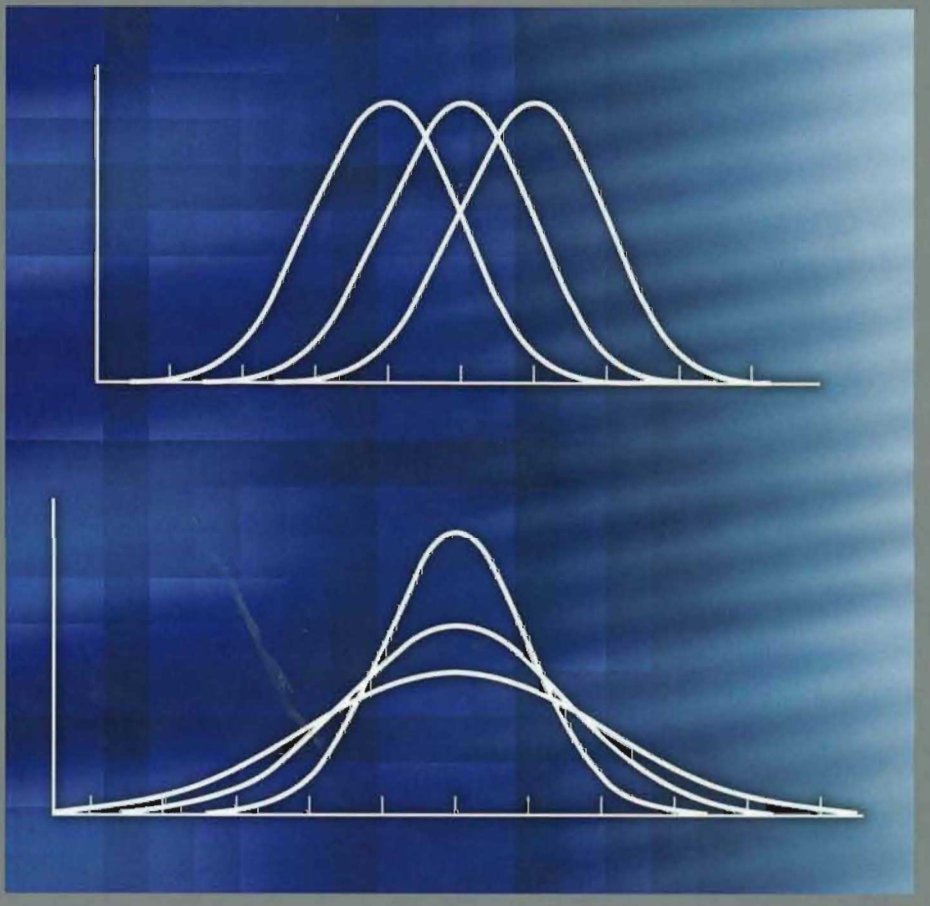
BIOSTATISTICAL ANALYSIS

FIFTH EDITION



JERROLD H. ZAR

STATISTICAL TABLES AND GRAPHS

INTERPOLATION

BIOSTATISTICAL ANALYSIS

Fifth Edition

Jerrold H. Zar
Department of Biological Sciences
Northern Illinois University

Prentice Hall
is an imprint of

Upper Saddle River, New Jersey 07458

Library of Congress Cataloging-in-Publication Data

Zar, Jerrold H.
Biostatistical Analysis / Jerrold H. Zar—5th ed.
p. cm.
1. Biometry. I. Title.
ISBN: 978-0-13-100846-5
1. Biostatistical analysis I. Title
QH323.5.Z37 2010
570.1'5195—dc22 2009035156

Editor-in-Chief, Statistics and Mathematics: *Deirdre Lynch*
Acquisitions Editor: *Christopher Cummings*
Editorial Project Manager: *Christine O'Brien*
Assistant Editor: *Christina Lepre*
Project Manager: *Raegan Keida Heerema*
Associate Managing Editor: *Bayani Mendoza de Leon*
Senior Managing Editor: *Linda Mihatov Behrens*
Senior Operations Supervisor: *Diane Peirano*
Marketing Manager: *Alex Gay*
Marketing Assistant: *Kathleen DeChavez*
Art Director: *Jayne Conte*
Cover Designer: *Bruce Kenselaar*

Prentice Hall is an imprint of

PEARSON

Pearson Prentice Hall
Pearson Education, Inc.
Upper Saddle River, New Jersey 07458

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

ISBN-10: 0-13-100846-3
ISBN-13: 978-0-13-100846-5

Pearson Education Ltd., *London*
Pearson Education Singapore, Pte. Ltd.
Pearson Education Canada, Inc.
Pearson Education—Japan
Pearson Education Australia PTY, Limited
Pearson Education North Asia Ltd., *Hong Kong*
Pearson Education de Mexico, S.A. de C.V.
Pearson Education Malaysia, Pte. Ltd.
Pearson Education Upper Saddle River, New Jersey

Contents

Preface

Beginning with the first edition of this book, the goal has been to introduce a broad array of techniques for the examination and analysis of a wide variety of data that may be encountered in diverse areas of biological studies. As such, the book has been called upon to fulfill two purposes. First, it has served as an introductory textbook, assuming no prior knowledge of statistics. Second, it has functioned as a reference work consulted long after formal instruction has ended.

Colleges and universities have long offered an assortment of introductory statistics courses. Some of these courses are without concentration on a particular field in which quantitative data might be collected (and often emphasize mathematics and statistical theory), and some focus on statistical methods of utility to a specific field (such as this book, which has an explicit orientation to the biological sciences). Walker (1929: 148–163) reported that, although the teaching of probability has a much longer history, the first statistics course at a U.S. university or college probably was at Columbia College (renamed Columbia University in 1896) in 1880 in the economics department; followed in 1887 by the second—the first in psychology—at the University of Pennsylvania; in 1889 by the first in anthropology, at Clark University; in 1897 by the first in biology, at Harvard University; in 1898 by the first in mathematics, at the University of Illinois; and in 1900 by the first in education, at Teachers College, Columbia University. In biology, the first courses with statistical content were probably taught by Charles B. Davenport at Harvard (1887–1899), and his *Statistical Methods in Biological Variation*, first published in 1899, may have been the first American book focused on statistics (ibid.: 159).

The material in this book requires no mathematical competence beyond very elementary algebra, although the discussions include many topics that appear seldom, if at all, in other general texts. Some statistical procedures are mentioned though not recommended. This is done for the benefit of readers who may encounter them in research reports or computer software.

Many literature references and footnotes are given throughout most chapters, to provide support for material discussed, to provide historical points, or to direct the reader to sources of additional information. More references are given for controversial and lesser-known topics.

The data in the examples and exercises are largely fictional, though generally realistic, and are intended to demonstrate statistical procedures, not to present actual research conclusions. The exercises at the end of chapters can serve as additional examples of statistical methods, and the answers are given at the back of the book. The sample sizes of most examples and exercises are small in order to conserve space and to enhance the ease of presentation and computation. Although the examples and exercises represent a variety of areas within the biological sciences, they are intended to be understood by biology students and researchers across a diversity of fields.

There are important statistical procedures that involve computations so demanding that they preclude practical execution without appropriate computer software. Basic principles and aspects of the underlying calculations are presented to show how results may be obtained; for even if laborious calculations will be performed by computer, the biologist should be informed enough to interpret properly the computational results. Many statistical packages are available, commercially or otherwise, addressing various subsets of the procedures in this book; but no single package is promoted herein.

A final contribution toward achieving a book with self-sufficiency for most biostatistical needs is the inclusion of a comprehensive set of statistical tables, more extensive than those found in similar texts.

To be useful as a reference, and to allow for differences in content among courses for which it might be used, this book contains much more material than would be covered during one academic term. Therefore, I am sometimes asked to recommend what I consider to be the basic topics for an introduction to the subject. I suggest these book sections (though not necessarily in their entirety) as a core treatment of biostatistical methods, to be augmented or otherwise amended with others of the instructor's preference: 1.1–1.4, 2.1–2.4, 3.1–3.3, 4.1, 4.4–4.6, 6.1–6.4, 7.1–7.4, 7.6–7.7, 8.1–8.5, 8.10–8.11, 9.1–9.3, 10.1–10.4, 11.1–11.4, 12.1–12.4, 14.1, 15.1, 17.1–17.7, 18.1–18.3, 19.1–19.3, 19.9, 20.2–20.4, 22.1–22.3, 22.5, 23.1–23.4; and the introductory paragraph(s) to each of these chapters.

Jerrold H. Zar
DeKalb, Illinois

Acknowledgments

A book of this nature requires, and benefits from, the assistance of many. For the preparation of the early editions, the outstanding library collections of the University of Illinois at Urbana–Champaign were invaluable, and for all editions I am greatly indebted to the library materials and services of Northern Illinois University and its vast book- and journal-collection networks. I also gratefully acknowledge the cooperation of the computer services at Northern Illinois University, which assisted in executing many of the computer programs I prepared to generate some of the appendix tables. For the tables taken from previously published sources, thanks are given for the permission to reprint them, and acknowledgment of each source is given immediately following the appearance of the reprinted material. Additionally, I am pleased to recognize the editorial and production staff at Pearson Education and Laserwords for their valued professional assistance in transforming my manuscript into the published product in hand.

Over many years, teachers, students, and colleagues have aided in leading me to important biostatistical questions and to the material presented in this volume. Special recognition must be made of S. Charles Kendeigh (1904–1986) of the University of Illinois at Urbana–Champaign, who, through considerate mentorship, first made me aware of the value of quantitative analysis of biological data, leading me to produce the first edition; Edward Batschelet (1914–1979), of the University of Zurich, who provided me with kind encouragement and inspiration on statistical matters during the preparation of much of the first two editions; Arthur W. Ghent (1927–2001), University of Illinois at Urbana–Champaign, who was constantly supportive through the first four editions, offering statistical and biological commentary both stimulating and challenging; and Carol J. Feltz (1956–2001), Northern Illinois University, who provided substantive consultation on some major new material for the fourth edition. Prior to publication, the book drew upon the expertise and insights of reviewers, including Raid Amin, University of West Florida; Franklin R. Ampy, Howard University; Sulekha Anand, San José State University; Roxanne Barrows, Hocking College; Thomas Beitinger, University of North Texas; Mark Butler, Old Dominion University; William Caire, University of Central Oklahoma; Gary Cobbs, University of Louisville; Loveday L. Conquest, University of Washington; Todd A. Crowl, Utah State University; Matthew Edwards, University of California, Santa Cruz; Todd Fearer, University of Arkansas-Monticello; Avshalom Gamliel, Wavemetrics, Inc.; Peter Homann, Western Washington University; Robert Huber, Bowling Green State University; James W. Kirchner, University of California, Berkeley; Gary A. Lamberti, University of Notre Dame; David Larsen, University of Missouri, Columbia; David R. McConville, Saint Mary's University of Minnesota; J. Kelly McCoy, Angelo State University; David J. Moriarty, California State Polytechnic University; Mark Rizzardi, Humboldt State University; Michael A. Romano, Western Illinois University; and Thomas A. Wolosz, the State University of New York, Plattsburgh. I also acknowledge my wife, Carol, for her prolonged and consistent patience during the thirty-five years of production of the five editions of this book.

Jerrold H. Zar
DeKalb, Illinois

CHAPTER 1

Data: Types and Presentation

Scientific study involves the systematic collection, organization, analysis, and presentation of knowledge. Many investigations in the biological sciences are quantitative, where knowledge is in the form of numerical observations called *data*. (One numerical observation is a *datum*.*) In order for the presentation and analysis of data to be valid and useful, we must use methods appropriate to the type of data obtained, to the design of the data collection, and to the questions asked of the data; and the limitations of the data, of the data collection, and of the data analysis should be appreciated when formulating conclusions. This chapter, and those that follow, will introduce many concepts relevant to this goal.

The word *statistics* is derived from the Latin for "state," indicating the historical importance of governmental data gathering, which related principally to demographic information (including census data and "vital statistics") and often to their use in military recruitment and tax collecting.†

The term *statistics* is often encountered as a synonym for *data*: One hears of college enrollment statistics (such as the numbers of newly admitted students, numbers of senior students, numbers of students from various geographic locations), statistics of a basketball game (such as how many points were scored by each player, how many fouls were committed), labor statistics (such as numbers of workers unemployed, numbers employed in various occupations), and so on. Hereafter, this use of the word *statistics* will not appear in this book. Instead, it will be used in its other common manner: to refer to the *orderly collection, analysis, and interpretation of data with a view to objective evaluation of conclusions based on the data.* (Section 2.4 will introduce another fundamentally important use of the term *statistic*.)

Statistics applied to biological problems is simply called *biostatistics* or, sometimes, *biometry*‡ (the latter term literally meaning "biological measurement"). Although

*The term *data* is sometimes seen as a singular noun meaning "numerical information." This book refrains from that use.

†Peters (1987: 79) and Walker (1929: 32) attribute the first use of the term *statistics* to a German professor, Gottfried Achenwall (1719–1772), who used the German word *Statistik* in 1749, and the first published use of the English word to John Sinclair (1754–1835) in 1791.

‡The word *biometry*, which literally means "biological measurement," had, since the nineteenth century, been found in several contexts (such as demographics and, later, quantitative genetics: Armitage, 1985; Stigler, 2000), but using it to mean the application of statistical methods to biological information apparently was conceived between 1892 and 1901 by Karl Pearson, along with the name *Biometrika* for the still-important English journal he helped found; and it was first published in the inaugural issue of this journal in 1901 (Snedecor, 1954). The Biometrics Section of the American

their magnitudes relative to each other; or success in learning to run a maze may be recorded as *A*, *B*, or *C*.

It is often true that biological data expressed on the ordinal scale could have been expressed on the interval or ratio scale had exact measurements been obtained (or obtainable). Sometimes data that were originally on interval or ratio scales will be changed to ranks; for example, examination grades of 99, 85, 73, and 66% (ratio scale) might be recorded as A, B, C, and D (ordinal scale), respectively.

Ordinal-scale data contain and convey less information than ratio or interval data, for only relative magnitudes are known. Consequently, quantitative comparisons are impossible (e.g., we cannot speak of a grade of C being half as good as a grade of A, or of the difference between cell sizes 1 and 2 being the same as the difference between sizes 3 and 4). However, we will see that many useful statistical procedures are, in fact, applicable to ordinal data.

(d) Data in Nominal Categories. Sometimes the variable being studied is classified by some qualitative measure it possesses rather than by a numerical measurement. In such cases the variable may be called an *attribute*, and we are said to be dealing with *nominal*, or *categorical*, data. Genetic phenotypes are commonly encountered biological attributes: The possible manifestations of an animal's eye color might be brown or blue; and if human hair color were the attribute of interest, we might record black, brown, blond, or red. As other examples of nominal data (*nominal* is from the Latin word for "name"), people might be classified as male or female, or right-handed or left-handed. Or, plants might be classified as dead or alive, or as with or without fertilizer application. Taxonomic categories also form a nominal classification scheme (for example, plants in a study might be classified as pine, spruce, or fir).

Sometimes, data that might have been expressed on an ordinal, interval, or ratio scale of measurement may be recorded in nominal categories. For example, heights might be recorded as tall or short, or performance on an examination as pass or fail, where there is an arbitrary cut-off point on the measurement scale to separate tall from short and pass from fail.

As will be seen, statistical methods useful with ratio, interval, or ordinal data generally are not applicable to nominal data, and we must, therefore, be able to identify such situations when they occur.

(e) Continuous and Discrete Data. When we spoke previously of plant heights, we were dealing with a variable that could be any conceivable value within any observed range; this is referred to as a *continuous variable*. That is, if we measure a height of 35 cm and a height of 36 cm, an infinite number of heights is possible in the range from 35 to 36 cm: a plant might be 35.07 cm tall or 35.988 cm tall, or 35.3263 cm tall, and so on, although, of course, we do not have devices sensitive enough to detect this infinity of heights. A continuous variable is one for which there is a possible value between any other two values.

However, when speaking of the number of leaves on a plant, we are dealing with a variable that can take on only certain values. It might be possible to observe 27 leaves, or 28 leaves, but 27.43 leaves and 27.9 leaves are values of the variable that are impossible to obtain. Such a variable is termed a *discrete* or *discontinuous variable* (also known as a *meristic variable*). The number of white blood cells in 1 mm^3 of blood, the number of giraffes visiting a water hole, and the number of eggs laid by a grasshopper are all discrete variables. The possible values of a discrete variable generally are consecutive integers, but this is not necessarily so. If the leaves on our

plants are always formed in pairs, then only even integers are possible values of the variable. And the ratio of number of wings to number of legs of insects is a discrete variable that may only have the value of 0, 0.3333..., or 0.6666... (i.e., $\frac{0}{6}$, $\frac{2}{6}$, or $\frac{4}{6}$, respectively).*

Ratio-, interval-, and ordinal-scale data may be either continuous or discrete. Nominal-scale data by their nature are discrete.

1.2 ACCURACY AND SIGNIFICANT FIGURES

Accuracy is the nearness of a measurement to the true value of the variable being measured. *Precision* is not a synonymous term but refers to the closeness to each other of repeated measurements of the same quantity. Figure 1.1 illustrates the difference between accuracy and precision of measurements.

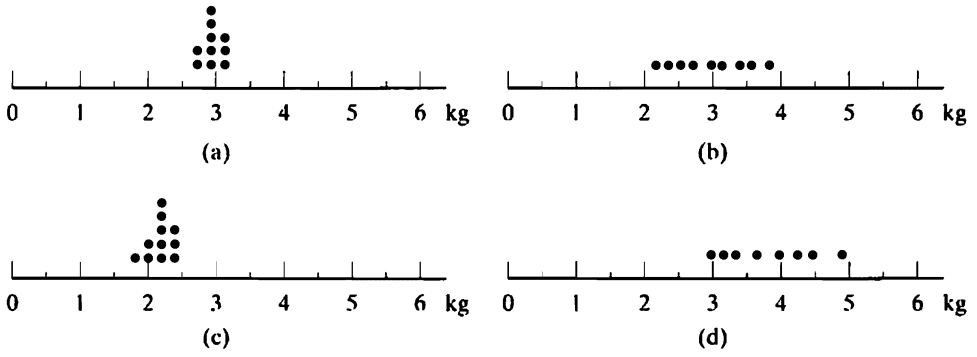


FIGURE 1.1: Accuracy and precision of measurements. A 3-kilogram animal is weighed 10 times. The 10 measurements shown in sample (a) are relatively accurate and precise; those in sample (b) are relatively accurate but not precise; those of sample (c) are relatively precise but not accurate; and those of sample (d) are relatively inaccurate and imprecise.

Human error may exist in the recording of data. For example, a person may miscount the number of birds in a tract of land or misread the numbers on a heart-rate monitor. Or, a person might obtain correct data but record them in such a way (perhaps with poor handwriting) that a subsequent data analyst makes an error in reading them. We shall assume that such errors have not occurred, but there are other aspects of accuracy that should be considered.

Accuracy of measurement can be expressed in numerical reporting. If we report that the hind leg of a frog is 8 cm long, we are stating the number 8 (a value of a continuous variable) as an estimate of the frog's true leg length. This estimate was made using some sort of a measuring device. Had the device been capable of more accuracy, we might have declared that the leg was 8.3 cm long, or perhaps 8.32 cm long. When recording values of continuous variables, it is important to designate the accuracy with which the measurements have been made. By convention, the value 8 denotes a measurement in the range of 7.50000 ... to 8.49999..., the value 8.3 designates a range of 8.25000 ... to 8.34999..., and the value 8.32 implies that the true value lies within the range of 8.31500... to 8.32499.... That is, the reported value is the midpoint of the implied range, and the size of this range is designated by the last decimal place in the measurement. The value of 8 cm implies an ability to

*The ellipsis marks (...) may be read as "and so on." Here, they indicate that $\frac{2}{6}$ and $\frac{4}{6}$ are repeating decimal fractions, which could just as well have been written as 0.3333333333333 ... and 0.6666666666666..., respectively.

determine length within a range of 1 cm, 8.3 cm implies a range of 0.1 cm, and 8.32 cm implies a range of 0.01 cm. Thus, to record a value of 8.0 implies greater accuracy of measurement than does the recording of a value of 8, for in the first instance the true value is said to lie between 7.95000 ... and 8.049999... (i.e., within a range of 0.1 cm), whereas 8 implies a value between 7.50000 ... and 8.49999... (i.e., within a range of 1 cm). To state 8.00 cm implies a measurement that ascertains the frog's limb length to be between 7.99500 ... and 8.00499 ... cm (i.e., within a range of 0.01 cm). Those digits in a number that denote the accuracy of the measurement are referred to as *significant figures*. Thus, 8 has one significant figure, 8.0 and 8.3 each have two significant figures, and 8.00 and 8.32 each have three.

In working with exact values of discrete variables, the preceding considerations do not apply. That is, it is sufficient to state that our frog has four limbs or that its left lung contains thirteen flukes. The use of 4.0 or 13.00 would be inappropriate, for as the numbers involved are exactly 4 and 13, there is no question of accuracy or significant figures.

But there are instances where significant figures and implied accuracy come into play with discrete data. An entomologist may report that there are 72,000 moths in a particular forest area. In doing so, it is probably not being claimed that this is the exact number but an estimate of the exact number, perhaps accurate to two significant figures. In such a case, 72,000 would imply a range of accuracy of 1000, so that the true value might lie anywhere from 71,500 to 72,500. If the entomologist wished to convey the fact that this estimate is believed to be accurate to the nearest 100 (i.e., to three significant figures), rather than to the nearest 1000, it would be better to present the data in the form of *scientific notation*,* as follows: If the number $7.2 \times 10^4 (= 72{,}000)$ is written, a range of accuracy of $0.1 \times 10^4 (= 1000)$ is implied, and the true value is assumed to lie between 71,500 and 72,500. But if 7.20×10^4 were written, a range of accuracy of $0.01 \times 10^4 (= 100)$ would be implied, and the true value would be assumed to be in the range of 71,950 to 72,050. Thus, the accuracy of large values (and this applies to continuous as well as discrete variables) can be expressed succinctly using scientific notation.

Calculators and computers typically yield results with more significant figures than are justified by the data. However, it is good practice—to avoid rounding error—to retain many significant figures until the last step in a sequence of calculations, and on attaining the result of the final step to round off to the appropriate number of figures. A suggestion for the number of figures to report is given at the end of Section 6.2.

1.3 FREQUENCY DISTRIBUTIONS

When collecting and summarizing large amounts of data, it is often helpful to record the data in the form of a *frequency table*. Such a table simply involves a listing of all the observed values of the variable being studied and how many times each value is observed. Consider the tabulation of the frequency of occurrence of sparrow nests in each of several different locations. This is illustrated in Example 1.1, where the observed kinds of nest sites are listed, and for each kind the number of nests observed is recorded. The distribution of the total number of observations among the various categories is termed a *frequency distribution*. Example 1.1 is a frequency table for nominal data, and these data may also be presented graphically by means of a *bar graph* (Figure 1.2), where the height of each bar is proportional to the frequency in the class represented. The widths of all bars in a bar graph should be equal so

*The use of scientific notation—by physicists—can be traced back to at least the 1860s (Miller, 2004b).

EXAMPLE 1.1 The Location of Sparrow Nests: A Frequency Table of Nominal Data

The variable is nest site, and there are four recorded categories of this variable. The numbers recorded in these categories constitute the frequency distribution.

Nest Site	*Number of Nests Observed*
A. Vines	56
B. Building eaves	60
C. Low tree branches	46
D. Tree and building cavities	49

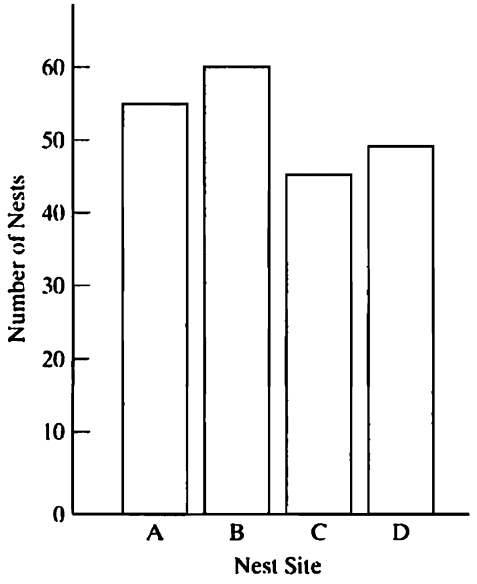


FIGURE 1.2: A bar graph of the sparrow nest data of Example 1.1. An example of a bar graph for nominal data.

that the eye of the reader is not distracted from the differences in bar heights; this also makes the area of each bar proportional to the frequency it represents. Also, the frequency scale on the vertical axis should begin at zero to avoid the apparent differences among bars. If, for example, a bar graph of the data of Example 1.1 were constructed with the vertical axis representing frequencies of 45 to 60 rather than 0 to 60, the results would appear as in Figure 1.3. Huff (1954) illustrates other techniques that can mislead the readers of graphs. It is good practice to leave space between the bars of a bar graph of nominal data, to emphasize the distinctness among the categories represented.

A frequency tabulation of ordinal data might appear as in Example 1.2, which presents the observed numbers of sunfish collected in each of five categories, each category being a degree of skin pigmentation. A bar graph (Figure 1.4) can be prepared for this frequency distribution just as for nominal data.

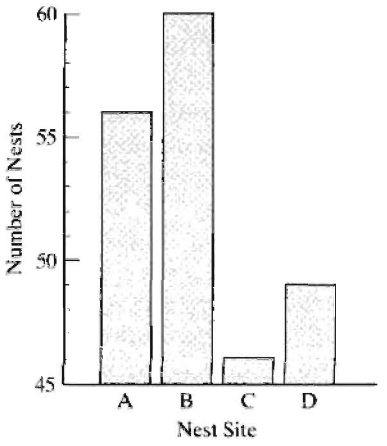


FIGURE 1.3: A bar graph of the sparrow nest data of Example 1.1, drawn with the vertical axis starting at 45. Compare this with Figure 1.1, where the axis starts at 0.

EXAMPLE 1.2 Numbers of Sunfish, Tabulated According to Amount of Black Pigmentation: A Frequency Table of Ordinal Data

The variable is amount of pigmentation, which is expressed by numerically ordered classes. The numbers recorded for the five pigmentation classes compose the frequency distribution.

Pigmentation Class	*Amount of Pigmentation*	*Number of Fish*
0	No black pigmentation	13
1	Faintly speckled	68
2	Moderately speckled	44
3	Heavily speckled	21
4	Solid black pigmentation	8

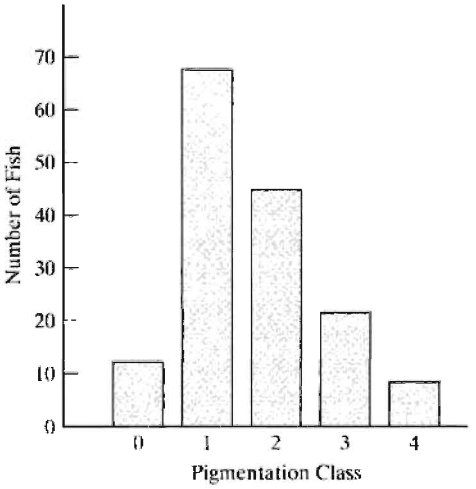


FIGURE 1.4: A bar graph of the sunfish pigmentation data of Example 1.2. An example of a bar graph for ordinal data.

In preparing frequency tables of interval- and ratio-scale data, we can make a procedural distinction between discrete and continuous data. Example 1.3 shows discrete data that are frequencies of litter sizes in foxes, and Figure 1.5 presents this frequency distribution graphically.

EXAMPLE 1.3 Frequency of Occurrence of Various Litter Sizes in Foxes: A Frequency Table of Discrete, Ratio-Scale Data

The variable is litter size, and the numbers recorded for the five litter sizes make up frequency distribution.

Litter Size	*Frequency*
3	10
4	27
5	22
6	4
7	1

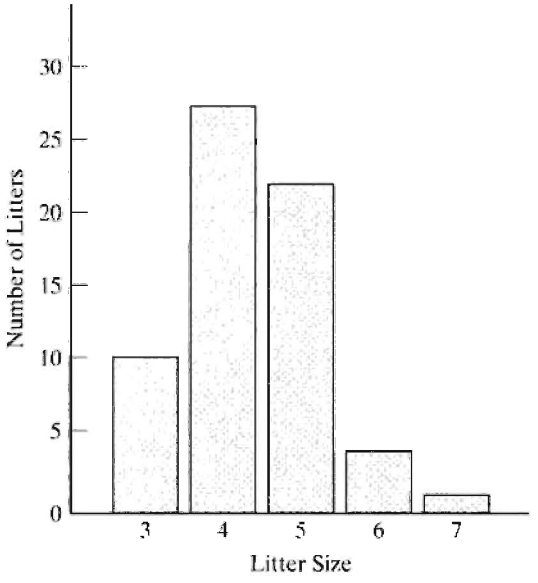


FIGURE 1.5: A bar graph of the fox litter data of Example 1.3. An example of a bar graph for discrete, ratio-scale data.

Example 1.4a shows discrete data that are the numbers of aphids found per clover plant. These data create quite a lengthy frequency table, and it is not difficult to imagine sets of data whose tabulation would result in an even longer list of frequencies. Thus, for purposes of preparing bar graphs, we often cast data into a frequency table by grouping them.

Example 1.4b is a table of the data from Example 1.4a arranged by grouping the data into size classes. The bar graph for this distribution appears as Figure 1.6. Such grouping results in the loss of some information and is generally utilized only to make frequency tables and bar graphs easier to read, and not for calculations performed on

the data. There have been several "rules of thumb" proposed to aid in deciding into how many classes data might reasonably be grouped, for the use of too few groups will obscure the general shape of the distribution. But such "rules" or recommendations are only rough guides, and the choice is generally left to good judgment, bearing in mind that from 10 to 20 groups are useful for most biological work. (See also Doane, 1976.) In general, groups should be established that are equal in the size interval of the variable being measured. (For example, the group size interval in Example 1.4b is four aphids per plant.)

EXAMPLE 1.4a Number of Aphids Observed per Clover Plant: A Frequency Table of Discrete, Ratio-Scale Data

Number of Aphids on a Plant	*Number of Plants Observed*	*Number of Aphids on a Plant*	*Number of Plants Observed*
0	3	20	17
1	1	21	18
2	1	22	23
3	1	23	17
4	2	24	19
5	3	25	18
6	5	26	19
7	7	27	21
8	8	28	18
9	11	29	13
10	10	30	10
11	11	31	14
12	13	32	9
13	12	33	10
14	16	34	8
15	13	35	5
16	14	36	4
17	16	37	1
18	15	38	2
19	14	39	1
		40	0
		41	1
Total number of observations = 424			

Because continuous data, contrary to discrete data, can take on an infinity of values, one is essentially always dealing with a frequency distribution tabulated by groups. If the variable of interest were a weight, measured to the nearest 0.1 mg, a frequency table entry of the number of weights measured to be 48.6 mg would be interpreted to mean the number of weights grouped between 48.5500... and 48.6499... mg (although in a frequency table this class interval is usually written as 48.55–48.65). Example 1.5 presents a tabulation of 130 determinations of the amount of phosphorus, in milligrams per gram, in dried leaves. (Ignore the last two columns of this table until Section 1.4.)

EXAMPLE 1.4b Number of Aphids Observed per Clover Plant: A Frequency Table Grouping the Discrete, Ratio-Scale Data of Example 1.4a

Number of Aphids on a Plant	*Number of Plants Observed*
0–3	6
4–7	17
8–11	40
12–15	54
16–19	59
20–23	75
24–27	77
28–31	55
32–35	32
36–39	8
40–43	1

Total number of observations = 424

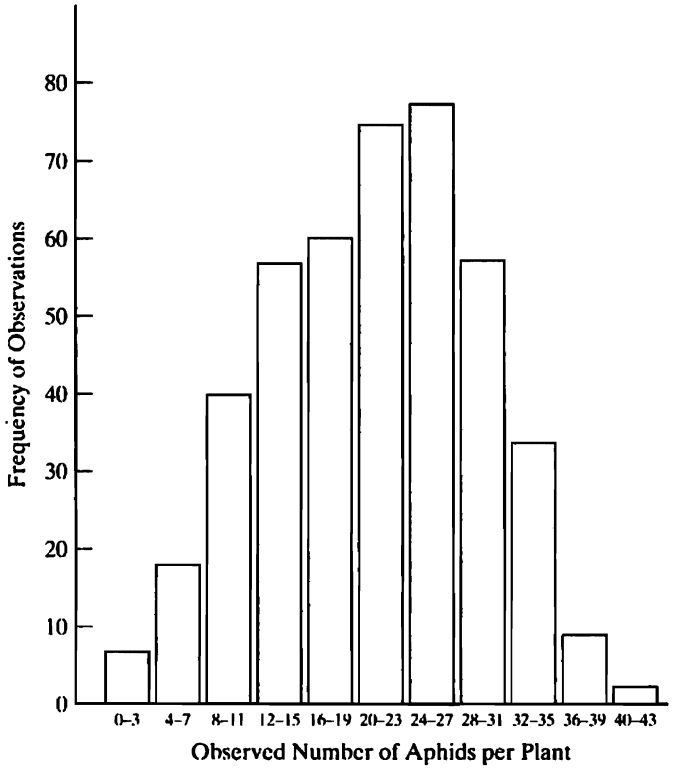


FIGURE 1.6: A bar graph of the aphid data of Example 1.4b. An example of a bar graph for grouped discrete, ratio-scale data.

EXAMPLE 1.5 Determinations of the Amount of Phosphorus in Leaves: A Frequency Table of Continuous Data

Phosphorus (mg/g of leaf)	*Frequency (i.e., number of determinations)*	**Cumulative frequency**	
		Starting with Low Values	*Starting with High Values*
8.15–8.25	2	2	130
8.25–8.35	6	8	128
8.35–8.45	8	16	122
8.45–8.55	11	27	114
8.55–8.65	17	44	103
8.65–8.75	17	61	86
8.75–8.85	24	85	69
8.85–8.95	18	103	45
8.95–9.05	13	116	27
9.05–9.15	10	126	14
9.15–9.25	4	130	4

Total frequency = 130 = n

In presenting this frequency distribution graphically, one can prepare a *histogram*,* which is the name given to a bar graph based on continuous data. This is done in Figure 1.7; note that rather than indicating the range on the horizontal axis, we indicate only the midpoint of the range, a procedure that results in less crowded printing on the graph. Note also that adjacent bars in a histogram are often drawn touching each other, to emphasize the continuity of the scale of measurement, whereas in the other bar graphs discussed they generally are not.

