words,

$$t = \frac{(sample\ mean\ difference) - (hypothesized\ population\ mean\ difference)}{estimated\ standard\ error}$$

Expressed in symbols,

t RATIO FOR TWO POPULATION MEANS (TWO RELATED SAMPLES)

$$t = \frac{\overline{D} - \mu_{D_{hyp}}}{s_{\overline{D}}} \qquad (15.3)$$

which has a *t* sampling distribution with $n - 1$ degrees of freedom. In Formula 15.3, $\overline{D}$ represents the sample mean of the difference scores; $\mu_{D_{hyp}}$ represents the hypothesized population mean (of zero) for the difference scores; and $s_{\overline{D}}$ represents the estimated standard error of $\overline{D}$, as defined later in Formula 15.5.

Finding Critical *t* Values

To find a critical *t* in Table B in Appendix C, follow the usual procedure. Read the entry in the cell intersected by the row for the correct number of degrees of freedom and the column for the test specifications. To find the critical *t* for the current EPO experiment, go to the right-hand panel for a one-tailed test in Table B, then locate the row corresponding to 5 degrees of freedom (from $df = n - 1 = 6 - 1 = 5$), and locate the column for a one-tailed test at the .05 level of significance. The intersected cell specifies 2.015.

Summary for EPO Experiment

The boxed hypothesis test summary for the current EPO experiment indicates that since the calculated *t* of 7.35 exceeds the critical *t* of 2.015, we're able to reject H_0.

It's important to mention the use of repeated measures (or any matching) in the conclusion of the report. Repeated measures eliminates one important source of variability among endurance scores—the variability due to individual differences—that otherwise inflates the standard error term and causes an increase in β, the probability of a type II error.

Because of the smaller standard error for repeated measures, the calculated *t* of 7.35, with $df = 5$, permits us to claim a much smaller *p*-value ($p < .001$) than that ($p < .05$) for a *t* test based on the same data in the original EPO experiment with two independent samples.

HYPOTHESIS TEST SUMMARY
t Test for Two Population Means: Repeated Measures (EPO Experiment)

Research Problem

When patients are measured twice, once with and once without EPO, does the population mean difference score show greater endurance due to EPO?

Statistical Hypotheses

$$H_0: \mu_D \leq 0$$
$$H_1: \mu_D > 0$$

Decision Rule

Reject H_0 at the .05 level of significance if $t \geq 2.015$ (from Table B in Appendix C, given that $df = n - 1 = 6 - 1 = 5$).

Calculations

$$t = \frac{5 - 0}{0.68} = 7.35 = \text{(See Table 15.1 for all computations.)}$$

Decision

Reject H_0 at the .05 level of significance because the calculated t of 7.35 exceeds 2.015.

Interpretation

There is evidence that when patients are measured twice, EPO is found to increase the mean endurance score.

15.5 DETAILS: CALCULATIONS FOR THE *t* TEST

The three panels in **Table 15.2** show the computational steps that produce a t of 7.35 in the current experiment.

Panel I

Panel I involves most of the computational labor, and it generates values for the sample mean difference, $\overline{D}$, and the sample standard deviation for the difference scores, s_D. To obtain the sample standard deviation, first use a variation on the computation formula for the sum of squares (Formula 4.4 on page 69), where X has been replaced with D, that is,

$$SS_D = \Sigma D^2 - \frac{(\Sigma D)^2}{n}$$

Table 15.2
CALCULATIONS FOR THE *t* TEST: REPEATED MEASURES (EPO EXPERIMENT)

I. FINDING THE MEAN AND STANDARD DEVIATION, $\bar{D}$ AND s_D

(a) Computational sequence:
Assign a value to *n*, the number of difference scores (1)
Subtract X_2 from X_1 to obtain *D* (2)
Sum all *D* scores (3)
Substitute numbers in the formula (4) and solve for $\bar{D}$
Square each *D* score (5), one at a time, and then add all squared *D* scores (6)
Substitute numbers in the formula (7) for SS_D, and then solve for s_D (8)

(b) Data and computations:

ENDURANCE SCORES (MINUTES)

	EPO	CONTROL	DIFFERENCE SCORES	
PATIENT	X_1	X_2	2 D	5 D^2
1	12	7	5	25
2	5	3	2	4
3	11	4	7	49
4	11	6	5	25
5	9	3	6	36
6	18	13	5	25
1 $n = 6$			3 $\Sigma D = 30$	6 $\Sigma D^2 = 164$

$$4\ \bar{D} = \frac{\Sigma D}{n} = \frac{30}{6} = 5$$

$$7\ SS_D = \Sigma D^2 - \frac{(\Sigma D)^2}{n} = 164 - \frac{(30)^2}{6} = 164 - 150 = 14$$

$$8\ s_D = \sqrt{\frac{SS_D}{n-1}} = \sqrt{\frac{14}{6-1}} = \sqrt{2.8} = 1.67$$

II. FINDING THE STANDARD ERROR, $s_{\bar{D}}$

(a) Computational sequence:
Substitute numbers obtained above in the formula 9 and solve for $s_{\bar{D}}$

(b) Computations:

$$9\ s_{\bar{D}} = \frac{s_D}{\sqrt{n}} = \frac{1.67}{\sqrt{6}} = \frac{1.67}{2.45} = 0.68$$

III. FINDING THE OBSERVED *t* RATIO

(a) Computational sequence:
Substitute numbers obtained above in the formula 10, as well as a value of 0 for $\mu_{D_{hyp}}$, and solve for *t*.

(b) Computations:

$$10\ t = \frac{\bar{D} - \mu_{D_{hyp}}}{s_{\bar{D}}} = \frac{5-0}{0.68} = 7.35$$

and then, after dividing the sum of squares, SS_D, by its degrees of freedom, $n - 1$, extract the square root, that is,

SAMPLE STANDARD DEVIATION, s_D

$$s_D = \sqrt{\frac{SS_D}{n-1}} \tag{15.4}$$

Panel II

Dividing the sample standard deviation, s_D, by the square root of its sample size, n, gives the estimated standard error, $s_{\bar{D}}$, that is,

ESTIMATED STANDARD ERROR, $s_{\bar{D}}$

$$s_{\bar{D}} = \frac{s_D}{\sqrt{n}} \tag{15.5}$$

Panel III

Finally, as defined in Formula 15.3, dividing the difference between the sample mean, $\bar{D}$, and the null hypothesized value, $\mu_{D_{hyp}}$ (of zero), by the estimated standard error, $s_{\bar{D}}$, culminates in the value for the t ratio.

Progress Check *15.2 An investigator tests a claim that vitamin C reduces the severity of common colds. To eliminate the variability due to different family environments, pairs of children from the same family are randomly assigned to either a treatment group that receives vitamin C or a control group that receives fake vitamin C. Each child estimates, on a 10-point scale, the severity of their colds during the school year. The following scores are obtained for ten pairs of children:

	ESTIMATED SEVERITY	
PAIR NUMBER	VITAMIN C (X_1)	FAKE VITAMIN C (X_2)
1	2	3
2	5	4
3	7	9
4	0	3
5	3	5
6	7	7
7	4	6
8	5	8
9	1	2
10	3	5

Using t, test the null hypothesis at the .05 level of significance.

Answer on pages 440 and 441.

15.6 ESTIMATING EFFECT SIZE

Confidence Interval for μ_D

Given that two samples are related, as when patients were measured twice in the EPO experiment, a confidence interval for μ_D can be constructed from the following expression:

CONFIDENCE INTERVAL FOR μ_D (TWO RELATED SAMPLES)

$$\bar{D} \pm (t_{conf})(s_{\bar{D}}) \qquad (15.6)$$

where $\bar{D}$ represents the sample mean of the difference scores; t_{conf} represents a number (distributed with $n - 1$ degrees of freedom) from the t tables, which satisfies the confidence specifications; and $s_{\bar{D}}$ represents the estimated standard error defined in Formula 15.5.

Finding t_{conf}

To find the appropriate value of t_{conf} in Formula 15.6, refer to Table B in Appendix C and follow the usual procedure for obtaining confidence intervals. If a 95 percent confidence interval is desired for the EPO experiment with repeated measures, first locate the row corresponding to 5 degrees of freedom, and then locate the column for the 95 percent level of confidence, that is, the column heading identified with a single asterisk. The intersected cell specifies a value of 2.571 for t_{conf}.

Given a value of 2.571 for t_{conf}, and (from Table 15.1) values of 5 for $\bar{D}$, the sample mean of the difference scores, and 0.68 for $s_{\bar{D}}$, the estimated standard error, Formula 15.6 becomes

$$5 \pm (2.571)(0.68) = 5 \pm 1.75 = \begin{cases} 6.75 \\ 3.25 \end{cases}$$

It can be claimed, with 95 percent confidence, that the interval between 3.25 minutes and 6.75 minutes includes the true mean for the population of difference endurance scores.

Interpreting Confidence Intervals for μ_D

Because both limits have similar (positive) signs, a single interpretation describes all of the possibilities included in this confidence interval. The appearance of only positive differences indicates that when patients are measured twice, EPO facilitates endurance. Furthermore, we can be *reasonably confident* that, on average, the true facilitative effect is neither less than 3.25 minutes nor more than 6.75 minutes.

Compare the confidence limits of the current interval for two related samples, 3.25 to 6.75, to those of the previous interval for two independent samples, –0.17 to 10.17. Although both intervals are based on the same data with identical mean differences of 5, the more precise interval for repeated measures, with both limits positive, reflects a reduction in the standard error caused by the elimination of variability due to individual differences.

Progress Check 15.3 Referring to the vitamin C experiment in Question 15.2, construct and interpret a 95 percent confidence interval for the population mean difference score.

Answer on page 441.

Standardized Effect Size, Cohen's *d*

Having rejected the null hypothesis for the EPO experiment with repeated measures, we can claim that the sample mean difference of 5 minutes is statistically significant. As has been noted in Chapter 14, one way to gauge the importance of a statistically significant result is to calculate Cohen's d. When the two samples are related, the formula for Cohen's d is:

STANDARDIZED EFFECT SIZE, COHEN'S *d* (TWO RELATED SAMPLES)

$$d = \frac{\overline{D}}{s_D} \tag{15.7}$$

where d refers to the standardized estimate of effect size, while $\overline{D}$ and s_D represent the sample mean and standard deviation, respectively, of the difference scores.

When the mean difference of 5 is divided by the standard deviation of 1.67 (from Table 15.1), Cohen's d equals 2.99, a very large value equivalent to three standard deviations. (According to Cohen's guidelines, mentioned previously, the estimated effect size is small, medium, or large, depending on whether d is .20 or less, .50, or .80 or more, respectively.)

Progress Check *15.4 For the vitamin C experiment in Question 15.2, estimate and interpret the standardized effect size, d, given a mean, $\overline{D}$, of –1.50 days and a standard deviation, s_D, of 1.27 days.

Answer on page 441.

15.7 ASSUMPTIONS

Whether testing a hypothesis or constructing a confidence interval, t assumes that the population of difference scores is normally distributed. You need not be too concerned about violations of this assumption as long as the sample size is fairly large (greater than about ten pairs). Otherwise, in the *unlikely* event that you encounter conspicuous departures from normality, consider either increasing the sample size or using the less sensitive but more assumption-free Wilcoxon T test described in Chapter 20.

15.8 OVERVIEW: THREE *t* TESTS FOR POPULATION MEANS

In Chapters 13, 14, and 15, three t tests for population means have been described, and their more distinctive features are summarized in **Table 15.3**. Given a hypothesis test for one or two population means, a t test is appropriate if, as almost always is the case, the population standard deviation must be estimated. You must decide whether to use a t test for one sample, two independent samples, or two related samples. This decision is fairly straightforward if you proceed, step by step, as follows:

One or Two Samples?

First, decide whether there are one or two samples. If there is only one sample, because the study deals with a single set of observations, then, of course, you need not search any further: The appropriate t is that for one sample.

Are the Two Samples Paired?

Second, if there are two samples, decide whether or not there is any pairing. If each observation is paired, on a one-to-one basis, with a single observation in the other sample (because of either repeated measures or matched pairs of different subjects), then the appropriate t is that for two related samples.

Finally, if there is no evidence of pairing between individual observations, then the appropriate t is that for two independent samples.

Table 15.3
SUMMARY OF *t* TESTS FOR POPULATION MEANS

TYPE OF SAMPLE	SAMPLE MEAN	NULL HYPOTHESIS*	STANDARD ERROR	*t* RATIO	DEGREES OF FREEDOM
One sample	$\overline{X}$	$H_0: \mu =$ some number	$s_{\overline{X}}$ (Formula 13.3)	$\dfrac{\overline{X} - \mu_{hyp}}{s_{\overline{X}}}$	$n - 1$
Two independent samples (no pairing)	$\overline{X}_1 - \overline{X}_2$	$H_0: \mu_1 - \mu_2 = 0$	$s_{\overline{X}_1 - \overline{X}_2}$ (Formula 14.3)	$\dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)_{hyp}}{s_{\overline{X}_1 - \overline{X}_2}}$	$n_1 + n_2 - 2$
Two related samples (pairs of observations)	$\overline{D}$	$H_0: \mu_D = 0$	$s_{\overline{D}}$ (Formula 15.5)	$\dfrac{\overline{D} - \mu_{D_{hyp}}}{s_{\overline{D}}}$	$n - 1$ (where n refers to pairs of observations)

* *For a two-tailed test.*

Examples

Let's identify the appropriate *t* test for each of several similar studies where, with the aid of radar guns, investigators clock the speeds of randomly selected motorists on a dangerous section of a state highway. Follow the recommended decision procedure to arrive at your own answer before checking the answer in the book.

Study A

Research Problem: Clocked speeds of randomly selected trucks are compared with clocked speeds of randomly selected cars.

Answer: Because there are two sets of observations (speeds for trucks and speeds for cars), there are two samples. Furthermore, since there is no indication of pairing among individual observations, the appropriate *t* test is that for two independent samples.

Study B

Research Problem: Clocked speeds of randomly selected motorists are compared with the posted speed limit of 65 miles per hour.

Answer: Because there is a single set of observations, the appropriate *t* test is that for one sample (where the null hypothesis equals 65 miles per hour).

Study C

Research Problem: Clocked speeds of randomly selected motorists are compared at two different locations: one mile before and one mile after a large sign listing the number of fatalities on that stretch of highway during the previous year.

Answer: Because there are two sets of observations (speeds before and speeds after the sign), there are two samples. Furthermore, since each observation in one sample (the speed of a particular motorist one mile before the sign) is paired with a single observation in the other sample (the speed of the *same* motorist one mile after the sign), the appropriate *t* test is that for two related samples, with repeated measures.