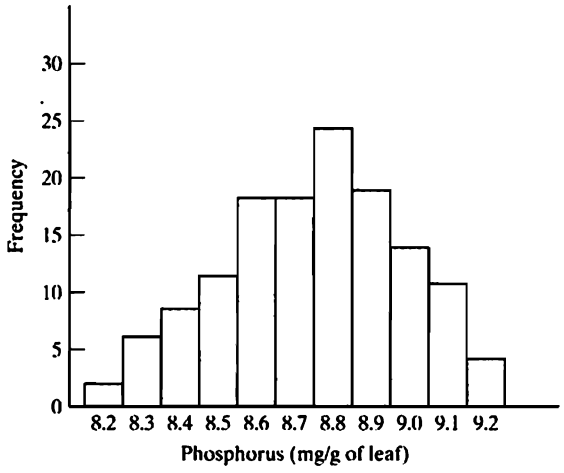


FIGURE 1.7: A histogram of the leaf phosphorus data of Example 1.5. An example of a histogram for continuous data.

*The term *histogram* is from Greek roots (referring to a pole-shaped drawing) and was first published by Karl Pearson in 1895 (David 1995).

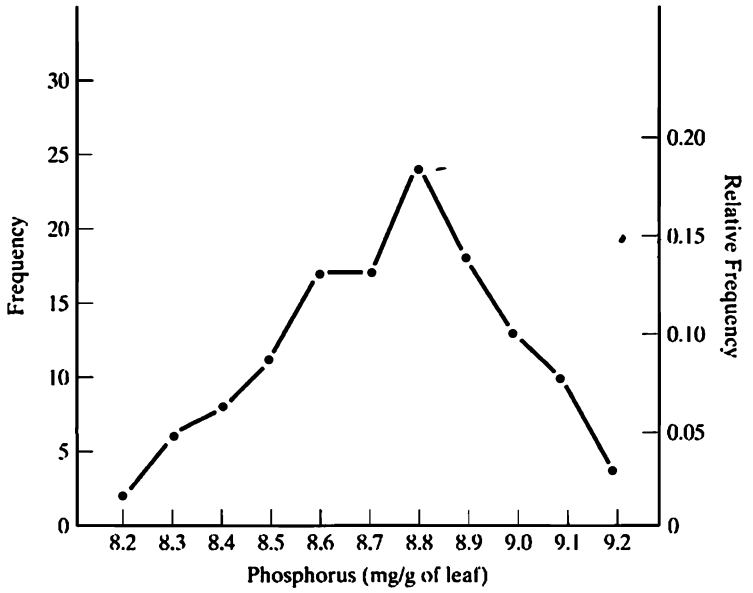


FIGURE 1.8: A frequency polygon for the leaf phosphorus data of Example 1.5.

Often a *frequency polygon* is drawn instead of a histogram. This is done by plotting the frequency of each class as a dot (or other symbol) at the class midpoint and then connecting each adjacent pair of dots by a straight line (Figure 1.8). It is, of course, the same as if the midpoints of the tops of the histogram bars were connected by straight lines. Instead of plotting frequencies on the vertical axis, one can plot *relative frequencies*, or proportions of the total frequency. This enables different distributions to be readily compared and even plotted on the same axes. Sometimes, as in Figure 1.8, frequency is indicated on one vertical axis and the corresponding relative frequency on the other. (Using the data of Example 1.5, the relative frequency for 8.2 mg/g is $2/130 = 0.015$, that for 8.3 mg/g is $6/130 = 0.046$, that for 9.2 mg/g is $4/130 = 0.030$, and so on. The total of all the frequencies is n, and the total of all the relative frequencies is 1.)

Frequency polygons are also commonly used for discrete distributions, but one can argue against their use when dealing with ordinal data, as the polygon implies to the reader a constant size interval horizontally between points on the polygon. Frequency polygons should not be employed for nominal-scale data.

If we have a frequency distribution of values of a continuous variable that falls into a large number of class intervals, the data may be grouped as was demonstrated with discrete variables. This results in fewer intervals, but each interval is, of course, larger. The midpoints of these intervals may then be used in the preparation of a histogram or frequency polygon. The user of frequency polygons is cautioned that such a graph is simply an aid to the eye in following trends in frequency distributions, and one should not attempt to read frequencies between points on the polygon. Also note that the method presented for the construction of histograms and frequency polygons requires that the class intervals be equal. Lastly, the vertical axis (e.g., the frequency scale) on frequency polygons and bar graphs generally should begin with zero, especially if graphs are to be compared with one another. If this is not done, the eye may be misled by the appearance of the graph (as shown for nominal-scale data in Figures 1.2 and 1.3).

1.4 CUMULATIVE FREQUENCY DISTRIBUTIONS

A frequency distribution informs us how many observations occurred for each value (or group of values) of a variable. That is, examination of the frequency table of Example 1.3 (or its corresponding bar graph or frequency polygon) would yield information such as, "How many fox litters of four were observed?", the answer being 27. But if it is desired to ask questions such as, "How many litters of four or more were observed?", or "How many fox litters of five or fewer were observed?", we are speaking of *cumulative frequencies*. To answer the first question, we sum all frequencies for litter sizes four and up, and for the second question, we sum all frequencies from the smallest litter size up through a size of five. We arrive at answers of 54 and 59, respectively.

In Example 1.5, the phosphorus concentration data are cast into two cumulative frequency distributions, one with cumulation commencing at the low end of the measurement scale and one with cumulation being performed from the high values toward the low values. The choice of the direction of cumulation is immaterial, as can be demonstrated. If one desired to calculate the number of phosphorus determinations less than 8.55 mg/g, namely 27, a cumulation starting at the low end might be used, whereas the knowledge of the frequency of determinations greater than 8.55 mg/g, namely 103, can be readily obtained from the cumulation commencing from the high end of the scale. But one can easily calculate any frequency from a low-to-high cumulation (e.g., 27) from its complementary frequency from a high-to-low cumulation (e.g., 103), simply by knowing that the sum of these two frequencies is the total frequency (i.e., $n = 130$); therefore, in practice it is not necessary to calculate both sets of cumulations.

Cumulative frequency distributions are useful in determining medians, percentiles, and other quantiles, as discussed in Sections 3.2 and 4.2. They are not often presented in bar graphs, but *cumulative frequency polygons* (sometimes called *ogives*) are not

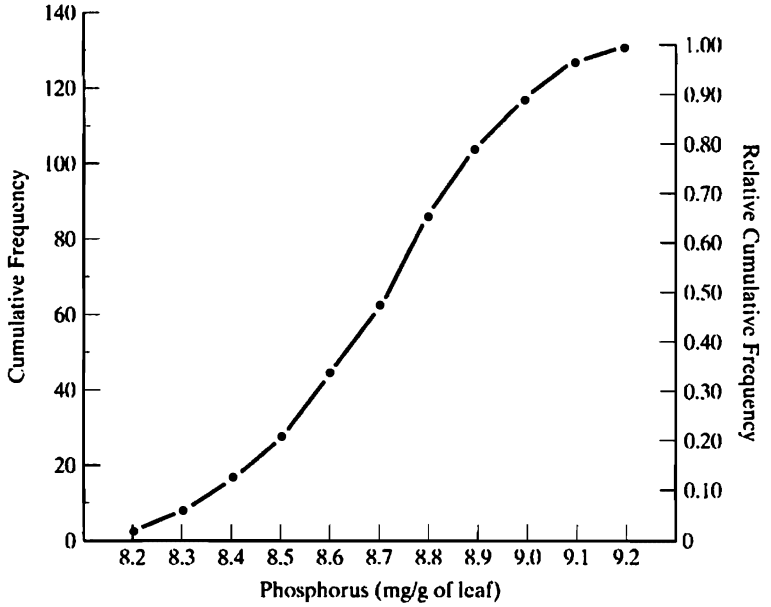


FIGURE 1.9: Cumulative frequency polygon of the leaf phosphorus data of Example 1.5, with cumulation commencing from the lowest to the highest values of the variable.

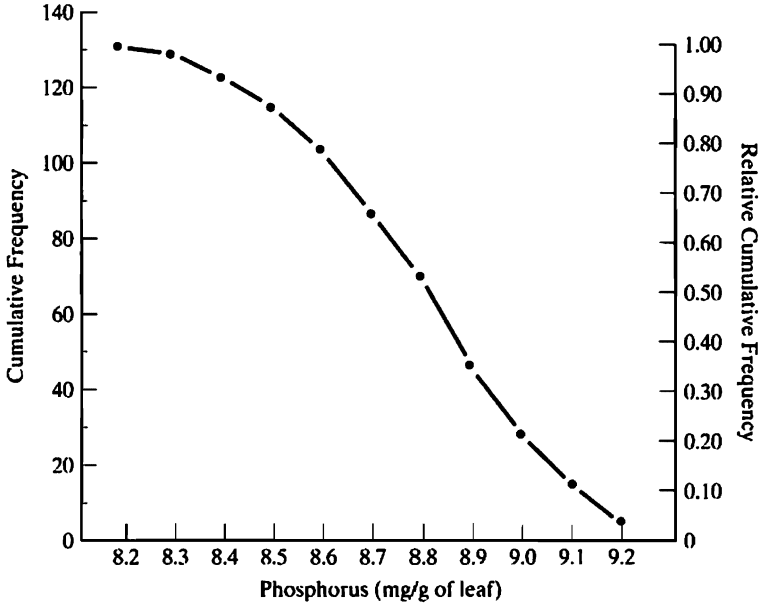


FIGURE 1.10: Cumulative frequency polygon of the leaf phosphorus data of Example 1.5, with cumulation commencing from the highest to the lowest values of the variable.

uncommon. (See Figures 1.9 and 1.10.) Relative frequencies (proportions of the total frequency) can be plotted instead of (or, as in Figures 1.9 and 1.10, in addition to) frequencies on the vertical axis of a cumulative frequency polygon. This enables different distributions to be readily compared and even plotted on the same axes. (Using the data of Example 1.5 for Figure 1.9, the relative cumulative frequency for 8.2 mg/g is $2/130 = 0.015$, that for 8.3 mg/g is $8/130 = 0.062$, and so on. For Figure 1.10, the relative cumulative frequency for 8.2 mg/g is $130/130 = 1.000$, that for 8.3 mg/g is $128/130 = 0.985$, and so on.)

CHAPTER 2

Populations and Samples

The primary objective of a statistical analysis is to infer characteristics of a group of data by analyzing the characteristics of a small sampling of the group. This generalization from the part to the whole requires the consideration of such important concepts as population, sample, parameter, statistic, and random sampling. These topics are discussed in this chapter.

2.1 POPULATIONS

Basic to statistical analysis is the desire to draw conclusions about a group of measurements of a variable being studied. Biologists often speak of a "population" as a defined group of humans or of another species of organisms. Statisticians speak of a *population* (also called a *universe*) as a group of measurements (not organisms) about which one wishes to draw conclusions. It is the latter definition, the statistical definition of *population*, that will be used throughout this book. For example, an investigator may desire to draw conclusions about the tail lengths of bobcats in Montana. All Montana bobcat tail lengths are, therefore, the population under consideration. If a study is concerned with the blood-glucose concentration in three-year-old children, then the blood-glucose levels in all children of that age are the population of interest.

Populations are often very large, such as the body weights of all grasshoppers in Kansas or the eye colors of all female New Zealanders, but occasionally populations of interest may be relatively small, such as the ages of men who have traveled to the moon or the heights of women who have swum the English Channel.

2.2 SAMPLES FROM POPULATIONS

If the population under study is very small, it might be practical to obtain all the measurements in the population. If one wishes to draw conclusions about the ages of all men who have traveled to the moon, it would not be unreasonable to attempt to collect all the ages of the small number of individuals under consideration. Generally, however, populations of interest are so large that obtaining all the measurements is unfeasible. For example, we could not reasonably expect to determine the body weight of every grasshopper in Kansas. What can be done in such cases is to obtain a subset of all the measurements in the population. This subset of measurements constitutes a *sample*, and from the characteristics of samples we can

draw conclusions about the characteristics of the populations from which the samples came.*

Biologists may sample a population that does not physically exist. Suppose an experiment is performed in which a food supplement is administered to 40 guinea pigs, and the sample data consist of the growth rates of these 40 animals. Then the population about which conclusions might be drawn is the growth rates of all the guinea pigs that conceivably might have been administered the same food supplement under identical conditions. Such a population is said to be "imaginary" and is also referred to as "hypothetical" or "potential."

2.3 RANDOM SAMPLING

Samples from populations can be obtained in a number of ways; however, for a sample to be representative of the population from which it came, and to reach valid conclusions about populations by induction from samples, statistical procedures typically assume that the samples are obtained in a *random* fashion. To sample a population randomly requires that each member of the population has an equal and independent chance of being selected. That is, not only must each measurement in the population have an equal chance of being chosen as a member of the sample, but the selection of any member of the population must in no way influence the selection of any other member. Throughout this book, "sample" will always imply "random sample."†

It is sometimes possible to assign each member of a population a unique number and to draw a sample by choosing a set of such numbers at random. This is equivalent to having all members of a population in a hat and drawing a sample from them while blindfolded. Appendix Table B.41 provides 10,000 random digits for this purpose. In this table, each digit from 0 to 9 has an equal and independent chance of appearing anywhere in the table. Similarly, each combination of two digits, from 00 to 99, is found at random in the table, as is each three-digit combination, from 000 to 999, and so on.

Assume that a random sample of 200 names is desired from a telephone directory having 274 pages, three columns of names per page, and 98 names per column. Entering Table B.41 at random (i.e., do not always enter the table at the same place), one might decide first to arrive at a random combination of three digits. If this three-digit number is 001 to 274, it can be taken as a randomly chosen page number (if it is 000 or larger than 274, simply skip it and choose another three-digit number, e.g., the next one on the table). Then one might examine the next digit in the table; if it is a 1, 2, or 3, let it denote a page column (if a digit other than 1, 2, or 3 is encountered, it is ignored, passing to the next digit that is 1, 2, or 3). Then one could look at the next two-digit number in the table; if it is from 01 to 98, let it represent a randomly selected name within that column. This three-step procedure would be performed a total of 200 times to obtain the desired random sample. One can proceed in any direction in the random number table: left to right, right to left, upward, downward, or diagonally; but the direction should be decided on before looking at the table. Computers are capable of quickly generating random numbers (sometimes called "pseudorandom" numbers because the number generation is not perfectly random), and this is how Table B.41 was derived.

---

*This use of the terms *population* and *sample* was established by Karl Pearson (1903).

†This concept of random sampling was established by Karl Pearson between 1897 and 1903 (Miller, 2004a).

Very often it is not possible to assign a number to each member of a population, and random sampling then involves biological, rather than simply mathematical, considerations. That is, the techniques for sampling Montana bobcats or Kansas grasshoppers require knowledge about the particular organism to ensure that the sampling is random. Researchers consult relevant books, periodical articles, or reports that address the specific kind of biological measurement to be obtained.

2.4 PARAMETERS AND STATISTICS

Several measures help to describe or characterize a population. For example, generally a preponderance of measurements occurs somewhere around the middle of the range of a population of measurements. Thus, some indication of a population "average" would express a useful bit of descriptive information. Such information is called a *measure of central tendency* (also called a *measure of location*), and several such measures (e.g., the mean and the median) will be discussed in Chapter 3.

It is also important to describe how dispersed the measurements are around the "average." That is, we can ask whether there is a wide spread of values in the population or whether the values are rather concentrated around the middle. Such a descriptive property is called a *measure of variability* (or a *measure of dispersion*), and several such measures (e.g., the range and the standard deviation) will be discussed in Chapter 4.

A quantity such as a measure of central tendency or a measure of dispersion is called a *parameter* when it describes or characterizes a population, and we shall be very interested in discussing parameters and drawing conclusions about them. Section 2.2 pointed out, however, that one seldom has data for entire populations, but nearly always has to rely on samples to arrive at conclusions about populations. Thus, one rarely is able to calculate parameters. However, by random sampling of populations, parameters can be estimated well, as we shall see throughout this book. An estimate of a population parameter is called a *statistic*.* It is statistical convention to represent population parameters by Greek letters and sample statistics by Latin letters; the following chapters will demonstrate this custom for specific examples.

The statistics one calculates will vary from sample to sample for samples taken from the same population. Because one uses sample statistics as estimates of population parameters, it behooves the researcher to arrive at the "best" estimates possible. As for what properties to desire in a "good" estimate, consider the following.

First, it is desirable that if we take an indefinitely large number of samples from a population, the long-run average of the statistics obtained will equal the parameter being estimated. That is, for some samples a statistic may underestimate the parameter of interest, and for others it may overestimate that parameter; but in the long run the estimates that are too low and those that are too high will "average out." If such a property is exhibited by a statistic, we say that we have an *unbiased* statistic or an unbiased estimator.

Second, it is desirable that a statistic obtained from any single sample from a population be very close to the value of the parameter being estimated. This property of a statistic is referred to as *precision*,† *efficiency*, or *reliability*. As we commonly secure only one sample from a population, it is important to arrive at a close estimate of a parameter from a single sample.

---

*This use of the terms *parameter* and *statistic* was defined by R. A. Fisher as early as 1922 (Miller, 2004a; Savage, 1976).

†The precision of a sample statistic, as defined here, should not be confused with the precision of a measurement, defined in Section 1.2.

Third, consider that one can take larger and larger samples from a population (the largest sample being the entire population). As the sample size increases, a *consistent* statistic will become a better estimate of the parameter it is estimating. Indeed, if the sample were the size of the population, then the best estimate would be obtained: the parameter itself.

In the chapters that follow, the statistics recommended as estimates of parameters are "good" estimates in the sense that they possess a desirable combination of unbiasedness, efficiency, and consistency.

2.5 OUTLIERS

Occasionally, a set of data will have one or more observations that are so different, relative to the other data in the sample, that we doubt they should be part of the sample. For example, suppose a researcher collected a sample consisting of the body weights of nineteen 20-week-old mallard ducks raised in individual laboratory cages, for which the following 19 data were recorded:

1.87, 3.75, 3.79, 3.82, 3.85, 3.87, 3.90, 3.94, 3.96, 3.99,
3.99, 4.00, 4.03, 4.04, 4.05, 4.06, 4.09, 8.97, and 39.8 kilograms.

Visual inspection of these 19 recorded data casts doubt upon the smallest datum (1.87 kg) and the two largest data (8.97 kg and 39.8 kg) because they differ so greatly from the rest of the weights in the sample. Data in striking disagreement with nearly all the other data in a sample are often called *outliers* or *discordant data*, and the occurrence of such observations generally calls for closer examination.

Sometimes it is clear that an outlier is the result of incorrect recording of data. In the preceding example, a mallard duck weight of 39.8 kg is highly unlikely (to say the least!), for that is about the weight of a 12-year-old boy or girl (and such a duck would probably not fit in one of the laboratory cages). In this case, inspection of the data records might lead us to conclude that this body weight was recorded with a careless placement of the decimal point and should have been 3.98 kg instead of 39.8 kg. And, upon interrogation, the research assistant may admit to weighing the eighteenth duck with the scale set to pounds instead of kilograms, so the metric weight of that animal should have been recorded as 4.07 (not 8.97) kg.

Also, upon further examination of the data-collection process, we may find that the 1.87-kg duck was taken from a wrong cage and was, in fact, only 4 weeks old, not 20 weeks old, and therefore did not belong in this sample. Or, perhaps we find that it was not a mallard duck, but some other bird species (and, therefore, did not belong in this sample). Statisticians say a sample is *contaminated* if it contains a datum that does not conform to the characteristics of the population being sampled. So the weight of a 4-week-old duck, or of a bird of a different species, would be a statistical contaminant and should be deleted from this sample.

There are also instances where it is known that a measurement was faulty—for example, when a laboratory technician spills coffee onto an electronic measuring device or into a blood sample to be analyzed. In such a case, the measurements known to be erroneous should be eliminated from the sample.

However, outlying data can also be correct observations taken from an intended population, collected purely by chance. As we shall see in Section 6.1, when drawing a random sample from a population, it is relatively likely that a datum in the sample will be around the average of the population and very unlikely that a sample datum will be dramatically far from the average. But sample data very far from the average still may be possible.

It should also be noted that in some situations the examination of an outlier may reveal the effect of a previously unsuspected factor. For example, the 1.87-kg duck might, indeed, have been a 20-week-old mallard but suffering from a genetic mutation or a growth-impeding disease deserving of further consideration in additional research.

In summary, it is not appropriate to discard data simply because they appear (to someone) to be unreasonably extreme. However, if there is a very obvious reason for correcting or eliminating a datum, such as the situations described previously, the incorrect data should be corrected or eliminated. In some other cases questionable data can be *accommodated* in statistical analysis, perhaps by employing statistical procedures that give them less weight or analytical techniques that are *robust* in that they are resistant to effects of discrepant data. And in situations when this cannot be done, dubious data will have to remain in the sample (perhaps encouraging the researcher to repeat the experiment with a new set of data).

The idea of rejecting erroneous data dates back over 200 years; and recommendations for formal, objective methods for such rejection began to appear about 150 years ago. Major discussions of outliers, their origin, and treatment (rejection or accommodation) are those of Barnett and Lewis (1994), Beckman and Cook (1983), and Thode (2002: 123–142).

CHAPTER 3

Measures of Central Tendency

In samples, as well as in populations, one generally finds a preponderance of values somewhere around the middle of the range of observed values. The description of this concentration near the middle is an *average*, or a *measure of central tendency* to the statistician. It is also termed a *measure of location*, for it indicates where, along the measurement scale, the sample or population is located. Various measures of central tendency are useful population parameters, in that they describe an important property of populations. This chapter discusses the characteristics of these parameters and the sample statistics that are good estimates of them.

3.1 THE ARITHMETIC MEAN

The most widely used measure of central tendency is the *arithmetic mean*,* usually referred to simply as the *mean*,† which is the measure most commonly called an "average."

Each measurement in a population may be referred to as an X_i (read "X sub i") value. Thus, one measurement might be denoted as X_1, another as X_2, another as X_3, and so on. The subscript i might be any integer value up through N, the total number of X values in the population.‡ The mean of the population is denoted by the Greek letter μ (lowercase mu) and is calculated as the sum of all the X_i values divided by the size of the population.

The calculation of the population mean can be abbreviated concisely by the formula

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}. \tag{3.1}$$

*As an adjective, *arithmetic* is pronounced with the accent on the third syllable. In early literature on the subject, the adjective *arithmetical* was employed.

†The term *mean* (as applied to the arithmetic mean, as well as to the geometric and harmonic means of Section 3.4) dates from ancient Greece (Walker, 1929: 183), with its current statistical meaning in use by 1755 (Miller, 2004a; Walker, 1929: 176); *central tendency* appeared by the late 1920s (Miller, 2004a).

‡Charles Babbage (1791–1871) (O'Connor and Robertson, 1998) was an English mathematician and inventor who conceived principles used by modern computers—well before the advent of electronics—and who, in 1832, proposed the modern convention of italicizing Latin (also called Roman) letters to denote quantities; nonitalicized letters had already been employed for this purpose for more than six centuries (Miller, 2001).

The Greek letter Σ (capital sigma) means "summation"* and $\sum_{i=1}^{N} X$ means "summation of all X_i values from X_1 through X_N." Thus, for example, $\sum_{i=1}^{4} X_i = X_1 + X_2 + X_3 + X_4$ and $\sum_{i=3}^{5} X_i = X_3 + X_4 + X_5$. Since, in statistical computations, summations are nearly always performed over the entire set of X_i values, this book will assume $\sum X_i$ to mean "sum X_i's over all values of i," simply as a matter of printing convenience, and $\mu = \sum X_i / N$ would therefore designate the same calculation as would $\mu = \sum_{i=1}^{N} X_i / N$.

The most efficient, unbiased, and consistent estimate of the population mean, μ, is the sample mean, denoted as $\overline{X}$ (read as "X bar"). Whereas the size of the population (which we generally do not know) is denoted as N, the size of a sample is indicated by n, and $\overline{X}$ is calculated as

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{or} \quad \overline{X} = \frac{\sum X_i}{n}, \tag{3.2}$$

which is read "the sample mean equals the sum of all measurements in the sample divided by the number of measurements in the sample."† Example 3.1 demonstrates the calculation of the sample mean. Note that the mean has the same units of measurement as do the individual observations. The question of how many decimal places should be reported for the mean will be answered at the end of Section 6.2; until then we shall simply record the mean with one more decimal place than the data.

EXAMPLE 3.1 A Sample of 24 from a Population of Butterfly Wing Lengths

X_i (in centimeters): 3.3, 3.5, 3.6, 3.6, 3.7, 3.8, 3.8, 3.8, 3.9, 3.9, 3.9, 4.0, 4.0, 4.0, 4.0, 4.1, 4.1, 4.1, 4.2, 4.2, 4.3, 4.3, 4.4, 4.5.

$$\sum X_i = 95.0 \text{ cm}$$
$$n = 24$$
$$\overline{X} = \frac{\sum X_i}{n} = \frac{95.0 \text{ cm}}{24} = 3.96 \text{ cm}$$

*Mathematician Leonhard Euler (1707–1783; born in Switzerland, worked mostly in Russia), in 1755, was the first to use Σ to denote summation (Cajori, 1928/9, Vol. II: 61).

†The modern symbols for plus and minus ("+" and "−") appear to have first appeared in a 1456 unpublished manuscript by German mathematician and astronomer Regiomontanus (Johannes Müller, 1436–1476), with Bohemia-born Johann (Johannes) Widman (1562–1498) the first, in 1489, to use them in print (Cajori, 1928/9, Vol. I: 128, 231–232). The modern equal sign ("=") was invented by Welsh physician and mathematician Robert Recorde (1510–1558), who published it in 1557 (though its use then disappeared in print until 1618), and it was well recognized starting in 1631 (Cajori, ibid.: 298; Gullberg, 1997: 107). Recorde also was the first to use the plus and minus symbols in an English work (Miller, 2004b). Using a horizontal line to express division derives from its use, in denoting fractions, by Arabic author Al-Ḥaṣṣâr in the twelfth century, though it was not consistently employed for several more centuries (Cajori, ibid. I: 269, 310). The slash mark ("/"; also known as a solidus, virgule, or diagonal) was recommended to denote division by the English logician and mathematician Augustus De Morgan (1806–1871) in 1845 (ibid. I: 312–313), and the India-born Swiss author Johann Heinhirch Rahn (1622–1676) proposed, in 1659, denoting division by the symbol "÷", which previously was often used by authors as a minus sign (ibid.: 211, 270; Gullberg, 1997: 105). Many other symbols were used for mathematical operations, before and after these introductions (e.g., Cajori, ibid.: 229–245).

If, as in Example 3.1, a sample contains multiple identical data for several values of the variable, then it may be convenient to record the data in the form of a frequency table, as in Example 3.2. Then X_i can be said to denote each of k different measurements and f_i can denote the frequency with which that X_i occurs in the sample. The sample mean may then be calculated, using the sums of the products of f_i and X_i, as*

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i X_i}{n}. \tag{3.3}$$

Example 3.2 demonstrates this calculation for the same data as in Example 3.1.

EXAMPLE 3.2 The Data from Example 3.1 Recorded as a Frequency Table

X_i (cm)	f_i	$f_i X_i$ (cm)
3.3	1	3.3
3.4	0	0
3.5	1	3.5
3.6	2	7.2
3.7	1	3.7
3.8	3	11.4
3.9	3	11.7
4.0	4	16.0
4.1	3	12.3
4.2	2	8.4
4.3	2	8.6
4.4	1	4.4
4.5	1	4.5
	$\sum f_i = 24$	$\sum f_i X_i = 95.0$ cm

$$k = 13$$

$$\sum_{i=1}^{k} f_i = n = 24$$

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i X_i}{n} = \frac{95.0 \text{ cm}}{24} = 3.96 \text{ cm}$$

$$\text{median} = 3.95 \text{ cm} + \left(\frac{1}{4}\right)(0.1 \text{ cm})$$
$$= 3.95 \text{ cm} + 0.025 \text{ cm}$$
$$= 3.975 \text{ cm}$$

A similar procedure is computing what is called a *weighted mean*, an expression of the average of several means. For example, we may wish to combine the mean of 3.96 cm from the sample of 24 measurements in Example 3.1 with a mean of 3.78 cm from a sample of 30 measurements and a mean of 4.02 cm from a sample of 15. These three means would be from a total of 24 + 30 + 15 = 69 data; and if we had all 69 of the data we could sum them and divide the sum by 69 to obtain the overall mean length. However, that overall mean can be obtained without knowing the 69

*Denoting the multiplication of two quantities (e.g., a and b) by their adjacent placement (i.e., ab) derives from practices in Hindu manuscripts of the seventh century (Cajori, 1928/9, Vol. I: 77, 250). Modern multiplication symbols include a raised dot (as in $a \cdot b$), which was suggested in a 1631 posthumous publication of Thomas Harriot (1560?–1621) and prominently adopted in 1698 by the outstanding mathematician Gottfried Wilhelm Leibniz (1646–1716, in what is now Germany); the St. Andrew's cross (as in $a \times b$), which was used in 1631 by English mathematician William Oughtred (1574–1660) though it was not in general use until more than 200 years later; and the letter X, which was used, perhaps by Oughtred, as early as 1618 (Cajori, ibid.: 251; Gullberg, 1997: 104; Miller 2004b). Johann Rahn's 1659 use of an asterisk-like symbol (as in $a * b$) (Cajori, ibid: 212–213) did not persist but resurfaced in electronic computer languages of the latter half of the twentieth century.

individual measurements, by employing Equation 3.3 with $f_1 = 24$, $X_1 = 3.96$ cm, $f_2 = 30$, $X_2 = 3.78$ cm, $f_3 = 15$, $X_3 = 4.02$ cm, and $n = 69$. This would yield a weighted mean of $\overline{X} = [(24)(3.96\text{ cm}) + (30)(3.78\text{ cm}) + (15)(4.02\text{ cm})]/69 = (268.74\text{ cm})/69 = 3.89\text{ cm}$.

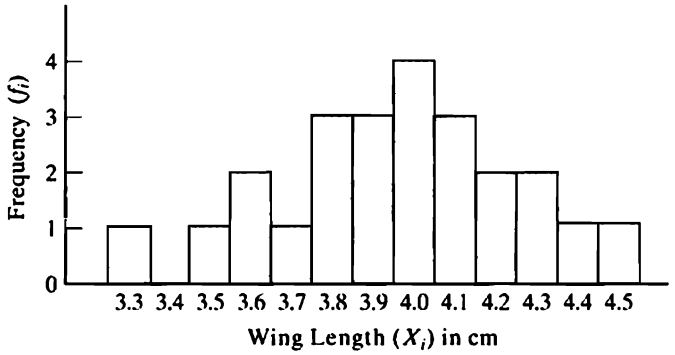


FIGURE 3.1: A histogram of the data in Example 3.2. The mean (3.96 cm) is the center of gravity of the histogram, and the median (3.975 cm) divides the histogram into two equal areas.

If data are plotted as a histogram (Figure 3.1), the mean is the *center of gravity* of the histogram.* That is, if the histogram were made of a solid material, it would balance horizontally with the fulcrum at $\overline{X}$. The mean is applicable to both ratio- and interval-scale data; it should not be used for ordinal data and cannot be used for nominal data.

3.2 THE MEDIAN

The median is typically defined as the middle measurement in an ordered set of data.† That is, there are just as many observations larger than the median as there are smaller. The sample median is the best estimate of the population median. In a symmetrical distribution (such as Figures 3.2a and 3.2b) the sample median is also an unbiased and consistent estimate of μ, but it is not as efficient a statistic as $\overline{X}$ and should not be used as a substitute for $\overline{X}$. If the frequency distribution is asymmetrical, the median is a poor estimate of the mean.

The median of a sample of data may be found by first arranging the measurements in order of magnitude. The order may be either ascending or descending, but ascending order is most commonly used as is done with the samples in Examples 3.1, 3.2, and 3.3. Then, we define the sample median as

$$\text{sample median} = X_{(n+1)/2}. \tag{3.4}$$

*The concept of the mean as the center of gravity was used by L. A. J. Quetelet in 1846 (Walker, 1929: 73).

†The concept of the median was conceived as early as 1816, by K. F. Gauss; enunciated and reinforced by others, including F. Galton in 1869 and 1874; and independently discovered and promoted by G. T. Fechner beginning in 1874 (Walker, 1929: 83–88, 184). It received its name, in English, from F. Galton in 1882 (David, 1995) and, in French, from A. A. Cournot in 1843 (David, 1998a).

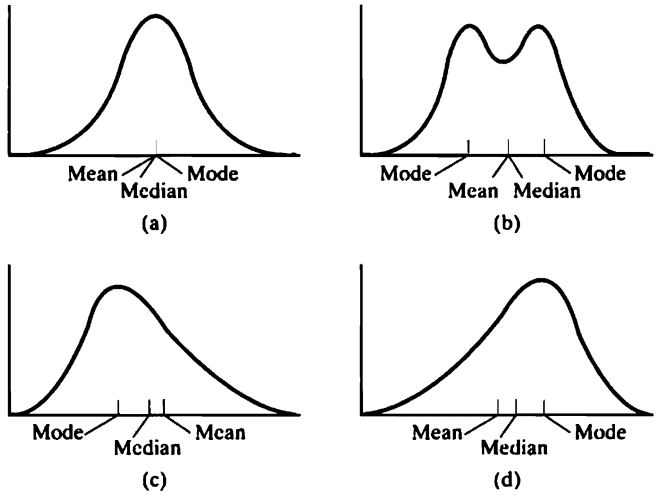


FIGURE 3.2: Frequency distributions showing measures of central tendency. Values of the variable are along the abscissa (horizontal axis), and the frequencies are along the ordinate (vertical axis). Distributions (a) and (b) are symmetrical, (c) is asymmetrical and said to be positively skewed, and (d) is asymmetrical and said to be negatively skewed. Distributions (a), (c), and (d) are unimodal, and distribution b is bimodal. In a unimodal asymmetric distribution, the median lies about one-third the distance between the mean and the mode.*

EXAMPLE 3.3 Life Span for Two Species of Birds in Captivity

The data for each species are arranged in order of magnitude

Species A X_i (mo)	*Species B* X_i (mo)
16	34
32	36
37	38
39	45
40	50
41	54
42	56
50	59
82	69
	91
$n = 9$	$n = 10$
median $= X_{(n+1)/2} = X_{(9+1)/2}$	median $= X_{(n+1)/2} = X_{(10+1)/2}$
$= X_5 = 40$ mo	$= X_{5.5} = 52$ mo
$\overline{X} = 42.11$ mo	$\overline{X} = 53.20$ mo

*An interesting relationship among the mean, median, and standard deviation is shown in Equation 4.21.

If the sample size (n) is odd, then the subscript in Equation 3.4 will be an integer and will indicate which datum is the middle measurement in the ordered sample. For the data of species A in Example 3.3, $n = 9$ and the sample median is $X_{(n+1)/2} = X_{(9+1)/2} = X_5 = 40$ mo. If n is even, then the subscript in Equation 3.4 will be a number midway between two integers. This indicates that there is not a middle value in the ordered list of data; instead, there are two middle values, and the median is defined as the midpoint between them. For the species B data in Example 3.3, $n = 10$ and $X(n + 1)/2 = X_{(10+1)/2} = X_{5.5}$, which signifies that the median is midway between X_5 and X_6, namely a median of (50 mo + 54 mo)/2 = 52 mo.

Note that the median has the same units as each individual measurement. If data are plotted as a frequency histogram (e.g., Figure 3.1), the median is the value of X that divides the area of the histogram into two equal parts. In general, the sample median is a more efficient estimate of the population median when the sample size is large.

If we find the middle value(s) in an ordered set of data to be among identical observations (referred to as *tied* values), as in Example 3.1 or 3.2, a difficulty arises. If we apply Equation 3.4 to these 24 data, then we conclude the median to be $X_{12.5} = 4.0$ cm. But four data are tied at 4.0 cm, and eleven measurements are less than 4.0 cm and nine are greater. Thus, 4.0 cm does not fit the definition above of the median as that value for which there is the same number of data larger and smaller. Therefore, a better definition of the median of a set of data is that value for which no more than half the data are smaller and no more than half are larger.

When the sample median falls among tied observations, we may interpolate to better estimate the population median. Using the data of Example 3.2, we desire to estimate a value below which 50% of the observations in the population lie. Fifty percent of the observations in the sample would be 12 observations. As the first 7 classes in the frequency table include 11 observations and 4 observations are in class 4.0 cm, we know that the desired sample median lies within the range of 3.95 to 4.05 cm. Assuming that the four observations in class 4.0 cm are distributed evenly within the 0.1-cm range of 3.95 to 4.05 cm, then the median will be $\left(\frac{1}{4}\right)(0.1\text{ cm}) = 0.025$ cm into this class. Thus, the median = 3.95 cm + 0.025 cm = 3.975 cm. In general, for the sample median within a class interval containing tied observations,

$$\text{median} = \begin{pmatrix}\text{lower limit}\\ \text{of interval}\end{pmatrix} + \left(\frac{0.5n - \text{cum. freq.}}{\text{no. of observations in interval}}\right)\begin{pmatrix}\text{interval}\\ \text{size}\end{pmatrix}, \quad (3.5)$$

where "cum. freq." refers to the cumulative frequency of the previous classes.* By using this procedure, the calculated median will be the value of X that divides the area of the histogram of the sample into two equal parts. As another example, refer back to Example 1.5, where, by Equation 3.5, median = 8.75 mg/g + {[(0.5)(130) − 61]/24}{0.10 mg/g} = 8.75 mg/g + 0.02 mg/g = 8.77 mg/g.

The median expresses less information than does the mean, for it does not take into account the actual value of each measurement, but only considers the rank of each measurement. Still, it offers advantages in some situations. For example, extremely high or extremely low measurements ("outliers"; Section 2.5) do not affect the median as much as they affect the mean (causing the sample median to be called a "resistant" statistic). Distributions that are not symmetrical around the mean (such as in Figures 3.2c and 3.2d) are said to be *skewed*.† When we deal with skewed

*This procedure was enunciated in 1878 by the German psychologist Gustav Theodor Fechner (1801–1887) (Walker, 1929: 86).

†This term, applied to a distribution and to a curve, was used as early as 1895 by Karl Pearson (Miller, 2004a).

populations and do not want the strong influence of outliers, we may prefer the median to the mean to express central tendency.

Note that in Example 3.3 the researcher would have to wait 82 months to compute a mean life expectancy for species *A* and 91 months for species *B*, whereas the median for species *A* could be determined in only 40 months and in only 52 months for species *B*. Also, to calculate a median one does not need to have accurate data for all members of the sample. If, for example, we did not have the first three data for species *A* accurately recorded, but could state them as "less than 39 months," then the median could have been determined just as readily as if we had all 9 data fully recorded, while calculation of the mean would not have been possible.

The expression "LD fifty" (LD_{50}), used in some areas of biological research, is simply the median lethal dose (and is so named because the median is the 50th percentile, as we shall see in Section 4.2).

The median can be determined not only for interval-scale and ratio-scale data, but also for data on an ordinal scale, data for which the use of the mean usually would not be considered appropriate. But neither the median nor the mean is applicable to nominal data.

3.3 THE MODE

The *mode* is commonly defined as the most frequently occurring measurement in a set of data.* In Example 3.2, the mode is 4.0 cm. But it is perhaps better to define a mode as a measurement of relatively great concentration, for some frequency distributions may have more than one such point of concentration, even though these concentrations might not contain precisely the same frequencies. Thus, a sample consisting of the data 6, 7, 7, 8, 8, 8, 8, 8, 8, 9, 9, 10, 11, 12, 12, 12, 12, 12, 13, 13, and 14 mm would be said to have two modes: at 8 mm and 12 mm. (Some authors would refer to 8 mm as the "major mode" and call 12 mm the "minor mode.") A distribution in which each different measurement occurs with equal frequency is said to have no mode. If two consecutive values of X have frequencies great enough to declare the X values modes, the mode of the distribution may be said to be the midpoint of these two X's; for example, the mode of 3, 5, 7, 7, 7, 8, 8, 8, and 10 liters is 7.5 liters. A distribution with two modes is said to be *bimodal* (e.g., Figure 3.2b) and may indicate a combination of two distributions with different modes (e.g., heights of men and women). Modes are often discerned from histograms or frequency polygons; but we should be aware that the shape of such graphs (such as Figures 1.6, 1.7, and 1.8), and therefore the appearance of modes, may be influenced by the measurement intervals on the horizontal axis.

The sample mode is the best estimate of the population mode. When we sample a symmetrical unimodal population, the mode is an unbiased and consistent estimate of the mean and median (Figure 3.2a), but it is relatively inefficient and should not be so used. As a measure of central tendency, the mode is affected by skewness less than is the mean or the median, but it is more affected by sampling and grouping than these other two measures. The mode, but neither the median nor the mean, may be used for data on the nominal, as well as the ordinal, interval, and ratio scales of measurement. In a unimodal asymmetric distribution (Figures 3.2c and 3.2d), the median lies about one-third the distance between the mean and the mode.

The mode is not often used in biological research, although it is often interesting to report the number of modes detected in a population, if there are more than one.

*The term *mode* was introduced by Karl Pearson in 1895 (David, 1995).

3.4 OTHER MEASURES OF CENTRAL TENDENCY

(a) The Geometric Mean. The *geometric mean* is the nth root* of the product of the n data:

$$\overline{X}_G = \sqrt[n]{X_1 X_2 X_3 \ldots X_n} = \sqrt[n]{\prod_{i=1}^{n} X_i}. \tag{3.6}$$

Capital Greek pi, Π, means "take the product"† in an analogous fashion as Σ indicates "take the sum." The geometric mean may also be calculated as the antilogarithm of the arithmetic mean of the logarithms of the data (where the logarithms may be in any base); this is often more feasible computationally:

$$\overline{X}_G = \text{antilog}\left(\frac{\log X_1 + \log X_2 + \cdots + \log X_n}{n}\right) = \text{antilog}\frac{\sum_{i=1}^{n} \log X_i}{n}. \tag{3.7}$$

The geometric mean is appropriate to use only for ratio-scale data and only when all of the data are positive (that is, greater than zero). If the data are all equal, then the geometric mean, $\overline{X}_G$, is equal to the arithmetic mean, $\overline{X}$ (and also equal to the harmonic mean described below); if the data are not all equal, then‡ $\overline{X}_G < \overline{X}$.

$\overline{X}_G$ is sometimes used as a measure of location when the data are highly skewed to the right (i.e., when there are many more data larger than the arithmetic mean than there are data smaller than the arithmetic mean).

$\overline{X}_G$ is also useful when dealing with data that represent ratios of change. As an illustration of this, Example 3.4 considers changes in the size of a population of organisms over four decades. Each of the original data (population size at the end of a decade) is expressed as a ratio, X_i, of the population size to the population size of the previous decade. The geometric mean of those ratios is computed and may be thought of as representing the average rate of growth per decade (which is the same as a constant rate of compound interest). This example demonstrates that the arithmetic mean of those ratios is $\overline{X} = 1.1650$ (i.e., 16.50% growth) per decade. But over the four decades of population change, this mean would have us calculate a final population size of (10,000)(1.1650)(1.1650)(1.1650)(1.1650) = 18,421, which is *not* the population size recorded at the end of the fourth decade. However, using the geometric mean, $\overline{X}_G$, to indicate the average rate of growth, the final population size would be computed to be (10,000)(1.608)(1.608)(1.608)(1.608) = 18,156, which is the fourth-decade population size that was observed.

*The second footnote in Section 4.5 outlines the origin of the square-root symbol, $\surd$; indicating the cube root as $\sqrt[3]{}$ was suggested by Albert Girard (1595–1632, French-born but studied and worked in the Netherlands) as early as 1629, but this symbol was not generaly used until well into the eighteenth century (Cajori, 1928/9, Vol. I: 371–372). The cube-root symbol eventually was expanded to $\sqrt[n]{}$ to denote the nth root.

†Use of this symbol to indicate taking the product was introduced by René Descartes (Gullberg, 1997: 105).

‡The symbols "<" and ">" (meaning "less than" and "greater than") were inserted by someone else into a 1631 posthumous publication by the English mathematician and astronomer Thomas Harriot (1560?–1621), (Cajori, 1928/9, Vol. I: 199; Gullberg, 1997: 109; Miller, 2004b). The symbols for "less than or equal to" ($\leq$) and "greater than or equal to" ($\geq$) were written as $\leqq$ and $\geqq$ when introduced by the French scientist Pierre Bouguere (1698–1758) in 1734. (Gullberg, 1997: 109).

EXAMPLE 3.4 The Geometric Mean of Ratios of Change

Decade	*Population Size*	*Ratio of Change* X_i
0	10,000	
1	10,500	$\frac{10{,}500}{10{,}000} = 1.05$
2	11,550	$\frac{11{,}550}{10{,}500} = 1.10$
3	13,860	$\frac{13{,}860}{11{,}550} = 1.20$
4	18,156	$\frac{18{,}156}{13{,}860} = 1.31$

$$\overline{X} = \frac{1.05 + 1.10 + 1.20 + 1.31}{4} = \frac{4.66}{4} = 1.1650$$

and $(10{,}000)(0.1650)(1.650)(1.650)(1.650) = 18{,}421$

But,

$$\overline{X}_G = \sqrt[4]{(1.05)(1.10)(1.20)(1.31)} = \sqrt[4]{1.8157} = 1.1608$$

or

$$\overline{X}_G = \text{antilog}\left[\frac{\log(1.05) + \log(1.10) + \log(1.20) + \log(1.31)}{4}\right]$$

$$= \frac{\text{antilog}(0.0212 + 0.0414 + 0.0792 + 0.1173)}{4} = \frac{\text{antilog}(0.2591)}{4}$$

$$= \text{antilog}\, 0.0648 = 1.1608$$

and $(10{,}000)(1.1608)(1.1608)(1.1608)(1.1608) = 18{,}156$

(b) The Harmonic Mean. The *harmonic mean* is the reciprocal of the arithmetic mean of the reciprocals of the data:

$$\overline{X}_H = \frac{1}{\frac{1}{n}\sum \frac{1}{X_i}} = \frac{n}{\sum \frac{1}{X_i}}. \tag{3.8}$$

It may be used for ratio-scale data when no datum is zero. If all of the data are identical, then the harmonic mean, $\overline{X}_H$, is equal to the arithmetic mean, $\overline{X}$ (and equal to the geometric mean, $\overline{X}_G$). If the data are all positive and not identical, then $\overline{X}_H < \overline{X}_G < \overline{X}$.

$\overline{X}_H$ finds use when desiring an average of rates, as described by Croxton, Cowden, and Klein (1967: 182–188). For example, consider that a flock of birds flies from a roosting area to a feeding area 20 km away, flying at a speed of 40 km/hr (which

takes 0.5 hr). The flock returns to the roosting area along the same route (20 km), flying at 20 km/hr (requiring 1 hr of flying time). To ask what the average flying speed was, we might employ Equation 3.2 and calculate the arithmetic mean as $\overline{X} = (40 \text{ km/hr} + 20 \text{ km/hr})/2 = 30 \text{ km/hr}$. However, this answer may not be satisfying, because a total of 40 km was traveled in 1.5 hr, indicating a speed of $(40 \text{ km})/(1.5 \text{ hr}) = 26.7 \text{ km/hr}$. Example 3.5 shows that the harmonic mean ($\overline{X}_G$) is 26.7 km/hr.

EXAMPLE 3.5 The Harmonic Mean of Rates

$$X_1 = 40 \text{ km/hr}, \; X_2 = 20 \text{ km/hr}$$

$$\overline{X} = \frac{40 \text{ km/hr} + 20 \text{ km/hr}}{2} = \frac{60 \text{ km/hr}}{2} = 30 \text{ km/hr}$$

But

$$\overline{X}_H = \frac{2}{\frac{1}{40 \text{ km/hr}} + \frac{1}{20 \text{ km/hr}}} = \frac{2}{0.0250 \text{ hr/km} + 0.0500 \text{ hr/km}}$$

$$= \frac{2}{0.075 \text{ hr/km}} = 26.67 \text{ km/hr}$$

(c) The Range Midpoint. The *range midpoint*, or *midrange*, is a measure of location defined as the point halfway between the minimum and the maximum values in the set of data. It may be used with data measured on the ratio, interval, or ordinal scale; but it is not generally a good estimate of location, for it utilizes relatively little information from the data. (However, the so-called mean daily temperature is often reported as the mean of the minimum and maximum and is, therefore, a range midpoint.)

The midpoint of any two symmetrically located percentiles (see Section 4.2), such as the point midway between the first and third quartiles (i.e., the 25th and 75th percentiles), may be used as a location measure in the same fashion as the range midpoint is used (see Dixon and Massey, 1969: 133–134). Such measures are not as adversely affected by aberrantly extreme values as is the range midpoint, and they may be applied to ratio or interval data. If used with ordinal data, they (and the range midpoint) would be the same as the median.

3.5 CODING DATA

Often in the manipulation of data, considerable time and effort can be saved if *coding* is employed. Coding is the conversion of the original measurements into easier-to-work-with values by simple arithmetic operations. Generally coding employs a *linear transformation* of the data, such as multiplying (or dividing) or adding (or subtracting) a constant. The addition or subtraction of a constant is sometimes termed a translation of the data (i.e., changing the origin), whereas the multiplication or division by a constant causes an expansion or contraction of the scale of measurement.

EXAMPLE 3.6 Coding Data to Facilitate Calculations

Sample 1 (Coding by Subtraction: $A = -840$ g)		**Sample 2 (Coding by Division: $M = 0.001$ liters/ml)**	
X_i (g)	coded $X_i = X_i - 840$ g	X_i (ml)	coded $X_i = (X_i)(0.001 \text{ liters/ml}) = X_i$ liters
842	2	8,000	8.000
844	4	9,000	9.000
846	6	9,500	9.500
846	6	11,000	11.000
847	7	12,500	12.500
848	8	13,000	13.000
849	9		
$\sum X_i = 5922$ g	coded $\sum X_i = 42$ g	$\sum X_i = 63{,}000$ ml	coded $\sum X_i = 63.000$ liters
$\overline{X} = \frac{5922 \text{ g}}{7} = 846$ g	coded $\overline{X} = \frac{42 \text{ g}}{7} = 6$ g	$\overline{X} = 10{,}500$ ml	coded $\overline{X} = 10.500$ liters
$\overline{X} = \text{coded } \overline{X} - A = 6 \text{ g} - (-840 \text{ g}) = 846$ g		$\overline{X} = \text{coded } \frac{\overline{X}}{M} = \frac{10.500 \text{ liters}}{0.001 \text{ liters/ml}} = 10{,}500$ ml	

The first set of data in Example 3.6 are coded by subtracting a constant value of 840 g. Not only is each coded value equal to $X_i - 840$ g, but the mean of the coded values is equal to $\overline{X} - 840$ g. Thus, the easier-to-work-with coded values may be used to calculate a mean that then is readily converted to the mean of the original data, simply by adding back the coding constant.

In Sample 2 of Example 3.6, the observed data are coded by dividing each observation by 1000 (i.e., by multiplying by 0.001).* The resultant mean only needs to be multiplied by the coding factor of 1000 (i.e., divided by 0.001) to arrive at the mean of the original data. As the other measures of central tendency have the same units as the mean, they are affected by coding in exactly the same fashion.

Coding affects the median and mode in the same way as the mean is affected. The widespread use of computers has greatly diminished the need for researchers to

*In 1593, mathematician Christopher Clavius (1538–1612, born in what is now Germany but spent most of his life in what is now Italy; also credited with proposing the currently used Gregorian calendar rules regarding leap years: O'Connor and Robertson, 1996) became the first to use a decimal point to separate units from tenths; in 1617, the Scottish mathematician John Napier (1550–1617) used both points and commas for this purpose (Cajori, 1928/9. Vol. I: 322–323), and the comma is still so used in some parts of the world. In some countries a raised dot has been used—a symbol Americans sometimes employ to denote multiplication.

utilize coding (although computer software may use it). Appendix C presents coding for a variety of statistics.

EXERCISES

3.1. If $X_1 = 3.1\,\text{kg}$, $X_2 = 3.4\,\text{kg}$, $X_3 = 3.6\,\text{kg}$, $X_4 = 3.7\,\text{kg}$, and $X_5 = 4.0\,\text{kg}$, calculate the value of

(a) $\sum_{i=1}^{4} X_i$.

(b) $\sum_{i=2}^{4} X_i$.

(c) $\sum_{i=1}^{5} X_i$.

(d) $\sum X_i$.

3.2. **(a)** Calculate the mean of the five weights in Exercise 3.1.

(b) Calculate the median of those weights.

3.3. The ages, in years, of the faculty members of a university biology department are 32.2, 37.5, 41.7, 53.8, 50.2, 48.2, 46.3, 65.0, and 44.8.

(a) Calculate the mean age of these nine faculty members.

(b) Calculate the median of the ages.

(c) If the person 65.0 years of age retires and is replaced on the faculty with a person 46.5 years old, what is the new mean age?

(d) What is the new median age?

3.4. Consider the following frequency tabulation of leaf weights (in grams):

X_i	f_i
1.85–1.95	2
1.95–2.05	1
2.05–2.15	2
2.15–2.25	3
2.25–2.35	5
2.35–2.45	6
2.45–2.55	4
2.55–2.65	3
2.65–2.75	1

Using the midpoints of the indicated ranges of X_i,

(a) Calculate the mean leaf weight using Equation 3.2, and

(b) Calculate the mean leaf weight using Equation 3.3.

(c) Calculate the median leaf weight using Equation 3.4, and

(d) Calculate the median using Equation 3.5.

(e) Determine the mode of the frequency distribution.

3.5. A fruit was collected from each of eight lemon trees, with the intent of measuring the calcium concentration in the rind (grams of calcium per 100 grams of dry rind). The analytical method used could only detect a concentration of at least 0.80 g/100 g of dry weight. Six of the eight concentrations were measured to be 1.02, 0.98, 0.91, 0.84, 0.87, 1.04 g/100 g of dry weight, and two of the concentrations were known to be less than 0.80 g/100 g of dry weight. What is the median of this sample of eight data?

CHAPTER 4

Measures of Variability and Dispersion

In addition to a description of the central tendency of a set of data, it is generally desirable to have a description of the *variability*, or of the *dispersion*,* of the data. A measure of variability (or measure of dispersion, as it is often called) is an indication of the spread of measurements around the center of the distribution. Measurements that are concentrated around the center of a distribution of data have low variability (low dispersion), whereas data that are very spread out along the measurement scale have high variability (high dispersion). Measures of variability of a population are population parameters, and sample measures of variability are statistics that estimate those parameters.

4.1 THE RANGE

The difference between the highest and lowest measurements in a group of data is termed the *range*.† If sample measurements are arranged in increasing order of magnitude, as if the median were about to be determined, then

$$\text{sample range} = X_n - X_1, \tag{4.1}$$

which is

$$\text{sample range} = \text{largest } X - \text{smallest } X.$$

Sample 1 in Example 4.1 is a hypothetical set of ordered data in which $X_1 = 1.2$ g and $X_n = 2.4$ g. Thus, the range may be expressed as 1.2 to 2.4 g, or as 2.4 g − 1.2 g = 1.2 g. Note that the range has the same units as the individual measurements. Sample 2 in Example 4.1 has the same range as Sample 1.

*The statistical use of this term first appeared in an 1876 publication by Francis Galton (David, 1998a).

†This statistical term dates from an 1848 paper by H. Lloyd (David, 1995). It was already used by the Greek astronomer Hipparchus as a measure of dispersion in the second century B.C.E. (David, 1998b).

EXAMPLE 4.1 Calculation of Measures of Dispersion for Two Hypothetical Samples of 7 Insect Body Weights

Sample 1

X_i (g)	$X_i - \overline{X}$ (g)	$\lvert X_i - \overline{X}\rvert$ (g)	$(X_i - \overline{X})^2$ (g^2)
1.2	−0.6	0.6	0.36
1.4	−0.4	0.4	0.16
1.6	−0.2	0.2	0.04
1.8	0.0	0.0	0.00
2.0	0.2	0.2	0.04
2.2	0.4	0.4	0.16
2.4	0.6	0.6	0.36
$\sum X_i$ = 12.6 g	$\sum(X_i - \overline{X})$ = 0.0 g	$\sum\lvert X_i - \overline{X}\rvert$ = 2.4 g	$\sum(X_i - \overline{X})^2$ = 1.12 g^2

= sum of squared deviations from the mean
= "sum of squares"

$$n = 7;\ \overline{X} = \frac{\sum X_i}{n} = \frac{12.6\text{ g}}{7} = 1.8\text{ g}$$

$$\text{range} = X_7 - X_1 = 2.4\text{ g} - 1.2\text{ g} = 1.2\text{ g}$$

$$\text{interquartile range} = Q_3 - Q_1 = 2.2\text{ g} - 1.4\text{ g} = 0.8\text{ g}$$

$$\text{mean deviation} = \frac{\sum\lvert X_i - \overline{X}\rvert}{n} = \frac{2.4\text{ g}}{7} = 0.34\text{ g}$$

$$\text{variance} = s^2 = \frac{\sum(X_i - \overline{X})^2}{n - 1} = \frac{1.12\text{ g}^2}{6} = 0.1867\text{ g}^2$$

$$\text{standard deviation} = s = \sqrt{0.1867\text{ g}^2} = 0.43\text{ g}$$

Sample 2

X_i (g)	$X_i - \overline{X}$ (g)	$\lvert X_i - \overline{X}\rvert$ (g)	$(X_i - \overline{X})^2$ (g^2)
1.2	−0.6	0.6	0.36
1.6	−0.2	0.2	0.04
1.7	−0.1	0.1	0.01
1.8	0.0	0.0	0.00
1.9	0.1	0.1	0.01
2.0	0.2	0.2	0.04
2.4	0.6	0.6	0.36
$\sum X_i$ = 12.6 g	$\sum(X_i - \overline{X})$ = 0.0 g	$\sum\lvert X_i - \overline{X}\rvert$ = 1.8 g	$\sum(X_i - \overline{X})^2$ = 0.82 g^2

= sum of squared deviations from the mean
= "sum of squares"

$$n = 7;\ \overline{X} = \frac{\sum X_i}{n} = \frac{12.6\text{ g}}{7} = 1.8\text{ g}$$

$$\text{range} = X_7 - X_1 = 2.4\text{ g} - 1.2\text{ g} = 1.2\text{ g}$$

$$\text{interquartile range} = Q_3 - Q_1 = 2.0\,\text{g} - 1.6\,\text{g} = 0.4\,\text{g}$$

$$\text{mean deviation} = \frac{\sum|X_i - \overline{X}|}{n} = \frac{1.8\,\text{g}}{7} = 0.26\,\text{g}$$

$$\text{variance} = s^2 = \frac{\sum(X_i - \overline{X})^2}{n-1} = \frac{0.82\,\text{g}^2}{6} = 0.1367\,\text{g}^2$$

$$\text{standard deviation} = s = \sqrt{0.1367\,\text{g}^2} = 0.37\,\text{g}$$

The range is a relatively crude measure of dispersion, inasmuch as it does not take into account any measurements except the highest and the lowest. Furthermore, it is unlikely that a sample will contain both the highest and lowest values in the population, so the sample range usually underestimates the population range; therefore, it is a biased and inefficient estimator. Nonetheless, it is considered useful by some to present the sample range as an estimate (although a poor one) of the population range. For example, taxonomists are often concerned with having an estimate of what the highest and lowest values in a population are expected to be. Whenever the range is specified in reporting data, however, it is usually a good practice to report another measure of dispersion as well. The range is applicable to ordinal-, interval-, and ratio-scale data.

4.2 DISPERSION MEASURED WITH QUANTILES

Because the sample range is a biased and inefficient estimate of the population range, being sensitive to extremely large and small measurements, alternative measures of dispersion may be desired. Just as the median (Section 3.2) is the value above and below which lies half the set of data, one can define measures, called *quantiles*, above or below which lie other fractional portions of the data.

For example, if the data are divided into four equal parts, we speak of *quartiles*. One-fourth of all the ranked observations are smaller than the first quartile, one-fourth lie between the first and second quartiles, one-fourth lie between the second and third quartiles, and one-fourth are larger than the third quartile. The second quartile is identical to the median. As with the median, the first and third quartiles might be one of the data or the midpoint between two of the data. The first quartile, Q_1, is

$$Q_1 = X_{(n+1)/4}; \tag{4.2}$$

if the subscript, $(n + 1)/4$, is not an integer or half-integer, then it is rounded up to the nearest integer or half-integer. The second quartile is the median, and the subscript on X for the third quartile, Q_3, is

$$n + 1 - (\text{subscript on } X \text{ for } Q_1, \text{ after any rounding}). \tag{4.3}$$

Examining the data in Example 3.3: For species A, $n = 9$, $(n + 1)/4 = 2.5$, and $Q_1 = X_{2.5} = 34.5$ mo; and $Q_3 = X_{10-2.5} = X_{7.5} = 46$ mo. For species B, $n = 10$, $(n + 1)/4 = 2.75$ (which we round up to 3), and $Q_1 = X_3 = 38$ mo, and $Q_3 = X_{11-3} = X_8 = 59$ mo.

The distance between Q_1 and Q_3, the first and third quartiles (i.e., the 25th and 75th percentiles), is known as the *interquartile range* (or *semiquartile range*):

$$\text{interquartile range} = Q_3 - Q_1. \tag{4.4}$$

One may also encounter the *semi-interquartile range*:

$$\text{semi-interquartile range} = \frac{Q_3 - Q_1}{2}, \tag{4.5}$$

also known as the *quartile deviation*.*

If the distribution of data is symmetrical, then 50% of the measurements lie within one quartile deviation above and below the median. For Sample 1 in Example 4.1, $Q_1 = 1.4$ g, $Q_3 = 2.2$ g, and the interquartile range is 2.2 g − 1.4 g = 0.8 g. And for Sample 2, $Q_1 = 1.6$ g, $Q_3 = 2.0$ g, and the interquartile range is 2.0 g − 1.6 g = 0.4 g.

Similarly, values that partition the ordered data set into eight equal parts (or as equal as n will allow) are called *octiles*. The first octile, $\mathbb{O}_1$, is

$$\mathbb{O}_1 = X_{(n+1)/8}; \tag{4.6}$$

and if the subscript, $(n + 1)/8$, is not an integer or half-integer, then it is rounded up to the nearest integer or half-integer. The second, fourth, and sixth octiles are the same as quartiles; that is, $\mathbb{O}_2 = Q_1$, $\mathbb{O}_4 = Q_2$ = median and $\mathbb{O}_6 = Q_3$. The subscript on X for the third octile, $\mathbb{O}_3$, is

$$2(\text{subscript on } X \text{ for } Q_1) - \text{subscript on } X \text{ for } \mathbb{O}_1; \tag{4.7}$$

the subscript on X for the fifth octile, $\mathbb{O}_5$, is

$$n + 1 - \text{subscript on } X \text{ for } \mathbb{O}_3; \tag{4.8}$$

and the subscript on X for the seventh octile, $\mathbb{O}_7$, is

$$n + 1 - \text{subscript on } X \text{ for } \mathbb{O}_1. \tag{4.9}$$

Thus, for the data of Example 3.3: For species A, $n = 9$, $(n + 1)/8 = 1.5$ and $\mathbb{O}_1 = X_{1.5} = 35$ mo; $2(2.5) - 1.5 = 3.5$, so $\mathbb{O}_3 = X_{3.5} = 38$ mo; $n + 1 - 3.5 = 6.5$, so $\mathbb{O}_5 = X_{6.5} = 41.5$ mo; and $n + 1 - 1.5 = 8.5$, so $\mathbb{O}_7 = 61$. For species B, $n = 10$, $(n + 1)/8 = 1.25$ (which we round up to 1.5) and $\mathbb{O}_1 = X_{1.5} = 35$ mo; $2(3) - 1.5 = 4.5$, so $\mathbb{O}_3 = X_{4.5} = 39.5$ mo; $n + 1 - 4.5 = 6.5$, so $\mathbb{O}_5 = X_{6.5} = 41.5$ mo; and $n + 1 - 1.5 = 9.5$, so $\mathbb{O}_7 = 44.5$ mo.

Besides the median, quartiles, and octiles, ordered data may be divided into fifths, tenths, or hundredths by quantities that are respectively called *quintiles*, *deciles*, and *centiles* (the latter also called *percentiles*). Measures that divide a group of ordered data into equal parts are collectively termed *quantiles*.† The expression "LD_{50}," used in some areas of biological research, is simply the 50th percentile of the lethal doses, or the median lethal dose. That is, 50% of the experimental subjects survived this dose, whereas 50% did not. Likewise, "LC_{50}" is the median lethal concentration, or the 50th percentile of the lethal concentrations.

Instead of distance between the 25th and 75th percentiles, distances between other quantiles (e.g., 10th and 90th percentiles) may be used as a dispersion measure. Quantile-based measures of dispersion are valid for ordinal-, interval-, or ratio-scale data, and they do not exhibit the bias and inefficiency of the range.

*This measure was proposed in 1846 by L. A. J. Quetelet (1796–1874); Sir Francis Galton (1822–1911) later called it the "quartile deviation" (Walker, 1929: 84) and, in 1882, used the terms "quartile" and "interquartile range" (David, 1995).

†Sir Francis Galton developed the concept of percentiles, quartiles, deciles, and other quantiles in writings from 1869 to 1885 (Walker, 1929: 86–87, 177, 179). The term *quantile* was introduced in 1940 by M. G. Kendall (David, 1995).

4.3 THE MEAN DEVIATION

As is evident from the two samples in Example 4.1, the range conveys no information about how clustered about the middle of the distribution the measurements are. As the mean is so useful a measure of central tendency, one might express dispersion in terms of deviations from the mean. The sum of all deviations from the mean, that is, $\sum(X_i - \overline{X})$, will always equal zero, however, so such a summation would be useless as a measure of dispersion (as seen in Example 4.1).

Using the absolute values of the deviations from the mean eliminates the negative signs of the deviations, and summing those absolute values results in a quantity that is an expression of dispersion about the mean. Dividing this quantity by n yields a measure known as the *mean deviation*, or *mean absolute deviation*,* of the sample; this measure has the same units as do the data. In Example 4.1, Sample 1 is more variable (or more dispersed, or less concentrated) than Sample 2. Although the two samples have the same range, the mean deviations, calculated as

$$\text{sample mean deviation} = \frac{\sum|X_i - \overline{X}|}{n}, \tag{4.10}$$

express the differences in dispersion.† A different kind of mean deviation can be defined by using the sum of the absolute deviations from the median instead of from the mean.

Mean deviations are seldom encountered, because their utility is far less than that of the statistics in Sections 4.4 and 4.5.

4.4 THE VARIANCE

Another method of eliminating the negative signs of deviations from the mean is to square the deviations. The sum of the squares of the deviations from the mean is often simply called the *sum of squares*, abbreviated SS, and is defined as follows:‡

$$\text{population SS} = \sum(X_i - \mu)^2 \tag{4.11}$$

$$\text{sample SS} = \sum(X_i - \overline{X})^2. \tag{4.12}$$

It can be seen from the above two equations that as a measure of variability, or dispersion, the sum of squares considers how far the X_i's deviate from the mean. In

*The term *mean deviation* is apparently due to Karl Pearson (1857–1936) (Walker, 1929: 55) and *mean absolute deviation*, in 1972, to D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey (David, 1995).

†Karl Weierstrass, in 1841, was the first to denote the absolute value of a quantity by enclosing it within two vertical lines (Cajori, 1928/9, Vol. II: p. 123); that is, $|a| = a$ and $|-a| = a$.

‡The modern notation using raised numerals as exponents was introduced by René Descartes in 1637, and many other kinds of notation for exponents were employed before and after that (Cajori, 1928/9, Vol. I: 358; Gullberg, 1997: 134). An 1845 notation of Augustus De Morgan, $a \wedge b$ to indicate a^b (Cajori, ibid.: 358), has reemerged in modern computer use. Nicolas Chuquet (1445–1488) was the first to use negative exponents, and Nicole (also known as Nicolaus) Oresme (1323–1382) was the first to use fractional exponents, though neither of these French mathematicians employed the modern notation of Isaac Newton (1642–1727), the colossal English mathematician, physicist, and astronomer (Cajori, ibid.: 91, 102, 354–355):

$$X^{-a} = \frac{1}{X^a}; \quad X^{\frac{1}{a}} = \sqrt[a]{X}.$$

Using parentheses or brackets to group quantities dates from the mid-sixteenth century, though it was not common mathematical notation until more than two centuries later (ibid.: 392).

Sample 1 of Example 4.1, the sample mean is 1.8 g and it is seen (in the last column) that

$$\begin{aligned}\text{Sample SS} &= (1.2 - 1.8)^2 + (1.4 - 1.8)^2 + (1.6 - 1.8)^2 + (1.8 - 1.8)^2 \\ &\quad + (2.0 - 1.8)^2 + (2.2 - 1.8)^2 + (2.4 - 1.8)^2 \\ &= 0.36 + 0.16 + 0.04 + 0.00 + 0.04 + 0.16 + 0.36 \\ &= 1.12\end{aligned}$$

(where the units are grams^2).* The sum of squares may also be visualized as a measure of the average extent to which the data deviate from each other, for (using the same seven data from Sample 1 in Example 4.1):

$$\begin{aligned}\text{SS} &= [(1.2 - 1.4)^2 + (1.2 - 1.6)^2 + (1.2 - 1.8)^2 + (1.2 - 2.0)^2 \\ &\quad + (1.2 - 2.2)^2 + (1.2 - 2.4)^2 + (1.4 - 1.6)^2 + (1.4 - 1.8)^2 \\ &\quad + (1.4 - 2.0)^2 + (1.4 - 2.2)^2 + (1.4 - 2.4)^2 + (1.6 - 1.8)^2 \\ &\quad + (1.6 - 2.0)^2 + (1.6 - 2.2)^2 + (1.6 - 2.4)^2 + (1.8 - 2.0)^2 \\ &\quad + (1.8 - 2.2)^2 + (1.8 - 2.4)^2 + (2.0 - 2.2)^2 + (2.0 - 2.4)^2 \\ &\quad + (2.2 - 2.4)^2]/7 \\ &= [0.04 + 0.16 + 0.36 + 0.64 + 1.00 + 1.44 + 0.04 + \cdots + 0.04 + 0.16 \\ &\quad + 0.04]/7 \\ &= 7.84/7 = 1.12\end{aligned}$$

(again in grams^2).