Beginning with the next set of exercises, you will be exposed to a variety of studies for which you must identify the appropriate statistical test. By following a step-by-step procedure, such as the one described here, you will be able to make this identification not only for textbook studies, but also for those encountered in everyday practice.

Progress Check *15.5 Each of the following studies requires a *t* test for one or more population means. Specify whether the appropriate *t* test is for one sample, two independent samples, or two related samples, and in the last case, whether it involves repeated measures or matched pairs of different subjects.

(a) College students are randomly assigned to receive either behavioral or cognitive therapy. After twenty therapeutic sessions, each student earns a score on a mental health questionnaire.

(b) A researcher wishes to determine whether attendance at a day-care center increases the scores of three-year-old children on a motor skill test. Random assignment dictates which twin from each pair of twenty twins attends the day-care center and which twin stays at home. (Such a draconian experiment doubtless would incur great resistance from the parents, not to mention the twins!)

(c) One hundred college freshmen are randomly assigned to sophomore roommates who have either similar or dissimilar vocational goals. At the end of their first year, the mean GPAs of these two groups are to be analyzed.

(d) According to the U.S. Department of Health, the average 16-year-old male can do 23 push-ups. A physical education instructor finds that in his school district, 30 randomly selected 16-year-old males can do an average of 28 pushups.

(e) A child psychologist assigns aggression scores to each of 10 children during two 60-minute observation periods separated by an intervening exposure to a series of violent TV cartoons.

Answers on page 441.

15.9 *t* TEST FOR THE POPULATION CORRELATION COEFFICIENT, ρ

In Chapter 6, .80 describes the sample correlation coefficient, *r*, between the number of cards sent and the number of cards received by five friends. Any conclusions about the correlation coefficient in the underlying population—for instance, the population of all friends—must consider chance sampling variability as described by the sampling distribution of *r*.

Population Correlation Coefficient, ρ

A number between +1.00 and −1.00 that describes the linear relationship between pairs of quantitative variables for some population.

Null Hypothesis

Let's view the greeting card data for the five friends as if they were a random sample of pairs of observations from the population of all friends. Then it's possible to test the null hypothesis that the **population correlation coefficient,** symbolized by the Greek letter ρ (rho), equals zero. In other words, it is possible to test the hypothesis that in the population of all friends, there is no correlation between the number of cards sent and the number of cards received.

Focus on Relationships Instead of Mean Differences

These five pairs of observations also can be viewed as two related samples, since each observation in one sample is paired with a single observation in the other sample. Now, however, we wish to determine whether there is a *relationship* between the number of cards sent and received, not whether there is a *mean difference* between the number of cards sent and received. Accordingly, the appropriate measure is the correlation coefficient, not the sample mean difference, and the appropriate *t* test is for the population correlation coefficient, not for the population mean of difference scores.

t Test

A new t test must be used to determine whether the observed r of .80 qualifies as a common or a rare outcome under the null hypothesis that r equals zero. To obtain a value for the t ratio, use the following formula:

t RATIO FOR A SINGLE POPULATION CORRELATION COEFFICIENT

$$t = \frac{r - \rho_{hyp}}{\sqrt{\frac{1 - r^2}{n - 2}}} \qquad (15.8)$$

where r refers to the sample correlation coefficient (Formula 6.2); ρ_{hyp} refers to the hypothesized population correlation coefficient (which always must equal zero); and n refers to the number of pairs of observations. The expression in the denominator represents the estimated standard error of the sample correlation coefficient. As implied by the term at the bottom of this expression, the sampling distribution of t has $n - 2$ degrees of freedom. When pairs of observations are represented as points in a scatterplot, r assumes that the cluster of points approximates a straight line. Two degrees of freedom are lost because only $n - 2$ points are free to vary about some straight line that, itself, always depends on two points.

Importance of Sample Size

According to the present hypothesis test, the population correlation coefficient *could* equal zero.* This conclusion might seem surprising, given that an r of .80 was observed for the greeting card exchange. When the value of r is based on only five pairs of observations, as in the present example, its sampling variability is huge, and in fact, an r of .88 would have been required to reject the null hypothesis at the .05 level. Ordinarily, a serious investigation would use a larger sample size, preferably one that, with the aid of power curves similar to those described in Section 11.11, reflects the investigator's judgment about what would be the smallest important correlation.

If, in fact, a larger sample size had permitted the rejection of the null hypothesis, an r of .80 would have indicated a strong relationship. As mentioned in Chapter 6, values of r in the general vicinity of .10, .30, and .50 indicate weak, moderate, or strong relationships, respectively, according to Cohen's guidelines.

Progress Check *15.6 A random sample of 27 California taxpayers reveals an r of .43 between years of education and annual income. Use t to test the null hypothesis at the .05 level of significance that there is no relationship between educational level and annual income for the population of California taxpayers.

Answer on page 441.

Assumptions

When using the t test for the population correlation coefficient, you must assume that the relationship between the two variables, X and Y, can be described with a straight line and that the sample originates from a *normal bivariate population.* The

*Strictly speaking, since it could have originated from a population with a zero correlation, the current r of .80 should have been ignored in Chapters 6 and 7, where it was featured because of its computational simplicity.

HYPOTHESIS TEST SUMMARY
t TEST FOR A POPULATION CORRELATION COEFFICIENT
(Greeting Card Exchange)

Problem
Could there be a correlation between the number of cards sent and the number of cards received for the population of all friends?

Statistical Hypotheses

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$

Decision Rule
Reject H_0 at the .05 level of significance if $t \geq 3.182$ or if $t \leq -3.182$ (from Table B in Appendix C, given that $df = n - 2 = 5 - 2 = 3$).

Calculations
Given that $r = 0.80$ and $n = 5$:

$$t = \frac{.80 - 0}{\sqrt{\frac{1 - (.80)^2}{5 - 2}}} = \frac{.80}{\sqrt{\frac{1 - .64}{3}}} = \frac{.80}{\sqrt{\frac{.36}{3}}} = \frac{.80}{\sqrt{.12}} = \frac{.80}{.35} = 2.29$$

Decision
Retain H_0 at the .05 level of significance because $t = 2.29$ is less positive than 3.182.

Interpretation
The population correlation coefficient *could* equal zero. There is no evidence of a relationship between the number of cards sent and the number of cards received in the population of friends.

latter term means that the separate population distributions for each variable (X and Y) should be normal. When these assumptions are suspect—for instance, if the observed distribution for one variable appears to be extremely non-normal—the test results are only approximate and should be interpreted accordingly.

Summary

Variability due to individual differences can be eliminated by using repeated measures, that is, measuring the same subject twice. Whether because of repeated measures for the same subject or matched pairs of different subjects, two samples are related whenever each observation in one sample is paired, on a one-to-one basis, with a single observation in the other sample.

When using repeated measures, be aware of potential complications due to inadvertent interactions between conditions or to a lack of counterbalancing.

The statistical hypotheses must be selected from among the following three possibilities, where μ_D represents the population mean for all difference scores:

Nondirectional:

$$H_0: \mu_D = 0$$
$$H_1: \mu_D \neq 0$$

Directional, lower tail critical:

$$H_0: \mu_D \geq 0$$
$$H_1: \mu_D < 0$$

Directional, upper tail critical:

$$H_0: \mu_D \leq 0$$
$$H_1: \mu_D > 0$$

The t ratio for two related samples has a sampling distribution with $n - 1$ degrees of freedom, given that n equals the number of paired observations.

A confidence interval can be constructed for estimating μ_D. A single interpretation is possible only if the limits of the interval have the same sign, either both positive or both negative.

If the t test is statistically significant, Cohen's d can be used as a standardized estimate of effect size.

When using t for two related samples, you must assume that the population of difference scores is normally distributed. You need not be too concerned about violations of this assumption as long as sample sizes are relatively large.

To test the hypothesis that the population correlation coefficient equals zero, use a new t ratio with $n - 2$ degrees of freedom.

Important Terms

difference score (D)
Repeated measures
Two related samples
Counterbalancing
Population correlation coefficient (ρ)

Key Equations

t RATIO

$$t = \frac{\bar{D} - \mu_{D_{hyp}}}{s_{\bar{D}}}$$

where $s_{\bar{D}} = \frac{s_D}{\sqrt{n}}$

REVIEW QUESTIONS

***15.7** An educational psychologist wants to check the claim that regular physical exercise improves academic achievement. To control for academic aptitude, pairs of college students with similar GPAs are randomly assigned to either a treatment group that

NOTE: The relatively modest value of .07 for ϕ_c^2 compensates for the role of the very large sample size of 1291 in generating the highly significant χ^2 value of 88.65 and a minuscule approximate p-value of .000.

(c) $OR = \dfrac{299/280}{186/526} = \dfrac{1.07}{.35} = 3.06$

A cabin passenger is 3.06 times more likely to have survived than a steerage passenger.

Chapter 20

20.1 **(a)** 1, 2, 3, 4.5, 4.5, 6, 7, 9, 9, 9, 11
(b) 1, 2.5, 2.5, 4.5, 4.5, 6, 7, 8, 9, 10
(c) 1, 3.5, 3.5, 3.5, 3.5, 6, 7, 8.5, 8.5, 10

20.2 **(a)** *Research Problem*

Do therapy groups with directive leaders (1) produce more or less growth (in members) than therapy groups with nondirective leaders (2)?

Statistical Hypotheses

H_0: Population distribution 1 = Population distribution 2
H_1: Population distribution 1 ≠ Population distribution 2

Decision Rule

Reject H_0 at the .05 level of significance if $U \leq 5$, given $n_1 = 6$ and $n_2 = 6$.

Calculations

$U_1 = 31$
$U_2 = 5$
$U = 5$

Decision

Reject H_0 at the .05 level of significance because $U = 5$ equals 5.

Interpretation

Therapy groups with directive leaders produce less growth than those with nondirective leaders.

(b) $p = .05$ (since the calculated U equals the critical U for $p = .05$)

20.3 **(a)** Each distribution of difference scores tends to be non-normal, with "heavy" tails and a "light" middle.

(b) *Research Problem*

Does a quit-smoking workshop cause a decline in cigarette smoking?

Statistical Hypotheses

H_0: Population distribution 1 ≤ Population distribution 2
H_1: Population distribution 1 > Population distribution 2

NOTE: The directional H_1 assumes that both population distributions have roughly similar shapes.

Decision Rule

Reject H_0 at the .05 level (directional test) if T equals or is less than 5 given $n = 8$.

Calculations

$R_+ = 34$
$R_- = 2$
$T = 2$

Decision

Reject H_0 at the .05 level of significance because $T = 2$ is less than 5.

Interpretation

A quit-smoking workshop causes a decline in smoking.

(c) $p < .05$

(d) Smoking is significantly less after a quit-smoking workshop [$T\ (n = 8) = 2$, $p < .05$].

20.4 (a) Observed scores tend not to be normally distributed. There is no obvious cluster of scores in the middle range of each of the five groups.

(b) *Research Problem*

Are motion picture ratings associated with the number of violent or sexually explicit scenes in films?

Statistical Hypotheses

H_0: Population dist. NC-17 = Population dist. R = Population dist. PG-13 = Population dist. PG = Population dist. G

H_1: H_0 is false.

Decision Rule

Reject H_0 at the .05 level if $H \geq 9.49$, given $df = 4$.

Calculations

$$H = \frac{12}{25(25+1)}\left[\frac{(114)^2}{5} + \frac{(75.5)^2}{5} + \frac{(69)^2}{5} + \frac{(49.5)^2}{5} + \frac{(17)^2}{5}\right] - 3(25+1)$$

$$= .02[5239.30] - 78.00$$

$$= 26.79$$

Decision

Reject H_0 at the .05 level of significance because $H = 26.79$ exceeds 9.49.

Interpretation

Motion picture ratings are associated with the number of violent or sexually explicit scenes in films.

(c) $p < .001$

Chapter 21

21.1 Two-variable χ^2

21.2 t for two independent samples

21.3 t for two related samples (repeated measures)

21.5 One-variable χ^2

21.6 One-factor F

21.9 Two-factor F

21.10 t for two independent samples

21.11 t for correlation coefficient

APPENDIX C

Tables

Table A entries were computed by the second author.

Table B is taken from Table 12 of E. Pearson and H. Hartley (Eds.), *Biometrika Tables for Statisticians*, Vol. 1, 3rd ed. Cambridge: University Press, 1966, with permission of the Biometrika Trustees.

Table C is taken from *Statistical Methods*, by George W. Snedecor and William G. Cochran, 8th ed. Ames: Iowa State University Press, 1989, with permission of Wiley-Blackwell, Inc., a subsidiary of John Wiley & Sons, Inc.

Table D is taken from Table 8 of E. Pearson and H. Hartley (Eds.), *Biometrika Tables For Statisticians*, Vol. 1, 3rd. ed. Cambridge: University Press, 1966, with permission of the Biometrika Trustees.

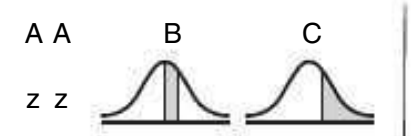
Table E is taken from the Bulletin of the Institute of Educational Research, 1953, Vol. No. 2, Indiana University, with permission of the publishers.

Table F is taken from F. Wilcoxon and R. A. Wilcox. *Some Rapid Approximate Statistical Procedures*, 2nd edition. Pearl River, New York: Lederle Laboratories. 1964, with permission of the American Cyanamid Company.

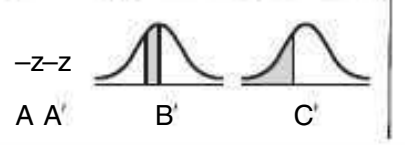
Table G is taken from Table 29 of E. Pearson and H. Hartley (Eds.), *Biometrika Tables for Statisticians*, Vol. 1, 3rd ed. Cambridge: University Press, 1966, with permission of the Biometrika Trustees.

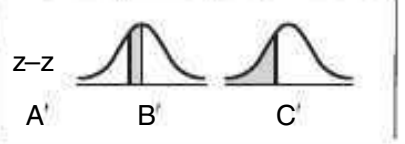
Table H reprinted from page 1 of *A. Million Random Digits with 100,000 Normal Deviates*, Rand, 1994. RP-295, 200 pp. Used by permission.