The mean sum of squares is called the *variance* (or *mean square*,† the latter being short for *mean squared deviation*), and for a population is denoted by σ^2 ("sigma squared," using the lowercase Greek letter):

$$\sigma^2 = \frac{\sum(X_i - \mu)^2}{N}. \tag{4.14}$$

The best estimate of the population variance, σ^2, is the sample variance, s^2:

$$s^2 = \frac{\sum(X_i - \overline{X})^2}{n - 1}. \tag{4.15}$$

If, in Equation 4.14, we replace μ by $\overline{X}$ and N by n, the result is a quantity that is a biased estimate of σ^2 in that it underestimates σ^2. Dividing the sample sum of squares

*Owing to an important concept in statistics, known as *least squares*, the sum of squared deviations from the mean is smaller than the sum of squared deviations from any other quantity (e.g., the median). Indeed, if Equation 4.12 is applied using some quantity in place of the mean, the resultant "sum of squares" would be

$$SS + nd^2, \tag{4.13}$$

where d is the difference between the mean and the quantity used. For the population sum of squares (defined in Equation 4.11), the relationship would be $SS + Nd^2$.

†The term *mean square* dates back at least to an 1875 publication of Sir George Biddel Airy (1801–1892), Astronomer Royal of England (Walker, 1929: 54). The term *variance* was introduced in 1918 by English statistician Sir Ronald Aylmer Fisher (1890–1962) (ibid.: 189; David, 1995).

by $n - 1$ (called the *degrees of freedom*,* often abbreviated DF), rather than by n, yields an unbiased estimate, and it is Equation 4.15 that should be used to calculate the sample variance.

If all observations in a sample are equal, then there is no variability (that is, no dispersion) and $s^2 = 0$. And s^2 becomes increasingly large as the amount of variability, or dispersion, increases. Because s^2 is a mean sum of squares, it can never be a negative quantity.

The variance expresses the same type of information as does the mean deviation, but it has certain very important mathematical properties relative to probability and hypothesis testing that make it superior. Thus, the mean deviation is very seldom encountered in biostatistical analysis.

The calculation of s^2 can be tedious for large samples, but it can be facilitated by the use of the equality

$$\text{sample SS} = \sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}. \tag{4.16}$$

This formula is equivalent to Equation 4.12 but is much simpler to work with. Example 4.2 demonstrates its use to obtain a sample sum of squares.

Because the sample variance equals the sample SS divided by DF,

$$s^2 = \frac{\sum X_i^2 - \dfrac{\left(\sum X_i\right)^2}{n}}{n - 1}. \tag{4.17}$$

This last formula is often referred to as a "working formula," or "machine formula," because of its computational advantages. There are, in fact, two major advantages in calculating SS by Equation 4.16 rather than by Equation 4.12. First, fewer computational steps are involved, a fact that decreases chance of error. On many calculators the summed quantities, $\sum X_i$ and $\sum X_i^2$, can both be obtained with only one pass through the data, whereas Equation 4.12 requires one pass through the data to calculate $\overline{X}$ and at least one more pass to calculate and sum the squares of the deviations, $X_i - \overline{X}$. Second, there may be a good deal of rounding error in calculating each $X_i - \overline{X}$, a situation that leads to decreased accuracy in computation, but that is avoided by the use of Equation 4.16.†

For data recorded in frequency tables,

$$\text{sample SS} = \sum f_i X_i^2 - \frac{\left(\sum f_i X_i\right)^2}{n}, \tag{4.18}$$

*Given the sample mean ($\overline{X}$) and sample size (n) in Example 4.1, *degrees of freedom* means that the data could have been weights different from those shown, but when any six (i.e., $n - 1$) of the seven weights are specified, then the seventh weight is also known. The term was first used, though in a different context, by Ronald Aylmer Fisher in 1922 (David, 1955).

†Computational formulas advantageous on calculators may not prove accurate on computers (Wilkinson and Dallal, 1977), largely because computers may use fewer significant figures. (Also see Ling, 1974.) Good computer programs use calculation techniques designed to help avoid rounding errors.

where f_i is the frequency of observations with magnitude X_i. But with a calculator or computer it is often faster to use Equation 4.18 for the individual observations, disregarding the class groupings.

The variance has square units. If measurements are in grams, their variance will be in grams squared, or if the measurements are in cubic centimeters, their variance will be in terms of cubic centimeters squared, even though such squared units have no physical interpretation. The question of how many decimal places to report for the variance will be considered at the end of Section 6.2.

EXAMPLE 4.2 "Machine Formula" Calculation of Variance, Standard Deviation, and Coefficient of Variation (These are the data of Example 4.1)

Sample 1		Sample 2	
X_i (g)	X_i^2 (g^2)	X_i (g)	X_i^2 (g^2)
1.2	1.44	1.2	1.44
1.4	1.96	1.6	2.56
1.6	2.56	1.7	2.89
1.8	3.24	1.8	3.24
2.0	4.00	1.9	3.61
2.2	4.84	2.0	4.00
2.4	5.76	2.4	5.76

$\sum X_i = 12.6\text{ g} \quad \sum X_i^2 = 23.80\text{ g}^2$

$n = 7$

$$\overline{X} = \frac{12.6\text{ g}}{7} = 1.8\text{ g}$$

$$SS = \sum X_i^2 - \frac{(\sum X_i)^2}{n}$$

$$= 23.80\text{ g}^2 - \frac{(12.6\text{ g})^2}{7}$$

$$= 23.80\text{ g}^2 - 22.68\text{ g}^2$$

$$= 1.12\text{ g}^2$$

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

$$= \frac{1.12\text{ g}^2}{6} = 0.1867\text{ g}^2$$

$$s = \sqrt{0.1867\text{ g}^2} = 0.43\text{ g}$$

$$V = \frac{s}{\overline{X}} = \frac{0.43\text{ g}}{1.8\text{ g}} = 0.24 = 24\%$$

$\sum X_i = 12.6\text{ g} \quad \sum X_i^2 = 23.50\text{ g}^2$

$n = 7$

$$\overline{X} = \frac{12.6\text{ g}}{7} = 1.8\text{ g}$$

$$SS = 23.50\text{ g}^2 - \frac{(12.6\text{ g})^2}{7}$$

$$= 0.82\text{ g}^2$$

$$s^2 = \frac{0.82\text{ g}^2}{6} = 0.1367\text{ g}^2$$

$$s = \sqrt{0.1367\text{ g}^2} = 0.37\text{ g}$$

$$V = \frac{0.37\text{ g}}{1.8\text{ g}} = 0.21 = 21\%$$

4.5 THE STANDARD DEVIATION

The *standard deviation** is the positive square root† of the variance; therefore, it has the same units as the original measurements. Thus, for a population,

$$\sigma = \sqrt{\frac{\sum X_i^2 - \frac{(\sum X_i)^2}{N}}{N}}. \tag{4.19}$$

And for a sample,‡

$$s = \sqrt{\frac{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}{n-1}}. \tag{4.20}$$

Examples 4.1 and 4.2 demonstrate the calculation of s. This quantity frequently is abbreviated SD, and on rare occasions is called the *root mean square deviation* or *root mean square*. Remember that the standard deviation is, by definition, always a nonnegative quantity.§ The end of Section 6.2 will explain how to determine

*It was the great English statistician Karl Pearson (1857–1936) who coined the term *standard deviation* and its symbol, σ, in 1893, prior to which this quantity was called the *mean error* (Eells, 1926; Walker, 1929: 54–55, 183, 188). In early literature (e.g., by G. U. Yule in 1919), it was termed *root mean square deviation* and acquired the symbol s, and (particularly in the fields of education and psychology) it was occasionally computed using deviations from the median (or even the mode) instead of from the mean (Eells, 1926).

†The square root sign ($\surd$) was introduced by Silesian-born Austrian mathematician Christoff Rudolff (1499–1545) in 1525; by 1637 René Descartes (1596–1650) combined this with a vinculum (a horizontal bar placed above quantities to group them as is done with parentheses or brackets) to obtain the symbol $\sqrt{}$, but Gottfried Wilhelm Leibniz (1646–1716) preferred $\surd(\quad)$, which is still occasionally seen (Cajori, 1928/9, Vol. I: 135, 208, 368, 372, 375). The first footnote in Section 3.4 speaks to the origin of the cube root symbol ($\sqrt[3]{}$).

‡The sample s is actually a slightly biased estimate of the population σ, in that on the average it is a slightly low estimate, especially in small samples. But this fact is generally considered to be offset by the statistic's usefulness. Correction for this bias is sometimes possible (e.g., Bliss, 1967: 131; Dixon and Massey, 1969: 136; Gurland and Tripathi, 1971; Tolman, 1971), but it is rarely employed.

§It can be shown that the median of a distribution is never more than one standard deviation away from the mean (μ); that is,

$$|\text{median} - \mu| \leq \sigma \tag{4.21}$$

(Hotelling and Solomon, 1932; O'Cinneide, 1990; Page and Murty, 1982; Watson, 1994). This is a special case, where $p = 50$, of the relationship

$$\mu - \sigma\sqrt{\frac{1 - p/100}{p/100}} \leq X_p \leq \mu + \sigma\sqrt{\frac{p/100}{1 - p/100}}, \tag{4.22}$$

where X_p is the pth percentile of the distribution (Dharmadhikari, 1991). Also, Page and Murty (1982) have shown these population-parameter relationships between the standard deviation and the range and between the standard deviation and the mean, median, and mode:

$$\text{range}/\sqrt{2n} \leq \sigma \leq \text{range}/2; \tag{4.22a}$$

$$|\text{mode} - \mu| \leq \sigma\sqrt{n/m} \text{ and } |\text{mode} - \text{median}| \leq \sigma(n/m), \tag{4.22b}$$

where m is the number of data at the modal value.

the number of decimal places that may appropriately be recorded for the standard deviation.

4.6 THE COEFFICIENT OF VARIATION

The *coefficient of variation** or *coefficient of variability*, is defined as

$$V = \frac{s}{\overline{X}} \quad \text{or} \quad V = \frac{s}{\overline{X}} \cdot 100\%. \tag{4.23}$$

As $s/\overline{X}$ is generally a small quantity, it is frequently multiplied by 100% in order to express V as a percentage. (The coefficient of variation is often abbreviated as CV.)

As a measure of variability, the variance and standard deviation have magnitudes that are dependent on the magnitude of the data. Elephants have ears that are perhaps 100 times larger than those of mice. If elephant ears were no more variable, relative to their size, than mouse ears, relative to their size, the standard deviation of elephant ear lengths would be 100 times as great as the standard deviation of mouse ear lengths (and the variance of the former would be $100^2 = 10{,}000$ times the variance of the latter). The sample coefficient of variation expresses sample variability relative to the mean of the sample (and is on rare occasion referred to as the "relative standard deviation"). It is called a measure of *relative variability* or *relative dispersion*.

Because s and $\overline{X}$ have identical units, V has no units at all, a fact emphasizing that it is a relative measure, divorced from the actual magnitude or units of measurement of the data. Thus, had the data in Example 4.2 been measured in pounds, kilograms, or tons, instead of grams, the calculated V would have been the same. The coefficient of variation of a sample, namely V, is an estimate of the coefficient of variation of the population from which the sample came (i.e., an estimate of σ/μ). The coefficient of variation may be calculated only for ratio scale data; it is, for example, not valid to calculate coefficients of variation of temperature data measured on the Celsius or Fahrenheit temperature scales. Simpson, Roe, and Lewontin (1960: 89–95) present a good discussion of V and its biological application, especially with regard to zoomorphological measurements.

4.7 INDICES OF DIVERSITY

For nominal-scale data there is no mean or median or ordered measurements to serve as a reference for discussion of dispersion. Instead, we can invoke the concept of *diversity*, the distribution of observations among categories. Consider that sparrows are found to nest in four different types of location (vines, eaves, branches, and cavities). If, out of twenty nests observed, five are found at each of the four locations, then we would say that there was great diversity in nesting sites. If, however, seventeen nests were found in cavities and only one in each of the other three locations, then we would consider the situation to be one of very low nest-site diversity. In other words, observations distributed evenly among categories display high diversity, whereas a set of observations where most of the data occur in very few of the categories is one exhibiting low diversity.

A large number of diversity measures have been introduced, especially for ecological data (e.g., Brower, Zar, and von Ende, 1998: 177–184; Magurran, 2004), a few of which are presented here.

*The term *coefficient of variation* was introduced by the statistical giant Karl Pearson (1857–1936) in 1896 (David, 1995). In early literature the term was variously applied to the ratios of different measures of dispersion and different measures of central tendency (Eells, 1926).

Among the quantitative descriptions of diversity available are those based on a field known as *information theory*.* The underlying considerations of these measures can be visualized by considering *uncertainty* to be synonymous with diversity. If seventeen out of twenty nest sites were to be found in cavities, then one would be relatively certain of being able to predict the location of a randomly encountered nest site. However, if nests were found to be distributed evenly among the various locations (a situation of high nest-site diversity), then there would be a good deal of uncertainty involved in predicting the location of a nest site selected at random. If a set of nominal scale data may be considered to be a random sample, then a quantitative expression appropriate as a measure of diversity is that of Shannon (1948):

$$H' = -\sum_{i=1}^{k} p_i \log p_i \tag{4.24}$$

(often referred to as the Shannon-Wiener diversity index or the Shannon-Weaver index). Here, k is the number of categories and p_i is the proportion of the observations found in category i. Denoting n to be sample size and f_i to be the number of observations in category i, then $p_i = f_i/n$; and an equivalent equation for H' is

$$H' = \frac{n \log n - \sum_{i=1}^{k} f_i \log f_i}{n}, \tag{4.25}$$

a formula that is easier to use than Equation 4.24 because it eliminates the necessity of calculating the proportions (p_i). Published tables of $n \log n$ and $f_i \log f_i$ are available (e.g., Brower, Zar, and von Ende, 1998: 181; Lloyd, Zar, and Karr, 1968). Any logarithmic base may be used to compute H'; bases 10, e, and 2 (in that order of commonness) are the most frequently encountered. A value of H' (or of any other measure of this section except evenness measures) calculated using one logarithmic base may be converted to that of another base; Table 4.1 gives factors for doing this for bases 10, e, and 2. Unfortunately, H' is known to be an underestimate of the diversity in the sampled population (Bowman et al., 1971). However, this bias decreases with increasing sample size. Ghent (1991) demonstrated a relationship between H' and testing hypotheses for equal abundance among the k categories.

The magnitude of H' is affected not only by the distribution of the data but also by the number of categories, for, theoretically, the maximum possible diversity for a set of data consisting of k categories is

$$H'_{\max} = \log k. \tag{4.26}$$

Therefore, some users of Shannon's index prefer to calculate

$$J' = \frac{H'}{H'_{\max}} \tag{4.27}$$

instead of (or in addition to) H', thus expressing the observed diversity as a proportion of the maximum possible diversity. The quantity J' has been termed *evenness* (Pielou, 1966) and may also be referred to as *homogeneity* or *relative diversity*. The measure

*Claude Elwood Shannon (1916–2001) founded what he first called "a mathematical theory of communication" and has become known as "information theory."

TABLE 4.1: Multiplication Factors for Converting among Diversity Measures (H, H', H_{max}, or H'_{max}) Calculated Using Different Logarithmic Bases*

To convert to:	To convert from:		
	Base 2	**Base *e***	**Base 10**
Base 2	1.0000	1.4427	3.3219
Base *e*	0.6931	1.0000	2.3026
Base 10	0.3010	0.4343	1.0000

For example, if $H' = 0.255$ using base 10; H' would be $(0.255)(3.3219) = 0.847$ using base 2.

*The measures J and J' are unaffected by change in logarithmic base.

$1 - J'$ may then be viewed as a measure of *heterogeneity*; it may also be considered a measure of *dominance*, for it reflects the extent to which frequencies are concentrated in a small number of categories. The number of categories in a sample (k) is typically an underestimate of the number of categories in the population from which the sample came, because some categories (especially the rarer ones) are likely to be missed in collecting the sample. Therefore, the sample evenness, J', is typically an overestimate of the population evenness. (That is, J' is a biased statistic.) Example 4.3 demonstrates the calculation of H' and J'.

If a set of data may not be considered a random sample, then Equation 4.24 (or 4.25) is not an appropriate diversity measure (Pielou, 1966). Examples of such

EXAMPLE 4.3 Indices of Diversity for Nominal Scale Data: The Nesting Sites of Sparrows

Category (i)	Observed Frequencies (f_i)
	Sample 1
Vines	5
Eaves	5
Branches	5
Cavities	5

$$H' = \frac{n\log n - \sum f_i \log f_i}{n} = [20\log 20 - (5\log 5 + 5\log 5 + 5\log 5 + 5\log 5)]/20$$

$$= [26.0206 - (3.4949 + 3.4949 + 3.4949 + 3.4949)]/20$$

$$= 12.0410/20 = 0.602$$

$$H'_{max} = \log 4 = 0.602$$

$$J' = \frac{0.602}{0.602} = 1.00$$

	Sample 2
Vines	1
Eaves	1
Branches	1
Cavities	17

$$H' = \frac{n\log n - \sum f_i \log f_i}{n} = [20\log 20 - (1\log 1 + 1\log 1 + 1\log 1 + 17\log 17)]/20$$
$$= [26.0206 - (0 + 0 + 0 + 20.9176)]/20$$
$$= 5.1030/20 = 0.255$$
$$H'_{\max} = \log 4 = 0.602$$
$$J' = \frac{0.255}{0.602} = 0.42$$

	Sample 3
Vines	2
Eaves	2
Branches	2
Cavities	34

$$H' = \frac{n\log n - \sum f_i \log f_i}{n} = [40\log 40 - (2\log 2 + 2\log 2 + 2\log 2 + 34\log 34)]/40$$
$$= [64.0824 - (0.6021 + 0.6021 + 0.6021 + 52.0703)]/40$$
$$= 10.2058/40 = 0.255$$
$$H'_{\max} = \log 4 = 0.602$$
$$J' = \frac{0.255}{0.602} = 0.42$$

situations may be when we have, in fact, data composing an entire population, or data that are a sample obtained nonrandomly from a population. In such a case, one may use the information-theoretic diversity measure of Brillouin (1962: 7–8):*

$$H = \frac{\log\left(\dfrac{n!}{\prod_{i=1}^{k} f_i!}\right)}{n}, \tag{4.28}$$

*The notation $n!$ is read as "n factorial" and signifies the product $(n)(n-1)(n-2)\cdots(2)(1)$. It was proposed by French physician and mathematician Christian Kramp (1760–1826) around 1798; he originally called this function *faculty* ("facultés" in French) but in 1808 accepted the term *factorial* ("factorielle" in French) used by Alsatian mathematician Louis François Antoine Arbogast (1759–1803) (Cajori, 1928/9, Vol. II: 72; Gullberg, 1997: 106; Miller, 2004a; O'Connor and Robertson, 1997). English mathematician Augustus De Morgan (1806–1871) decried the adoption of this symbol as a "barbarism" because it introduced into mathematics a symbol that already had an established meaning in written language, thus giving "the appearance of expressing surprise or admiration" in a mathematical result (Cajori, ibid.: 328).

where Π (capital Greek pi) means to take the product, just as Σ means to take the sum. Equation 4.28 may be written, equivalently, as

$$H = \frac{\log \dfrac{n!}{f_1!f_2!\ldots f_k!}}{n} \tag{4.29}$$

or as

$$H = \frac{(\log n! - \sum \log f_i!)}{n}. \tag{4.30}$$

Table B.40 gives logarithms of factorials to ease this calculation. Other such tables are available, as well (e.g., Brower, Zar, and von Ende 1998: 183; Lloyd, Zar, and Karr, 1968; Pearson and Hartly, 1966: Table 51).* Ghent (1991) discussed the relationship between H and the test of hypotheses about equal abundance among k categories.

The maximum possible Brillouin diversity for a set of n observations distributed among k categories is

$$H_{\max} = \frac{\log n! - (k - d)\log c! - d\log(c + 1)!}{n}, \tag{4.35}$$

where c is the integer portion of n/k, and d is the remainder. (For example, if $n = 17$ and $k = 4$, then $n/k = 17/4 = 4.25$ and $c = 4$ and $d = 0.25$.) The Brillouin-based evenness measure is, therefore,

$$J = \frac{H}{H_{\max}}, \tag{4.36}$$

with $1 - J$ being a dominance measure. When we consider that we have data from an entire population, k is a population measurement, rather than an estimate of one, and J is not a biased estimate as is J'.

For further considerations of these and other diversity measures, see Brower, Zar, and von Ende (1998: Chapter 5B) and Magguran (2004: 100–121).

4.8 CODING DATA

Section 3.5 showed how coding data may facilitate statistical computations of measures of central tendency. Such benefits are even more apparent when calculating SS, s^2,

*For moderate to large n (or f_i), "Stirling's approximation" is excellent (see note after Table B.40):

$$n! = \sqrt{2\pi n}(n/e)^n = \sqrt{2\pi}\sqrt{n}e^{-n}n^n, \tag{4.31}$$

of which this is an easily usable derivation:

$$\log n! = (n + 0.5)\log n - 0.434294n + 0.399090. \tag{4.32}$$

An approximation with only half the error of the above is

$$n! = \sqrt{2\pi}\left(\frac{n + 0.5}{e}\right)^{n+0.5} \tag{4.33}$$

and

$$\log n! = (n + 0.5)\log(n + 0.5) - 0.434294(n + 0.5) + 0.399090. \tag{4.34}$$

This is named for James Stirling, who published something similar to the latter approximation formula in 1730, making an arithmetic improvement in the approximation earlier known by Abraham de Moivre (Kemp, 1989; Pearson, 1924; Walker, 1929: 16).

and s, because of the labor, and concomitant chances of error, involved in the unwieldy squaring of large or small numbers.

When data are coded by adding or subtracting a constant (call it A), the measures of dispersion of Sections 4.1 through 4.5 are not changed from what they were for the data before coding. This is because these measures are based upon deviations, and deviations are not changed by moving the data along the measurement scale (e.g., the deviation between 1 and 10 is the same as the deviation between 11 and 20). Sample 1 in Example 4.4 demonstrates this.

However, when coding by multiplying by a constant (call it M), the measures of dispersion are affected, for the magnitudes of the deviations will be changed. With such coding, the range, mean deviation, and standard deviation are changed by a factor of M, in the same manner as the arithmetic mean and the median are, whereas the sum of squares and variance are changed in accordance with the square of the coding constant (i.e., M^2), and the coefficient of variance is not affected. This is demonstrated in Sample 2 of Example 4.4.

Appendix C presents the results of coding these and many other statistics, where a coded datum is described as

$$[X_i] = MX_i + A. \tag{4.37}$$

EXAMPLE 4.4 Coding Data to Facilitate the Calculation of Measures of Dispersion

Sample 1 (Coding by Subtraction: $A = -840$ g)

Without Coding X_i		*Using Coding* $[X_i]$	
X_i (g)	X_i^2 (g^2)	$[X_i]$ (g)	$[X_i]^2$ (g^2)
842	708,964	2	4
843	710,649	3	9
844	712,336	4	16
846	715,716	6	36
846	715,716	6	36
847	717,409	7	49
848	719,104	8	64
849	720,801	9	81
$\sum X_i = 6765$ g	$\sum X_i^2 = 5{,}720{,}695$ g^2	$\sum[X_i] = 45$ g	$\sum[X_i]^2 = 295$ g^2

$$s^2 = \frac{5720695\ \text{g}^2 - \frac{(6765\ \text{g})^2}{8}}{7}$$
$$= 5.98\ \text{g}^2$$
$$s = 2.45\ \text{g}$$
$$\overline{X} = 845.6\ \text{g}$$
$$V = \frac{s}{\overline{X}} = \frac{2.45\ \text{g}}{845.6\ \text{g}}$$
$$= 0.0029 = 0.29\%$$

$$[s^2] = \frac{295\ \text{g}^2 - \frac{(45\ \text{g})^2}{8}}{7}$$
$$= 5.98\ \text{g}^2$$
$$[s] = 2.44\ \text{g}$$
$$[\overline{X}] = 5.6\ \text{g}$$

Sample 2 (Coding by Division: $M = 0.01$)

Without Coding X_i		*Using Coding* $[X_i]$	
X_i (sec)	X_i^2 (sec^2)	$[X_i]$ (sec)	$[X_i]^2$ (sec^2)
800	640,000	8.00	64.00
900	810,000	9.00	81.00
950	902,500	9.50	90.25
1100	1,210,000	11.00	121.00
1250	1,562,500	12.50	156.25
1300	1,690,000	13.00	169.00
$\sum X_i = 6300$ sec	$\sum X_i^2 = 6{,}815{,}000$ sec^2	$\sum[X_i] = 63.00$ sec	$\sum[X_i]^2 = 681.50$ sec^2

$$s^2 = \frac{6815000 \text{ sec}^2 - \frac{(6300 \text{ sec})^2}{6}}{5}$$

$$= 40{,}000 \text{ sec}^2$$

$$s = 200 \text{ sec}$$

$$\overline{X} = 1050 \text{ sec}$$

$$V = 0.19 = 19\%$$

$$[s^2] = \frac{681.50 \text{ sec}^2 - \frac{(63.00 \text{ sec})^2}{6}}{5}$$

$$= 4 \text{ sec}^2$$

$$[s] = 2.00 \text{ sec}$$

$$[\overline{X}] = 10.50 \text{ sec}$$

$$[V] = 0.19 = 19\%$$

EXERCISES

4.1. Five body weights, in grams, collected from a population of rodent body weights are

66.1, 77.1, 74.6, 61.8, 71.5.

(a) Compute the "sum of squares" and the variance of these data using Equations 4.12 and 4.15, respectively.

(b) Compute the "sum of squares" and the variance of these data by using Equations 4.16 and 4.17, respectively.

4.2. Consider the following data, which are a sample of amino acid concentrations (mg/100 ml) in arthropod hemolymph:

240.6, 238.2, 236.4, 244.8, 240.7, 241.3, 237.9.

(a) Determine the range of the data.

(b) Calculate the "sum of squares" of the data.

(c) Calculate the variance of the data.

(d) Calculate the standard deviation of the data.

(e) Calculate the coefficient of variation of the data.

4.3. The following frequency distribution of tree species was observed in a random sample from a forest:

Species	*Frequency*
White oak	44
Red oak	3
Shagbark hickory	28
Black walnut	12
Basswood	2
Slippery elm	8

(a) Use the Shannon index to express the tree species diversity.

(b) Compute the maximum Shannon diversity possible for the given number of species and individuals.

(c) Calculate the Shannon evenness for these data.

4.4. Assume the data in Exercise 4.3 were an entire population (e.g., all the trees planted around a group of buildings).

(a) Use the Brillouin index to express the tree species diversity.

(b) Compute the maximum Brillouin diversity possible for the given number of species and individuals.

(c) Calculate the Brillouin evenness measure for these data.

CHAPTER 5

Probabilities

Everyday concepts of "likelihood," "predictability," and "chance" are formalized by that branch of mathematics called *probability*. Although earlier work on the subject was done by writers such as Giralamo Cardano (1501–1576) and Galileo Galilei (1564–1642), the investigation of probability as a branch of mathematics sprang in earnest from 1654 correspondence between two great French mathematicians, Blaise Pascal (1623–1662) and Pierre Fermat (1601–1665). These two men were stimulated by the desire to predict outcomes in the games of chance popular among the French nobility of the mid-seventeenth century; we still use the devices of such games (e.g., dice and cards) to demonstrate the basic concepts of probability.*

A thorough discourse on probability is well beyond the scope and intent of this book, but aspects of probability are of biological interest and considerations of probability theory underlie the many procedures for statistical hypothesis testing discussed in the following chapters. Therefore, this chapter will introduce probability concepts that bear the most pertinence to biology and biostatistical analysis. Although mastery of this chapter is not essential to apply the statistical procedures in the remainder of the book, occasionally later reference will be made to it.

Worthwhile presentations of probability specifically for the biologist are found in Batschelet (1976: 441–474); Eason, Coles, and Gettinby (1980: 395–414); and Mosimann (1968).

5.1 COUNTING POSSIBLE OUTCOMES

Suppose a phenomenon can occur in any one of k different ways, but in only one of those ways at a time. For example, a coin has two sides and when tossed will land

*The first published work on the subject of probability and gaming was by the Dutch astronomer, physicist, and mathematician Christiaan (also known as Christianus) Huygens (1629–1695), in 1657 (Asimov, 1982: 138; David, 1962: 113, 133). This, in turn, aroused the interest of other major minds, such as Jacob (also known as Jacques, Jakob, and James) Bernoulli (1654–1705, whose 1713 book was the first devoted entirely to probability), several other members of the remarkable Bernoulli family of Swiss mathematicians, and others such as Abraham de Moivre (1667–1754), Pierre Rémond de Montmort (1678–1719), and Pierre-Simon Laplace (1749–1827) of France. The term *probability* in its modern mathematical sense was used as early as 1718 by de Moivre (Miller, 2004a). For more detailed history of the subject, see David (1962) and Walker (1928: 5–13).

with either the "head" side (H) up or the "tail" side (T) up, but not both. Or, a die has six sides and when thrown will land with either the 1, 2, 3, 4, 5, or 6 side up.* We shall refer to each possible outcome (i.e., H or T with the coin; or 1, 2, 3, 4, 5, or 6 with the die) as an *event*.

If something can occur in any one of k_1 different ways and something else can occur in any one of k_2 different ways, then the number of possible ways for both things to occur is $k_1 \times k_2$. For example, suppose that two coins are tossed, say a silver one and a copper one. There are two possible outcomes of the toss of the silver coin (H or T) and two possible outcomes of the toss of the copper coin (H or T). Therefore, $k_1 = 2$ and $k_2 = 2$ and there are $(k_1)(k_2) = (2)(2) = 4$ possible outcomes of the toss of both coins: both heads, silver head and copper tail, silver tail and copper head, and both tails (i.e., H,H; H,T; T,H; T,T).

Or, consider tossing of a coin together with throwing a die. There are two possible coin outcomes ($k_1 = 2$) and six possible die outcomes ($k_2 = 6$), so there are $(k_1)(k_2) = (2)(6) = 12$ possible outcomes of the two events together:

H,1; H,2; H,3; H,4; H,5; H,6; T,1; T,2; T,3; T,4; T,5; T,6.

If two dice are thrown, we can count six possible outcomes for the first die and six for the second, so there are $(k_1)(k_2) = (6)(6) = 36$ possible outcomes when two dice are thrown:

1,1; 1,2; 1,3; 1,4; 1,5; 1,6;	2,1; 2,2; 2,3; 2,4; 2,5; 2,6;
3,1; 3,2; 3,3; 3,4; 3,5; 3,6;	4,1; 4,2; 4,3; 4,4; 4,5; 4,6;
5,1; 5,2; 5,3; 5,4; 5,5; 5,6;	6,1; 6,2; 6,3; 6,4; 6,5; 6,6.

The preceding counting rule is extended readily to determine the number of ways more than two things can occur together. If one thing can occur in any one of k_1 ways, a second thing in any one of k_2 ways, a third thing in any of k_3 ways, and so on, through an nth thing in any one of k_n ways, then the number of ways for all n things to occur together is

$$(k_1)(k_2)(k_3)\cdots(k_n).$$

Thus, if three coins are tossed, each toss resulting in one of two possible outcomes, then there is a total of

$$(k_1)(k_2)(k_3) = (2)(2)(2) = 8$$

possible outcomes for the three tosses together:

H,H,H; H,H,T; H,T,H; H,T,T; T,H,H, T,H,T; T,T,H; T,T,T.

Similarly, if three dice are thrown, there are $(k_1)(k_2)(k_3) = (6)(6)(6) = 6^3 = 216$ possible outcomes; if two dice and three coins are thrown, there are

*What we recognize as metallic coins originated shortly after 650 B.C.E.—perhaps in ancient Lydia (located on the Aegean Sea in what is now western Turkey). From the beginning, the obverse and reverse sides of coins have had different designs, in earliest times with the obverse commonly depicting animals and, later, deities and rulers (Sutherland, 1992). Dice have long been used for both games and religion. They date from nearly 3000 years B.C.E., with the modern conventional arrangement of dots on the six faces of a cubic die (1 opposite 6, 2 opposite 5, and 3 opposite 4) becoming dominant around the middle of the fourteenth century B.C.E. (David, 1962: 10). Of course, the arrangement of the numbers 1 through 6 on the six faces has no effect on the outcome of throwing a die.

$(k_1)(k_2)(k_3)(k_4)(k_5) = (6)(6)(2)(2)(2) = (6^2)(2^3) = 288$ outcomes; and so on. Example 5.1 gives two biological examples of counting possible outcomes.

EXAMPLE 5.1 Counting Possible Outcomes

(a) A linear arrangement of three deoxyribonucleic acid (DNA) nucleotides is called a triplet. A nucleotide may contain any one of four possible bases: adenine (A), cytosine (C), guanine (G), and thymine (T). How many different triplets are possible?

As the first nucleotide in the triplet may be any one of the four bases (A; C; G; T), the second may be any one of the four, and the third may be any one of the four, there is a total of

$$(k_1)(k_2)(k_3) = (4)(4)(4) = 64 \text{ possible outcomes;}$$

that is, there are 64 possible triplets:

A, A, A; A, A, C; A, A, G; A, A, T;
A, C, A; A, C, C; A, C, G; A, C, T;
A, G, A; A, G, C; A, G, G; A, G, T;
and so on.

(b) If a diploid cell contains three pairs of chromosomes, and one member of each pair is found in each gamete, how many different gametes are possible?

As the first chromosome may occur in a gamete in one of two forms, as may the second and the third chromosomes,

$$(k_1)(k_2)(k_3) = (2)(2)(2) = 2^3 = 8.$$

Let us designate one of the pairs of chromosomes as "long," with the members of the pair being L_1 and L_2; one pair as "short," indicated as S_1 and S_2; and one pair as "midsized," labeled M_1 and M_2. Then the eight possible outcomes may be represented as

L_1, M_1, S_1; L_1, M_1, S_2; L_1, M_2, S_1; L_1, M_2, S_2;
L_2, M_1, S_1; L_2, M_1, S_2; L_2, M_2, S_1; L_2, M_2, S_2.

5.2 PERMUTATIONS

(a) Linear Arrangements. A *permutation** is an arrangement of objects in a specific sequence. For example, a horse (H), cow (C), and sheep (S) could be arranged linearly in six different ways: H,C,S; H,S,C; C,H,S; C,S,H; S,H,C; S,C,H. This set of outcomes may be examined by noting that there are three possible ways to fill the first position in the linear order; but once an animal is placed in this position, there are only two ways to fill the second position; and after animals are placed in the first two positions, there is only one possible way to fill the third position. Therefore, $k_1 = 3$, $k_2 = 2$, and $k_3 = 1$, so that by the method of counting of Section 5.1 there are $(k_1)(k_2)(k_3) = (3)(2)(1) = 6$ ways to align these three animals. We may say that there are six permutations of three distinguishable objects.

*The term *permutation* was invented by Jacob Bernoulli in his landmark posthumous 1713 book on probability (Walker, 1929: 9).

In general, if there are n linear positions to fill with n objects, the first position may be filled in any one of n ways, the second may be filled in any one of $n - 1$ ways, the third in any one of $n - 2$ ways, and so on until the last position, which may be filled in only one way. That is, the filling of n positions with n objects results in ${}_nP_n$ permutations, where

$$ {}_nP_n = n(n - 1)(n - 2)\cdots(3)(2)(1). \tag{5.1} $$

This equation may be written more simply in *factorial* notation as

$$ {}_nP_n = n!, \tag{5.2} $$

where "n factorial" is the product of n and each smaller positive integer*; that is,

$$ n! = n(n - 1)(n - 2)\cdots(3)(2)(1). \tag{5.3} $$

Example 5.2 demonstrates such computation of the numbers of permutations.

EXAMPLE 5.2 The Number of Permutations of Distinct Objects

In how many sequences can six photographs be arranged on a page?

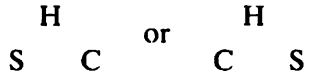
$$ {}_nP_n = 6! = (6)(5)(4)(3)(2)(1) = 720 $$

(b) Circular Arrangements. The numbers of permutations considered previously are for objects arranged on a line. If objects are arranged on a circle, there is no "starting position" as there is on a line, and the number of permutations is

$$ {}_nP'_n = \frac{n!}{n} = (n - 1)!. \tag{5.4} $$

(Observe that the notation ${}_nP'_n$ is used here for circular permutations to distinguish it from the symbol ${}_nP_n$ used for linear permutations.)

Referring again to a horse, a cow, and a sheep, there are ${}_nP'_n = \frac{n!}{n} = (n - 1)! = (3 - 1)! = 2! = 2$ distinct ways in which the three animals could be seated around a table, or arranged around the shore of a pond:



In this example, there is an assumed orientation of the observer, so clockwise and counterclockwise patterns are treated as different. That is, the animals are observed arranged around the top of the table, or observed from above the surface of the pond. But either one of these arrangements would look like the other one if observed from under the table or under the water; and if we did not wish to count the results of these two mirror-image observations as different, we would speak of there being one possible permutation, not two. For example, consider each of the preceding two diagrams to represent three beads on a circular string, one bead in the shape of a horse, one in the shape of a cow, and the other in the shape of a sheep. The two arrangements of H, C, and S shown are not really different, for there is no specific way of viewing the circle; one of the two arrangements turns into the other if the circle is turned over. If $n > 2$ and the orientation of the circle is not specified, then

*See the second footnote in Section 4.7.

the number of permutations of n objects on a circle is

$$ {}_nP''_n = \frac{n!}{2n} = \frac{(n-1)!}{2}. \tag{5.5} $$

(c) Fewer than n Positions. If one has n objects, but fewer than n positions in which to place them, then there would be considerably fewer numbers of ways to arrange the objects than in the case where there are positions for all n. For example, there are ${}_4P_4 = 4! = (4)(3)(2)(1) = 24$ ways of placing a horse (H), cow (C), sheep (S), and pig (P) in four positions on a line. However, there are only twelve ways of linearly arranging these four animals two at a time:

H,C; H,S; H,P; C,H; C,S; C,P; S,H; S,C; S,P; P,H; P,C; P,S.

The number of linear permutations of n objects taken X at a time is*

$$ {}_nP_X = \frac{n!}{(n-X)!}. \tag{5.6} $$

For the preceding example,

$$ {}_4P_2 = \frac{4!}{(4-2)!} = \frac{4!}{2!} = \frac{(4)(3)(2)(1)}{(2)(1)} = 12. $$

Equation 5.2 is a special case of Equation 5.6, where $X = n$; it is important to know that 0! is defined to be 1.†

If the arrangements are circular, instead of linear, then the number of them possible is

$$ {}_nP'_X = \frac{n!}{(n-X)!X}. \tag{5.7} $$

So, for example, there are only $4!/[(4-2)!2] = 6$ different ways of arranging two out of our four animals around a table:

H	H	H	C	C	S
C	S	P	S	P	P

for C seated at the table opposite H is the same arrangement as H seated across from C, S seated with H is the same as H with S, and so on. Example 5.3 demonstrates this further. Equation 5.4 is a special case of Equation 5.7, where $X = n$; and recall that 0! is defined as 1.

EXAMPLE 5.3 The Number of Permutations of n Objects Taken X at a Time: In How Many Different Ways Can a Sequence of Four Slides Be Chosen from a Collection of Six Slides?

$$ {}_nP_X = {}_6P_4 = \frac{6!}{(6-4)!} = \frac{6!}{2!} = \frac{(6)(5)(4)(3)(2)(1)}{(2)(1)} $$
$$ = (6)(5)(4)(3) = 360 $$

*Notation in the form of ${}_nP_X$ to indicate permutations of n items taken X at a time was used prior to 1869 by Harvey Goodwin (Cajori, 1929: 79).

†Why is 0! defined to be 1? In general, $n! = n[(n-1)!]$; for example, $5! = 5(4!)$, $4! = 4(3!)$, $3! = 3(2!)$, and $2! = 2(1!)$. Thus, $1! = 1(0!)$, which is so only if $0! = 1$.

If $n > 2$, then for every circular permutation viewed from above there is a mirror image of that permutation, which would be observed from below. If these two mirror images are not to be counted as different (e.g., if we are dealing with beads of different shapes or colors on a string), then the number of circular permutations is

$$ {}_nP''_X = \frac{n!}{2(n - X)!X}. \tag{5.8} $$

(d) If Some of the Objects Are Indistinguishable. If our group of four animals consisted of two horses (H), a cow (C), and a sheep (S), the number of permutations of the four animals would be twelve:

H,H,C,S; H,H,S,C; H,C,H,S; H,C,S,H; H,S,H,C; H,S,C,H;
C,H,H,S; C,H,S,H; C,S,H,H; S,H,H,C; S,H,C,H; S,C,H,H.

If n_i represents the number of like individuals in category i (in this case the number of animals in species i), then in this example $n_1 = 2$, $n_2 = 1$, and $n_3 = 1$, and we can write the number of permutations as

$$ {}_nP_{n_1,n_2,n_3} = \frac{n!}{n_1!n_2!n_3!} = \frac{4!}{2!1!1!} = 12. $$

If the four animals were two horses (H) and two cows (C), then there would be only six permutations:

H,H,C,C; C,C,H,H; H,C,H,C; C,H,C,H; H,C,C,H; C,H,H,C.

In this case, $n = 4, n_1 = 2$, and $n_2 = 2$, and the number of permutations is calculated to be ${}_nP_{n_1,n_2} = n!/(n_1!n_2!) = 4!/(2!2!) = (4)(3)(2)/[(2)(2)] = 6$.

In general, if n_1 members of the first category of objects are indistinguishable, as are n_2 of the second category, n_3 of the third category, and so on through n_k members of the kth category, then the number of different permutations is

$$ {}_nP_{n_1,n_2,\cdots,n_k} = \frac{n!}{n_1!n_2!\cdots n_k!} \text{ or } \frac{n!}{\prod_{i=1}^{k} n_i!}, \tag{5.9} $$

where the capital Greek letter pi (Π) denotes taking the product just as the capital Greek sigma (Σ, introduced in Section 3.1) indicates taking the sum. This is shown further in Example 5.4.

EXAMPLE 5.4 Permutations with Categories Containing Indistinguishable Members

There are twelve potted plants, six of one species, four of a second species, and two of a third species. How many different linear sequences of species are possible (for example, if arranging the pots on a shelf)?

$$ {}_nP_{n_1,n_2,n_3} = \frac{n}{\prod n_i!} $$

$$ = {}_{12}P_{6,4,2} = \frac{12!}{6!4!2!} $$

$$ = \frac{(12)(11)(10)(9)(8)(7)(6)(5)(4)(3)(2)(1)}{(6)(5)(4)(3)(2)(1)(4)(3)(2)(1)(2)(1)} = 13{,}860. $$

Note that the above calculation could have been simplified by writing

$$\frac{12!}{6!4!2!} = \frac{(12)(11)(10)(9)(8)(7)6!}{6!(4)(3)(2)(2)} = \frac{(12)(11)(10)(9)(8)(7)}{(4)(3)(2)(2)} = 13{,}860.$$

Here, "(1)" is dropped; also, "6!" appears in both the numerator and denominator, thus canceling out.

5.3 COMBINATIONS

In Section 5.2 we considered groupings of objects where the sequence within the groups was important. In many instances, however, only the components of a group, not their arrangement within the group, are important. We saw that if we select two animals from among a horse (H), cow (C), sheep (S), and pig (P), there are twelve ways of arranging the two on a line:

H,C; H,S; H,P; C,H; C,S; C,P; S,H; S,C; S,P; P,H; P,C; P,S.

However, some of these arrangements contain exactly the same kinds of animals, only in different order (e.g., H,C and C,H; H,S and S,H). If the groups of two are important to us, but not the sequence of objects within the groups, then we are speaking of *combinations*,* rather than permutations. Designating the number of combinations of n objects taken X at a time as ${}_nC_X$, we have†

$${}_nC_X = \frac{{}_nP_X}{X!} = \frac{n!}{X!(n - X)!}. \tag{5.10}$$

So for the present example, $n = 4$, $\overline{X} = 2$, and

$${}_4C_2 = \frac{4!}{2!(4 - 2)!} = \frac{4!}{2!2!} = \frac{(4)(3)(2)(1)}{(2)(1)(2)(1)} = \frac{(4)(3)}{2} = 6,$$

the six combinations of the four animals taken two at a time being

H,C; H,S; H,P; C,S; C,P; S,P.

Example 5.5 demonstrates the determination of numbers of combinations for another set of data.

It may be noted that

$${}_nC_n = 1, \tag{5.11}$$

meaning that there is only one way of selecting all n items; and

$${}_nC_1 = n, \tag{5.12}$$

indicating that there are n ways of selecting n items one at a time. Also,

$${}_nC_X = {}_nC_{n-X}, \tag{5.13}$$

*The word *combination* was used in this mathematical sense by Blaise Pascal (1623–1662) in 1654 (Smith, 1953: 528).

†Notation in the form of ${}_nC_X$ to indicate combinations of n items taken X at a time was used by G. Chrystal in 1899 (Cajori, 1929: 80).

EXAMPLE 5.5 Combinations of *n* Objects Taken *X* at a Time

Of a total of ten dogs, eight are to be used in a laboratory experiment. How many different combinations of eight animals may be formed from the ten?

$$_nC_X = {}_{10}C_8 = \frac{10!}{8!(10-8)!} = \frac{10!}{8!2!} = \frac{(10)(9)(8)(7)(6)(5)(4)(3)(2)(1)}{(8)(7)(6)(5)(4)(3)(2)(1)(2)(1)}$$
$$= 45.$$

It should be noted that the above calculations with factorials could have been simplified by writing

$$_{10}C_8 = \frac{10!}{8!2!} = \frac{(10)(9)8!}{8!2!} = \frac{(10)(9)}{2} = 45,$$

so that "8!" appears in both the numerator and denominator, thus canceling each other out.

which means that if we select X items from a group of n, we have at the same time selected the remaining $n - X$ items; that is, an exclusion is itself a selection. For example, if we selected two out of five persons to write a report, we have simultaneously selected three of the five to refrain from writing. Thus,

$$_5C_2 = \frac{5!}{2!(5-2)!} = \frac{5!}{2!3!} = 10 \quad \text{and} \quad {}_5C_{5-2} = {}_5C_3 = \frac{5!}{3!(5-3)!} = \frac{5!}{3!2!} = 10,$$

meaning that there are ten ways to select two out of five persons to perform a task and ten ways to select three out of five persons to be excluded from that task. This question may be addressed by applying Equation 5.9, reasoning that we are asking how many distinguishable arrangements there are of two writers and three nonwriters: $_5P_{2,3} = 5!/(2!3!) = 10$.

The product of combinatorial outcomes may also be employed to address questions such as in Example 5.4. This is demonstrated in Example 5.6.

EXAMPLE 5.6 Products of Combinations

This example provides an alternate method of answering the question of Example 5.4.

There are twelve potted plants, six of one species, four of a second species, and two of a third. How many different linear sequences of species are possible?

There are twelve positions in the sequence, which may be filled by the six members of the first species in this many ways:

$$_{12}C_6 = \frac{12!}{(12-6)!6!} = 924.$$

The remaining six positions in the sequence may be filled by the four members of the second species in this many ways:

$$_6C_4 = \frac{6!}{(6-4)!4!} = 15.$$

And the remaining two positions may be filled by the two members of the third species in only one way:

$$_2C_2 = \frac{2!}{(2-2)!2!} = 1.$$

As each of the ways of filling positions with members of one species exists in association with each of the ways of filling positions with members of each other species, the total different sequences of species is

$$(924)(15)(1) = 13{,}860.$$

From Equation 5.10 it may be noted that, as $_nC_X = {}_nP_X/X!$,

$$_nP_X = X!{}_nC_X. \tag{5.14}$$

It is common mathematical convention to indicate the number of combinations of n objects taken X at a time as $\binom{n}{X}$ instead of $_nC_X$, so for the problem at the beginning of Section 5.3 we could have written*

$$\binom{n}{X} = \binom{4}{2} = \frac{4!}{2!(4-2)!} = 6.$$

Binomial coefficients, which are discussed in Section 24.1, take this form.

5.4 SETS

A *set* is a defined collection of items. For example, a set may be a group of four animals, a collection of eighteen amino acids, an assemblage of twenty-five students, or a group of three genetic traits. Each item in a set is termed an *element*. If a set of animals includes these four elements: horse (H), cow (C), sheep (S), and pig (P), and a second set consists of the elements P, S, H, and C, then we say that the two sets are *equal*, as they contain exactly the same elements. The sequence of elements within sets is immaterial in defining equality or inequality of sets.

If a set consisted of animals H and P, it would be declared a *subset* of the above set (H, C, S, P). A subset is a set, all of whose elements are elements of a larger set.† Therefore, the determination of combinations of X items taken from a set of n items (Section 5.3) is really the counting of possible subsets of items from the set of n items.

In an experiment (or other phenomenon that yields results to observe), there is a set (usually very large) of possible outcomes. Let us refer to this set as the *outcome set*.‡

Each element of the set is one of the possible outcomes of the experiment. For example, if an experiment consists of tossing two coins, the outcome set consists of four elements: H,H; H,T; T,H; T,T, as these are all of the possible outcomes.

A subset of the outcome set is called an *event*. If the outcome set were the possible rolls of a die: 1, 2, 3, 4, 5, 6, an event might be declared to be "even-numbered rolls" (i.e., 2, 4, 6), and another event might be defined as "rolls greater than 4"

*This parenthetical notation for combinations was introduced by Andreas von Ettingshausen in 1826 (Miller, 2004c). Some authors have used a symbol in the form of C_X^n (or ${}^nC_X$) instead of $_nC_X$ for combinations and P_X^n (or ${}^nP_X$) instead of $_nP_X$ for permutations; those symbols will not be used in this book, in order to avoid confusing n with an exponent.

†Utilizing the terms *set* and *subset* in this fashion dates from the last half of the nineteenth century (Miller, 2004a).

‡Also called the *sample space*.

(i.e., 5, 6). In tossing two coins, one event could be "the two coins land differently" (i.e., T,H; H,T), and another event could be "heads do not appear" (i.e., T,T). If the two events in the same outcome set have some elements in common, the two events are said to intersect; and the *intersection* of the two events is that subset composed of those common elements. For example, the event "even-numbered rolls" of a die (2, 4, 6) and the event "rolls greater than 4" (5, 6) have an element in common (namely, the roll 6); therefore 6 is the intersection of the two events. For the events "even-numbered rolls" (2, 4, 6) and "rolls less than 5" (1, 2, 3, 4), the intersection subset consists of those elements of the events that are both even-numbered and less than 5 (namely, 2, 4).*

If two events have no elements in common, they are said to be *mutually exclusive*, and the two sets are said to be *disjoint*. The set that is the intersection of disjoint sets contains no elements and is often called the *empty set* or the *null set*. For example, the events "odd-numbered rolls" and "even-numbered rolls" are mutually exclusive and there are no elements common to both of them.

If we ask what elements are found in either one event or another, or in both of them, we are speaking of the *union* of the two events. The union of the events "even-numbered rolls" and "rolls less than 5" is that subset of the outcome set that contains elements found in either set (or both sets), namely 1, 2, 3, 4, 6.†

Once a subset has been defined, all other elements in the outcome set are said to be the *complement* of that subset. So, if an event is defined as "even-numbered rolls" of a die (2, 4, 6), the complementary subset consists of "odd-numbered rolls" (1, 3, 5). If subset is "rolls less than 5" (1, 2, 3, 4), the complement is the subset consisting of rolls 5 or greater (5, 6).

The above considerations may be presented by what are known as *Venn diagrams*,‡ shown in Figure 5.1.

The rectangle in this diagram denotes the outcome set, the set of all possible outcomes from an experiment or other producer of observations. The circle on the

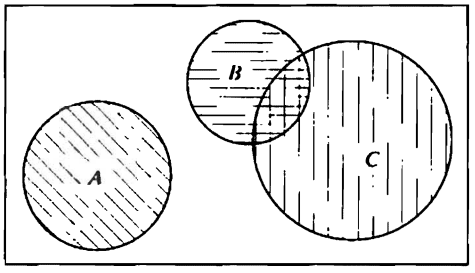


FIGURE 5.1: A Venn diagram showing the relationships among the outcome set represented by the rectangle and the subsets represented by circles *A*, *B*, and *C*. Subsets *B* and *C* intersect, with no intersection with *A*.

*The term *intersection* had been employed in this manner by 1909 (Miller, 2004a). The mathematical symbol for intersection is "∩", first used by Italian mathematician Giuseppe Peano (1858–1932) in 1888 (Miller, 2004a); so, for example, the intersection of set *A* (consisting of 2, 4, 6) and set *B* (consisting of 5, 6) is set $A \cap B$ (consisting of 6).

†The term *union* had been employed in this way by 1912 (Miller, 2004a). The mathematical symbol for union is "∪", first used by Giuseppe Peano in 1888 (Miller, 2004a); so, for example, if set *A* is composed of even-numbered rolls of a die (2, 4, 6), and set *B* is odd-numbered rolls (1, 3, 5), the union of the two sets, namely $A \cup B$, is 2, 4, 6, 1, 3, 5.

‡Named for English mathematical logician John Venn (1834–1923), who in 1880 greatly improved and popularized the diagrams (sometimes called "Euler diagrams") devised by Leonhard Euler (1707–1783) (Gullberg, 1997: 242; O'Connor and Robertson, 2003).

left represents a subset of the outcome set that we shall refer to as event A, the circle in the center signifies a second subset of the outcome set that we shall refer to as event B, and the circle on the right depicts a third subset of the outcome set that we shall call event C. If, for example, an outcome set (the rectangle) is the number of vertebrate animals in a forest, subset A might be animals without legs (namely, snakes), subset B might be mammals, and subset C might be flying animals. Figure 5.1 demonstrates graphically what is meant by union, intersection, mutually exclusive, and complementary sets: The union of B and C (the areas with any horizontal or vertical shading) represents all birds and mammals; the intersection of B and C (the area with both horizontal and vertical shading) represents flying mammals (i.e., bats); the portion of C with only vertical shading represents birds; A is mutually exclusive relative to the union of B and C, and the unshaded area (representing all other vertebrates—namely, amphibians and turtles) is complementary to A, B, and C (and is also mutually exclusive of A, B, and C).

5.5 PROBABILITY OF AN EVENT

As in Section 1.3, we shall define the *relative frequency* of an event as the proportion of the total observations of outcomes that event represents. Consider an outcome set with two elements, such as the possible results from tossing a coin (H; T) or the sex of a person (male; female). If n is the total number of coin tosses and f is the total number of heads observed, then the relative frequency of heads is f/n. Thus, if heads are observed 52 times in 100 coin tosses, the relative frequency is $52/100 = 0.52$ (or 52%). If 275 males occur in 500 human births, the relative frequency of males is $f/n = 275/500 = 0.55$ (or 55%). In general, we may write

$$\text{relative frequency of an event} = \frac{\text{frequency of that event}}{\text{total number of all events}} = \frac{f}{n}. \tag{5.15}$$

The value of f may, of course, range from 0 to n, and the relative frequency may, therefore, range from 0 to 1 (or 0% to 100%). A biological example is given as Example 5.7.

EXAMPLE 5.7 Relative Frequencies

A sample of 852 vertebrate animals is taken randomly from a forest. The sampling was done *with replacement*, meaning that the animals were taken one at a time, returning each one to the forest before the next one was selected. This is done to prevent the sampling procedure from altering the relative frequency in the sampled population. If the sample size is very small compared to the population size, replacement is not necessary. (Recall that random sampling assumes that each individual animal is equally likely to become a part of the sample.)

Vertebrate Subset	*Number*	*Relative Frequency*
amphibians	53	53/852 = 0.06
turtles	41	41/852 = 0.05
snakes	204	204/852 = 0.24
birds	418	418/852 = 0.49
mammals	136	136/852 = 0.16
total	852	1.00

The *probability* of an event is the likelihood of that event expressed either by the relative frequency observed from a large number of data or by knowledge of the system under study. In Example 5.7 the relative frequencies of vertebrate groups have been observed from randomly sampling forest animals. If, for the sake of the present example, we assume that each animal has the same chance of being caught as part of our sample (an unrealistic assumption in nature), we may estimate the probability, P, that the next animal captured will be a snake ($P = 0.24$). Or, using the data of the preceding paragraph, we can estimate that the probability that a human birth will be a male is 0.55, or that the probability of tossing a coin that lands head side up is 0.52. A probability may sometimes be predicted on the basis of knowledge about the system (e.g., the structure of a coin or of a die, or the Mendelian principles of heredity). If we assume that there is no reason why a tossed coin should land "heads" more or less often than "tails," we say there is an equal probability of each outcome: $P(H) = \frac{1}{2}$ and $P(T) = \frac{1}{2}$ states that "the probability of heads is 0.5 and the probability of tails is 0.5."

Probabilities, like relative frequencies, can range from 0 to 1. A probability of 0 means that the event is impossible. For example, in tossing a coin, P(neither H nor T) = 0, or in rolling a die, P(number > 6) = 0. A probability of 1 means that an event is certain. For example, in tossing a coin, P(H or T) = 1; or in rolling a die, $P(1 \leq \text{number} \leq 6) = 1$.*

5.6 ADDING PROBABILITIES

(a) If Events Are Mutually Exclusive. If two events (call them A and B) are mutually exclusive (e.g., legless vertebrates and mammals are disjoint sets in Figure 5.1), then the probability of either event A or event B is the sum of the probabilities of the two events:

$$P(A \text{ or } B) = P(A) + P(B). \tag{5.16}$$

For example, if the probability of a tossed coin landing head up is $\frac{1}{2}$ and the probability of its landing tail up is $\frac{1}{2}$, then the probability of either head or tail up is

$$P(H \text{ or } T) = P(H) + P(T) = \frac{1}{2} + \frac{1}{2} = 1. \tag{5.17}$$

And, for the data in Example 5.7, the probability of selecting, at random, a reptile would be P(turtle or snake) = P(turtle) + P(snake) = 0.05 + 0.24 = 0.29.

This rule for adding probabilities may be extended for more than two mutually exclusive events. For example, the probability of rolling a 2 on a die is $\frac{1}{6}$, the probability of rolling a 4 is $\frac{1}{6}$, and the probability of rolling a 6 is $\frac{1}{6}$; so the probability of rolling an even number is

$$\begin{aligned} P(\text{even number}) = P(2 \text{ or } 4 \text{ or } 6) &= P(2) + P(4) + P(6) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}. \end{aligned}$$

*A concept related to probability is the *odds* for an event, namely the ratio of the probability of the event occurring and the probability of that event not occurring. For example, if the probability of a male birth is 0.55 (and, therefore, the probability of a female birth is 0.45), then the odds in favor of male births are 0.55/0.45, expressed as "11 to 9."

And, for the data in Example 5.7, the probability of randomly selecting a reptile or amphibian would be $P(\text{turtle}) + P(\text{snake}) + P(\text{amphibian}) = 0.05 + 0.24 + 0.06 = 0.35$.

(b) If Events Are Not Mutually Exclusive. If two events are not mutually exclusive—that is, they intersect (e.g., mammals and flying vertebrates are not disjoint sets in Figure 5.1)—then the addition of the probabilities of the two events must be modified. For example, if we roll a die, the probability of rolling an odd number is

$$P(\text{odd number}) = P(1 \text{ or } 3 \text{ or } 5) = P(1) + P(3) + P(5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2};$$

and the probability of rolling a number less than 4 is

$$P(\text{number} < 4) = P(1 \text{ or } 2 \text{ or } 3) = P(1) + P(2) + P(3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}.$$

The probability of rolling either an odd number or a number less than 4 obviously is *not* calculated by Equation 5.16, for that equation would yield

$$\begin{aligned} &P(\text{odd number or number} < 4) \\ &\overset{?}{=} P(\text{odd}) + P(\text{number} < 4) \\ &= P[(1 \text{ or } 3 \text{ or } 5) \text{ or } (1 \text{ or } 2 \text{ or } 3)] \\ &= [P(1) + P(3) + P(5)] + [P(1) + P(2) + P(3)] \\ &= \left(\frac{1}{6} + \frac{1}{6} + \frac{1}{6}\right) + \left(\frac{1}{6} + \frac{1}{6} + \frac{1}{6}\right) = 1, \end{aligned}$$

and that would mean that we are certain ($P = 1$) to roll either an odd number or a number less than 4, which would mean that a roll of 4 or 6 is impossible!

The invalidity of the last calculation is due to the fact that the two elements (namely 1 and 3) that lie in both events are counted twice. The subset of elements consisting of rolls 1 and 3 is the intersection of the two events and its probability needs to be subtracted from the preceding computation so that $P(1 \text{ or } 3)$ is counted once, not twice. Therefore, for two intersecting events, A and B, the probability of either A or B is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B). \tag{5.18}$$

In the preceding example,

$$\begin{aligned} &P(\text{odd number or number} < 4) \\ &\quad = P(\text{odd number}) + P(\text{number} < 4) \\ &\qquad - P(\text{odd number and number} < 4) \\ &\quad = P[(1 \text{ or } 3 \text{ or } 5) \text{ or } (1 \text{ or } 2 \text{ or } 3)] - P(1 \text{ or } 3) \\ &\quad = [P(1) + P(3) + P(5)] + [P(1) + P(2) + P(3)] - [P(1) + P(3)] \\ &\quad = \left(\frac{1}{6} + \frac{1}{6} + \frac{1}{6}\right) + \left(\frac{1}{6} + \frac{1}{6} + \frac{1}{6}\right) - \left(\frac{1}{6} + \frac{1}{6}\right) = \frac{4}{6} = \frac{2}{3}. \end{aligned}$$

It may be noted that Equation 5.16 is a special case of Equation 5.18, where $P(A \text{ and } B) = 0$. Example 5.8 demonstrates these probability calculations with a different set of data.

EXAMPLE 5.8 Adding Probabilities of Intersecting Events

A deck of playing cards is composed of 52 cards, with thirteen cards in each of four suits called clubs, diamonds, hearts, and spades. In each suit there is one card each of the following thirteen denominations: ace (A), 2, 3, 4, 5, 6, 7, 8, 9, 10, jack (J), queen (Q), king (K). What is the probability of selecting at random a diamond from the deck of 52 cards?

The event in question (diamonds) is a subset with thirteen elements; therefore,

$$P(\text{diamond}) = \frac{13}{52} = \frac{1}{4} = 0.250.$$

What is the probability of selecting at random a king from the deck?

The event in question (king) has four elements; therefore,

$$P(\text{king}) = \frac{4}{52} = \frac{1}{13} = 0.077.$$

What is the probability of selecting at random a diamond or a king?

The two events (diamonds and kings) intersect, with the intersection having one element (the king of diamonds); therefore,

$$\begin{aligned} P(\text{diamond or king}) &= P(\text{diamond}) + P(\text{king}) - P(\text{diamond and king}) \\ &= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} \\ &= \frac{16}{52} = \frac{4}{13} = 0.308. \end{aligned}$$

If three events are not mutually exclusive, the situation is more complex, yet straightforward. As seen in Figure 5.2, there may be three two-way intersections, shown with vertical shading (A and B; A and C; and B and C), and a three-way

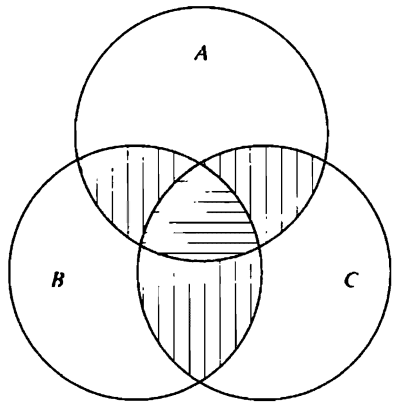


FIGURE 5.2: A Venn diagram showing three intersecting sets: *A*, *B*, and *C*. Here there are three two-way intersections (vertical shading) and one three-way intersection (horizontal shading).

intersection, shown with horizontal shading (A and B and C). If we add the probabilities of the three events, A, B, and C, as $P(A) + P(B) + P(C)$, we are adding the two-way intersections twice. So, we can subtract $P(A \text{ and } B)$, $P(A \text{ and } C)$, and $P(B \text{ and } C)$. Also, the three-way intersection is added three times in $P(A) + P(B) + P(C)$, and subtracted three times by subtracting the three two-way intersections; thus, $P(A \text{ and } B \text{ and } C)$ must be added back into the calculation. Therefore, for three events, not mutually exclusive,

$$\begin{aligned} P(A \text{ or } B \text{ or } C) = {} & P(A) + P(B) + P(C) \\ & - P(A \text{ and } B) - P(A \text{ and } C) - P(B \text{ and } C) \\ & + P(A \text{ and } B \text{ and } C). \end{aligned} \tag{5.19}$$

5.7 MULTIPLYING PROBABILITIES

If two or more events intersect (as A and B in Figure 5.1 and A, B, and C in Figure 5.2), the probability associated with the intersection is the product of the probabilities of the individual events. That is,

$$P(A \text{ and } B) = [P(A)][P(B)], \tag{5.20}$$

$$P(A \text{ and } B \text{ and } C) = [P(A)][P(B)][P(C)], \tag{5.21}$$

and so on.

For example, the probability of a tossed coin landing heads is $\frac{1}{2}$. If two coins are tossed, the probability of *both* coins landing heads is

$$P(H,H) = [P(H)][P(H)] = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \left(\frac{1}{4}\right) = 0.25.$$

This can be verified by examining the outcome set:

H,H; H,T; T,H; T,T,

where $P(H,H)$ is one outcome out of four equally likely outcomes. The probability that 3 tossed coins will land heads is

$$P(H,H,H) = [P(H)][P(H)][P(H)] = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \left(\frac{1}{8}\right) = 0.125.$$

Note, however, that if one or more coins have already been tossed, the probability that the next coin toss (of the same or a different coin) will be heads is simply $\frac{1}{2}$.

5.8 CONDITIONAL PROBABILITIES

There are occasions when our interest will be in determining a *conditional probability*, which is the probability of one event with the stipulation that another event also occurs. An illustration of this, using a deck of 52 playing cards (as described in Example 5.8), would be the probability of selecting a queen, given that the card is a spade. In general, a conditional probability is

$$P(\text{event } A, \text{ given event } B) = \frac{P(A \text{ and } B \text{ jointly})}{P(B)}, \tag{5.22}$$

which can also be calculated as

$$P(\text{event } A, \text{ given event } B) = \frac{\text{frequency of events } A \text{ and } B \text{ jointly}}{\text{frequency of event } B}. \tag{5.23}$$

So, the probability of randomly selecting a queen, with the specification that the card is a spade, is (using Equation 5.22)

$$P(\text{queen, given it is a spade}) = \frac{P(\text{queen of spades})}{P(\text{spade})}$$
$$= (1/52)/(13/52) = 0.02/0.25 = 0.08,$$

which (by Equation 5.23) would be calculated as

$$P(\text{queen, given it is a spade}) = \frac{\text{frequency of queen of spades}}{\text{frequency of spades}}$$
$$= 1/13 = 0.8.$$

Note that this conditional probability is quite different from the probability of selecting a spade, given that the card is a queen, for that would be (by Equation 5.23)

$$P(\text{spade, given it is a queen}) = \frac{\text{frequency of queen of spades}}{\text{frequency of queens}}$$
$$= 1/4 = 0.25.$$

EXERCISES

5.1. A person may receive a grade of either high (H), medium (M), or low (L) on a hearing test, and a grade of either good (G) or poor (P) on a sight test.
(a) How many different outcomes are there if both tests are taken?
(b) What are these outcomes?

5.2. A menu lists three meats, four salads, and two desserts. In how many ways can a meal of one meat, one salad, and one dessert be selected?

5.3. If an organism (e.g., human) has 23 pairs of chromosomes in each diploid cell, how many different gametes are possible for the individual to produce by assortment of chromosomes?

5.4. In how many ways can five animal cages be arranged on a shelf?

5.5. In how many ways can 12 different amino acids be arranged into a polypeptide chain of five amino acids?

5.6. An octapeptide is known to contain four of one amino acid, two of another, and two of a third. How many different amino-acid sequences are possible?

5.7. Students are given a list of nine books and told that they will be examined on the contents of five of them. How many combinations of five books are possible?

5.8. The four human blood types below are genetic phenotypes that are mutually exclusive events. Of 5400 individuals examined, the following frequency of each blood type is observed. What is the relative frequency of each blood type?

Blood Type	*Frequency*
O	2672
A	2041
B	486
AB	201

5.9. An aquarium contains the following numbers of tropical freshwater fishes. What is the relative frequency of each species?

Species	*Number*
Paracheirodon innesi, neon tetra	11
Cheirodon axelrodi, cardinal tetra	6
Pterophyllum scalare, angelfish	4
Pterophyllum altum, angelfish	2
Pterophyllum dumerilii, angelfish	2
Nannostomus marginatus, one-lined pencilfish	2
Nannostomus anomalus golden pencilfish	2

5.10. Use the data of Exercise 5.8, assuming that each of the 5400 has an equal opportunity of being encountered.
- **(a)** Estimate the probability of encountering a person with type A blood.
- **(b)** Estimate the probability of encountering a person who has either type A or type AB blood.

5.11. Use the data of Exercise 5.9, assuming that each individual fish has the same probability of being encountered.
- **(a)** Estimate the probability of encountering an angelfish of the species *Pterophyllum scalare*.
- **(b)** Estimate the probability of encountering a fish belonging to the angelfish genus *Pterophyllum*.

5.12. Either allele A or a may occur at a particular genetic locus. An offspring receives one of its alleles from each of its parents. If one parent possesses alleles A and a and the other parent possesses a and a:
- **(a)** What is the probability of an offspring receiving an A and an a?
- **(b)** What is the probability of an offspring receiving two a alleles?
- **(c)** What is the probability of an offspring receiving two A alleles?

5.13. In a deck of playing cards (see Example 5.8 for a description),
- **(a)** What is the probability of selecting a queen of clubs?
- **(b)** What is the probability of selecting a black (i.e., club or spade) queen?
- **(c)** What is the probability of selecting a black face card (i.e., a black jack, queen, or king)?