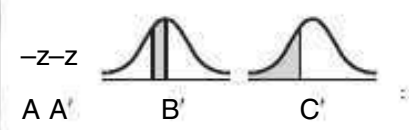
Table A[a]
PROPORTIONS (OF AREA) UNDER THE STANDARD NORMAL CURVE FOR VALUES OF *z*



z	A	B	C
0.00	.0000	.5000
0.01	.0040	.4960
0.02	.0080	.4920
0.03	.0120	.4880
0.04	.0160	.4840
0.05	.0199	.4801
0.06	.0239	.4761
0.07	.0279	.4721
0.08	.0319	.4681
0.09	.0359	.4641
0.10	.0398	.4602
0.11	.0438	.4562
0.12	.0478	.4522
0.13	.0517	.4483
0.14	.0557	.4443
0.15	.0596	.4404
0.16	.0636	.4364
0.17	.0675	.4325
0.18	.0714	.4286
0.19	.0753	.4247
0.20	.0793	.4207
0.21	.0832	.4168
0.22	.0871	.4129
0.23	.0910	.4090
0.24	.0948	.4052
0.25	.0987	.4013
0.26	.1026	.3974
0.27	.1064	.3936
0.28	.1103	.3897
0.29	.1141	.3859
0.30	.1179	.3821
0.31	.1217	.3783
0.32	.1255	.3745
0.33	.1293	.3707
0.34	.1331	.3669
0.35	.1368	.3632
0.36	.1406	.3594
0.37	.1443	.3557
0.38	.1480	.3520
0.39	.1517	.3483
0.40	.1554	.3446
0.41	.1591	.3409
0.42	.1628	.3372
0.43	.1664	.3336
0.44	.1700	.3300
0.45	.1736	.3264
0.46	.1772	.3228
0.47	.1808	.3192
0.48	.1844	.3156
0.49	.1879	.3121
0.50	.1915	.3085
0.51	.1950	.3050
0.52	.1985	.3015
0.53	.2019	.2981
0.54	.2054	.2946
0.55	.2088	.2912
0.56	.2123	.2877
0.57	.2157	.2843
0.58	.2190	.2810
0.59	.2224	.2776
0.60	.2257	.2743
0.61	.2291	.2709
0.62	.2324	.2676
0.63	.2357	.2643
0.64	.2389	.2611
0.65	.2422	.2578
0.66	.2454	.2546
0.67	.2486	.2514
0.68	.2517	.2483
0.69	.2549	.2451
0.70	.2580	.2420
0.71	.2611	.2389
0.72	.2642	.2358
0.73	.2673	.2327
0.74	.2704	.2296
0.75	.2734	.2266
0.76	.2764	.2236
0.77	.2794	.2206
0.78	.2823	.2177
0.79	.2852	.2148
0.80	.2881	.2119
0.81	.2910	.2090
0.82	.2939	.2061
0.83	.2967	.2033
0.84	.2995	.2005
0.85	.3023	.1977
0.86	.3051	.1949
0.87	.3078	.1922
0.88	.3106	.1894
0.89	.3133	.1867
0.90	.3159	.1841
0.91	.3186	.1814
0.92	.3212	.1788
0.93	.3238	.1762
0.94	.3264	.1736
0.95	.3289	.1711
0.96	.3315	.1685
0.97	.3340	.1660
0.98	.3365	.1635
0.99	.3389	.1611
1.00	.3413	.1587
1.01	.3438	.1562
1.02	.3461	.1539
1.03	.3485	.1515
1.04	.3508	.1492
1.05	.3531	.1469
1.06	.3554	.1446
1.07	.3577	.1423
1.08	.3599	.1401
1.09	.3621	.1379
1.10	.3643	.1357
1.11	.3665	.1335
1.12	.3686	.1314
1.13	.3708	.1292
1.14	.3729	.1271
1.15	.3749	.1251
1.16	.3770	.1230
1.17	.3790	.1210
1.18	.3810	.1190
1.19	.3830	.1170
1.20	.3849	.1151
1.21	.3869	.1131
1.22	.3888	.1112
1.23	.3907	.1093
1.24	.3925	.1075
1.25	.3944	.1056
1.26	.3962	.1038
1.27	.3980	.1020
1.28	.3997	.1003
1.29	.4015	.0985
1.30	.4032	.0968
1.31	.4049	.0951
1.32	.4066	.0934
1.33	.4082	.0918
1.34	.4099	.0901
1.35	.4115	.0885
1.36	.4131	.0869
1.37	.4147	.0853
1.38	.4162	.0838
1.39	.4177	.0823
1.40	.4192	.0808
1.41	.4207	.0793
1.42	.4222	.0778
1.43	.4236	.0764
1.44	.4251	.0749
1.45	.4265	.0735
1.46	.4279	.0721
1.47	.4292	.0708
1.48	.4306	.0694
1.49	.4319	.0681
1.50	.4332	.0668
1.51	.4345	.0655
1.52	.4357	.0643
1.53	.4370	.0630
1.54	.4382	.0618
1.55	.4394	.0606
1.56	.4406	.0594
1.57	.4418	.0582
1.58	.4429	.0571
1.59	.4441	.0559
1.60	.4452	.0548
1.61	.4463	.0537
1.62	.4474	.0526
1.63	.4484	.0516
1.64	.4495	.0505
1.65	.4505	.0495
1.66	.4515	.0485
1.67	.4525	.0475







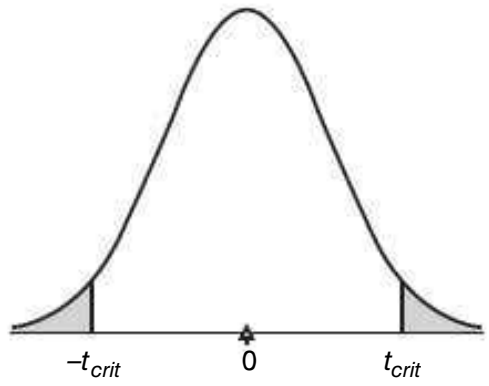
[a] *Discussed in Section 5.3.*

Table A[a] (Continued)
PROPORTIONS (OF AREA) UNDER THE STANDARD NORMAL CURVE FOR VALUES OF *z*

z	A	B	C
1.68	.4535	.0465
1.69	.4545	.0455
1.70	.4554	.0446
1.71	.4564	.0436
1.72	.4573	.0427
1.73	.4582	.0418
1.74	.4591	.0409
1.75	.4599	.0401
1.76	.4608	.0392
1.77	.4616	.0384
1.78	.4625	.0375
1.79	.4633	.0367
1.80	.4641	.0359
1.81	.4649	.0351
1.82	.4656	.0344
1.83	.4664	.0336
1.84	.4671	.0329
1.85	.4678	.0322
1.86	.4686	.0314
1.87	.4693	.0307
1.88	.4699	.0301
1.89	.4706	.0294
1.90	.4713	.0287
1.91	.4719	.0281
1.92	.4726	.0274
1.93	.4732	.0268
1.94	.4738	.0262
1.95	.4744	.0256
1.96	.4750	.0250
1.97	.4756	.0244
1.98	.4761	.0239
1.99	.4767	.0233
2.00	.4772	.0228
2.01	.4778	.0222
2.02	.4783	.0217
2.03	.4788	.0212
2.04	.4793	.0207
2.05	4798	.0202
2.06	.4803	.0197
2.07	.4808	.0192
2.08	.4812	.0188
2.09	.4817	.0183
2.10	.4821	.0179
2.11	.4826	.0174
2.12	.4830	.0170
2.13	.4834	.0166
2.14	.4838	.0162
2.15	.4842	.0158
2.16	.4846	.0154
2.17	.4850	.0150
2.18	.4854	.0146
2.19	.4857	.0143
2.20	.4861	.0139
2.21	.4864	.0136
2.22	4868	.0132
2.23	.4871	.0129
2.24	4875	.0125
2.25	.4878	.0122
2.26	.4881	.0119
2.27	.4884	.0116
2.28	.4887	.0113
2.29	.4890	.0110
2.30	.4893	.0107
2.31	.4896	.0104
2.32	.4898	.0102
2.33	.4901	.0099
2.34	.4904	.0096
2.35	.4906	.0094
2.36	.4909	.0091
2.37	.4911	.0089
2.38	.4913	.0087
2.39	.4916	.0084
2.40	.4918	.0082
2.41	.4920	.0080
2.42	.4922	.0078
2.43	.4925	.0075
2.44	.4927	.0073
2.45	.4929	.0071
2.46	.4931	.0069
2.47	.4932	.0068
2.48	.4934	.0066
2.49	.4936	.0064
2.50	.4938	.0062
2.51	.4940	.0060
2.52	.4941	.0059
2.53	.4943	.0057
2.54	.4945	.0055
2.55	.4946	.0054
2.56	.4948	.0052
2.57	.4949	.0051
2.58	.4951	.0049
2.59	.4952	.0048
2.60	.4953	.0047
2.61	.4955	.0045
2.62	.4956	.0044
2.63	.4957	.0043
2.64	.4959	.0041
2.65	.4960	.0040
2.66	.4961	.0039
2.67	.4962	.0038
2.68	.4963	.0037
2.69	.4964	.0036
2.70	.4965	.0035
2.71	.4966	.0034
2.72	.4967	.0033
2.73	.4968	.0032
2.74	.4969	.0031
2.75	.4970	.0030
2.76	.4971	.0029
2.77	.4972	.0028
2.78	.4973	.0027
2.79	.4974	.0026
2.80	.4974	.0026
2.81	.4975	.0025
2.82	.4976	.0024
2.83	.4977	.0023
2.84	.4977	.0023
2.85	.4978	.0022
2.86	.4979	.0021
2.87	.4979	.0021
2.88	.4980	.0020
2.89	.4981	.0019
2.90	.4981	.0019
2.91	.4982	.0018
2.92	.4982	.0018
2.93	.4983	.0017
2.94	.4984	.0016
2.95	.4984	.0016
2.96	.4985	.0015
2.97	.4985	.0015
2.98	.4986	.0014
2.99	.4986	.0014
3.00	.4987	.0013
3.01	.4987	.0013
3.02	.4987	.0013
3.03	.4988	.0012
3.04	.4988	.0012
3.05	.4989	.0011
3.06	.4989	.0011
3.07	.4989	.0011
3.08	.4990	.0010
3.09	.4990	.0010
3.10	.4990	.0010
3.11	.4991	.0009
3.12	.4991	.0009
3.13	.4991	.0009
3.14	.4992	.0008
3.15	.4992	.0008
3.16	.4992	.0008
3.17	.4992	.0008
3.18	.4993	.0007
3.19	.4993	.0007
3.20	.4993	.0007
3.21	.4993	.0007
3.22	.4994	.0006
3.23	.4994	.0006
3.24	.4994	.0006
3.25	.4994	.0006
3.30	.4995	.0005
3.35	.4996	.0004
3.40	.4997	.0003
3.45	.4997	.0003
3.50	.4998	.0002
3.60	.4998	.0002
3.70	.4999	.0001
3.80	.4999	.0001
3.90	.49995	.00005
4.00	.49997	.00003

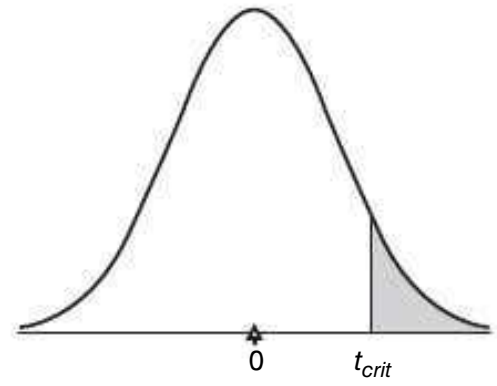
−z−z	A′	B′	C′

Table B[a]
CRITICAL VALUES OF *t*



Two-tailed or Nondirectional Test
LEVEL OF SIGNIFICANCE

	$p > .05$	$p < .05$	$p < .01$	$p < .001$
df	.05*	.01**	.001
1	12.706	63.657	636.62
2	4.303	9.925	31.598
3	3.182	5.841	12.924
4	2.776	4.604	8.610
5	2.571	4.032	6.869
6	2.447	3.707	5.959
7	2.365	3.499	5.408
8	2.306	3.355	5.041
9	2.262	3.250	4.781
10	2.228	3.169	4.587
11	2.201	3.106	4.437
12	2.179	3.055	4.318
13	2.160	3.012	4.221
14	2.145	2.977	4.140
15	2.131	2.947	4.073
16	2.120	2.921	4.015
17	2.110	2.898	3.965
18	2.101	2.878	3.922
19	2.093	2.861	3.883
20	2.086	2.845	3.850
21	2.080	2.831	3.819
22	2.074	2.819	3.792
23	2.069	2.807	3.767
24	2.064	2.797	3.745
25	2.060	2.787	3.725
26	2.056	2.779	3.707
27	2.052	2.771	3.690
28	2.048	2.763	3.674
29	2.045	2.756	3.659
30	2.042	2.750	3.646
40	2.021	2.704	3.551
60	2.000	2.660	3.460
120	1.980	2.617	3.373
∞	1.960	2.576	3.291

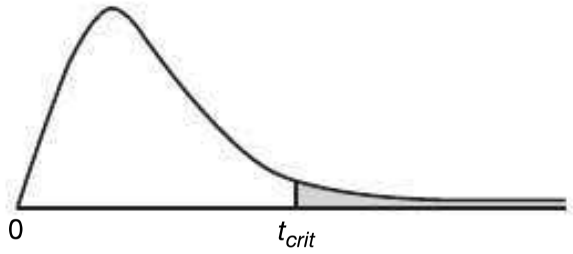


One-tailed or Directional Test
LEVEL OF SIGNIFICANCE

	$p > .05$	$p < .05$	$p < .01$	$p < .001$
df	.05	.01	.001
1	6.314	31.821	318.31
2	2.920	6.965	22.326
3	2.353	4.541	10.213
4	2.132	3.747	7.173
5	2.015	3.365	5.893
6	1.943	3.143	5.208
7	1.895	2.998	4.785
8	1.860	2.896	4.501
9	1.833	2.821	4.297
10	1.812	2.764	4.144
11	1.796	2.718	4.025
12	1.782	2.681	3.930
13	1.771	2.650	3.852
14	1.761	2.624	3.787
15	1.753	2.602	3.733
16	1.746	2.583	3.686
17	1.740	2.567	3.646
18	1.734	2.552	3.610
19	1.729	2.539	3.579
20	1.725	2.528	3.552
21	1.721	2.518	3.527
22	1.717	2.508	3.505
23	1.714	2.500	3.485
24	1.711	2.492	3.467
25	1.708	2.485	3.450
26	1.706	2.479	3.435
27	1.703	2.473	3.421
28	1.701	2.467	3.408
29	1.699	2.462	3.396
30	1.697	2.457	3.385
40	1.684	2.423	3.307
60	1.671	2.390	3.232
120	1.658	2.358	3.160
∞	1.645	2.326	3.090

[a] *Discussed in Section 13.2.*
** 95% level of confidence.*
*** 99% level of confidence.*

Table C[a]
CRITICAL VALUES OF *F*



FINDING *p*-VALUE
If observed *F* is
. . .smaller than light number, $p > .05$
. . .between light and dark numbers, $p < .05$
. . .larger than dark number, $p < .01$

.05 level of significance (light numbers)
.01 level of significance (**dark numbers**)

DEGREES OF FREEDOM IN NUMERATOR

DEGREES OF FREEDOM IN DENOMINATOR	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248	249	250	251	252	253	253	254	254	254
	4,052	**4,999**	**5,403**	**5,625**	**5,764**	**5,859**	**5,928**	**5,981**	**6,022**	**6,056**	**6,082**	**6,106**	**6,142**	**6,169**	**6,208**	**6,234**	**6,258**	**6,286**	**6,302**	**6,323**	**6,334**	**6,352**	**6,361**	**6,366**
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41	19.42	19.43	19.44	19.45	19.46	19.47	19.47	19.48	19.49	19.49	19.50	19.50
	98.49	**99.00**	**99.17**	**99.25**	**99.30**	**99.33**	**99.34**	**99.36**	**99.38**	**99.40**	**99.41**	**99.42**	**99.43**	**99.44**	**99.45**	**99.46**	**99.47**	**99.48**	**99.48**	**99.49**	**99.49**	**99.49**	**99.50**	**99.50**
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71	8.69	8.66	8.64	8.62	8.60	8.58	8.57	8.56	8.54	8.54	8.53
	34.12	**30.82**	**29.46**	**28.71**	**28.24**	**27.91**	**27.67**	**27.49**	**27.34**	**27.23**	**27.13**	**27.05**	**26.92**	**26.83**	**26.69**	**26.60**	**26.50**	**26.41**	**26.35**	**26.27**	**26.23**	**26.18**	**26.14**	**26.12**
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87	5.84	5.80	5.77	5.74	5.71	5.70	5.68	5.66	5.65	5.64	5.63
	21.20	**18.00**	**16.69**	**15.98**	**15.52**	**15.21**	**14.98**	**14.80**	**14.66**	**14.54**	**14.45**	**14.37**	**14.24**	**14.15**	**14.02**	**13.93**	**13.83**	**13.74**	**13.69**	**13.61**	**13.57**	**13.52**	**13.48**	**13.46**
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.64	4.60	4.56	4.53	4.50	4.46	4.44	4.42	4.40	4.38	4.37	4.36
	16.26	**13.27**	**12.06**	**11.39**	**10.97**	**10.67**	**10.45**	**10.27**	**10.15**	**10.05**	**9.96**	**9.89**	**9.77**	**9.68**	**9.55**	**9.47**	**9.38**	**9.29**	**9.24**	**9.17**	**9.13**	**9.07**	**9.04**	**9.02**
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68	3.67
	13.74	**10.92**	**9.78**	**9.15**	**8.75**	**8.47**	**8.26**	**8.10**	**7.98**	**7.87**	**7.79**	**7.72**	**7.60**	**7.52**	**7.39**	**7.31**	**7.23**	**7.14**	**7.09**	**7.02**	**6.99**	**6.94**	**6.90**	**6.88**
7	5.59	4.47	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.52	3.49	3.44	3.41	3.38	3.34	3.32	3.29	3.28	3.25	3.24	3.23
	12.25	**9.55**	**8.45**	**7.85**	**7.46**	**7.19**	**7.00**	**6.84**	**6.71**	**6.62**	**6.54**	**6.47**	**6.35**	**6.27**	**6.15**	**6.07**	**5.98**	**5.90**	**5.85**	**5.78**	**5.75**	**5.70**	**5.67**	**5.65**
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.23	3.20	3.15	3.12	3.08	3.05	3.03	3.0	2.98	2.96	2.94	2.93
	11.26	**8.65**	**7.59**	**7.01**	**6.63**	**6.37**	**6.19**	**6.03**	**5.91**	**5.82**	**5.74**	**5.67**	**5.56**	**5.48**	**5.36**	**5.28**	**5.20**	**5.11**	**5.06**	**5.00**	**4.96**	**4.91**	**4.88**	**4.86**
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07	3.02	2.98	2.93	2.90	2.86	2.82	2.80	2.77	2.76	2.73	2.72	2.71
	10.56	**8.02**	**6.99**	**6.42**	**6.06**	**5.80**	**5.62**	**5.47**	**5.35**	**5.26**	**5.18**	**5.11**	**5.00**	**4.92**	**4.80**	**4.73**	**4.64**	**4.56**	**4.51**	**4.45**	**4.41**	**4.36**	**4.33**	**4.31**
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77	2.74	2.70	2.67	2.64	2.61	2.59	2.56	2.55	2.54
	10.04	**7.56**	**6.55**	**5.99**	**5.64**	**5.39**	**5.21**	**5.06**	**4.95**	**4.85**	**4.78**	**4.71**	**4.60**	**4.52**	**4.41**	**4.33**	**4.25**	**4.17**	**4.12**	**4.05**	**4.01**	**3.96**	**3.93**	**3.91**
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79	2.74	2.70	2.65	2.61	2.57	2.53	2.50	2.47	2.45	2.42	2.41	2.40
	9.65	**7.20**	**6.22**	**5.67**	**5.32**	**5.07**	**4.88**	**4.74**	**4.63**	**4.54**	**4.46**	**4.40**	**4.29**	**4.21**	**4.10**	**4.02**	**3.94**	**3.86**	**3.80**	**3.74**	**3.70**	**3.66**	**3.62**	**3.60**
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69	2.64	2.60	2.54	2.50	2.46	2.42	2.40	2.36	2.35	2.32	2.31	2.30
	9.33	**6.93**	**5.95**	**5.41**	**5.06**	**4.82**	**4.65**	**4.50**	**4.39**	**4.30**	**4.22**	**4.16**	**4.05**	**3.98**	**3.86**	**3.78**	**3.70**	**3.61**	**3.56**	**3.49**	**3.46**	**3.41**	**3.38**	**3.36**
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60	2.55	2.51	2.46	2.42	2.38	2.34	2.32	2.28	2.26	2.24	2.22	2.21
	9.07	**6.70**	**5.74**	**5.20**	**4.86**	**4.62**	**4.44**	**4.30**	**4.19**	**4.10**	**4.02**	**3.96**	**3.85**	**3.78**	**3.67**	**3.59**	**3.51**	**3.42**	**3.37**	**3.30**	**3.27**	**3.21**	**3.18**	**3.16**

[a] *Discussed in Section 16.6.*

Table C[a] (Continued)
CRITICAL VALUES OF *F*

FINDING *p*-VALUE
If observed *F* is
... smaller than light number, $p > .05$
... between light and dark numbers, $p < .05$
... larger than dark number, $p < .01$

DEGREES OF FREEDOM IN NUMERATOR

DEGREES OF FREEDOM IN DENOMINATOR	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13
	8.86	**5.51**	**5.56**	**5.03**	**4.69**	**4.46**	**4.28**	**4.14**	**4.03**	**3.94**	**3.86**	**3.80**	**3.70**	**3.62**	**3.51**	**3.43**	**3.34**	**3.26**	**3.21**	**3.14**	**3.11**	**3.06**	**3.02**	**3.00**
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48	2.43	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.12	2.10	2.08	2.07
	8.68	**6.36**	**5.42**	**4.89**	**4.56**	**4.32**	**4.14**	**4.00**	**3.89**	**3.80**	**3.73**	**3.67**	**3.56**	**3.48**	**3.36**	**3.29**	**3.20**	**3.12**	**3.07**	**3.00**	**2.97**	**2.9**	**2.89**	**2.80**
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42	2.37	2.33	2.28	2.24	2.20	2.16	2.13	2.09	2.07	2.04	2.02	2.01
	8.53	**6.23**	**5.29**	**4.77**	**4.44**	**4.20**	**4.03**	**3.89**	**3.78**	**3.69**	**3.61**	**3.55**	**3.45**	**3.37**	**3.25**	**3.18**	**3.10**	**3.01**	**2.96**	**2.89**	**2.86**	**2.80**	**2.77**	**2.75**
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38	2.33	2.29	2.23	2.19	2.15	2.11	2.08	2.04	2.02	1.99	1.97	1.96
	8.40	**6.11**	**5.18**	**4.67**	**4.34**	**4.10**	**3.93**	**3.79**	**3.68**	**3.59**	**3.52**	**3.45**	**3.35**	**3.27**	**3.16**	**3.08**	**3.00**	**2.92**	**2.86**	**2.79**	**2.76**	**2.70**	**2.67**	**2.65**
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29	2.25	2.19	2.15	2.11	2.07	2.04	2.00	1.98	1.95	1.93	1.92
	8.28	**6.01**	**5.09**	**4.58**	**4.25**	**4.01**	**3.85**	**3.71**	**3.60**	**3.51**	**3.44**	**3.37**	**3.27**	**3.19**	**3.07**	**3.00**	**2.91**	**2.83**	**2.78**	**2.71**	**2.68**	**2.62**	**2.59**	**2.57**
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31	2.26	2.21	2.15	2.11	2.07	2.02	2.00	1.96	1.94	1.91	1.90	1.88
	8.18	**5.93**	**5.01**	**4.50**	**4.17**	**3.94**	**3.77**	**3.63**	**3.52**	**3.43**	**3.36**	**3.30**	**3.19**	**3.12**	**3.00**	**2.92**	**2.84**	**2.76**	**2.70**	**2.63**	**2.60**	**2.54**	**2.51**	**2.49**
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28	2.23	2.18	2.12	2.08	2.04	1.99	1.96	1.92	1.90	1.87	1.85	1.84
	8.10	**5.85**	**4.94**	**4.43**	**4.10**	**3.87**	**3.71**	**3.56**	**3.45**	**3.37**	**3.30**	**3.23**	**3.13**	**3.05**	**2.94**	**2.86**	**2.77**	**2.69**	**2.63**	**2.56**	**2.53**	**2.47**	**2.44**	**2.42**
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.20	2.15	2.09	2.05	2.00	1.96	1.93	1.89	1.87	1.84	1.82	1.81
	8.02	**5.78**	**4.87**	**4.37**	**4.04**	**3.81**	**3.65**	**3.51**	**3.40**	**3.31**	**3.24**	**3.17**	**3.07**	**2.99**	**2.88**	**2.80**	**2.72**	**2.63**	**2.58**	**2.51**	**2.47**	**2.42**	**2.38**	**2.36**
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26	2.23	2.18	2.13	2.07	2.03	1.98	1.93	1.91	1.87	1.84	1.81	1.80	1.78
	7.94	**5.72**	**4.82**	**4.31**	**3.99**	**3.76**	**3.59**	**3.45**	**3.35**	**3.26**	**3.18**	**3.12**	**3.02**	**2.94**	**2.83**	**2.75**	**2.67**	**2.58**	**2.53**	**2.46**	**2.42**	**2.37**	**2.33**	**2.31**
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.20	2.14	2.10	2.04	2.00	1.96	1.91	1.88	1.84	1.82	1.79	1.77	1.76
	7.88	**5.66**	**4.76**	**4.26**	**3.94**	**3.71**	**3.54**	**3.41**	**3.30**	**3.21**	**3.14**	**3.07**	**2.97**	**2.89**	**2.78**	**2.70**	**2.62**	**2.53**	**2.48**	**2.41**	**2.37**	**2.32**	**2.28**	**2.26**
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18	2.13	2.09	2.02	1.98	1.94	1.89	1.86	1.82	1.80	1.76	1.74	1.73
	7.82	**5.61**	**4.72**	**4.22**	**3.90**	**3.67**	**3.50**	**3.36**	**3.25**	**3.17**	**3.09**	**3.03**	**2.93**	**2.85**	**2.74**	**2.66**	**2.58**	**2.49**	**2.44**	**2.36**	**2.33**	**2.27**	**2.23**	**2.21**
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16	2.11	2.06	2.00	1.96	1.92	1.87	1.84	1.80	1.77	1.74	1.72	1.71
	7.77	**5.57**	**4.68**	**4.18**	**3.86**	**3.63**	**3.46**	**3.32**	**3.21**	**3.13**	**3.05**	**2.99**	**2.89**	**2.81**	**2.70**	**2.62**	**2.54**	**2.45**	**2.40**	**2.32**	**2.29**	**2.23**	**2.19**	**2.17**
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.10	2.05	1.99	1.95	1.90	1.86	1.82	1.78	1.76	1.72	1.70	1.69
	7.72	**5.53**	**4.64**	**4.14**	**3.82**	**3.59**	**3.42**	**3.29**	**3.17**	**3.09**	**3.02**	**2.96**	**2.86**	**2.77**	**2.66**	**2.58**	**2.50**	**2.41**	**2.36**	**2.28**	**2.25**	**2.19**	**2.15**	**2.13**