5.14. A cage contains six rats, two of them white (W) and four of them black (B); a second cage contains four rats, two white and two black; and a third cage contains five rats, three white and two black. If one rat is selected randomly from each cage,
- **(a)** What is the probability that all three rats selected will be white?
- **(b)** What is the probability that exactly two of the three will be white?
- **(c)** What is the probability of selecting at least two white rats?

5.15. A group of dogs consists of three brown males, two brown females, four white males, four white females, five black males, and four black females. What is the probability of selecting at random
- **(a)** A brown female dog?
- **(b)** A female dog, if the dog is brown?
- **(c)** A brown dog, if the dog is a female?

CHAPTER 6

The Normal Distribution

Commonly, a distribution of interval- or ratio-scale data is observed to have a preponderance of values around the mean with progressively fewer observations toward the extremes of the range of values (see, e.g., Figure 1.5). If n is large, the frequency polygons of many biological data distributions are "bell-shaped"* and look something like Figure 6.1.

Figure 6.1 is a frequency curve for a *normal distribution.*† Not all bell-shaped curves are normal; although biologists are unlikely to need to perform calculations with this equation, it can be noted that a *normal distribution* is defined as one in which height of the curve at X_i is as expressed by the relation:

$$Y_i = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X_i-\mu)^2/2\sigma^2}. \tag{6.1}$$

The height of the curve, Y_i, is referred to as the *normal density*. It is not a frequency, for in a normally distributed population of continuous data the frequency of occurrence of a measurement *exactly* equal to X_i (e.g., exactly equal to 12.5000 cm, or exactly equal to 12.50001 cm) is zero. Equation 6.1 contains two mathematical constants:

*Comparing the curve's shape to that of a bell has been traced as far back as 1872 (Stigler, 1999: 405).

†The normal distribution is sometimes called the *Gaussian distribution*, after [Johann] Karl Friedrich Gauss (1777–1855), a phenomenal German mathematician contributing to many fields of mathematics and for whom the unit of magnetic induction ("gauss") is named. Gauss discussed this distribution in 1809, but the influential French mathematician and astronomer Pierre-Simon Laplace (1749–1827) mentioned it in 1774, and it was first announced in 1733 by mathematician Abraham de Moivre (1667–1754; also spelled De Moivre and Demoivre), who was born in France but emigrated to England at age 21 (after three years in prison) to escape religious persecution as a Protestant (David, 1962: 161–178; Pearson, 1924; Stigler, 1980; Walker, 1934). This situation has been cited as an example of "Stigler's Law of Eponymy," which states that "no scientific discovery is named after its original discoverer" (Stigler, 1980). The distribution was first used, by de Moivre, to approximate a binomial distribution (discussed in Section 24.1) (Stigler, 1999: 407). The adjective *normal* was first used for the distribution by Charles S. Peirce in 1873, and by Wilhelm Lexis and Sir Francis Galton in 1877 (Stigler, 1999: 404–415); Karl Pearson recommended the routine use of that term to avoid "an international question of priority" although it "has the disadvantage of leading people to believe that all other distributions of frequency are in one sense or another 'abnormal'" (Pearson, 1920).

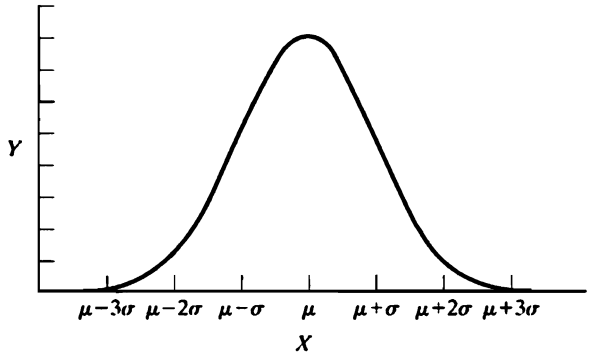


FIGURE 6.1: A normal distribution.

π (lowercase Greek pi),* which equals 3.14159...; and e (the base of Naperian, or natural, logarithms),† which equals 2.71828.... There are also two parameters (μ and σ^2) in the equation. Thus, for any given standard deviation, σ, there are an infinite number of normal curves possible, depending on μ. Figure 6.2a shows normal curves for $\sigma = 1$ and $\mu = 0, 1$, and 2. Likewise, for any given mean, μ, an infinity of normal curves is possible, each with a different value of σ. Figure 6.2b shows normal curves for $\mu = 0$ and $\sigma = 1, 1.5$, and 2.

A normal curve with $\mu = 0$ and $\sigma = 1$ is said to be a *standardized normal curve.* Thus, for a standardized normal distribution,

$$Y_i = 1\sqrt{2\pi}e^{-X_i^2/2}. \tag{6.2}$$

*The lowercase Greek letter pi, π, denotes the ratio between the circumference and the diameter of a circle. This symbol was advanced in 1706 by Wales-born William Jones (1675–1749), after it had been used for over 50 years to represent the circumference (Cajori, 1928/9, Vol. II: 9; Smith, 1953: 312); but it did not gain popularity for this purpose until Swiss Leonhard Euler (1707–1783) began using it in 1736 instead of p (Blatner, 1997: 78; Smith, 1953: 312). According to Gullberg (1997: 85), Jones probably selected this symbol because it is the first letter of the Greek word for "periphery." (See also Section 26.1.) Pi is an "irrational number," meaning that it cannot be expressed as the ratio of two integers. To 20 decimal places its value is 3.14159 26535 89792 33846 (and it may be noted that this number rounded to 10 decimal places is sufficient to obtain, from the diameter, the circumference of a circle as large as the earth's equator to within about a centimeter of accuracy). Beckmann (1977), Blatner (1997), and Dodge (1996) present the history of π and its calculation. By 2000 B.C.E., the Babylonians knew its value to within 0.02. Archimedes of Syracuse (287–212 B.C.E.) was the first to present a procedure to calculate π to any desired accuracy, and he computed it accurate to the third decimal place. Many computational methods were subsequently developed, and π was determined to six decimal places of accuracy by around 500 C.E., to 20 decimal places by around 1600, and to 100 in 1706; 1000 decimal places were reached, using a mechanical calculating machine, before electronic computers joined the challenge in 1949. In the computer era, with advancement of machines and algorithms, one million digits were achieved in 1973, by the end of the 1980s there were calculations accurate to more than a billion digits, and more than one trillion (1,000,000,000,000) digits have now been attained.

†e is an irrational number (as is π; see the preceding footnote). To 20 decimal places e is 2.71828 18284 59045 23536. The symbol, e, for this quantity was introduced by the great Swiss mathematician Leonhard Euler (1707–1783) in 1727 or 1728 and published by him in 1736 (Cajori, 1928/9, Vol. 2: 13; Gullberg, 1997: 85). Johnson and Leeming (1990) discussed the randomness of the digits of e, and Maor (1994) presented a history of this number and its mathematical ramifications. In 2000, e was calculated to 17 billion decimal places (Adrian, 2006: 63).

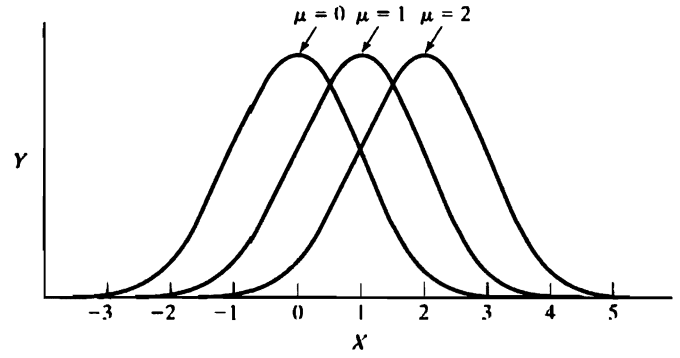


FIGURE 6.2a: Normal distribution with $\sigma = 1$, varying in location with different means (μ).

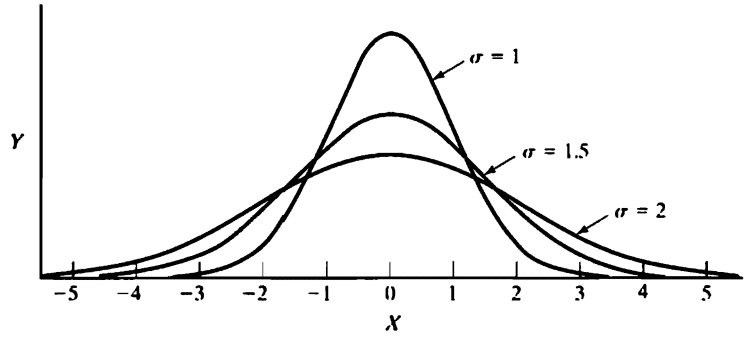


FIGURE 6.2b: Normal distributions with $\mu = 0$, varying in spread with different standard deviations (σ).

6.1 PROPORTIONS OF A NORMAL DISTRIBUTION

If a population of 1000 body weights is normally distributed and has a mean, μ, of 70 kg, one-half of the population (500 weights) is larger than 70 kg and one-half is smaller. This is true simply because the normal distribution is symmetrical. But if we desire to ask what portion of the population is larger than 80 kg, we need to know σ, the standard deviation of the population. If $\sigma = 10$ kg, then 80 kg is one standard deviation larger than the mean, and the portion of the population in question is the shaded area in Figure 6.3a. If, however, $\sigma = 5$ kg, then 80 kg is two standard deviations above μ, and we are referring to a relatively small portion of the population, as shown in Figure 6.3b.

Appendix Table B.2 enables us to determine proportions of normal distributions. For any X_i value from a normal population with mean μ, and standard deviation σ, the value

$$Z = \frac{X_i - \mu}{\sigma} \tag{6.3}$$

tells us how many standard deviations from the mean the X_i value is located. Carrying out the calculation of Equation 6.3 is known as *normalizing*, or *standardizing*, X_i; and Z is known as a *normal deviate*, or a standard score.* The mean of a set of standard scores is 0, and the variance is 1.

*This standard normal curve was introduced in 1899 by W. F. Sheppard (Walker, 1929: 188), and the term *normal deviate* was first used, in 1907, by F. Galton (David, 1995).

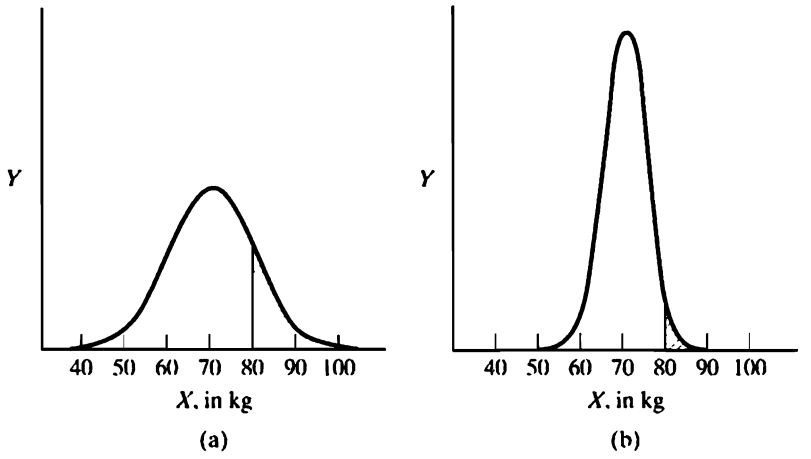


FIGURE 6.3: Two normal distributions with $\mu = 70$ kg. The shaded areas are the portions of the curves that lie above $X = 80$ kg. For distribution (a), $\mu = 70$ kg and $\sigma = 10$ kg; for distribution (b), $\mu = 70$ kg and $\sigma = 5$ kg.

Table B.2 tells us what proportion of a normal distribution lies beyond a given value of Z.* If $\mu = 70$ kg, $\sigma = 10$ kg, and $X_i = 70$ kg, then $Z = (70 \text{ kg} - 70 \text{ kg})/10 \text{ kg} = 0$, and by consulting Table B.2 we see that $P(X_i > 70 \text{ kg}) = P(Z > 0) = 0.5000$.† That is, 0.5000 (or 50.00%) of the distribution is larger than 70 kg. To determine the proportion of the distribution that is greater than 80 kg in weight, $Z = (80 \text{ kg} - 70 \text{ kg})/10 \text{ kg} = 1$, and $P(X_i > 80 \text{ kg}) = P(Z > 1) = 0.1587$ (or 15.87%). This could be stated as being the probability of drawing at random a measurement, X_i, greater than 80 kg from a population with a mean (μ) of 70 kg and a standard deviation (σ) of 10 kg. What, then, is the probability of obtaining, at random, a measurement, X_i, which is less than 80 kg? $P(X_i > 80 \text{ kg}) = 0.1587$, so $P(X_i < 80 \text{ kg}) = 1.0000 - 0.1587 = 0.8413$; that is, if 15.87% of the population is greater than X_i, then 100% − 15.87% (i.e., 84.13% of the population is less than X_i).‡ Example 6.1a presents calculations for determining proportions of a normal distribution lying between a variety of limits.

Note that Table B.2 contains no negative values of Z. However, if we are concerned with proportions in the left half of the distribution, we are simply dealing with areas of the curve that are mirror images of those present in the table. This is demonstrated in Example 6.1b.§

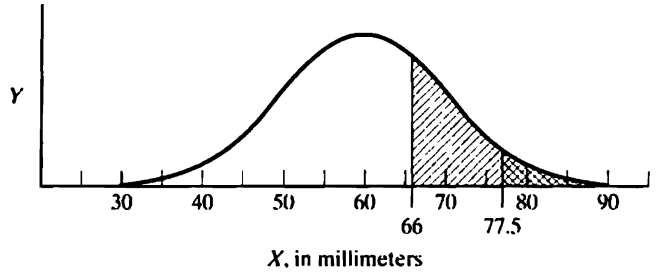
*The first tables of areas under the normal curve were published in 1799 by Christian Kramp (Walker, 1929: 58). Today, some calculators and many computer programs determine normal probabilities (e.g., see Boomsma and Molenaar, 1994).

†Read $P(X_i > 70 \text{ kg})$ as "the probability of an X_i greater than 70 kg"; $P(Z > 0)$ is read as "the probability of a Z greater than 0."

‡The statement that "$P(X_i > 80 \text{ kg}) = 0.1587$, therefore $P(X_i < 80) = 1.0000 - 0.1587$" does not take into account the case of $X_i = 80$ kg. But, as we are considering the distribution at hand to be a continuous one, the probability of X_i being *exactly* 80.000 ... kg (or being *exactly* any other stated value) is practically nil, so these types of probability statements offer no practical difficulties.

§Some old literature avoided referring to negative Z's by expressing the quantity, $Z + 5$, called a *probit*. This term was introduced in 1934 by C. I. Bliss (David, 1995).

EXAMPLE 6.1a Calculating Proportions of a Normal Distribution of Bone Lengths, Where $\mu = 60$ mm and $\sigma = 10$ mm



1. What proportion of the population of bone lengths is larger than 66 mm?

$$Z = \frac{X_i - \mu}{\sigma} = \frac{66\text{ mm} - 60\text{ mm}}{10\text{ mm}} = 0.60$$

$$P(X_i > 66\text{ mm}) = P(Z > 0.60) = 0.2743 \text{ or } 27.43\%$$

2. What is the probability of picking, at random from this population, a bone larger than 66 mm? This is simply another way of stating the quantity calculated in part (1). The answer is 0.2743.
3. If there are 2000 bone lengths in this population, how many of them are greater than 66 mm?

$$(0.2743)(2000) = 549$$

4. What proportion of the population is smaller than 66 mm?

$$P(X_i < 66\text{ mm}) = 1.0000 - P(X_i > 66\text{ mm}) = 1.0000 - 0.2743 = 0.7257$$

5. What proportion of this population lies between 60 and 66 mm? Of the total population, 0.5000 is larger than 60 mm and 0.2743 is larger than 66 mm. Therefore, $0.5000 - 0.2743 = 0.2257$ of the population lies between 60 and 66 mm. That is, $P(60\text{ mm} < X_i < 66\text{ mm}) = 0.5000 - 0.2743 = 0.2257$.
6. What portion of the area under the normal curve lies to the right of 77.5 mm?

$$Z = \frac{77.5\text{ mm} - 60\text{ mm}}{10\text{ mm}} = 1.75$$

$$P(X_i > 77.5\text{ mm}) = P(Z > 1.75) = 0.0401 \text{ or } 4.01\%$$

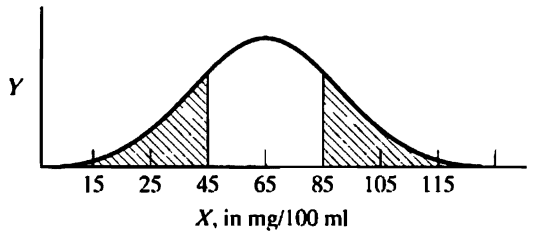
7. If there are 2000 bone lengths in the population, how many of them are larger than 77.5 mm?

$$(0.0401)(2000) = 80$$

8. What is the probability of selecting at random from this population a bone measuring between 66 and 77.5 mm in length?

$$P(66\text{ mm} < X_i < 77.5\text{ mm}) = P(0.60 < Z < 1.75) = 0.2743 - 0.0401 = 0.2342$$

EXAMPLE 6.1b Calculating Proportions of a Normal Distribution of Sucrose Concentrations, Where $\mu = 65$ mg/100 ml and $\sigma = 25$ mg/100 ml



1. What proportion of the population is greater than 85 mg/100 ml?

$$Z = \frac{(X_i - \mu)}{\sigma} = \frac{85 \text{ mg/100 ml} - 65 \text{ mg/100 ml}}{25 \text{ mg/100 ml}} = 0.8$$

$$P(X_i > 85 \text{ mg/100 ml}) = P(Z > 0.8) = 0.2119 \text{ or } 21.19\%$$

2. What proportion of the population is less than 45 mg/100 ml?

$$Z = \frac{45 \text{ mg/100 ml} - 65 \text{ mg/100 ml}}{25 \text{ mg/100 ml}} = -0.80$$

$$P(X_i < 45 \text{ mg/100 ml}) = P(Z < -0.80) = P(Z > 0.80) = 0.2119$$

That is, the probability of selecting from this population an observation less than 0.80 standard deviations below the mean is equal to the probability of obtaining an observation greater than 0.80 standard deviations above the mean.

3. What proportion of the population lies between 45 and 85 mg/100 ml?

$$\begin{aligned} P(45 \text{ mg/100 ml} < X_i < 85 \text{ mg/100 ml}) &= P(-0.80 < Z < 0.80) \\ &= 1.0000 - P(Z < -0.80 \\ &\quad \text{or } Z > 0.80) \\ &= 1.0000 - (0.2119 + 0.2119) \\ &= 1.0000 - 0.4238 \\ &= 0.5762 \end{aligned}$$

Using the preceding considerations of the table of normal deviates (Table B.2), we can obtain the following information for measurements in a normal population:

The interval of $\mu \pm \sigma$ will contain 68.27% of the measurements.*
The interval of $\mu \pm 2\sigma$ will contain 95.44% of the measurements.
The interval of $\mu \pm 2.5\sigma$ will contain 98.76% of the measurements.
The interval of $\mu \pm 3\sigma$ will contain 99.73% of the measurements.
50% of the measurements lie within $\mu \pm 0.67\sigma$.
95% of the measurements lie within $\mu \pm 1.96\sigma$.
97.5% of the measurements lie within $\mu \pm 2.24\sigma$.

*The symbol "±" indicates "plus or minus" and was first published by William Oughtred in 1631 (Cajori, 1928: 245).

99% of the measurements lie within $\mu \pm 2.58\sigma$.
99.5% of the measurements lie within $\mu \pm 2.81\sigma$.
99.9% of the measurements lie within $\mu \pm 3.29\sigma$.

6.2 THE DISTRIBUTION OF MEANS

If random samples of size n are drawn from a normal population, the means of these samples will conform to normal distribution. The distribution of means from a nonnormal population will not be normal but will tend to approximate a normal distribution as n increases in size.* Furthermore, the variance of the distribution of means will decrease as n increases; in fact, the variance of the population of all possible means of samples of size n from a population with variance σ^2 is

$$\sigma_{\overline{X}}^2 = \frac{\sigma^2}{n}. \tag{6.4}$$

The quantity $\sigma_{\overline{X}}^2$ is called the *variance of the mean*. A distribution of sample statistics is called a *sampling distribution*†; therefore, we are discussing the sampling distribution of means.

Since $\sigma_{\overline{X}}^2$ has square units, its square root, $\sigma_{\overline{X}}$, will have the same units as the original measurements (and, therefore, the same units as the mean, μ, and the standard deviation, σ). This value, $\sigma_{\overline{X}}$, is the *standard deviation of the mean*. The standard deviation of a statistic is referred to as a *standard error*; thus, $\sigma_{\overline{X}}$ is frequently called the *standard error of the mean* (sometimes abbreviated SEM), or simply the *standard error* (sometimes abbreviated SE)‡:

$$\sigma_{\overline{X}} = \sqrt{\frac{\sigma^2}{n}} \quad \text{or} \quad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}. \tag{6.5}$$

Just as $Z = (X_i - \mu)/\sigma$ (Equation 6.3) is a normal deviate that refers to the normal distribution of X_i values,

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} \tag{6.6}$$

is a normal deviate referring to the normal distribution of means ($\overline{X}$ values). Thus, we can ask questions such as: What is the probability of obtaining a random sample of nine measurements with a mean larger than 50.0 cm from a population having a mean of 47.0 cm and a standard deviation of 12.0 cm? This and other examples of the use of normal deviates for the sampling distribution of means are presented in Example 6.2.

As seen from Equation 6.5, to determine $\sigma_{\overline{X}}$ one must know σ^2 (or σ), which is a population parameter. Because we very seldom can calculate population parameters, we must rely on estimating them from random samples taken from the population. The best estimate of $\sigma_{\overline{X}}^2$, the population variance of the mean, is

$$s_{\overline{X}}^2 = \frac{s^2}{n}, \tag{6.7}$$

*This result is known as the *central limit theorem*.

†This term was apparently first used by Ronald Aylmer Fisher in 1922 (Miller, 2004a).

‡This relationship between the standard deviation of the mean and the standard deviation was published by Karl Friedrich Gauss in 1809 (Walker, 1929: 23). The term *standard error* was introduced in 1897 by G. U. Yule (David, 1995), though in a different context (Miller, 2004a).

EXAMPLE 6.2 Proportions of a Sampling Distribution of Means

1. A population of one-year-old children's chest circumferences has $\mu = 47.0$ cm and $\sigma = 12.0$ cm, what is the probability of drawing from it a random sample of nine measurements that has a mean larger than 50.0 cm?

$$\sigma_{\overline{X}} = \frac{12.0 \text{ cm}}{\sqrt{9}} = 4.0 \text{ cm}$$

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = \frac{50.0 \text{ cm} - 47.0 \text{ cm}}{4.0 \text{ cm}} = 0.75$$

$$P(\overline{X} > 50.0 \text{ cm}) = P(Z > 0.75) = 0.2266$$

2. What is the probability of drawing a sample of 25 measurements from the preceding population and finding that the mean of this sample is less than 40.0 cm?

$$\sigma_{\overline{X}} = \frac{12.0 \text{ cm}}{\sqrt{25}} = 2.4 \text{ cm}$$

$$Z = \frac{40.0 \text{ cm} - 47.0 \text{ cm}}{2.4 \text{ cm}} = -2.92$$

$$P(\overline{X} < 40.0 \text{ cm}) = P(Z < -2.92) = P(Z > 2.92) = 0.0018$$

3. If 500 random samples of size 25 are taken from the preceding population, how many of them would have means larger than 50.0 cm?

$$\sigma_{\overline{X}} = \frac{12.0 \text{ cm}}{\sqrt{25}} = 2.4 \text{ cm}$$

$$Z = \frac{50.0 \text{ cm} - 47.0 \text{ cm}}{2.4 \text{ g}} = 1.25$$

$$P(\overline{X} > 50.0 \text{ cm}) = P(Z > 1.25) = 0.1056$$

Therefore, $(0.1056)(500) = 53$ samples would be expected to have means larger than 50.0 cm.

the sample variance of the mean. Thus,

$$s_{\overline{X}} = \sqrt{\frac{s^2}{n}} \text{ or } s_{\overline{X}} = \frac{s}{\sqrt{n}} \tag{6.8}$$

is an estimate of $\sigma_{\overline{X}}$ and is the sample standard error of the mean. Example 6.3 demonstrates the calculation of $s_{\overline{X}}$.

The importance of the standard error in hypothesis testing and related procedures will be evident in Chapter 7. At this point, however, it can be noted that the magnitude of $s_{\overline{X}}$ is helpful in determining the precision to which the mean and some measures of variability may be reported. Although different practices have been followed by many, we shall employ the following (Eisenhart, 1968). We shall state the standard error to two significant figures (e.g., 2.7 mm in Example 6.3; see Section 1.2 for an explanation

of significant figures). Then the standard deviation and the mean will be reported with the same number of decimal places (e.g., $\overline{X}$ = 137.6 mm in Example 6.3*). The variance may be reported with twice the number of decimal places as the standard deviation.

EXAMPLE 6.3 The Calculation of the Standard Error of the Mean, $s_{\overline{X}}$

The Following are Data for Systolic Blood Pressures, in mm of Mercury, of 12 Chimpanzees.

121
125
128
134
136
138
139
141
144
145
149
151

$\sum X = 1651$ mm

$\sum X^2 = 228{,}111$ mm^2

$n = 12$

$$\overline{X} = \frac{1651 \text{ mm}}{12} = 137.6 \text{ mm}$$

$$SS = 228{,}111 \text{ mm}^2 - \frac{(1651 \text{ mm})^2}{12}$$

$$= 960.9167 \text{ mm}^2$$

$$s^2 = \frac{960.9167 \text{ mm}^2}{11} = 87.3561 \text{ mm}^2$$

$$s = \sqrt{87.3561 \text{ mm}^2} = 9.35 \text{ mm}$$

$$s_{\overline{X}} = \frac{s}{\sqrt{n}} = \frac{9.35 \text{ mm}}{\sqrt{12}} = 2.7 \text{ mm } \textit{or}$$

$$s_{\overline{X}} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{87.3561 \text{ mm}^2}{12}} = \sqrt{7.2797 \text{ mm}^2} = 2.7 \text{ mm}$$

6.3 INTRODUCTION TO STATISTICAL HYPOTHESIS TESTING

A major goal of statistical analysis is to draw inferences about a population by examining a sample from that population. A very common example of this is the desire to draw conclusions about one or more population means.

We begin by making a concise statement about the population mean, a statement called a *null hypothesis* (abbreviated H_0)† because it expresses the concept of "no difference." For example, a null hypothesis about a population mean (μ) might assert that μ is not different from zero (i.e., μ is equal to zero); and this would be written as

$$H_0: \mu = 0.$$

Or, we could hypothesize that the population mean is not different from (i.e., is equal to) 3.5 cm, or not different from 10.5 kg, in which case we would write H_0: μ = 3.5 cm or H_0: μ = 10.5 kg, respectively.

*In Example 6.3, s is written with more decimal places than the Eisenhart recommendations indicate because it is an intermediate, rather than a final, result; and rounding off intermediate computations may lead to serious rounding error. Indeed, some authors routinely report extra decimal places, even in final results, with the consideration that readers of the results may use them as intermediates in additional calculations.

†The term *null hypothesis* was first published by R. A. Fisher in 1935 (David, 1995; Miller, 2004a; Pearson, 1947). J. Neyman and E. S. Pearson were the first to use the symbol "H_0" and the term *alternate hypothesis*, in 1928 (Pearson, 1947; Miller, 2004a, 2004c). The concept of statistical testing of something akin to a null hypothesis was introduced 300 years ago by John Arbuthnot (1667–1725), a Scottish–English physician and mathematician (Stigler, 1986: 225–226).

If statistical analysis concludes that it is likely that a null hypothesis is false, then an *alternate hypothesis* (abbreviated H_A or H_1) is assumed to be true (at least tentatively). One states a null hypothesis and an alternate hypothesis for each statistical test performed, and all possible outcomes are accounted for by this pair of hypotheses. So, for the preceding examples.*

$$H_0: \mu = 0, \quad H_A: \mu \neq 0;$$
$$H_0: \mu = 3.5 \text{ cm}, \quad H_A: \mu \neq 3.5 \text{ cm};$$
$$H_0: \mu = 10.5 \text{ kg}, \quad H_A: \mu \neq 10.5 \text{ kg}.$$

It must be emphasized that statistical hypotheses are to be stated *before* data are collected to test them. To propose hypotheses after examination of data can invalidate a statistical test. One may, however, legitimately formulate hypotheses *after* inspecting data if a new set of data is then collected with which to test the hypotheses.

(a) Statistical Testing and Probability. Statistical testing of a null hypothesis about μ, the mean of a population, involves calculating $\overline{X}$, the mean of a random sample from that population. As noted in Section 2.1, $\overline{X}$ is the best estimate of μ; but it is only an estimate, and we can ask, What is the probability of an $\overline{X}$ at least as far from the hypothesized μ as is the $\overline{X}$ in the sample, *if* H_0 *is true*? Another way of visualizing this is to consider that, instead of obtaining one sample (of size n) from the population, a large number of samples (each sample of size n) could have been taken from that population. We can ask what proportion of those samples would have had means at least as far as our single sample's mean from the μ specified in the null hypothesis. This question is answered by the considerations of Section 6.2 and is demonstrated in Example 6.4.

EXAMPLE 6.4 Hypothesis Testing of H_0: $\mu = 0$ and H_A: $\mu \neq 0$

The variable, X_i, is the weight change of horses given an antibiotic for two weeks. The following measurements of X_i are those obtained from 17 horses (where a positive weight change signifies a weight gain and a negative weight change denotes a weight loss):

2.0, 1.1, 4.4, −3.1, −1.3, 3.9, 3.2 , −1.6 ,3.5
1.2, 2.5, 2.3, 1.9, 1.8, 2.9, −0.3, and −2.4 kg.

For these 17 data, the sample mean ($\overline{X}$) is 1.29 kg. Although the population variance (σ^2) is typically not known, for the demonstration purpose of this example, σ^2 is said to be 13.4621 kg^2. Then the population standard error of the mean would be

$$\sigma_{\overline{X}} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{13.4621 \text{ kg}^2}{17}} = \sqrt{0.7919 \text{ kg}^2} = 0.89 \text{ kg}$$

*The symbol "≠" denotes "is not equal to"; Ball (1935: 242) credits Leonhard Euler with its early, if not first, use (though it was first written with a vertical, not a diagonal, line through the equal sign).

and

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = \frac{1.29 \text{ kg} - 0}{0.89 \text{ kg}} = 1.45.$$

Using Table B.2,

$$P(\overline{X} \geq 1.29 \text{ kg}) = P(Z \geq 1.45) = 0.0735$$

and, because the distribution of Z is symmetrical,

$$P(\overline{X} \leq -1.29 \text{ kg}) = P(Z \leq -1.45) = 0.0735.$$

Therefore,

$$\begin{aligned} &P(\overline{X} \geq 1.29 \text{ kg } \textit{or } \overline{X} \leq -1.29 \text{ kg}) \\ &= P(Z \geq 1.45 \textit{ or } Z \leq -1.45) \\ &= 0.0735 + 0.0735 = 0.1470. \end{aligned}$$

As $0.1470 > 0.05$, do not reject H_0.

In Example 6.4, it is desired to ask whether treating horses with an experimental antibiotic results in a change in body weight. The data shown (X_i values) are the changes in body weight of 17 horses that received the antibiotic, and the statistical hypotheses to be tested are H_0: $\mu = 0$ kg and H_A: $\mu \neq 0$ kg. (As shown in this example, we can write "0" instead of "0 kg" in these hypotheses, because they are statements about *zero* weight change, and *zero* would have the same meaning regardless of whether the horses were weighed in kilograms, milligrams, pounds, ounces, etc.)

These 17 data have a mean of $\overline{X} = 1.29$ kg and they are considered to represent a random sample from a very large number of data, namely the body-weight changes that would result from performing this experiment with a very large number of horses. This large number of potential X_i's is the statistical population. Although one almost never knows the actual parameters of a sampled population, for this introduction to statistical testing let us suppose that the variance of the population sampled for this example is known to be $\sigma^2 = 13.4621 \text{ kg}^2$. Thus, for the population of means that could be drawn from this population of measurements, the standard error of the mean is $\sigma_{\overline{X}} = \sqrt{\sigma^2/n} = \sqrt{13.4621 \text{ kg}^2/17} = \sqrt{0.7919 \text{ kg}^2} = 0.89$ kg (by Equation 6.5). We shall further assume that the population of possible means follows a normal distribution, which is generally a reasonable assumption even when the individual data in the population are not normally distributed.

This hypothesis test may be conceived as asking the following:

> If we have a normal population with $\mu = 0$ kg, and $\sigma_{\overline{X}} = 0.89$ kg, what is the probability of obtaining a random sample of 17 data with a mean ($\overline{X}$) at least as far from 0 kg as 1.29 kg (i.e., at least 1.29 kg larger than 0 kg *or* at least 1.29 kg smaller than 0 kg)?

Section 6.2 showed that probabilities for a distribution of possible means may be ascertained through computations of Z (by Equation 6.6). The preceding null hypothesis is tested in Example 6.4, in which Z may be referred to as our *test statistic* (a computed quantity for which a probability will be determined). In this example, Z is calculated to be 1.45, and Appendix Table B.2 informs us that the probability of a

$Z \geq 1.45$ is 0.0735.* The null hypothesis asks about the deviation of the mean *in either direction* from 0 and, as the normal distribution is symmetrical, we can also say that $P(-Z \leq 1.45) = 0.0735$ and, therefore, $P(|Z| \geq 1.45) = 0.0735 + 0.0735 = 0.1470$. This tells us the probability associated with a $|Z|$ (absolute value of Z) at least as large as the $|Z|$ obtained; and this is the probability of a Z at least as extreme as that obtained, *if* the null hypothesis is true.

It should be noted that this probability,

$$P(|Z| \geq |\text{computed } Z|, \text{ if } H_0 \text{ is true}),$$

is *not* the same as

$$P(H_0 \text{ is true, if } |Z| \geq |\text{computed } Z|),$$

for these are *conditional probabilities*, discussed in Section 5.8. In addition to the playing-card example in that section, suppose a null hypothesis was tested 2500 times, with results as in Example 6.5. By Equation 5.23, the probability of rejecting H_0, *if* H_0 is true, is P(rejecting H_0, if H_0 is true) = (number of rejections of true H_0's)/(number of true H_0's) = 100/2000 = 0.05. And the probability that H_0 is true, *if* H_0 is rejected, is $P(H_0$ true, if H_0 is rejected) = (number of rejections of true H_0's)/(number of rejections of H_0's) = 100/550 = 0.18. These two probabilities (0.05 and 0.18) are decidedly not the same, for they are probabilities based on different conditions.

EXAMPLE 6.5 Probability of Rejecting a True Null Hypothesis

Hypothetical outcomes of testing the same null hypothesis for 2500 random samples of the same size from the same population (where the samples are taken with replacement).