Table C[a] (Continued)
CRITICAL VALUES OF *F*

27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13	2.08	2.03	1.97	1.93	1.88	1.84	1.80	1.76	1.74	1.71	1.68	1.67
	7.68	**5.49**	**4.60**	**4.11**	**3.79**	**3.56**	**3.39**	**3.26**	**3.14**	**3.06**	**2.98**	**2.93**	**2.83**	**2.74**	**2.63**	**2.55**	**2.47**	**2.38**	**2.33**	**2.25**	**2.21**	**2.16**	**2.12**	**2.10**
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12	2.06	2.02	1.96	1.91	1.87	1.81	1.78	1.75	1.72	1.69	1.67	1.65
	7.64	**5.45**	**4.57**	**4.07**	**3.76**	**3.53**	**3.36**	**3.23**	**3.11**	**3.03**	**2.95**	**2.90**	**2.80**	**2.71**	**2.60**	**2.52**	**2.44**	**2.35**	**2.30**	**2.22**	**2.18**	**2.13**	**2.09**	**2.06**
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.05	2.00	1.94	1.90	1.85	1.80	1.77	1.73	1.71	1.68	1.65	1.64
	7.60	**5.42**	**4.54**	**4.04**	**3.73**	**3.50**	**3.33**	**3.20**	**3.08**	**3.00**	**2.92**	**2.87**	**2.77**	**2.68**	**2.57**	**2.49**	**2.41**	**2.32**	**2.27**	**2.19**	**2.15**	**2.10**	**2.06**	**2.03**
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09	2.04	1.99	1.93	1.89	1.84	1.79	1.76	1.72	1.69	1.66	1.64	1.62
	7.56	**5.39**	**4.51**	**4.02**	**3.70**	**3.47**	**3.30**	**3.17**	**3.06**	**2.98**	**2.90**	**2.84**	**2.74**	**2.66**	**2.55**	**2.47**	**2.38**	**2.29**	**2.24**	**2.16**	**2.13**	**2.07**	**2.03**	**2.01**
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07	2.02	1.97	1.91	1.86	1.82	1.76	1.74	1.69	1.67	1.64	1.61	1.59
	7.50	**5.34**	**4.46**	**3.97**	**3.66**	**3.42**	**3.25**	**3.12**	**3.01**	**2.94**	**2.86**	**2.80**	**2.70**	**2.62**	**2.51**	**2.42**	**2.34**	**2.25**	**2.20**	**2.12**	**2.08**	**2.02**	**1.98**	**1.96**
34	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.17	2.12	2.08	2.05	2.00	1.95	1.89	1.84	1.80	1.74	1.71	1.67	1.64	1.61	1.59	1.57
	7.44	**5.29**	**4.42**	**3.93**	**3.61**	**3.38**	**3.21**	**3.08**	**2.97**	**2.89**	**2.82**	**2.76**	**2.66**	**2.58**	**2.47**	**2.38**	**2.30**	**2.21**	**2.15**	**2.08**	**2.04**	**1.98**	**1.94**	**1.91**
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.03	1.98	1.93	1.87	1.82	1.78	1.72	1.69	1.65	1.62	1.59	1.56	1.55
	7.39	**5.25**	**4.38**	**3.89**	**3.58**	**3.35**	**3.18**	**3.04**	**2.94**	**2.86**	**2.78**	**2.72**	**2.62**	**2.54**	**2.43**	**2.35**	**2.26**	**2.17**	**2.12**	**2.04**	**2.00**	**1.94**	**1.90**	**1.87**
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.96	1.92	1.85	1.80	1.76	1.71	1.67	1.63	1.60	1.57	1.54	1.53
	7.35	**5.21**	**4.34**	**3.86**	**3.54**	**3.32**	**3.15**	**3.02**	**2.91**	**2.82**	**2.75**	**2.69**	**2.59**	**2.51**	**2.40**	**2.32**	**2.22**	**2.14**	**2.08**	**2.00**	**1.97**	**1.90**	**1.86**	**1.84**
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00	1.95	1.90	1.84	1.79	1.74	1.69	1.66	1.61	1.59	1.55	1.53	1.51
	7.31	**5.18**	**4.31**	**3.83**	**3.51**	**3.29**	**3.12**	**2.99**	**2.88**	**2.80**	**2.73**	**2.66**	**2.56**	**2.49**	**2.37**	**2.29**	**2.20**	**2.11**	**2.05**	**1.97**	**1.94**	**1.88**	**1.84**	**1.81**
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99	1.94	1.89	1.82	1.78	1.73	1.68	1.64	1.60	1.57	1.54	1.51	1.49
	7.27	**5.15**	**4.29**	**3.80**	**3.49**	**3.26**	**3.10**	**2.96**	**2.86**	**2.77**	**2.70**	**2.64**	**2.54**	**2.46**	**2.35**	**2.26**	**2.17**	**2.08**	**2.02**	**1.94**	**1.91**	**1.85**	**1.80**	**1.78**
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98	1.92	1.88	1.81	1.76	1.72	1.66	1.63	1.58	1.56	1.52	1.50	1.48
	7.24	**5.12**	**4.26**	**3.78**	**3.46**	**3.24**	**3.07**	**2.94**	**2.84**	**2.75**	**2.68**	**2.62**	**2.52**	**2.44**	**2.32**	**2.24**	**2.15**	**2.06**	**2.00**	**1.92**	**1.88**	**1.82**	**1.78**	**1.75**
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97	1.91	1.87	1.80	1.75	1.71	1.65	1.62	1.57	1.54	1.51	1.48	1.46
	7.21	**5.10**	**4.24**	**3.76**	**3.44**	**3.22**	**3.05**	**2.92**	**2.82**	**2.73**	**2.66**	**2.60**	**2.50**	**2.42**	**2.30**	**2.22**	**2.13**	**2.04**	**1.98**	**1.90**	**1.85**	**1.80**	**1.76**	**1.72**
48	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	1.99	1.96	1.90	1.86	1.79	1.74	1.70	1.64	1.61	1.56	1.53	1.50	1.47	1.45
	7.19	**5.08**	**4.22**	**3.74**	**3.42**	**3.20**	**3.04**	**2.90**	**2.80**	**2.71**	**2.64**	**2.58**	**2.48**	**2.40**	**2.28**	**2.20**	**2.11**	**2.02**	**1.96**	**1.88**	**1.84**	**1.78**	**1.73**	**1.70**
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95	1.90	1.85	1.78	1.74	1.69	1.63	1.60	1.55	1.52	1.48	1.46	1.44
	7.17	**5.06**	**4.20**	**3.72**	**3.41**	**3.18**	**3.02**	**2.88**	**2.78**	**2.70**	**2.62**	**2.56**	**2.46**	**2.39**	**2.26**	**2.18**	**2.10**	**2.00**	**1.94**	**1.86**	**1.82**	**1.76**	**1.71**	**1.68**
55	4.02	3.17	2.78	2.54	2.38	2.27	2.18	2.11	2.05	2.00	1.97	1.93	1.88	1.83	1.76	1.72	1.67	1.61	1.58	1.52	1.50	1.46	1.43	1.41
	7.12	**5.01**	**4.16**	**3.68**	**3.37**	**3.15**	**2.98**	**2.85**	**2.75**	**2.66**	**2.59**	**2.53**	**2.43**	**2.35**	**2.23**	**2.15**	**2.06**	**1.96**	**1.90**	**1.82**	**1.78**	**1.71**	**1.66**	**1.64**
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.86	1.81	1.75	1.70	1.65	1.59	1.56	1.50	1.48	1.44	1.41	1.39
	7.08	**4.98**	**4.13**	**3.65**	**3.34**	**3.12**	**2.95**	**2.82**	**2.72**	**2.63**	**2.56**	**2.50**	**2.40**	**2.32**	**2.20**	**2.12**	**2.03**	**1.93**	**1.87**	**1.79**	**1.74**	**1.68**	**1.63**	**1.60**
65	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.02	1.98	1.94	1.90	1.85	1.80	1.73	1.68	1.63	1.57	1.54	1.49	1.46	1.42	1.39	1.37
	7.04	**4.95**	**4.10**	**3.62**	**3.31**	**3.09**	**2.93**	**2.79**	**2.70**	**2.61**	**2.54**	**2.47**	**2.37**	**2.30**	**2.18**	**2.09**	**2.00**	**1.90**	**1.84**	**1.76**	**1.71**	**1.64**	**1.60**	**1.56**

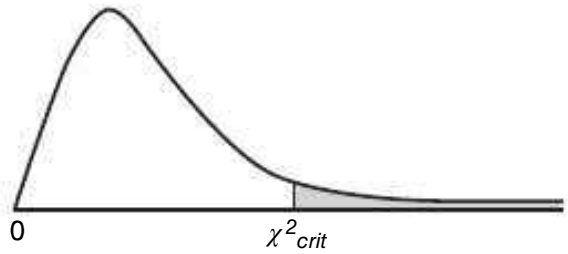
Table C[a] (Continued)
CRITICAL VALUES OF *F*

FINDING *p*-VALUE
If observed *F* is
... smaller than light number, $p > .05$
... between light and dark numbers, $p < .05$
... larger than dark number, $p < .01$

DEGREES OF FREEDOM IN NUMERATOR

DEGREES OF FREEDOM IN DENOMINATOR	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89	1.84	1.79	1.72	1.67	1.62	1.56	1.53	1.47	1.45	1.40	1.37	1.35
	7.01	**4.92**	**4.08**	**3.60**	**3.29**	**3.07**	**2.91**	**2.77**	**2.67**	**2.59**	**2.51**	**2.45**	**2.35**	**2.28**	**2.15**	**2.07**	**1.98**	**1.88**	**1.82**	**1.74**	**1.69**	**1.62**	**1.56**	**1.53**
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88	1.82	1.77	1.70	1.65	1.60	1.54	1.51	1.45	1.42	1.38	1.35	1.32
	6.96	**4.88**	**4.04**	**3.56**	**3.25**	**3.04**	**2.87**	**2.74**	**2.64**	**2.55**	**2.48**	**2.41**	**2.32**	**2.24**	**2.11**	**2.03**	**1.94**	**1.84**	**1.78**	**1.70**	**1.65**	**1.57**	**1.52**	**1.49**
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85	1.79	1.75	1.68	1.63	1.57	1.51	1.48	1.42	1.39	1.34	1.30	1.28
	6.90	**4.82**	**3.98**	**3.51**	**3.20**	**2.99**	**2.82**	**2.69**	**2.59**	**2.51**	**2.43**	**2.36**	**2.26**	**2.19**	**2.06**	**1.98**	**1.89**	**1.79**	**1.73**	**1.64**	**1.59**	**1.51**	**1.46**	**1.43**
125	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.95	1.90	1.86	1.83	1.77	1.72	1.65	1.60	1.55	1.49	1.45	1.39	1.36	1.31	1.27	1.25
	6.84	**4.78**	**3.94**	**3.47**	**3.17**	**2.95**	**2.79**	**2.65**	**2.56**	**2.47**	**2.40**	**2.33**	**2.23**	**2.15**	**2.03**	**1.94**	**1.85**	**1.75**	**1.68**	**1.59**	**1.54**	**1.46**	**1.40**	**1.37**
150	3.91	3.06	2.67	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82	1.76	1.71	1.64	1.59	1.54	1.47	1.44	1.37	1.34	1.29	1.25	1.22
	6.81	**4.75**	**3.91**	**3.44**	**3.14**	**2.92**	**2.76**	**2.62**	**2.53**	**2.44**	**2.37**	**2.30**	**2.20**	**2.12**	**2.00**	**1.91**	**1.83**	**1.72**	**1.66**	**1.56**	**1.51**	**1.43**	**1.37**	**1.33**
200	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.83	1.80	1.74	1.69	1.62	1.57	1.52	1.45	1.42	1.35	1.32	1.26	1.22	1.19
	6.76	**4.71**	**3.88**	**3.41**	**3.11**	**2.90**	**2.73**	**2.60**	**2.50**	**2.41**	**2.34**	**2.28**	**2.17**	**2.09**	**1.97**	**1.88**	**1.79**	**1.69**	**1.62**	**1.53**	**1.48**	**1.39**	**1.33**	**1.28**
400	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.78	1.72	1.67	1.60	1.54	1.49	1.42	1.38	1.32	1.28	1.22	1.16	1.13
	6.70	**4.66**	**3.83**	**3.36**	**3.06**	**2.85**	**2.69**	**2.55**	**2.46**	**2.37**	**2.29**	**2.23**	**2.12**	**2.04**	**1.92**	**1.84**	**1.74**	**1.64**	**1.57**	**1.47**	**1.42**	**1.32**	**1.24**	**1.19**
1000	3.85	3.00	2.61	2.38	2.22	2.10	2.02	1.95	1.89	1.84	1.80	1.76	1.70	1.65	1.58	1.53	1.47	1.41	1.36	1.30	1.26	1.19	1.13	1.08
	6.66	**4.62**	**3.80**	**3.34**	**3.04**	**2.82**	**2.66**	**2.53**	**2.43**	**2.34**	**2.26**	**2.20**	**2.09**	**2.01**	**1.89**	**1.81**	**1.71**	**1.61**	**1.54**	**1.44**	**1.38**	**1.28**	**1.19**	**1.11**
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.86	1.83	1.79	1.75	1.69	1.64	1.57	1.52	1.46	1.40	1.35	1.28	1.24	1.17	1.11	1.00
	6.64	**4.00**	**3.78**	**3.32**	**3.02**	**2.80**	**2.64**	**2.51**	**2.41**	**2.32**	**2.24**	**2.18**	**2.07**	**1.99**	**1.87**	**1.79**	**1.69**	**1.59**	**1.52**	**1.41**	**1.36**	**1.25**	**1.15**	**1.00**

Table D[a]
CRITICAL VALUES OF χ^2



LEVEL OF SIGNIFICANCE

	$p > .10$	$p < .10$	$p < .05$	$p < .01$	$p < .001$
df	**.10**	**.05**	**.01**	**.001**
1	2.71	3.84	6.64	10.83
2	4.60	5.99	9.21	13.82
3	6.25	7.81	11.34	16.27
4	7.78	9.49	13.28	18.47
5	9.24	11.07	15.09	20.52
6	10.64	12.59	16.81	22.46
7	12.02	14.07	18.48	24.32
8	13.36	15.51	20.09	26.12
9	14.68	16.92	21.67	27.88
10	15.99	18.31	23.21	29.59
11	17.28	19.68	24.72	31.26
12	18.55	21.03	26.22	32.91
13	19.81	22.36	27.69	34.53
14	21.06	23.68	29.14	36.12
15	22.31	25.00	30.58	37.70
16	23.54	26.30	32.00	39.25
17	24.77	27.59	33.41	40.79
18	25.99	28.87	34.80	42.31
19	27.20	30.14	36.19	43.82
20	28.41	31.41	37.57	45.32
21	29.62	32.67	38.93	46.80
22	30.81	33.92	40.29	48.27
23	32.01	35.17	41.64	49.73
24	33.20	36.42	42.98	51.18
25	34.38	37.65	44.31	52.62
26	35.56	38.88	45.64	54.05
27	36.74	40.11	46.96	55.48
28	37.92	41.34	48.28	56.89
29	39.09	42.56	49.59	58.30
30	40.26	43.77	50.89	59.70
40	51.80	55.76	63.69	73.40
50	63.17	67.50	76.15	86.66
60	74.40	79.08	88.38	99.61
70	85.53	90.53	100.42	112.32

[a] *Discussed in Section 19.4.*

Table E[a]
CRITICAL VALUES OF MANN–WHITNEY *U*

FINDING *p*-VALUE
If observed *U* is
…larger than light number, $p > .05$
…between light and dark numbers, $p < .05$
…smaller than dark numbers, $p < .01$

NONDIRECTIONAL TEST
.05 level of significance (light numbers)
.01 level of significance **(dark numbers)**

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2	—	—	—	—	—	—	—	0	0	0	0	1	1	1	1	1	2	2	2	2
	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	**0**	**0**
3	—	—	—	—	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
	—	—	—	—	—	—	—	—	**0**	**0**	**0**	**1**	**1**	**1**	**2**	**2**	**2**	**2**	**3**	**3**
4	—	—	—	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
	—	—	—	—	—	**0**	**0**	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**5**	**5**	**6**	**6**	**7**	**8**
5	—	—	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
	—	—	—	—	**0**	**1**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
6	—	—	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
	—	—	—	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**9**	**10**	**11**	**12**	**13**	**15**	**16**	**17**	**18**
7	—	—	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
	—	—	—	**0**	**1**	**3**	**4**	**6**	**7**	**9**	**10**	**12**	**13**	**15**	**16**	**18**	**19**	**21**	**22**	**24**
8	—	0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
	—	—	—	**1**	**2**	**4**	**6**	**7**	**9**	**11**	**13**	**15**	**17**	**18**	**20**	**22**	**24**	**26**	**28**	**30**
9	—	0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
	—	—	**0**	**1**	**3**	**5**	**7**	**9**	**11**	**13**	**16**	**18**	**20**	**22**	**24**	**27**	**29**	**31**	**33**	**36**
10	—	0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
	—	—	**0**	**2**	**4**	**6**	**9**	**11**	**13**	**16**	**18**	**21**	**24**	**26**	**29**	**31**	**34**	**37**	**39**	**42**
11	—	0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
	—	—	**0**	**2**	**5**	**7**	**10**	**13**	**16**	**18**	**21**	**24**	**27**	**30**	**33**	**36**	**39**	**42**	**45**	**48**
12	—	1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
	—	—	**1**	**3**	**6**	**9**	**12**	**15**	**18**	**21**	**24**	**27**	**31**	**34**	**37**	**41**	**44**	**47**	**51**	**54**
13	—	1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
	—	—	**1**	**3**	**7**	**10**	**13**	**17**	**20**	**24**	**27**	**31**	**34**	**38**	**42**	**45**	**49**	**53**	**56**	**60**
14	—	1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
	—	—	**1**	**4**	**7**	**11**	**15**	**18**	**22**	**26**	**30**	**34**	**38**	**42**	**46**	**50**	**54**	**58**	**63**	**67**
15	—	1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
	—	—	**2**	**5**	**8**	**12**	**16**	**20**	**24**	**29**	**33**	**37**	**42**	**46**	**51**	**55**	**60**	**64**	**69**	**73**
16	—	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	96
	—	—	**2**	**5**	**9**	**13**	**18**	**22**	**27**	**31**	**36**	**41**	**45**	**50**	**55**	**60**	**65**	**70**	**74**	**79**
17	—	2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
	—	—	**2**	**6**	**10**	**15**	**19**	**24**	**29**	**34**	**39**	**44**	**49**	**54**	**60**	**65**	**70**	**75**	**81**	**86**
18	—	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
	—	—	**2**	**6**	**11**	**16**	**21**	**26**	**31**	**37**	**42**	**47**	**53**	**58**	**64**	**70**	**75**	**81**	**87**	**92**
19	—	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
	—	**0**	**3**	**7**	**12**	**17**	**22**	**28**	**33**	**39**	**45**	**51**	**56**	**63**	**69**	**74**	**81**	**87**	**93**	**99**
20	—	2	8	13	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127
	—	**0**	**3**	**8**	**13**	**18**	**24**	**30**	**36**	**42**	**48**	**54**	**60**	**67**	**73**	**79**	**86**	**92**	**99**	**105**

[a] *Discussed in Section 20.3. To be significant, the observed U must equal or be less than the value shown in the table. Dashes in the table indicate that no decision is possible at the specified level of significance.*

Table E[a] (Continued)
CRITICAL VALUES OF MANN–WHITNEY *U*

DIRECTIONAL TEST
.05 level of significance (light numbers)
.01 level of significance **(dark numbers)**

n_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0	0
	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2	—	—	—	—	0	0	0	1	1	1	1	2	2	2	3	3	3	4	4	4
	—	—	—	—	—	—	—	—	—	—	—	—	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**
3	—	—	0	0	1	2	2	3	3	4	5	5	6	7	7	8	9	9	10	11
	—	—	—	—	—	—	**0**	**0**	**1**	**1**	**1**	**2**	**2**	**2**	**3**	**3**	**4**	**4**	**4**	**5**
4	—	—	0	1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
	—	—	—	—	**0**	**1**	**1**	**2**	**3**	**3**	**4**	**5**	**5**	**6**	**7**	**7**	**8**	**9**	**9**	**10**
5	—	0	1	2	4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
	—	—	—	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
6	—	0	2	3	5	7	8	10	12	14	16	17	19	21	23	25	26	28	30	32
	—	—	—	**1**	**2**	**3**	**4**	**6**	**7**	**8**	**9**	**11**	**12**	**13**	**15**	**16**	**18**	**19**	**20**	**22**
7	—	0	2	4	6	8	11	13	15	17	19	21	24	26	28	30	33	35	37	39
	—	—	**0**	**1**	**3**	**4**	**6**	**7**	**9**	**11**	**12**	**14**	**16**	**17**	**19**	**21**	**23**	**24**	**26**	**28**
8	—	1	3	5	8	10	13	15	18	20	23	26	28	31	33	36	39	41	44	47
	—	—	**0**	**2**	**4**	**6**	**7**	**9**	**11**	**13**	**15**	**17**	**20**	**22**	**24**	**26**	**28**	**30**	**32**	**34**
9	—	1	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51	54
	—	—	**1**	**3**	**5**	**7**	**9**	**11**	**14**	**16**	**18**	**21**	**23**	**26**	**28**	**31**	**33**	**36**	**38**	**40**
10	—	1	4	7	11	14	17	20	24	27	31	34	37	41	44	48	51	55	58	62
	—	—	**1**	**3**	**6**	**8**	**11**	**13**	**16**	**19**	**22**	**24**	**27**	**30**	**33**	**36**	**38**	**41**	**44**	**47**
11	—	1	5	8	12	16	19	23	27	31	34	38	42	46	50	54	57	61	65	69
	—	—	**1**	**4**	**7**	**9**	**12**	**15**	**18**	**22**	**25**	**28**	**31**	**34**	**37**	**41**	**44**	**47**	**50**	**53**
12	—	2	5	9	13	17	21	26	30	34	38	42	47	51	55	60	64	68	72	77
	—	—	**2**	**5**	**8**	**11**	**14**	**17**	**21**	**24**	**28**	**31**	**35**	**38**	**42**	**46**	**49**	**53**	**56**	**60**
13	—	2	6	10	15	19	24	28	33	37	42	47	51	56	61	65	70	75	80	84
	—	**0**	**2**	**5**	**9**	**12**	**16**	**20**	**23**	**27**	**31**	**35**	**39**	**43**	**47**	**51**	**55**	**59**	**63**	**67**
14	—	2	7	11	16	21	26	31	36	41	46	51	56	61	66	71	77	82	87	92
	—	**0**	**2**	**6**	**10**	**13**	**17**	**22**	**26**	**30**	**34**	**38**	**43**	**47**	**51**	**56**	**60**	**65**	**69**	**73**
15	—	3	7	12	18	23	28	33	39	44	50	55	61	66	72	77	83	88	94	100
	—	**0**	**3**	**7**	**11**	**15**	**19**	**24**	**28**	**33**	**37**	**42**	**47**	**51**	**56**	**61**	**66**	**70**	**75**	**80**
16	—	3	8	14	19	25	30	36	42	48	54	60	65	71	77	83	89	95	101	107
	—	**0**	**3**	**7**	**12**	**16**	**21**	**26**	**31**	**36**	**41**	**46**	**51**	**56**	**61**	**66**	**71**	**76**	**82**	**87**
17	—	3	9	15	20	26	33	39	45	51	57	64	70	77	83	89	96	102	109	115
	—	**0**	**4**	**8**	**13**	**18**	**23**	**28**	**33**	**38**	**44**	**49**	**55**	**60**	**66**	**71**	**77**	**82**	**88**	**93**
18	—	4	9	16	22	28	35	41	48	55	61	68	75	82	88	95	102	109	116	123
	—	**0**	**4**	**9**	**14**	**19**	**24**	**30**	**36**	**41**	**47**	**53**	**59**	**65**	**70**	**76**	**82**	**88**	**94**	**100**
19	0	4	10	17	23	30	37	44	51	58	65	72	80	87	94	101	109	116	123	130
	—	**1**	**4**	**9**	**15**	**20**	**26**	**32**	**38**	**44**	**50**	**56**	**63**	**69**	**75**	**82**	**88**	**94**	**101**	**107**
20	0	4	11	18	25	32	39	47	54	62	69	77	84	92	100	107	115	123	130	138
	—	**1**	**5**	**10**	**16**	**22**	**28**	**34**	**40**	**47**	**53**	**60**	**67**	**73**	**80**	**87**	**93**	**100**	**107**	**114**

Table F[a]
CRITICAL VALUES OF WILCOXON *T*

FINDING *p*-VALUE
If observed *T* is
...larger than .05 number, $p > .05$
...between .05 and .01 numbers, $p < .05$
...smaller than .01 number, $p < .01$

	LEVEL OF SIGNIFICANCE: NONDIRECTIONAL TEST					LEVEL OF SIGNIFICANCE: DIRECTIONAL TEST				
	.05	.01		.05	.01	.05	.01		.05	.01
n			*n*			*n*			*n*	
5	—	—	28	116	91	5	0	—	28	130	101
6	0	—	29	126	100	6	2	—	29	140	110
7	2	—	30	137	109	7	3	0	30	151	120
8	3	0	31	147	118	8	5	1	31	163	130
9	5	1	32	159	128	9	8	3	32	175	140
10	8	3	33	170	138	10	10	5	33	187	151
11	10	5	34	182	148	11	13	7	34	200	162
12	13	7	35	195	159	12	17	9	35	213	173
13	17	9	36	208	171	13	21	12	36	227	185
14	21	12	37	221	182	14	25	15	37	241	198
15	25	15	38	235	194	15	30	19	38	256	211
16	29	19	39	249	207	16	35	23	39	271	224
17	34	23	40	264	220	17	41	27	40	286	238
18	40	27	41	279	233	18	47	32	41	302	252
19	46	32	42	294	247	19	53	37	42	319	266
20	52	37	43	310	261	20	60	43	43	336	281
21	58	42	44	327	276	21	67	49	44	353	296
22	65	48	45	343	291	22	75	55	45	371	312
23	73	54	46	361	307	23	83	62	46	389	328
24	81	61	47	378	322	24	91	69	47	407	345
25	89	68	48	396	339	25	100	76	48	426	362
26	98	75	49	415	355	26	110	84	49	446	379
27	107	83	50	434	373	27	119	92	50	466	397

[a] *Discussed in Section 20.4. To be significant, the observed T must equal or be less than the value shown in the table. Dashes in the table indicate that no decision is possible at the specified level of significance.*

Table G[a]
CRITICAL VALUES OF *q* FOR TUKEY'S *HSD* TEST

.05 level of significance (light numbers)
.01 level of significance **(dark numbers)**

ERROR		NUMBER OF MEANS (*k*)									
df	α	2	3	4	5	6	7	8	9	10	11
2	.05	6.08	8.33	9.80	10.9	11.7	12.4	13.0	13.5	14.0	14.4
	.01	**14.0**	**19.0**	**22.3**	**24.7**	**26.6**	**28.2**	**29.5**	**30.7**	**31.7**	**32.6**
3	.05	4.50	5.91	6.82	7.50	8.04	8.48	8.85	9.18	9.46	9.72
	.01	**8.26**	**10.6**	**12.2**	**13.3**	**14.2**	**15.0**	**15.6**	**16.2**	**16.7**	**17.8**
4	.05	3.93	5.04	5.76	6.29	6.71	7.05	7.35	7.60	7.83	8.03
	.01	**6.51**	**8.12**	**9.17**	**9.96**	**10.6**	**11.1**	**11.5**	**11.9**	**12.3**	**12.6**
5	.05	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17
	.01	**5.70**	**6.98**	**7.80**	**8.42**	**8.91**	**9.32**	**9.67**	**9.97**	**10.24**	**10.48**
6	.05	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49	6.65
	.01	**5.24**	**6.33**	**7.03**	**7.56**	**7.97**	**8.32**	**8.61**	**8.87**	**9.10**	**9.30**
7	.05	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30
	.01	**4.95**	**5.92**	**6.54**	**7.01**	**7.37**	**7.68**	**7.94**	**8.17**	**8.37**	**8.55**
8	.05	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05
	.01	**4.75**	**5.64**	**6.20**	**6.62**	**6.96**	**7.24**	**7.47**	**7.68**	**7.86**	**8.03**
9	.05	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74	5.87
	.01	**4.60**	**5.43**	**5.96**	**6.35**	**6.66**	**6.91**	**7.13**	**7.33**	**7.49**	**7.65**
10	.05	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72
	.01	**4.48**	**5.27**	**5.77**	**6.14**	**6.43**	**6.67**	**6.87**	**7.05**	**7.21**	**7.36**
11	.05	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61
	.01	**4.39**	**5.15**	**5.62**	**5.97**	**6.25**	**6.48**	**6.67**	**6.84**	**6.99**	**7.13**
12	.05	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39	5.51
	.01	**4.32**	**5.05**	**5.50**	**5.84**	**6.10**	**6.32**	**6.51**	**6.67**	**6.81**	**6.94**
13	.05	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43
	.01	**4.26**	**4.96**	**5.40**	**5.73**	**5.98**	**6.19**	**6.37**	**6.53**	**6.67**	**6.79**
14	.05	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36
	.01	**4.21**	**4.89**	**5.32**	**5.63**	**5.88**	**6.08**	**6.26**	**6.41**	**6.54**	**6.66**
15	.05	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.31
	.01	**4.17**	**4.84**	**5.25**	**5.56**	**5.80**	**5.99**	**6.16**	**6.31**	**6.44**	**6.55**
16	.05	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26
	.01	**4.13**	**4.79**	**5.19**	**5.49**	**5.72**	**5.92**	**6.08**	**6.22**	**6.35**	**6.46**
17	.05	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11	5.21
	.01	**4.10**	**4.74**	**5.14**	**5.43**	**5.66**	**5.85**	**6.01**	**6.15**	**6.27**	**6.38**
18	.05	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17
	.01	**4.07**	**4.70**	**5.09**	**5.38**	**5.60**	**5.79**	**5.94**	**6.08**	**6.20**	**6.31**
19	.05	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14
	.01	**4.05**	**4.67**	**5.05**	**5.33**	**5.55**	**5.73**	**5.89**	**6.02**	**6.14**	**6.25**
20	.05	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11
	.01	**4.02**	**4.64**	**5.02**	**5.29**	**5.51**	**5.69**	**5.84**	**5.97**	**6.09**	**6.19**
24	.05	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01
	.01	**3.96**	**4.55**	**4.91**	**5.17**	**5.37**	**5.54**	**5.69**	**5.81**	**5.92**	**6.02**
30	.05	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	4.92
	.01	**3.89**	**4.45**	**4.80**	**5.05**	**5.24**	**5.40**	**5.54**	**5.65**	**5.76**	**5.85**
40	.05	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	4.82
	.01	**3.82**	**4.37**	**4.70**	**4.93**	**5.11**	**5.26**	**5.39**	**5.50**	**5.60**	**5.69**
60	.05	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73
	.01	**3.76**	**4.28**	**4.59**	**4.82**	**4.99**	**5.13**	**5.25**	**5.36**	**5.45**	**5.53**
120	.05	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56	4.64
	.01	**3.70**	**4.20**	**4.50**	**4.71**	**4.87**	**5.01**	**5.12**	**5.21**	**5.30**	**5.37**
∞	.05	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55
	.01	**3.64**	**4.12**	**4.40**	**4.60**	**4.76**	**4.88**	**4.99**	**5.08**	**5.16**	**5.23**