	If H_0 is true	If H_0 is false	Row total
If H_0 is rejected	100	450	550
If H_0 is not rejected	1900	50	1950
Column total	2000	500	2500

Probability that H_0 is rejected if H_0 is true = 100/2000 = 0.05.
Probability that H_0 is true if H_0 is rejected = 100/550 = 0.18.

In hypothesis testing, it is correct to say that the calculated probability (for example, using Z) is

$$P(\text{the data, given } H_0 \text{ is true})$$

and it is *not* correct to say that the calculated probability is

$$P(H_0 \text{ is true, given the data}).$$

Furthermore, in reality we may not be testing H_0: $\mu = 0$ kg in order to conclude that the population mean is *exactly* zero (which it probably is *not*). Rather, we

*Note that "≥" and "≤" are symbols for "greater than or equal to" and "less than or equal to," respectively.

are interested in concluding whether there is a very small difference between the population mean and 0 kg; and what is meant by *very small* will be discussed in Section 6.3(d).

(b) Statistical Errors in Hypothesis Testing. It is desirable to have an objective criterion for drawing a conclusion about the null hypothesis in a statistical test. Even if H_0 is true, random sampling might yield a sample mean ($\overline{X}$) far from the population mean (μ), and a large absolute value of Z would thereby be computed. However, such an occurrence is unlikely, and the larger the $|Z|$, the smaller the probability that the sample came from a population described by H_0. Therefore, we can ask how small a probability (which is the same as asking how large a $|Z|$) will be required to conclude that the null hypothesis is not likely to be true. The probability used as the criterion for rejection of H_0 is called the *significance level*, routinely denoted by α (the lowercase Greek letter alpha).* As indicated below, an α of 0.05 is commonly employed. The value of the test statistic (in this case, Z) corresponding to α is termed the *critical value* of the test statistic. In Appendix Table B.2 it is seen that $P(Z \geq 1.96) = 0.025$; and, inasmuch as the normal distribution is symmetrical, it is also the case that $P(Z \leq -1.96) = 0.025$. Therefore, the critical value for testing the above H_0 at the 0.05 level (i.e., 5% level) of significance is $Z = 1.96$ (see Figure 6.4). These values of Z may be denoted as $Z_{0.025(1)} = 1.96$ and $Z_{0.05(2)} = 1.96$, where the parenthetical number indicates whether one or two tails of the normal distribution are being referred to.

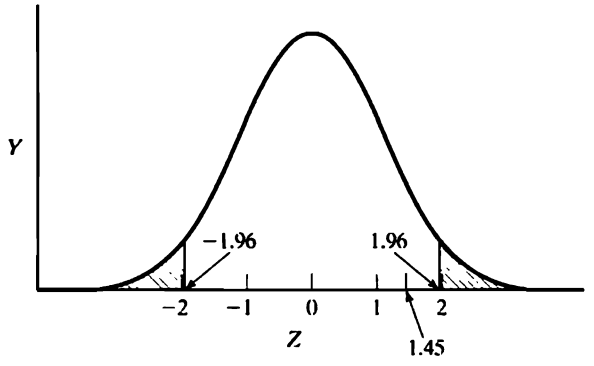


FIGURE 6.4: A normal curve showing (with shading) the 5% of the area under the curve that is the rejection region for the null hypothesis of Example 6.4. This rejection region consists of 2.5% of the curve in the right tail (demarcated by $Z_{0.05(2)} = 1.96$) and 2.5% in the left tail (delineated by $-Z_{0.05(2)} = -1.96$). The calculated test statistic in this example, $Z = 1.45$, does not lie within either tail; so H_0 is not rejected.

So, a calculated Z greater than or equal to 1.96, or less than or equal to -1.96, would be reason to reject H_0, and the shaded portion of Figure 6.4 is known as the "rejection region." The absolute value of the test statistic in Example 6.4 (namely, $|Z| = 1.45$) is not as large as the critical value (i.e., it is neither ≥ 1.9 nor ≤ -1.96), so in this example the null hypothesis is not rejected as a statement about the sampled population.

*David (1955) credits R. A. Fisher as the first to refer to "level of significance," in 1925. Fisher (1925b) also was the first to formally recommend use of the 5% significance level as guidance for drawing a conclusion about the propriety of a null hypothesis (Cowles and Davis, 1982), although he later argued that a fixed significance level should not be used. This use of the Greek "α" first appears in a 1936 publication of J. Neyman and E. S. Pearson (Miller, 2004c).

It is very important to realize that a true null hypothesis will sometimes be rejected, which of course means that an error has been committed in drawing a conclusion about the sampled population. Moreover, this error can be expected to be committed with a frequency of α. The rejection of a null hypothesis when it is in fact true is what is known as a *Type I error* (or "Type 1 error" or "alpha error" or "error of the first kind"). On the other hand, a statistical test will sometimes fail to detect that a H_0 is in fact false, and an erroneous conclusion will be reached by not rejecting H_0. The probability of committing this kind of error (that is, not rejecting H_0 when it is false) is represented by β (the lowercase Greek letter beta). This error is referred to as a *Type II error* (or "Type 2 error" or "beta error" or "error of the second kind"). The *power* of a statistical test is defined as $1 - \beta$: the probability of correctly rejecting the null hypothesis when it is false.* If H_0 is not rejected, some researchers refer to it as having been "accepted," but most consider it better to say "not rejected," for low statistical power often causes failure to reject, and "accept" sounds too definitive. Section 6.3(c) discusses how both the Type I and the Type II errors can be reduced.

Table 6.1 summarizes these two types of statistical errors, and Table 6.2 indicates their probabilities. Because, for a given n, a relatively small probability of a Type I error is associated with a relatively large probability of a Type II error, it is appropriate to ask what the acceptable combination of the two might be. By experience, and by convention, an α of 0.05 is typically considered to be a "small enough chance" of committing a Type I error while not being so small as to result in "too large a chance"of a Type II error (sometimes considered to be around 20%). But the 0.05 level of significance is not sacrosanct. It is an arbitrary, albeit customary, threshold for concluding that there is significant evidence against a null hypothesis. And caution should be exercised in emphatically rejecting a null hypothesis if $p = 0.049$ and not rejecting if $p = 0.051$, for in such borderline cases further examination—and perhaps repetition—of the experiment would be recommended.

TABLE 6.1: The Two Types of Errors in Hypothesis Testing

	If H_0 is true	If H_0 is false
If H_0 is rejected:	Type I error	No error
If H_0 is not rejected:	No error	Type II error

Although 0.05 has been the most widely used significance level, individual researchers may decide whether it is more important to keep one type of error

*The distinction between these two fundamental kinds of statistical errors, and the concept of power, date back to the pioneering work, in England, of Jerzy Neyman (1894–1981; Russian-born, of Polish roots, emigrating as an adult to Poland and then to England, and spending the last half of his life in the United States) and the English statistician Egon S. Pearson (1895–1980) (Lehmann and Reid, 1982; Neyman and Pearson, 1928a; Pearson, 1947). They conceived of the two kinds of errors in 1928 (Lehmann, 1999) and named them, and they formulated the concept of power in 1933 (David, 1995). With some influence by W. S. Gosset ("Student") (Lehmann, 1999), their modifications (e.g., Neyman and Pearson, 1933) of the ideas of the colossal British statistician (1890–1962) R. A. Fisher (1925b) provide the foundations of statistical hypothesis testing. However, from the mid-1930s until his death, Fisher disagreed intensely with the Neyman-Pearson approach, and the hypothesis testing commonly used today is a fusion of the Fisher and the Neyman-Pearson procedures (although this hybridization of philosophies has received criticism—e.g., by Hubbard and Bayarri, 2003). Over the years there has been further controversy regarding hypothesis testing, especially—but not entirely—within the social sciences (e.g., Harlow, Mulaik, and Steiger, 1997). The most extreme critics conclude that hypothesis tests should never be used, while most others advise that they may be employed but only with care to avoid abuse.

TABLE 6.2: The Long-Term Probabilities of Outcomes in Hypothesis Testing

	If H_0 is true	If H_0 is false
If H_0 is rejected	α	$1 - \beta$ ("power")
If H_0 is not rejected	$1 - \alpha$	β

or the other low. In some instances, we may be willing to test with an α greater than 0.05. An example of the later decision could be when there is an adverse health or safety implication if we incorrectly fail to reject a false null hypothesis. So in performing an experiment such as in Example 6.4, perhaps it is deemed important to the continued use of this antibiotic that it not cause a change in body weight; and we want to have a small chance of concluding that the drug causes no weight change when such a decision is incorrect. In other words, we may be especially desirous of avoiding a Type II error. In that case, an α of 0.10 (i.e., 10%) might be used, for that would decrease the probability of a Type II error, although it would concomitantly increase the likelihood of incorrectly rejecting a true H_0 (i.e., committing a Type I error). In other cases, such as indicated in Section 6.3(d), a 0.05 (i.e., 5%) chance of an incorrect rejection of H_0 may be felt to be unacceptably high, so a lower α would be employed in order to reduce the probability of a Type I error (even though that would increase the likelihood of a Type II error).

It is necessary, of course, to state the significance level used when communicating the results of a statistical test. Indeed, rather than simply stating whether the null hypothesis is rejected, it is good procedure to report also the sample size, the test statistic, and the best estimate of the exact probability of the statistic (and such probabilities are obtainable from many computer programs and some calculators, and may be estimated from tables such as those in Appendix B). Note that in Example 6.4, it is reported that $n = 17$, $Z = 1.45$, and $P = 0.1470$, in addition to expressing the conclusion that H_0 is not rejected. In this way, readers of the research results may draw their own conclusions, even if their choice of significance level is different from the author's. It is also good practice to report results regardless of whether H_0 is rejected. Bear in mind, however, that the choice of α is to be made before seeing the data. Otherwise there is a great risk of having the choice influenced by examination of the data, introducing bias instead of objectivity into the proceedings. The best practice generally is to decide on the null and alternate hypotheses, and the significance level, before commencing with data collection and, after performing the statistical test, to express the probability that the sample came from a population for which H_0 is true. It is conventional to refer to rejection of H_0 at the 5% significance level as denoting a "statistically significant" difference between $\overline{X}$ and the μ hypothesized in H_0 (e.g., in Example 6.4, between $\overline{X} = 1.45$ kg and $\mu = 0$ kg).* But, in analyzing biological data, we should consider whether a statistically detected difference reflects a *biologically significant* difference, as will be discussed in Section 6.3(d).

(c) One-Tailed versus Two-Tailed Testing. In Section 6.3(a), Example 6.4 tests whether a population mean was significantly different from a hypothesized value, where the alternate hypothesis embodies difference in either direction (i.e., greater than or less than) from that value. This is known as *two-sided*, or *two-tailed*, testing.

*In reporting research results, some authors have attached an asterisk (*) to a test statistic if it is associated with a probability ≤ 0.05 and two asterisks (**) if the probability is ≤ 0.01, sometimes referring to results at ≤ 0.01 as "highly significant"; but the latter term is best avoided, in preference to reporting the magnitude of p.

for we reject H_0 if Z (the test statistic in this instance) is within either of the two tails of the normal distribution demarcated by the positive and negative critical values of Z (the shaded areas in Figure 6.4).

However, there are cases where there is good scientific justification to test for a significant difference *specifically in one direction only*. That is, on occasion there is a good reason to ask whether a population mean is significantly *larger* than μ_0, and in other situations there is a good rationale for asking whether a population mean is significantly *smaller* than μ_0. Statistical testing that examines difference in only one of the two possible directions is called *one-sided*, or *one-tailed*, testing.

Example 6.4 involved a hypothesis test interested in whether a drug intended to be an antibiotic caused weight change as a side effect of its use. For such a test, H_0 is rejected if Z (the test statistic in this instance) is within the rejection region in either the right-hand *or* the left-hand tail of the normal distribution (i.e., within the shaded areas of Figure 6.4 and Figure 6.5a). However, consider a similar experiment where the purpose of the drug is to cause weight loss. In that case, the statistical hypotheses would be H_0: $\mu \geq 0$ versus H_A: $\mu < 0$. That is, if the drug works as intended and there

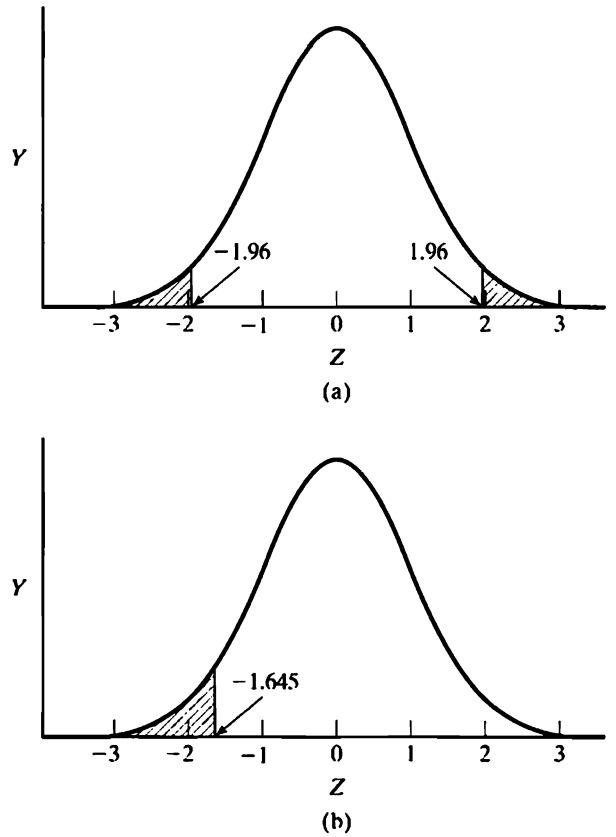


FIGURE 6.5: (a) As in Figure 6.4, a normal curve showing (with shading) the 5% of the area under the curve that is the rejection region for the two-tailed null hypotheses, H_0: $\mu = \mu_0$ versus H_A: $\mu \neq \mu_0$. This rejection region consists of 2.5% of the curve in the right tail (demarcated by $Z_{0.05(2)} = 1.96$), and 2.5% in the left tail (delineated by $-Z_{0.05(2)} = -1.96$). (b) A normal curve showing (with shading) the 5% of the area under the curve that is the rejection region for the one-tailed null hypotheses, H_0: $\mu \geq \mu_0$ vs. H_A: $\mu < \mu_0$. This rejection region consists of 5% of the curve in the left tail (demarcated by $Z_{0.05(1)} = 1.645$).

is a mean weight loss, then H_0 would be rejected; and if the drug does not work (that is, there is a mean weight gain or the mean weight did not change), H_0 would not be rejected. In such a situation, the rejection region would be entirely in one tail of the normal distribution, namely the left-hand tail. This is an example of a *one-tailed test*, whereas Example 6.4 represents a *two-tailed test.*

It can be seen in Appendix B.2 that, if one employs the 5% level of significance, the one-tailed Z value is 1.645. The normal distribution's tail defined by this one-tailed Z is the shaded area of Figure 6.5b. If the calculated Z is within this tail, H_0 is rejected as a correct statement about the population from which this sample came.

Figure 6.5a shows the rejection region of a normal distribution when performing two-tailed testing of H_0: $\mu = \mu_0$ at the 5% significance level (i.e., the same shaded area as in Figure 6.4, namely 2.5% in each tail of the curve); and Figure 6.5b shows the rejection region for one-tailed testing of H_0: $\mu \geq 0$ versus H_A: $\mu < 0$ at the 5% level. (If the experimental drug were intended to result in weight gain, not weight loss, then the rejection region would be in the right-hand tail instead of in the left-hand tail.)

In general, one-tailed hypotheses about a mean are

$$H_0: \mu \geq \mu_0 \text{ and } H_A: \mu < \mu_0,$$

in which case H_0 is rejected if the test statistic is in the left-hand tail of the distribution, *or*

$$H_0: \mu \leq \mu_0 \text{ and } H_A: \mu > \mu_0,$$

in which case H_0 is rejected if the test statistic is in the right-hand tail of the distribution.*

The one-tailed critical value (let's call it $Z_{\alpha(1)}$) is found in Appendix Table B.2. It is always smaller than the two-tailed critical value ($Z_{\alpha(2)}$); for example, at the 5% significance level $Z_{\alpha(1)} = 1.645$ and $Z_{\alpha(2)} = 1.96$. Thus, as will be noted in Section 6.3(d), for a given set of data a one-tailed test is more powerful than a two-tailed test. But it is inappropriate to employ a one-tailed test unless there is a scientific reason for expressing one-tailed, in preference to two-tailed, hypotheses. And recall that *statistical hypotheses are to be declared before examining the data.* Another example of one-tailed testing of a mean is found in Exercise 6.5(a).

(d) What Affects Statistical Power. The power of a statistical testing procedure was defined in Section 6.3(b) as the probability that a test correctly rejects the null hypothesis when that hypothesis is a false statement about the sampled population. It is useful to be aware of what affects the power of a test, and later chapters will show how to estimate the power a test will have and to estimate how small a difference will be detected between a population parameter (e.g., μ) and a hypothesized value (e.g., μ_0).

Figure 6.6a represents a normal distribution of sample means, where each sample was the same size and each sample mean estimates the same population mean. This mean of this distribution is μ_0, the population mean specified in the null hypothesis. This curve is the same as shown in Figure 6.5. As in Figure 6.5, the shaded area in each of the two tails denotes 0.025 of the area under the curve; so both shaded areas compose an area of 0.05, the probability of a Type I error (α).

*Some authors write the first of these two pairs of hypotheses as H_0: $\mu = \mu_0$ and H_A: $\mu < \mu_0$, and the second pair as H_0: $\mu = \mu_0$ and H_A: $\mu > \mu_0$, ignoring mention of the tail that is not of interest.

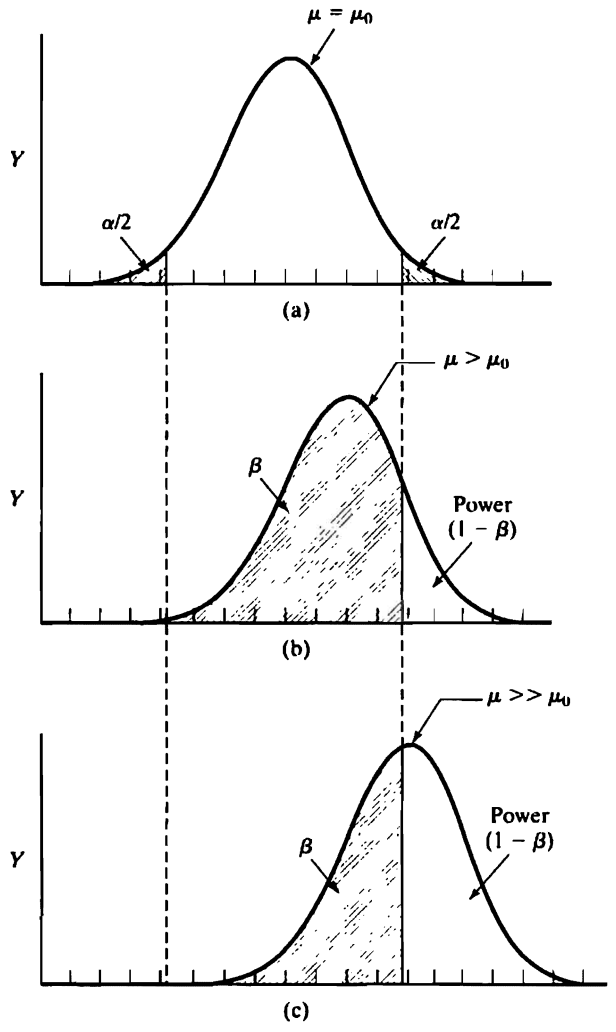


FIGURE 6.6: (a) A normal curve, such as that in Figure 6.4, where μ, the mean of the distribution, is μ_0, the value specified in the null and alternate hypotheses. The shaded area in each of the two tails is 0.025 of the area under the curve, so a total of 0.05 (i.e., 5%) of the curve is the shaded critical region, and α, the probability of a Type I error, is 0.05. (b) The same normal curve, but where μ is larger than μ_0 and the shaded area is the probability of a Type II error (β). (c) The same normal curve, but where μ is much larger than μ_0.

Figure 6.6b is the same normal curve, but with a population mean, μ, different from (i.e., larger than) μ_0. If H_0: $\mu = \mu_0$ is not a true statement about the population, yet we fail to reject H_0, then we have committed a Type II error, the probability of which is β, indicated by the shaded area between the vertical dashed lines in Figure 6.6b. The power of the hypothesis test is defined as $1 - \beta$, which is the unshaded area under this curve.

Figure 6.6c is the same depiction as in Figure 6.6b, but with a population mean, μ, even more different* from μ_0. An important result is that, the farther μ is from the μ_0 specified in H_0, the smaller β becomes and the larger the power becomes.

*The symbol ">" has been introduced as meaning "greater than," and "<" as meaning "less than." The symbols "≫" and "≪" mean "much greater than" and "much less than," respectively.

Figure 6.7 indicates the outcome if a larger α is used, namely 10% instead of 5% (meaning that 5%, instead of 2.5%, of the curve is in each tail). If the probability of a Type I error (α) is increased, then the probability of a Type II error (β) is decreased, and the power of the test is increased.

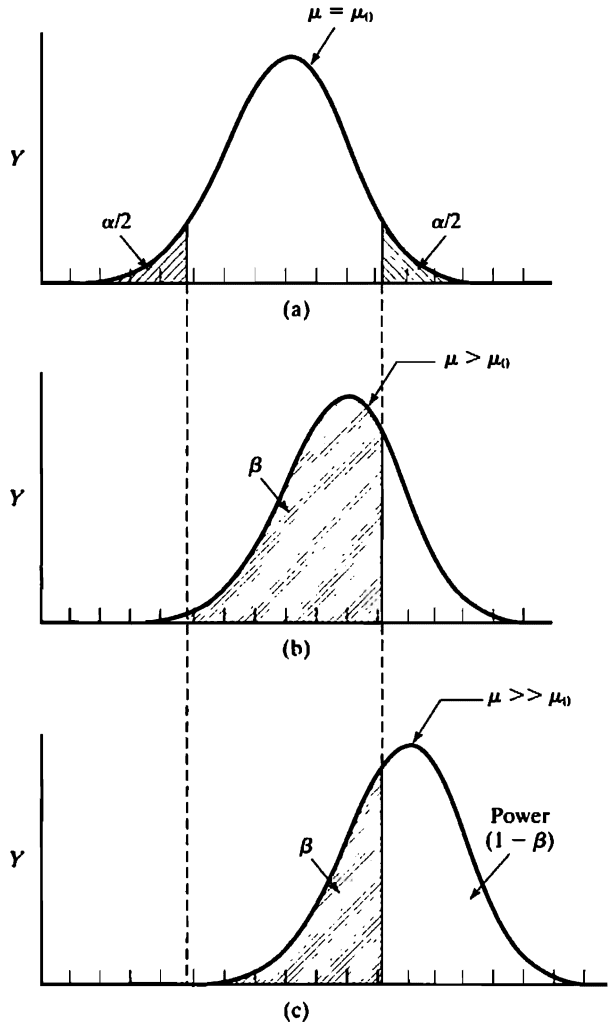


FIGURE 6.7: (a) A normal curve, such as that in Figure 6.6, where μ, the mean of the distribution, is μ_0, the value specified in the null and alternate hypotheses, but where the shaded area in each of the two tails is 0.05 of the area under the curve, so a total of 0.10 (i.e., 10%) of the curve is the shaded critical region, and α, the probability of a Type I error, is 0.10. (b) The same normal curve, but where μ is larger than μ_0 and the shaded area is the probability of a Type II error (β). (c) The same normal curve, but where μ is much larger than μ_0.

Another important outcome is seen by examining Equations 6.5 and 6.6. With larger sample size (n), or with smaller variance (σ^2), the standard error $\sigma_{\bar{X}}$ becomes smaller, which means that the shape of the normal distribution becomes narrower. Figure 6.3 shows an example of this narrowing as the variance decreases in a population of data, and the figures would appear similar if they were for a population of means. So, for a given value of α and of μ, either a smaller σ^2 or a larger n will result in a smaller $\sigma_{\bar{X}}$, which will result in a smaller β and greater power to reject H_0.

In some circumstances, a larger n can be used, but in other situations this would be difficult because of cost or effort. A smaller variance of the sampled population will result if the population is defined as a more homogeneous group of data. In Example 6.4, the experiment could have been performed using only female horses, or only horses of a specified age, or only horses of a specified breed. Then the hypothesis test would be about the specified sex, age, and/or breed, and the population variance would probably be smaller; and this would result in a greater power of the test.

To summarize what influences power,

- For given α, σ^2, and n, power is greater for larger difference between μ and μ_0.
- For given n, σ^2, and difference between μ and μ_0, power is greater for larger α.
- For given α, σ^2, and difference between μ and μ_0, power is greater for larger n.
- For given α, n, and difference between μ and μ_0, power is greater for smaller σ^2.
- For given α, n, σ^2, and difference between μ and μ_0, power is greater for one-tailed than for two-tailed tests (but one-tailed tests may be employed only when the hypotheses are appropriately one-tailed).

(e) Summary of Statistical Hypothesis Testing. Earlier portions of Section 6.3 introduced the principles and practice of testing hypotheses about population parameters, using sample statistics as estimates of those parameters. It is also good practice to report an estimate of the precision with which a parameter has been estimated, by expressing what are known as "confidence limits," which will be introduced in Section 6.4.

To summarize the steps for testing of statistical hypotheses,

1. State H_0 and H_A, using two-tailed or one-tailed hypotheses depending upon the objective of the data analysis.
2. Declare the level of significance, α, to be employed.
3. Collect the data and calculate the test statistic (Z in this chapter).
4. Compare the test statistic to the critical value(s) of that statistic (that is, the value(s) delimiting the rejection region of the statistical distribution of the test statistic). For the testing in this chapter, the critical values are both $Z_{\alpha(2)}$ and $-Z_{\alpha(2)}$ for a two-tailed test and the critical value is $Z_{\alpha(1)}$ for a one-tailed test. If the calculated Z exceeds a critical value, H_0 is rejected.
5. State P, the probability of the test statistic if H_0 is true.
6. State confidence limits (two-tailed or one-tailed) for the population parameter, as discussed in Section 6.4.
7. State conclusion in terms of biological or other practical significance.

6.4 CONFIDENCE LIMITS

Sections 6.3a and 6.3b discussed the distribution of all possible samples of size n from a population with mean μ. It was noted that 5% of the values of Z (by Equation 6.6) for those sample means will be at least as large as $Z_{0.05(2)}$ *or* no larger than $-Z_{0.05(2)}$. This can be expressed as

$$P\left[-Z_{0.05(2)} \leq \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} \leq Z_{0.05(2)}\right] = 95\%, \tag{6.9}$$

and this can be rearranged to read

$$P[\overline{X} - Z_{0.05(2)}\sigma_{\overline{X}} \leq \mu \leq \overline{X} + Z_{0.05(2)}\sigma_{\overline{X}}] = 0.95. \tag{6.10}$$

In general, we can say

$$P[\overline{X} - Z_{\alpha(2)}\sigma_{\overline{X}} \leq \mu \leq \overline{X} + Z_{\alpha(2)}\sigma_{\overline{X}}] = 1 - \alpha. \tag{6.11}$$

The *lower confidence limit* is defined as

$$L_1 = \overline{X} - Z_{\alpha(2)}\sigma_{\overline{X}}, \tag{6.12}$$

and the *upper confidence limit* is

$$L_2 = \overline{X} + Z_{\alpha(2)}\sigma_{\overline{X}}. \tag{6.13}$$

The distance between L_1 and L_2, namely

$$\overline{X} \pm Z_{\alpha(2)}\sigma_{\overline{X}} \tag{6.14}$$

(where "±" is read as "plus or minus"), is called a *confidence interval* (sometimes abbreviated CI).

When referring to a confidence interval, $1 - \alpha$ is known as the *confidence level* (or *confidence coefficient* or *confidence probability*).*

Although $\overline{X}$ is the best estimate of μ, it is only an estimate, and the calculation of a confidence interval for μ allows us to express the precision of this estimate. Example 6.6 demonstrates this for the data of Example 6.4, determining the confidence interval for the mean of the population from which the sample came. As the 95% confidence limits are computed to be −0.45 kg and 3.03 kg, the 95% confidence interval may be expressed as $P(-0.45 \text{ kg} \leq \mu \leq 3.03 \text{ kg}) = 95\%$. This means that, if all possible means of size n ($n = 17$ in this example) were taken from the population and a 95% confidence interval were calculated from each sample, 95% of those intervals would contain μ. (It does *not* mean that there is a 95% probability that the confidence interval computed from the one sample in Example 6.6 includes μ.)

EXAMPLE 6.6 Confidence Limits for the Mean

For the 17 data in Example 6.4, $\overline{X} = 1.29$ kg and $\sigma_{\overline{X}} = 0.89$ kg.
We can calculate the 95% confidence limits for μ using Equations 6.13 and 6.14 and $Z_{0.05(2)} = 1.96$:

$$\begin{aligned} L_1 &= \overline{X} - Z_{\alpha(2)}\sigma_{\overline{X}} \\ &= 1.29 \text{ kg} - (1.96)(0.89 \text{ kg}) \\ &= 1.29 \text{ kg} - 1.74 \text{ kg} = -0.45 \text{ kg} \end{aligned}$$

*We owe the development of confidence intervals to Jerzy Neyman, between 1928 and 1933 (Wang, 2000), although the concept had been enunciated a hundred years before. Neyman introduced the terms *confidence interval* and *confidence coefficient* in 1934 (David, 1995). On rare occasion, biologists may see reference to "fiducial intervals," a concept developed by R. A. Fisher beginning in 1930 and identical to confidence intervals in many, but not all, situations (Pfanzagl, 1978).

$$L_2 = \overline{X} + Z_{\alpha(2)}\sigma_{\overline{X}}$$
$$= 1.29 \text{ kg} + (1.96)(0.89 \text{ kg})$$
$$= 1.29 \text{ kg} + 1.74 \text{ kg} = 3.03 \text{ kg}.$$

So, the 95% confidence interval could be stated as

$$P(-0.45 \text{ kg} \leq \mu \leq 3.03 \text{ kg}).$$

Note that the μ_0 of Example 6.4 (namely 0) is included between L_1 and L_2, indicating that H_0 is not rejected.

As seen in Equation 6.15, a small $\sigma_{\overline{X}}$ will result in a smaller confidence interval, meaning that μ is estimated more precisely when $\sigma_{\overline{X}}$ is small. And, recall from Equation 6.5 that $\sigma_{\overline{X}}$ becomes small as n becomes large. So, in general, a parameter estimate from a large sample is more precise than an estimate of the same parameter from a small sample.

If, instead of a 95% confidence interval, we wished to state an interval that gave us 99% confidence in estimating μ, then $Z_{0.01(2)}$ (which is 2.575) would have been employed instead of $Z_{0.05(2)}$, and we would have computed $L_1 = 1.29 \text{ kg} - (2.575)(0.89 \text{ kg}) = 1.29 \text{ kg} - 2.29 = -1.00$ and $L_2 = 1.29 \text{ kg} + (2.575)(0.89 \text{ kg}) = 1.29 \text{ kg} + 2.29 \text{ kg} = 3.58 \text{ kg}$. It can be seen that a larger confidence level (e.g., 99% instead of 95%) results in a larger width of the confidence interval, evincing the trade-off between confidence and utility. Indeed, if we increase the confidence to 100%, then the confidence interval would be $-\infty$ to ∞, and we would have a statement of great confidence that was useless! Note, also, that it is a two-tailed value of Z (i.e., $Z_{0.05(2)}$) that is used in the computation of a confidence interval when we set confidence limits on both sides of μ.

In summary, a narrower confidence interval will be associated with a smaller standard error ($\sigma_{\overline{X}}$), a larger sample size (n), or a smaller confidence coefficient ($1 - \alpha$).

It is recommended that a $1 - \alpha$ confidence interval be reported for μ whenever results are presented from a hypothesis test at the α significance level. If H_0: $\mu = \mu_0$ is not rejected, then the confidence interval includes μ_0 (as is seen in Example 6.6, where $\mu_0 = 0$ is between L_1 and L_2).

(a) One-Tailed Confidence Limits. In the case of a one-tailed hypothesis test, it is appropriate to determine a one-tailed confidence interval; and, for this, a one-tailed critical value of Z (i.e., $Z_{\alpha(1)}$) is used instead of a two-tailed critical value ($Z_{\alpha(2)}$). For H_0: $\mu \leq \mu_0$ and H_A: $\mu > \mu_0$, the confidence limits for μ are $L_1 = \overline{X} - Z_{\alpha(1)}\sigma_{\overline{X}}$ and $L_2 = \infty$. For H_0: $\mu \geq \mu_0$ and H_A: $\mu < \mu_0$, the confidence limits are $L_1 = -\infty$ and $L_2 = \overline{X} + Z_{\alpha(1)}\sigma_{\overline{X}}$. An example of a one-sided confidence interval is Exercise 6.6(b). If a one-tailed null hypothesis is not rejected, then the associated one-tailed confidence interval includes μ_0.

6.5 SYMMETRY AND KURTOSIS

Chapters 3 and 4 showed how sets of data can be described by measures of central tendency and measures of variability. There are additional characteristics that help describe data sets, and they are sometimes used when we want to know whether a distribution resembles a normal distribution. Two basic features of a distribution of measurements are its *symmetry* and its *kurtosis*. A symmetric distribution (as in

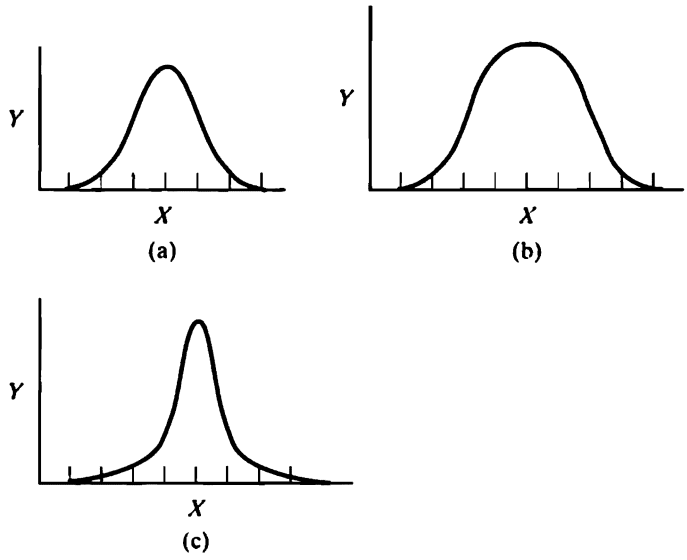


FIGURE 6.8: Symmetric frequency distributions. Distribution (a) is mesokurtic ("normal"), (b) is platykurtic, and (c) is leptokurtic.

distributed populations have $\beta_2 = 3$), some asymmetric distributions have a symmetry measure of 0 and some nonnormal distributions exhibit a kurtosis value of 3 (Thode, 2002: 43).