[a] *Discussed in Section 16.10.*

Table H[a]
RANDOM NUMBERS

ROW NUMBER										
00000	10097	32533	76520	13586	34673	54876	80959	09117	39292	74945
00001	37542	04805	64894	74296	24805	24037	20636	10402	00822	91665
00002	08422	68953	19645	09303	23209	02560	15953	34764	35080	33606
00003	99019	02529	09376	70715	38311	31165	88676	74397	04436	27659
00004	12807	99970	80157	36147	64032	36653	98951	16877	12171	76833
00005	66065	74717	34072	76850	36697	36170	65813	39885	11199	29170
00006	31060	10805	45571	82406	35303	42614	86799	07439	23403	09732
00007	85269	77602	02051	65692	68665	74818	73053	85247	18623	88579
00008	63573	32135	05325	47048	90553	57548	28468	28709	83491	25624
00009	73796	45753	03529	64778	35808	34282	60935	20344	35273	88435
00010	98520	17767	14905	68607	22109	40558	60970	93433	50500	73998
00011	11805	05431	39808	27732	50725	68248	29405	24201	52775	67851
00012	83452	99634	06288	98033	13746	70078	18475	40610	68711	77817
00013	88685	40200	86507	58401	36766	67951	90364	76493	29609	11062
00014	99594	67348	87517	64969	91826	08928	93785	61368	23478	34113
00015	65481	17674	17468	50950	58047	76974	73039	57186	40218	16544
00016	80124	35635	17727	08015	45318	22374	21115	78253	14385	53763
00017	74350	99817	77402	77214	43236	00210	45521	64237	96286	02655
00018	69916	26803	66252	29148	36936	87203	76621	13990	94400	56418
00019	09893	20505	14225	68514	46427	56788	96297	78822	54382	14598
00020	91499	14523	68479	27686	46162	83554	94750	89923	37089	20048
00021	80336	94598	26940	36858	70297	34135	53140	33340	42050	82341
00022	44104	81949	85157	47954	32979	26575	57600	40881	22222	06413
00023	12550	73742	11100	02040	12860	74697	96644	89439	28707	25815
00024	63606	49329	16505	34484	40219	52563	43651	77082	07207	31790
00025	61196	90446	26457	47774	51924	33729	65394	59593	42582	60527
00026	15474	45266	95270	79953	59367	83848	82396	10118	33211	59466
00027	94557	28573	67897	54387	54622	44431	91190	42592	92927	45973
00028	42481	16213	97344	08721	16868	48767	03071	12059	25701	46670
00029	23523	78317	73208	89837	68935	91416	26252	29663	05522	82562
00030	04493	52494	75246	33824	45862	51025	61962	79335	65337	12472
00031	00549	97654	64051	88159	96119	63896	54692	82391	23287	29529
00032	35963	15307	26898	09354	33351	35462	77974	50024	90130	39333
00033	59808	08391	45427	26842	83609	49700	13021	24892	78565	20106
00034	46058	85236	01390	92286	77281	44077	93910	83647	70617	42941
00035	32179	00597	87379	25241	05567	07007	86743	17157	85394	11838
00036	69234	61406	20117	45204	15956	60000	18743	92423	97188	96338
00037	19565	41430	01758	75379	40419	21585	66674	36806	84962	85207
00038	45155	14938	19476	07246	43667	94543	59047	90033	20826	69541
00039	94864	31994	36168	10851	34888	81553	01540	35456	05014	51176
00040	98086	24826	45240	28404	44999	08896	39094	73407	35441	31880
00041	33185	16232	41941	50949	89435	48581	88695	41944	37548	73043
00042	80951	00406	96382	70774	20151	23387	25016	25298	94624	61171
00043	79752	49140	71961	28296	69861	02591	74852	20539	00387	59579
00044	18633	32537	98145	06571	31010	24674	05455	61427	77938	91936
00045	74029	43902	77557	32270	97790	17119	52527	58021	80814	51748
00046	54178	45611	80993	37143	05335	12969	56127	19255	36040	90324
00047	11664	49883	52079	84827	59381	71539	09973	33440	88461	23356
00048	48324	77928	31249	64710	02295	36870	32307	57546	15020	09994
00049	69074	94138	87637	91976	35584	04401	10518	21615	01848	76938
00050	09188	20097	32825	39527	04220	86304	83389	87374	64278	58044
00051	90045	85497	51981	50654	94938	81997	91870	76150	68476	64659
00052	73189	50207	47677	26269	62290	64464	27124	67018	41361	82760
00053	75768	76490	20971	87749	90429	12272	95375	05871	93823	43178
00054	54016	44056	66281	31003	00682	27398	20714	53295	07706	17813
00055	08358	69910	78542	42785	13661	58873	04618	97553	31223	08420
00056	28306	03264	81333	10591	40510	07893	32604	60475	94119	01840
00057	53840	86233	81594	13628	51215	90290	28466	68795	77762	20791
00058	91757	53471	61613	62669	50263	90212	55781	76514	83483	47055
00059	89415	92694	00397	58391	12607	17646	48949	72306	94541	37408

[a] *Discussed in Section 8.4.*

APPENDIX D

Glossary

Glossary

Numbers in parentheses indicate the section in which the term is introduced.

Addition rule: Add together the separate probabilities of several mutually exclusive events to find the probability that any one of these events will occur. (8.8)

Alpha (α): The probability of a type I error, that is, the probability of rejecting a true null hypothesis. (11.7) Also see *level of significance.*

Alternative hypothesis (H_1): The opposite of the null hypothesis. Often identified with the research hypothesis. (10.6)

Analysis of variance (ANOVA): An overall test of the null hypothesis for more than two population means. (16.1)

Approximate numbers: Occur whenever numbers are rounded off, as is always the case with values for continuous variables. (1.6)

Bar graph: Bar-type graph for qualitative data with gaps between adjacent bars. (2.10)

Beta (β): The probability of a type II error, that is, the probability of retaining a false null hypothesis. (11.8)

Bimodal: Describes any distribution with two obvious peaks. (3.1)

Central limit theorem: Regardless of the population shape, the shape of the sampling distribution of the mean will approximate a normal curve if the sample size is sufficiently large. (9.6)

Conditional probability: The probability of one event, given the occurrence of another event. (8.9)

Confidence interval (CI): A range of values that, with a known degree of certainty, includes an unknown population characteristic, such as a population mean. (12.2)

Confidence interval for $\mu_1 - \mu_2$ (or μ_D): A range of values that, in the long run, includes the unknown effect (difference between population means) a certain percent of the time. (14.8)

Confounding variable: An uncontrolled variable that compromises the interpretation of a study. (1.6)

Constant: A characteristic or property that can take on only one value (1.6)

Continuous variable: A variable that consists of numbers whose values, at least in theory, have no restrictions. (1.6)

Correlation coefficient: See *Pearson correlation coefficient.*

Correlation matrix: A table showing correlations for all possible pairs of variables. (6.8)

Counterbalancing: Reversing the order of conditions for equal numbers of all subjects. (15.1)

Critical z score: A z score that separates common from rare outcomes and hence dictates whether the null hypothesis should be retained or rejected. (10.7)

Cumulative frequency distribution: A frequency distribution showing the total number of observations in each class and all lower-ranked classes. (2.5)

Curvilinear relationship: A relationship that can be described best with a curved line. (6.2)

Data: A collection of observations or scores from a survey or an experiment. (1.4)

Decision rule: Specifies precisely when the null hypothesis should be rejected (because the observed value qualifies as a rare outcome). (10.7)

Degrees of freedom (df): The number of values free to vary, given one or more mathematical restrictions. (4.6)

Dependent variable: A variable that is believed to have been influenced by the independent variable. (1.6)

Descriptive statistics: The area of statistics concerned with organizing and summarizing information about a collection of actual observations. (1.2)

Difference score (D): The arithmetic difference between each pair of scores in repeated measures or, more generally, in two related samples. (15.1)

Directional test: See *One-tailed test.*

Discrete variable: A variable that consists of isolated numbers separated by gaps. (1.6)

Distribution-free tests: Tests, such as U, T, and H, that make no assumptions about the form of the population distribution. (20.2)

Effect: Any difference between a true and a hypothesized population mean. (11.8) Also, any difference between two (or more) population means. (14.1) See also *Treatment effect*.

Estimated standard error of difference between sample means ($s_{\overline{X}_1 - \overline{X}_2}$): The standard deviation of the sampling distribution of difference between means used whenever the unknown variance common to both populations must be estimated. (14.5)

Estimated standard error of the mean ($s_{\overline{X}}$): The standard deviation of the sampling distribution of the mean used whenever the unknown population standard deviation must be estimated. (13.6)

Estimated standard error of the mean difference ($s_{\overline{D}}$): The standard deviation of the sampling distribution of the mean difference used whenever the unknown population standard deviation for difference scores must be estimated. (15.5)

Expected frequency (f_e): The hypothesized frequency for each category, given that the null hypothesis is true. Used with the chi-square test. (19.3)

Experiment: A study in which the investigator decides who receives the special treatment. (1.6)

File drawer effect: The publication only of statistically significant reports. (14.11)

Frequency distribution: A collection of observations produced by sorting observations into classes that show their frequency (f) of occurrence. (2.1)

Frequency distribution for grouped data: A frequency distribution produced whenever observations are sorted into classes of *more than one* value. (2.1)

Frequency distribution for ungrouped data: A frequency distribution produced whenever observations are sorted into classes of *single* values. (2.1)

Frequency polygon: A line graph for quantitative data that emphasizes the continuity of continuous variables. (2.8)

***F* ratio:** Ratio of the between-group mean square (for subjects treated differently) to the within-group mean square (for subjects treated similarly). (16.5)

F_{se} test for simple effects: A test of the effect of one factor on the dependent variable at a single level of another factor. (18.9)

Histogram: A bar-type graph for quantitative data, with no gaps between adjacent bars. (2.8)

Hypothesized sampling distribution: Centered about the hypothesized population mean, this distribution is used to generate the decision rule. (11.8)

Independent events: The occurrence of one event has no effect on the probability that the other event will occur. (8.9)

Independent variable: The treatment that is manipulated by the investigator in an experiment. (1.6)

Inferential statistics: The area of statistics concerned about generalizing beyond actual observations. (1.2)

Interaction: The product of inconsistent simple effects. (18.3)

Interquartile range (IQR): The range for the middle 50 percent of all scores. (4.7)

Interval/ratio measurement: Locates observations along a scale having equal intervals and a true zero. (1.5)

Kruskal-Wallis *H* test: A test for ranked data when there are more than two independent groups. (20.5)

Least squares regression equation: The equation that minimizes the total of all squared predictive errors for known Y scores in the original correlation analysis. (7.3)

Level of confidence: The percent of time that a series of confidence intervals includes the unknown population characteristic, such as the population mean. (12.4)

Level of measurement: Rules that specify the extent to which a number actually represents some attribute. (1.5)

Level of significance (α): The degree of rarity required of an observed outcome to reject the null hypothesis. (H_0) (10.7)

Linear relationship: A relationship that can be described best with a straight line. (6.2)

Main effect: The effect of a single factor when any other factor is ignored. (18.1)

Mann-Whitney *U* test: A test for ranked data when there are two independent groups. (20.3)

Margin of error: That which is added to and subtracted from some sample value, such as the sample proportion or sample mean, to obtain the limits of a confidence interval. (12.7)

Mean: See *Population mean* or *Sample mean.*

Mean of the sampling distribution of the mean ($\mu_{\bar{X}}$): The mean of all sample means always equals the population mean. (9.4)

Mean square (*MS*): A variance estimate obtained by dividing a sum of squares by its degrees of freedom. (16.4)

Measures of central tendency: A general term for the various averages that attempt to describe the middle or typical value in a distribution. (3.1)

Measures of variability: A general term for various measures of the amount by which scores are dispersed or scattered. (4.1)

Median: The middle value when observations are ordered from least to most. (3.2)

Meta-analysis: A set of data-collecting and statistical procedures designed to summarize the various effects reported by groups of similar studies. (14.10)

Mode: The value of the most frequent observation or score. (3.1)

Multiple comparisons: The possible comparisons whenever more than two population means are involved. (16.10)

Multiple regression equation: A least squares equation that contains more than one predictor or X variable. (7.7)

Multiplication rule: Multiply together the separate probabilities of several independent events to find the probability that these events will occur together. (8.9)

Mutually exclusive events: Events that cannot occur together. (8.8)

Negative relationship: Occurs insofar as pairs of observations tend to occupy dissimilar and opposite relative positions in their respective distributions. (6.1)

Negatively skewed distribution: A distribution that includes a few extreme observations in the negative direction. (2.9)

Nominal measurement: Sorts observations into different classes or categories. (1.5)

Nondirectional test: See *Two-tailed test*.

Nonparametric tests: Tests, such as U, T, and H, that evaluate entire population distributions rather than specific population characteristics. (20.2)

Normal curve: A theoretical curve noted for its symmetrical bell-shaped form. (5.1)

Null hypothesis (H_0): A statistical hypothesis that usually asserts that nothing special is happening with respect to some characteristic of the underlying population. (10.5)

Observational study: A study that focuses on the detection of relationships between variables not manipulated by the investigator. (1.6)

Observed frequency (f_0): The obtained frequency for each category. Used with the chi-square test. (19.3)

Odds ratio (OR): Indicates the relative occurrence of one value of the dependent variable across the two categories of the independent variable. (19.12)

One-factor ANOVA: The simplest type of analysis of variance that tests for differences among population means categorized by only one independent variable. (16.1)

One-tailed (or directional) test: The rejection region is located in just one tail of the sampling distribution. (11.3)

One-variable χ^2 test: Evaluates whether observed frequencies for a single qualitative variable are adequately described by hypothesized or expected frequencies. (19.1)

Ordinal measurement: Arranges observations in terms of order. (1.5)

Outlier: A very extreme observation. (2.3)

Partial squared curvilinear correlation (η_p^2): The proportion of explained variance in the dependent variable after one or more sources have been eliminated from the total variance. (17.8)

Pearson correlation coefficient (r): A number between –1.00 and 1.00 that describes the linear relationship between pairs of quantitative variables. (6.3)

Percentile rank of an observation: Percentage of scores in the entire distribution with similar or smaller values than that score. (2.5)

Point estimate: A single value that represents some unknown population characteristic, such as the population mean. (12.1)

Pooled variance estimate (s_p^2): The most accurate estimate of the population variance (assumed to be the same for both populations) based on a combination of two sample sums of squares and their degrees of freedom. (14.5)

Population: Any complete set of observations. (1.3)

Population correlation coefficient (ρ): A number between +1.00 and –1.00 that describes the linear relationship for all paired observations in a population. (15.9)

Population mean (μ): The balance point for a population, found by dividing the total value of all scores in the population by the number of scores in the population. (3.3)

Population size (N): The total number of scores in the population. (3.3)

Population standard deviation (σ): A rough measure of the average amount by which scores deviate from the population mean. (4.5)

Positively skewed distribution: A distribution that includes a few extreme observations in the positive direction. (2.9)

Positive relationship: Occurs insofar as pairs of observations tend to occupy similar relative positions in their respective distributions. (6.1)

Power ($1 - \beta$): The probability of detecting a particular effect. (11.11)

Power curve: Shows how the likelihood of detecting any possible effect varies for a fixed sample size. (11.11)

Probability: The proportion or fraction of times that a particular event is likely to occur. (8.7)

***p*-Value:** The degree of rarity of a test result, given that the null hypothesis is true. (14.6)

Qualitative data: A set of observations where any single observation is a word, letter, or code that represents a class or category. (1.4)

Quantitative data: A set of observations where any single observation is a number that represents an amount or a count. (1.4)

Random assignment: A procedure designed to ensure that each person has an equal chance of being assigned to any group in an experiment. (1.3)

Random error: The combined effects of all uncontrolled factors on the scores of individual subjects. (16.2)

Random sampling: A sample produced when all potential observations in the population have an equal chance of being selected. (1.3)

Range: The difference between the largest and smallest scores. (4.2)

Ranked data: A set of observations where any single observation is a number that indicates relative standing. (1.4)

Ratio measurement: See *Interval/ratio measurement.*

Real limits: Located at the mid-point of the gap between adjacent tabled boundaries. (2.2)

Regression equation: See *Least squares regression equation.*

Regression fallacy: Occurs whenever regression toward the mean is interpreted as a real, rather than a chance, effect. (7.7)

Regression toward the mean: A tendency for scores, particularly extreme scores, to shrink toward the mean. (7.8)

Relative frequency distribution: A frequency distribution showing the frequency of each class as a part or fraction of the total frequency for the entire distribution. (2.4)

Repeated measures: Whenever the same subject is measured more than once. (15.1)

Repeated-measures ANOVA: A type of analysis that tests whether differences exist among population means with measures on the same subjects. (17.1)

Research hypothesis: Usually identified with the alternative hypothesis, this is the informal hypothesis or hu nch that inspires the entire investigation. (10.6) See *Alternative hypothesis*.

Sample: Any subset of scores from a population. (1.3)

Sample correlation coefficient (r): A number between +1.00 and –1.00 that describes the linear relationship between paired observations in a sample. (6.3)

Sample mean ($\bar{X}$): The balance point for the sample, found by dividing the total value of all scores in the sample by the number of scores. (3.3)

Sample size (n): The total number of scores in the sample. (3.3)

Sample standard deviation (s): A rough measure of the average amount by which scores in the sample deviate from their mean. (4.5)

Sample standard deviation of difference scores (s_D): A rough measure of the average amount by which difference scores in the sample deviate from the mean difference score. (14.5)

Sampling distribution of the mean: The probability distribution of means for all possible random samples of a given size from some population. (9.1)

Sampling distribution of t: The distribution that would be obtained if a value of t were calculated for each sample mean for all possible random samples of a given size from some population. (13.2)

Sampling distribution of $\bar{X}_1 - \bar{X}_2$: Differences between sample means based on all possible pairs of random samples from two underlying populations. (14.3)

Sampling distribution of z: The distribution of z values that would be obtained if a value of z were calculated for each sample mean for all possible random samples of a given size from some population. (10.2)

Scatterplot: A special graph containing a cluster of dots that represents all pairs of observations. (6.2)

Simple effect: The effect of one factor on the dependent variable at a single level of another factor. (18.3)

Squared correlation coefficient (r^2): The proportion of variability in one variable that is predictable from its relationship with another variable. (7.6)

Squared Cramer's phi coefficient (ϕ_c^2): A very rough estimate of the proportion of explained variance (or predictability) between two qualitative variables. (19.11)

Squared curvilinear correlation (η^2): The proportion of variance in the dependent variable that can be explained by or attributed to the independent variable. (16.9)

Squared partial curvilinear correlation. See *Partial squared curvilinear correlation.*

Standard deviation: A rough measure of the average amount by which observations deviate from their mean. (4.4)

Standard error of estimate ($s_{y|x}$): A rough measure of the average amount of predictive error. (7.4)

Standard error of the difference between means, ($\sigma_{\bar{X}_1 - \bar{X}_2}$): A rough measure of the average amount by which any sample mean difference deviates from the difference between population means. (14.3)

Standard error of the mean ($\sigma_{\bar{X}}$): A rough measure of the average amount by which sample means deviate from the population mean. (9.5)

Standard error of the mean difference ($\sigma_{\bar{D}}$): The standard deviation of the sampling distribution for the mean difference. (15.3)

Standard normal curve: The one tabled normal curve for z scores with a mean of 0 and a standard deviation of 1. (5.3)

Standard score: Unit-free score expressed relative to a known mean and a known standard deviation. (5.7)

Standardized effect estimate, Cohen's d: Describes effect size by expressing the observed mean difference in standard deviation units. (14.9)

Statistical significance: Implies only that the null hypothesis is probably false, but not whether it's false because of a large or small effect. (14.7)

Stem and leaf display: A device for sorting quantitative data on the basis of leading and trailing digits. (2.8)

Sum of squares (SS): The sum of squared deviation scores. (4.5)

***t* ratio:** A replacement for the z ratio whenever the unknown population standard deviation must be estimated. (13.3)

Transformed standard score (z'): A standard score that, unlike a z score, usually lacks negative signs and decimal points. (5.7)

Treatment effect: The existence of at least one difference between the population means. (16.2)

True sampling distribution: Centered about the true population mean, this distribution produces the one observed mean (or z). (11.8)

Tukey's *HSD* test: A multiple comparison test for which the cumulative probability of at least one type I error never exceeds the specified level of significance. (16.10)

Two independent samples: Observations in each sample are based on different (and unmatched) subjects. (14.1)

Two-factor ANOVA: A more complex type of analysis of variance that tests whether differences exist among population means categorized by two factors or independent variables. (18.1)

Two related samples: Each observation in one sample is paired, on a one-to-one basis, with a single observation in the other sample. (15.1)

Two-tailed (or nondirectional) test: Rejection regions are located in both tails of the sampling distribution. (11.3)

Two-variable χ^2 test: Evaluates whether observed frequencies reflect the independence of two qualitative variables. (19.7)

Type I error: Rejecting a true null hypothesis. (11.6)

Type II error: Retaining a false null hypothesis. (11.6)

Unit of measurement: The smallest possible difference between scores. (2.2)

Variability between groups (in ANOVA): Variability among scores of subjects who, being in different groups, receive different experimental treatments. (16.2)

Variability within groups (in ANOVA): Variability among scores of subjects who, being in the same group, receive the same experimental treatment. (16.2)

Variable: A characteristic or property that can take on different values. (1.6)

Variance: The mean of all squared deviation scores. (4.3)

Variance estimate (in ANOVA): See *Mean square.*

Wilcoxon *T* test: A test for ranked data when there are two related groups. (20.4)

z score: A unit-free score that indicates how many standard deviations an observation is above or below the mean of its distribution. (5.2)

z test for a population mean: A hypothesis test that evaluates how far the observed sample mean deviates, in standard error units, from the hypothesized population mean. (10.2)

Index

N

O

P

Q

R

S

T

U

V

W

Z

Important Symbols

Numbers indicate the section where each symbol is introduced and defined.

Greek Letters

α **(alpha)** level of significance **10.7**; probability of a type I error **11.7**

β **(beta)** probability of a type II error **11.8**

η^2 **(eta)** squared curvilinear correlation **16.9**

η_p^2 **(partial eta)** squared partial curvilinear correlation **17.8**

μ **(mu)** population mean **3.3**

$\mu_1 - \mu_2$ difference between two population means **14.2**

μ_D population mean of difference scores **15.2**

ρ **(rho)** population correlation coefficient **15.9**

$\sum$ **(summation)** take the sum of **3.3**

σ **(sigma)** population standard deviation **4.5**

σ^2 population variance **4.5**

$\sigma_{\bar{X}}$ standard error of mean **9.5**

ϕ_c^2 **(phi)** squared Cramer's phi coefficient **19.11**

χ^2 chi-square ratio **19.3**

English Letters

CI confidence interval **12.2**

$\underline{D}$ difference between paired scores **15.1**

$\overline{D}$ sample mean of difference scores **15.1**

d Cohen's estimate of effect size **14.9**

df number of degrees of freedom **4.6**

F F ratio **16.5**

F_{se} F ratio for a simple effect **18.9**

f frequency **2.1**

f_e expected frequency in a sample **19.3**

f_o observed frequency in a sample **19.3**

H Kruskal-Wallis H test for ranked data **20.5**

H_0 null hypothesis **10.5**

H_1 alternative hypothesis **10.6**

HSD Tukey's critical value **16.10**

IQR interquartile range **4.7**

MS mean square **16.5**

MSE mean square error **16.12**

N population size **3.3**

n sample size **3.3**

OR odds ratio **19.12**

p probability of some outcome given that the null hypothesis is true **14.6**

Pr( ) probability of the outcome in parentheses **8.7**

r sample correlation coefficient **6.3**

r^2 squared correlation coefficient **7.6**

$s_{y|x}$ standard error of estimate **7.4**

SP_{xy} sum of products **6.5**

SS sum of squares **4.5**

s sample standard deviation **4.5**

s_D sample standard deviation of difference scores **15.5**

$s_{\bar{x}}$ estimated standard error of the mean **13.6**

$s_{\bar{x}_1 - \bar{x}_2}$ estimated standard error of the difference between two sample means **14.5**

$s_{\bar{D}}$ estimated standard error of the mean difference scores **15.5**

s^2 sample variance **4.5**

s_p^2 pooled sample variance **14.5**

T Wilcoxon T test for ranked data **20.4**

t t ratio **13.3**

U Mann-Whitney U test for ranked data **20.3**

X any unspecified observation or score **3.3**

$\overline{X}$ sample mean **3.3**

$\overline{X}_1 - \overline{X}_2$ difference between two sample means **14.3**

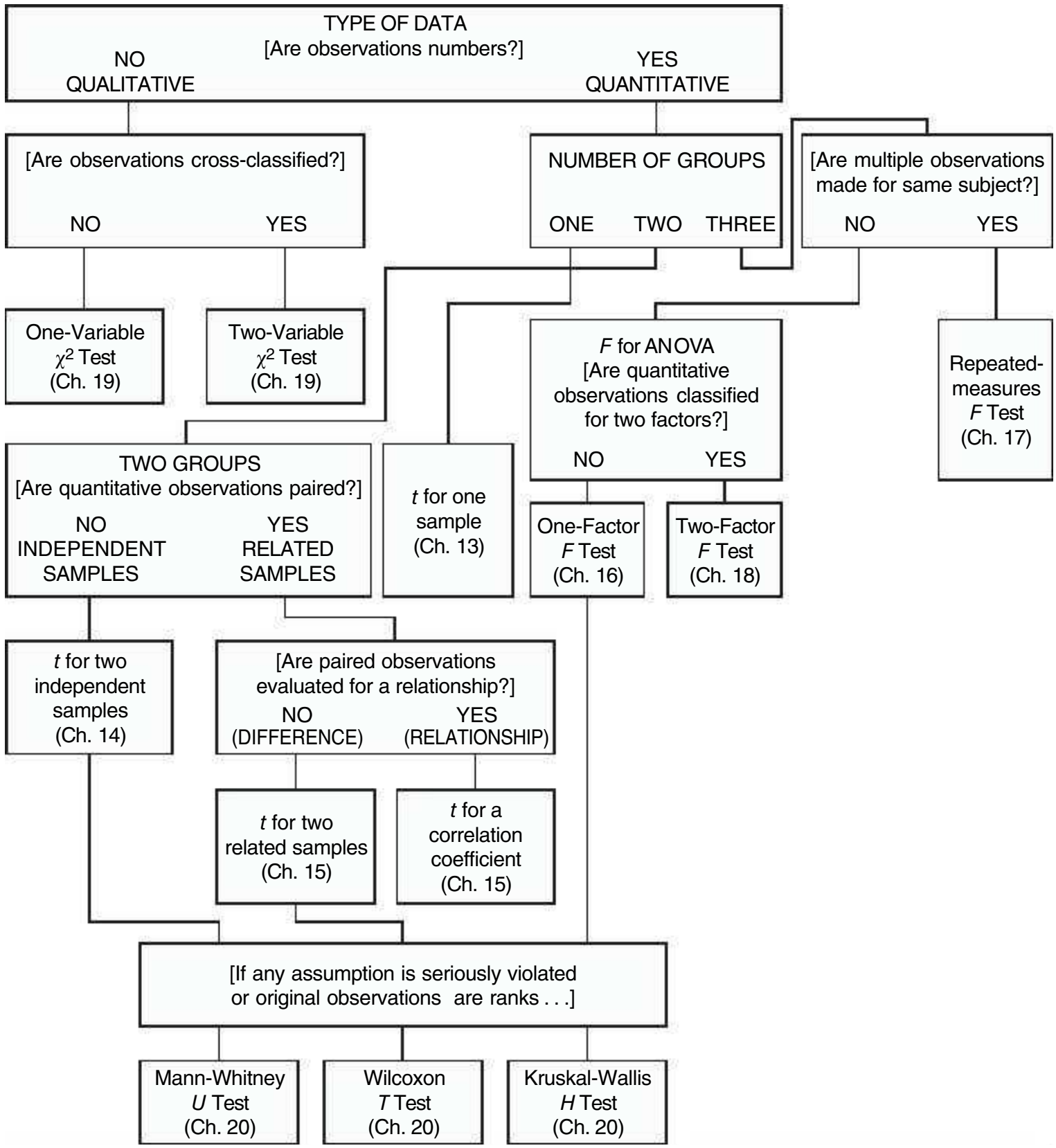
Y a score paired with X **6.2**

Y' predicted score **7.3**

z standard score **5.2**; z ratio **10.2**

z' transformed standard score **5.7**

Guidelines for Selecting the Appropriate Hypothesis Test



WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.