In practice, researchers seldom calculate these symmetry and kurtosis measures. When they do, however, they should be mindful that using the third and fourth powers of numbers can lead to very serious rounding errors, and they should employ computer programs that use procedures minimizing this problem.

(c) Quantile Measures of Symmetry and Kurtosis. Denoting the ith quartile as Q_i (as in Section 4.2), Q_1 is the first quartile (i.e., the 25% percentile), Q_3 is the third quartile (the 75% percentile), and Q_2 is the second quartile (the 50% percentile, namely the median). A quantile-based expression of skewness (Bowley, 1920: 116; Groeneveld and Meeden, 1984) considers the distance between Q_3 and Q_2 and that between Q_2 and Q_1:

$$\text{Quantile skewness measure} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}, \tag{6.18}$$

which is a measure, without units, that may range from -1, for a distribution with extreme left skewness; to 0, for a symmetric distribution; to 1, for a distribution with extreme right skewness. Because Equation 6.18 measures different characteristics of a set of data than $\sqrt{b_1}$ does, these two numerical measures can be very different (and, especially if the skewness is not great, one of the measures can be positive and the other negative).

Instead of using quartiles Q_1 and Q_3, any other symmetric quantiles could be used to obtain a skewness coefficient (Groeneveld and Meeden, 1984), though the

numerical value of the coefficient would not be the same as that of Equation 6.18. For example, the 10th and 90th percentiles could replace Q_1 and Q_3, respectively, in Equation 6.18, along with Q_2 (the median).

A kurtosis measure based on quantiles was proposed by Moors (1988), using octiles: $\mathcal{O}_1$, the first octile, is the 12.5th percentile; $\mathcal{O}_3$, the third octile, is the 37.5th percentile; $\mathcal{O}_5$ is the 62.5th percentile; and $\mathcal{O}_7$ is the 87.5th percentile. Also, $\mathcal{O}_2 = Q_1$, $\mathcal{O}_4 = Q_2$, and $\mathcal{O}_6 = Q_3$. The measure is

$$\text{Quantile kurtosis measure} = \frac{(\mathcal{O}_7 - \mathcal{O}_5) + (\mathcal{O}_3 - \mathcal{O}_1)}{(\mathcal{O}_6 - \mathcal{O}_2)} = \frac{(\mathcal{O}_7 - \mathcal{O}_5) + (\mathcal{O}_3 - \mathcal{O}_1)}{(Q_3 - Q_1)}, \tag{6.19}$$

which has no units and may range from zero, for extreme platykurtosis, to 1.233, for mesokurtosis; to infinity, for extreme leptokurtosis.

Quantile-based measures of symmetry and kurtosis are rarely encountered.

6.6 ASSESSING DEPARTURES FROM NORMALITY

It is sometimes desired to test the hypothesis that a sample came from a population whose members follow a normal distribution. Example 6.7 and Figure 6.9 present a frequency distribution of sample data, and we may desire to know whether the data are likely to have come from a population that had a normal distribution. Comprehensive examinations of statistical methods applicable to such a question have been reported (e.g., by D'Agostino, 1986; Landry and Lepage, 1992; Shapiro, 1986; and Thode, 2002), and a brief overview of some of these techniques will be given here. The latter author discusses about 40 methods for normality testing and notes (ibid.: 143–157) that the power of a testing procedure depends upon the sample size and the nature of the nonnormality that is to be detected (e.g., asymmetry, long-tailedness, short-tailedness).

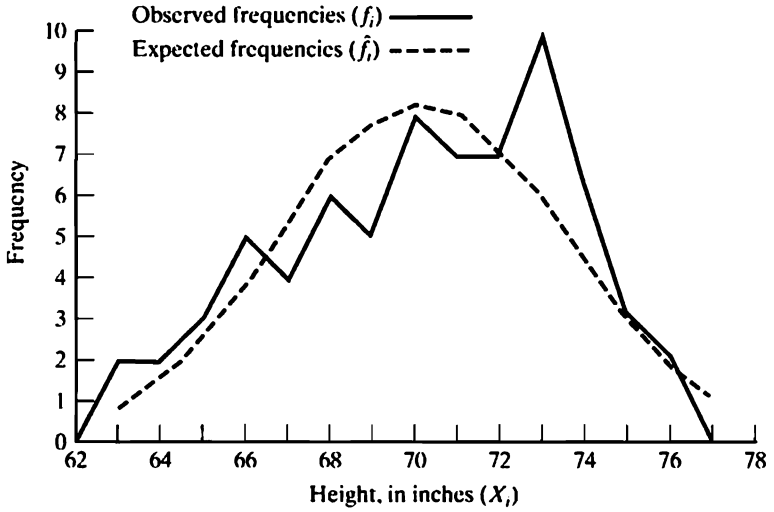


FIGURE 6.9: The frequency polygon for the student height data in Example 6.7 (solid line) with the frequency curve that would be expected if the data followed a normal distribution (broken line).

EXAMPLE 6.7 The Heights of the First 70 Graduate Students in My Biostatistics Course

Height (X_i) (in.)	Observed Frequency (f_i)	Cumulative Frequency (cum. f_i)	f_iX_i (in.)	$f_iX_i^2$ (in.^{2})
63	2	2	126	7,938
64	2	4	128	8,192
65	3	7	195	12,675
66	5	12	330	21,780
67	4	16	268	17,956
68	6	22	408	27,744
69	5	27	345	23,805
70	8	35	560	39,200
71	7	42	497	35,287
72	7	49	504	36,288
73	10	59	730	53,290
74	6	65	444	32,856
75	3	68	225	16,875
76	2	70	152	11,552
	$\Sigma f_i =$ $n = 70$		$\Sigma f_iX_i =$ 4,912 in.	$\Sigma f_iX_i^2 =$ 345,438 in.^{2}

$$SS = \Sigma f_i X_i^2 - \frac{\Sigma(f_i X_i)^2}{n} = 345{,}438 \text{ in.}^2 - \frac{(14{,}912 \text{ in.})^2}{70} = 755.9429 \text{ in.}^2$$

$$s^2 = \frac{SS}{n-1} = \frac{755.9429 \text{ in.}^2}{69} = 10.9557 \text{ in.}^2$$

(a) Graphical Assessment of Normality. Many methods have been used to assess graphically the extent to which a frequency distribution of observed data resembles a normal distribution (e.g., Thode, 2002: 15–40). Recall the graphical representation of a normal distribution as a frequency curve, shown in Figure 6.1. A frequency polygon for the data in Example 6.7 is shown in Figure 6.9, and superimposed on that figure is a dashed curve showing what a normal distribution, with the same number of data (n) mean ($\overline{X}$), and standard deviation (s), would look like. We may wish to ask whether the observed frequencies deviate significantly from the frequencies expected from a normally distributed sample.

Figure 6.10 shows the data of Example 6.7 plotted as a cumulative frequency distribution. A cumulative frequency graph of a normal distribution will be S-shaped (called "sigmoid"). The graph in Figure 6.10 is somewhat sigmoid in shape, but in this visual presentation it is difficult to conclude whether that shape is pronounced enough to reflect normality. So, a different approach is desired. Note that the vertical axis on the left side of the graph expresses cumulative frequencies and the vertical axis

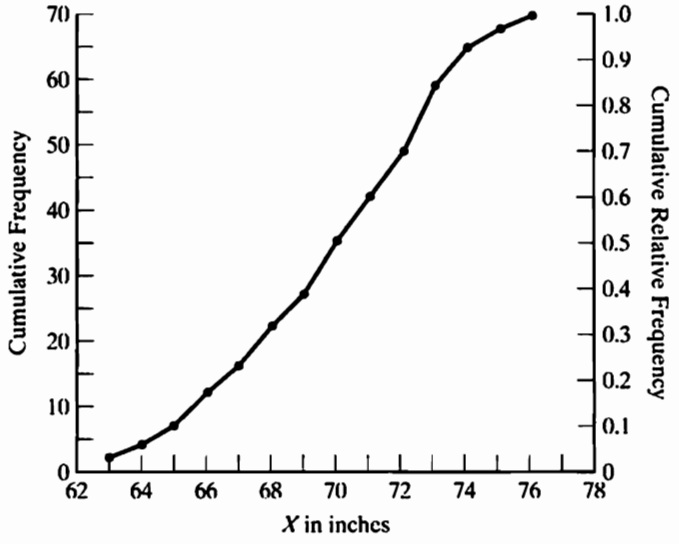


FIGURE 6.10: The cumulative frequency polygon of the student-height data of Example 6.7.

on the right side displays relative frequencies (as introduced in Figure 1.9), and the latter may be thought of as percentiles. For instance, the sample of 70 measurements in Example 6.7 contains 22 data, where $X_i \leq 68$ inches, so 68 in. on the horizontal axis is associated with a cumulative frequency of 22 on the left axis and a cumulative relative frequency of $22/70 = 0.31$ on the right axis; thus, we could say that a height of 68 in. is at the 31st percentile of this sample.

Examination of the relative cumulative frequency distribution is aided greatly by the use of the *normal probability scale*, as in Figure 6.11, rather than the linear scale of Figure 6.10. As the latter figure shows, a given increment in X_i (on the abscissa, the horizontal axis) near the median is associated with a much larger change in relative frequency (on the ordinate, the vertical axis) than is the same increment in X_i at very high or very low relative frequencies. Using the normal-probability scale on the ordinate expands the scale for high and low percentiles and compresses it for percentiles toward the median (which is the 50th percentile). The resulting cumulative frequency plot will be a straight line for a normal distribution. A leptokurtic distribution will appear as a sigmoid (S-shaped) curve on such a plot, and a platykurtic distribution will appear as a reverse S-shape. A negatively skewed distribution will show an upward curve, as the lower portion of an S, and a positively skewed distribution will manifest itself in a shape resembling the upper portion of an S. Figure 6.11 shows the data of Example 6.7 plotted as a cumulative distribution on a normal-probability scale. The curve appears to tend slightly toward leptokurtic.

Graph paper with the normal-probability scale on the ordinate is available commercially, and such graphs are produced by some computer software. One may also encounter graphs with a normal-probability scale on the abscissa and X_i on the ordinate. The shape of the plotted curves will then be converse of those described previously.

(b) Assessing Normality Using Symmetry and Kurtosis Measures. Section 6.5 indicated that a normally distributed population has symmetry and kurtosis parameters

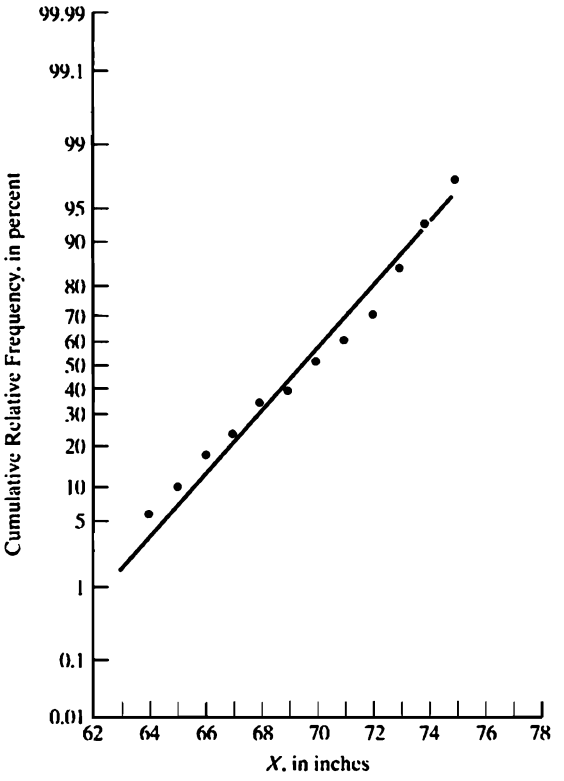


FIGURE 6.11: The cumulative relative frequency distribution for the data of Example 6.7, plotted with the normal probability scale as the ordinate. The expected frequencies (i.e., the frequencies from a normal distribution) would fall on the straight line shown.

of $\sqrt{\beta_1} = 0$ and $\beta_2 = 3$, respectively. Therefore, we can ask whether a sample of data came from a normal population by testing the null hypothesis H_0: $\sqrt{\beta_1} = 0$ (versus the alternate hypothesis, H_A: $\sqrt{\beta_1} \neq 0$) and the hypothesis H_0: $\beta_2 = 3$ (versus H_A: $\beta_2 \neq 3$), as shown in Section 7.16. There are also procedures that employ the symmetry and kurtosis measures simultaneously, to test H_0: The sample came from a normally distributed population versus H_A: The sample came from a population that is not normally distributed (Bowman and Shenton. 1975, 1986; D'Agostino and Pearson, 1973; Pearson, D'Agostino, and Bowman, 1977; Thode, 2002: 54–55, 283).

Statistical testing using these symmetry and kurtosis measures, or the procedure of Section 6.6(d), is generally the best for assessing a distribution's departure from normality (Thode, 2002: 2).

(c) Goodness-of-Fit Assessment of Normality. As will be discussed in Chapter 22, procedures called goodness-of-fit tests are applicable when asking whether a sample of data is likely to have come from a population with a specified distribution. Goodness-of-fit procedures known as chi-square, log-likelihood, and Kolmogorov-Smirnov, or modifications of them, have been used to test the hypothesis of normality (e.g., Zar, 1984: 88–93); and Thode (2002) notes that other goodness-of-fit tests, such as that of Kuiper (1960, which is alluded to in Section 27.18 for other purposes) may also be used. These methods perform poorly, however, in that they possess very low power; and they are not recommended for addressing hypotheses of normality (D'Agostino, 1986; D'Agostino, Belanger, and D'Agostino, 1990; Moore, 1986; Thode, 2002: 152).

(d) Other Methods of Assessing Normality. Shapiro and Wilk (1965) presented a test for normality involving the calculation of a statistic they called W. This computation requires an extensive table of constants, because a different set of $n/2$ constants is needed for each sample size, n. The authors provided a table of these constants and also of critical values of W, but only for n as large as 50. The power of this test has been shown to be excellent when testing for departures from normality (D'Agostino, 1986; Shapiro, Wilk, and Chen, 1968). Royston (1982a, 1982b) provided an approximation that extends the W test to n as large as 2000. Shapiro and Francia (1972) presented a modified procedure (employing a statistic they called W') that allows n to be as large as 99; but Pearson, D'Agostino, and Bowman (1977) noted errors in the published critical values. Among other modifications of W, Rahman and Govindarajulu (1997) offered one (with a test statistic they called $\tilde{W}$) declared to be applicable to any sample size, with critical values provided for n up to 5000. Calculation of W or its modifications is cumbersome and will most likely be done by computer; this test is unusual in that it involves rejection of the null hypothesis of normality if the test statistic is *equal to or less than* the one-tailed critical value.

The performance of the Shapiro-Wilk test is adversely affected by the common situation where there are tied data (i.e., data that are identical, as occur in Example 6.7, where there is more than one observation at each height) (Pearson, D'Agostino, and Bowman, 1977), but modifications of it have addressed that problem (e.g., Royston, 1986, 1989). Statistical testing using the Shapiro-Wilk test, or using symmetry and kurtosis measures (Section 6.6(b)), is generally the preferred method for inquiring whether an underlying population is normally distributed (Thode, 2002: 2).

EXERCISES

6.1. The following body weights were measured in 37 animals:

Weight (X_i) (kg)	*Frequency* (f_i)
4.0	2
4.3	3
4.5	5
4.6	8
4.7	6
4.8	5
4.9	4
5.0	3
5.1	1

(a) Calculate the symmetry measure, $\sqrt{b_1}$.
(b) Calculate the kurtosis measure, b_2.
(c) Calculate the skewness measure based on quantiles.
(d) Calculate the kurtosis measure based on quantiles.

6.2. A normally distributed population of lemming body weights has a mean of 63.5 g and a standard deviation of 12.2 g.
(a) What proportion of this population is 78.0 g or larger?
(b) What proportion of this population is 78.0 g or smaller?
(c) If there are 1000 weights in the population, how many of them are 78.0 g or larger?
(d) What is the probability of choosing at random from this population a weight smaller than 41.0 g?

6.3. **(a)** Considering the population of Exercise 6.2, what is the probability of selecting at random a body weight between 60.0 and 70.0 g?
(b) What is the probability of a body weight between 50.0 and 60.0 g?

6.4. **(a)** What is the standard deviation of all possible means of samples of size 10 which could be drawn from the population in Exercise 6.2?
(b) What is the probability of selecting at random from this population a sample of 10 weights that has a mean greater than 65.0 g?
(c) What is the probability of the mean of a sample of 10 being between 60.0 and 62.0 g?

6.5. The following 18 measurements are obtained of a pollutant in a body of water: 10.25, 10.37, 10.66, 10.47, 10.56, 10.22, 10.44, 10.38, 10.63, 10.40, 10.39, 10.26, 10.32, 10.35, 10.54, 10.33, 10.48, 10.68 milligrams per liter. Although we would not know this in practice, for the sake of this example let us say

we know that the standard error of the mean is $\sigma_{\overline{X}} = 0.24$ mg/liter in the population from which this sample came. The legal limit of this pollutant is 10.00 milligrams per liter.

(a) Test whether the mean concentration in this body of water exceeds the legal limit (i.e., test H_0: $\mu \leq 10.00$ mg/L versus H_A: $\mu > 10.00$ mg/L), using the 5% level of significance.

(b) Calculate the 95% confidence interval for μ.

6.6. The incubation time was measured for 24 alligator eggs. Let's say that these 24 data came from a population with a variance of $\sigma^2 = 89.06$ days^2, and the sample mean is $\overline{X} = 61.4$ days.

(a) Calculate the 99% confidence limits for the population mean.

(b) Calculate the 95% confidence limits for the population mean.

(c) Calculate the 90% confidence limits for the population mean.

CHAPTER 7

One-Sample Hypotheses

This chapter will continue the discussion of Section 6.3 on how to draw inferences about population parameters by testing hypotheses about them using appropriate sample statistics. It will consider hypotheses about each of several population parameters, including the population mean, median, variance, standard deviation, and coefficient of variation. The chapter will also discuss procedures (introduced in Section 6.4) for expressing the confidence one can have in estimating population parameters from sample statistics.

7.1 TWO-TAILED HYPOTHESES CONCERNING THE MEAN

Section 6.4 introduced the concept of statistical testing using a pair of statistical hypotheses, the null and alternate hypotheses, as statements that a population mean (μ) is equal to some specified value (let's call it μ_0):

$$H_0: \quad \mu = \mu_0;$$
$$H_A: \quad \mu \neq \mu_0.$$

For example, let us consider the body temperatures of 25 intertidal crabs that we exposed to air at 24.3° C (Example 7.1). We may wish to ask whether the mean body temperature of members of this species of crab is the same as the ambient air temperature of 24.3° C. Therefore,

$$H_0: \quad \mu = 24.3° \text{ C, and}$$
$$H_A: \quad \mu \neq 24.3° \text{ C.}$$

where the null hypothesis states that the mean of the population of data from which this sample of 25 came is 24.3°C (i.e., μ is "no different from 24.3°C"), and the alternate hypothesis is that the population mean is not equal to (i.e., μ is different from) 24.3°C.

EXAMPLE 7.1 The Two-Tailed *t* Test for Difference between a Population Mean and a Hypothesized Population Mean

Body temperatures (measured in °C) of 25 intertidal crabs placed in air at 24.3°C: 25.8, 24.6, 26.1, 22.9, 25.1, 27.3, 24.0, 24.5, 23.9, 26.2, 24.3, 24.6, 23.3, 25.5, 28.1, 24.8, 23.5, 26.3, 25.4, 25.5, 23.9, 27.0, 24.8, 22.9, 25.4.

$$H_0: \ \mu = 24.3°\text{C}$$

$$H_A: \ \mu \neq 24.3°\text{C}$$

$$\alpha = 0.05$$

$$n = 25$$

$$\overline{X} = 25.03°\text{C}$$

$$s^2 = 1.80(°\text{C})^2$$

$$s_{\overline{X}} = \sqrt{\frac{1.80(°\text{C})^2}{25}} = 0.27°\text{C}$$

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}} = \frac{25.03°\text{C} - 24.3°\text{C}}{0.27°\text{C}} = \frac{0.73°\text{C}}{0.27°\text{C}} = 2.704$$

$$\nu = 24$$

$$t_{0.05(2),24} = 2.064$$

As $|t| > t_{0.05(2),24}$, reject H_0 and conclude that the sample of 25 body temperatures came from a population whose mean is not 24.3°C.

$$0.01 < P < 0.02 \ [P = 0.012]^*$$

In Section 6.1 (Equation 6.6), $Z = (\overline{X} - \mu)/\sigma_{\overline{X}}$ was introduced as a *normal deviate*, and it was shown how one can determine the probability of obtaining a sample with mean $\overline{X}$ from a population with a specified mean μ. And Section 6.3 discussed how the normal deviate can be used to test hypotheses about a population mean. Note, however, that the calculation of Z requires the knowledge of $\sigma_{\overline{X}}$, which we typically do not have. The best we can do is to calculate $s_{\overline{X}}$ as an estimate of $\sigma_{\overline{X}}$. If n is very, very large, then $s_{\overline{X}}$ is a good estimate of $\sigma_{\overline{X}}$, and we can be

*Throughout the examples in this book, the exact probability of a calculated test statistic (such as t), as determined by computer software, is indicated in brackets. It should not be assumed that the many decimal places given by computer programs are all accurate (McCullough, 1998, 1999); therefore, the book's examples will routinely express these probabilities to only two or three (occasionally four) decimal places. The term "software" was coined by John Wilder Tukey (Leonhardt, 2000).

tempted to calculate Z using this estimate. However, for most biological situations n is insufficiently large to do this; but we can use, in place of the normal distribution (Z), a distribution known as t, the development of which was a major breakthrough in statistical methodology:*

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}}. \tag{7.1}$$

Because the t-testing procedure is so readily employed, we need not wonder whether n is large enough to use Z; and, in fact, Z is almost never used for hypothesis testing about means.

As do some other distributions to be encountered among statistical methods, the t distribution has different shapes for different values of what is known as *degrees of freedom* (denoted by ν, the lowercase Greek nu).† For hypotheses concerning a mean,

$$\nu = n - 1. \tag{7.2}$$

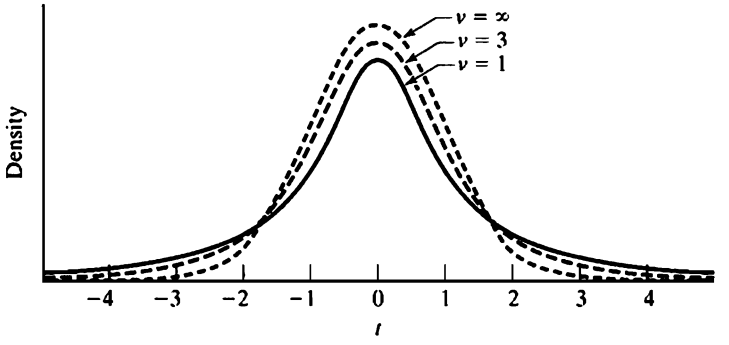


FIGURE 7.1: The t distribution for various degrees of freedom, ν. For $\nu = \infty$, the t distribution is identical to the normal distribution.

Recall that n is the size of the sample (i.e., the number of data from which $\overline{X}$ has been calculated). The influence of ν on the shape of the t distribution is shown in Figure 7.1.

*The t statistic is also referred to as "Student's t" because of William Sealy Gosset (1876–1937), who was an English statistician with the title "brewer" in the Guinness brewery of Dublin. He used the pen name "Student" (under his employer's policy requiring anonymity) to publish noteworthy developments in statistical theory and practice, including ("Student," 1908) the introduction of the distribution that bears his pseudonym. (See Boland, 1984, 2000; Irwin, 1978; Lehmann, 1999; Pearson, 1939; Pearson, Plackett, and Barnard, 1990; Zabell, 2008.) Gosset originally referred to his distribution as z; and, especially between 1922 and 1925, R. A. Fisher (e.g., 1925a, 1925b: 106–113, 117–125; 1928) helped develop its potential in statistical testing while modifying it; Gosset and Fisher then called the modification "t" (Eisenhart, 1979). Gosset was a modest man, but he was referred to as "one of the most original minds in contemporary science" by Fisher (1939a), himself one of the most insightful and influential statisticians of all time. From his first discussions of the t distribution, Gosset was aware that it was strictly applicable only if sampling normally distributed populations, though he surmised that only large deviations from normality would invalidate the use of t (Lehmann, 1999).

†In early writings of the t distribution (during the 1920s and 1930s), the symbol n or f was used for degrees of freedom. This was often confusing because these letters had commonly been used to denote other quantities in statistics so Maurice G. Kendall (1943: 292) recommended ν.

This distribution is leptokurtic (see Section 6.5), having a greater concentration of values around the mean and in the tails than does a normal distribution; but as n (and, therefore, ν) increases, the t distribution tends to resemble a normal distribution more closely, and for $\nu = \infty$ (i.e., for an infinitely large sample*), the t and normal distributions are identical; that is, $t_{\alpha,\infty} = Z_\alpha$.

The mean of the sample of 25 data (body temperatures) shown in Example 7.1 is 25.03°C, and the sample variance is 1.80(°C)^2. These statistics are estimates of the mean and variance of the population from which this sample came. However, this is only one of a very large number of samples of size 25 that could have been taken at random from the population. The distribution of the means of all possible samples with $n = 25$ is the t distribution for $\nu = 24$, which is represented by the curve of Figure 7.2. In this figure, the mean of the t distribution (i.e., $t = 0$) represents the mean hypothesized in H_0 (i.e., $\mu = \mu_0 = 24.3$°C), for, by Equation 7.1, $t = 0$ when $\overline{X} = \mu$. The shaded areas in this figure represent the extreme 5% of the total area under the curve (2.5% in each tail). Thus, an $\overline{X}$ so far from μ that it lies in either of the shaded areas has a probability of less than 5% of occurring by chance alone, and we assume that it occurred because H_0 is, in fact, false. As explained in Section 6.3 regarding the Z distribution, because an extreme t value in either direction from μ will cause us to reject H_0, we are said to be considering a "two-tailed" (or "two-sided") test.

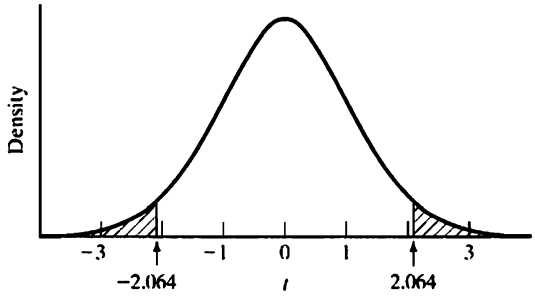


FIGURE 7.2: The t distribution for $\nu = 24$, showing the critical region (shaded area) for a two-tailed test using $\alpha = 0.05$. (The critical value of t is 2.064.)

For $\nu = 24$, we can consult Appendix Table B.3 to find the following two-tailed probabilities (denoted as "$\alpha(2)$") of various values of t:

ν	$\alpha(2)$:	0.50	0.20	0.10	0.05	0.02	0.01
24		0.685	1.318	1.711	2.064	2.492	2.797

Thus, for example, for a two-tailed α of 0.05, the shaded areas of the curve begin at 2.064 t units on either side of μ. Therefore, we can state:

$$P(|t| \geq 2.064) = 0.05.$$

That is, 2.064 and −2.064 are the *critical values* of t; and if t (calculated from Equation 7.1) is equal to or greater than 2.064, or is equal to or less than −2.064, that will be considered reasonable cause to reject H_0 and consider H_A to be a true

*The modern symbol for infinity (∞) was introduced in 1655 by influential English mathematician John Wallis (1616–1703) (Cajori, 1928/9, Vol. 2: 44), but it did not appear again in print until a work by Jacob Bernoulli was published posthumously in 1713 by his nephew Nikolaus Bernoulli (Gullberg, 1997: 30).

statement. That portion of the t distribution beyond the critical values (i.e., the shaded areas in the figure) is called the *critical region*,* or *rejection region*. For the sample of 25 body temperatures (see Example 7.1), $t = 2.704$. As 2.704 lies within the critical region (i.e., $2.704 > 2.064$), H_0 is rejected, and we conclude that the mean body temperature of crabs under the conditions of our experiment is not 24.3°C.

To summarize, the hypotheses for the two-tailed test are

$$H_0: \mu = \mu_0 \quad \text{and} \quad H_A: \mu \neq \mu_0,$$

where μ_0 denotes the hypothesized value to which we are comparing the population mean. (In the above example, $\mu_0 = 24.3$°C.) The test statistic is calculated by Equation 7.1, and if its absolute value is larger than the two-tailed critical value of t from Appendix Table B.3, we reject H_0 and assume H_A to be true. The critical value of t can be abbreviated as $t_{\alpha(2),\nu}$, where $\alpha(2)$ refers to the two-tailed probability of α. Thus, for the preceding example, we could write $t_{0.05(2),24} = 2.064$. In general, for a two-tailed t test,

$$\text{if } |t| \geq t_{\alpha(2),\nu}, \text{ then reject } H_0.$$

Example 7.1 presents the computations for the analysis of the crab data. A t of 2.704 is calculated, which for 24 degrees of freedom lies between the tabled critical values of $t_{0.02(2),24} = 2.492$ and $t_{0.01(2),24} = 2.797$. Therefore, if the null hypothesis, H_0, is a true statement about the population we sampled, the probability of $\overline{X}$ being at least this far from μ is between 0.01 and 0.02; that is, $0.01 < P(|t| \geq 2.704) < 0.02$.† As this probability is less than 0.05, we reject H_0 and declare it is not a true statement. For a consideration of the types of errors involved in rejecting or accepting the null hypothesis, refer to Section 6.4.

Frequently, the hypothesized value in the null and alternate hypotheses is zero. For example, the weights of twelve rats might be measured before and after the animals are placed on a regimen of forced exercise for one week. The change in weight of the animals (i.e., weight after minus weight before) could be recorded, and it might have been found that the mean weight change was −0.65 g (i.e., the mean weight change is a 0.65 g weight loss). If we wished to infer whether such exercise causes any significant change in rat weight, we could state $H_0: \mu = 0$ and $H_A: \mu \neq 0$; Example 7.2 summarizes the t test for this H_0 and H_A. This test is two tailed, for a large $\overline{X} - \mu$ difference in either direction will constitute grounds for rejecting the veracity of H_0.‡

EXAMPLE 7.2 A Two-Tailed Test for Significant Difference between a Population Mean and a Hypothesized Population Mean of Zero

Weight change of twelve rats after being subjected to a regimen of forced exercise. Each weight change (in g) is the weight after exercise minus the weight before.

1.7	$H_0: \mu = 0$
0.7	$H_A: \mu \neq 0$
−0.4	$\alpha = 0.05$

*David (1995) traces the first use of this term to J. Neyman and E. S. Pearson in 1933.

†Some calculators and many computer programs have the capability of determining the probability of a given t (e.g., see Boomsma and Molenaar, 1994). For the present example, we would thereby find that $P(|t| \geq 2.704) = 0.012$.

‡Data that result from the differences between pairs of data (such as measurements before and after an experimental treatment) are discussed further in Chapter 9.

−1.8
0.2
0.9
−1.2
−0.9
−1.8
−1.4
−1.8
−2.0

$n = 12$

$\overline{X} = -0.65\ \text{g}$

$s^2 = 1.5682\ \text{g}^2$

$$s_{\overline{X}} = \sqrt{\frac{1.5682\ \text{g}^2}{12}} = 0.36\ \text{g}$$

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}} = \frac{-0.65\ \text{g}}{0.36\ \text{g}} = -1.81$$

$\nu = n - 1 = 11$

$t_{0.05(2),11} = 2.201$

Since $|t| < t_{0.05(2),11}$, do not reject H_0.

$$0.05 < P < 0.10\ [P = 0.098]$$

Therefore, we conclude that the exercise does not cause a weight change in the population from which this sample came.

It should be kept in mind that concluding statistical significance is not the same as determining biological significance. In Example 7.1, statistical significance was achieved for a difference of 0.73°C between the mean crab body temperature (25.03°C) and the air temperature (24.3°C). The statistical question posed is whether that magnitude of difference is likely to occur by chance if the null hypothesis of no difference is true. The answer is that it is unlikely (there is only a 0.012 probability) and, therefore, we conclude that H_0 is not true. Now the biological question is whether a difference of 0.73°C is of significance (with respect to the crabs' physiology, to their ecology, or otherwise). If the sample of body temperatures had a smaller standard error, $s_{\overline{X}}$, an even smaller difference would have been declared statistically significant. But is a difference of, say, 0.1°C or 0.01°C of biological importance (even if it is statistically significant)? In Example 7.2, the mean weight change, 0.36 g, was determined not to be significant statistically. But if the sample mean weight change had been 0.8 g (and the standard error had been the same), t would have been calculated to be 2.222 and H_0 would have been rejected. The statistical conclusion would have been that the exercise regime does result in weight change in rats, but the biological question would then be whether a weight change as small as 0.8 g has significance biologically. Thus, assertion of statistical difference should routinely be followed by an assessment of the significance of that difference to the objects of the study (in these examples, to the crabs or to the rats).

(a) Assumptions. The theoretical basis of t testing assumes that sample data came from a normal population, assuring that the mean at hand came from a normal distribution of means. Fortunately, the t test is *robust*,* meaning that its validity is not seriously affected by moderate deviations from this underlying assumption. The test also assumes—as other statistical tests typically do—that the data are a random sample (see Section 2.3).

The adverse effect of nonnormality is that the probability of a Type I error differs substantially from the stated α. Various studies (e.g., Cicchitelli, 1989; Pearson and Please, 1975; and Ractliffe, 1968) have shown that the detrimental effect of nonnormality is greater for smaller α but less for larger n, that there is little effect if

*The term *robustness* was introduced by G. E. P. Box in 1953 (David, 1995).

the distribution is symmetrical, and that for asymmetric distributions the effect is less with strong leptokurtosis than with platykurtosis or mesokurtosis; and the undesirable effect of nonnormality is much less for two-tailed testing than for one-tailed testing (Section 7.2).

It is important to appreciate that a sample used in statistical testing such as that discussed here must consist of truly replicated data, where a *replicate** is defined as the smallest experimental unit to which a treatment is independently applied. In Example 7.1, we desired to draw conclusions about a population of measurements representing a large number of animals (i.e., crabs). Therefore, the sample must consist of measurements (i.e., body temperatures) from n (i.e., 25) animals; it would *not* be valid to obtain 25 body temperatures from a single animal. And, in Example 7.2, 12 individual rats must be used; it would *not* be valid to employ data obtained from subjecting the same animal to the experiment 12 times. Such invalid attempts at replication are discussed by Hurlbert (1984), who named them *pseudoreplication*.

7.2 ONE-TAILED HYPOTHESES CONCERNING THE MEAN

In Section 7.1, we spoke of the hypotheses $H_0: \mu = \mu_0$ and $H_A: \mu \neq \mu_0$, because we were willing to consider a large deviation of $\overline{X}$ in either direction from μ_0 as grounds for rejecting H_0. However, in some instances, our interest lies only in whether $\overline{X}$ is significantly larger (or significantly smaller) than μ_0, and this is termed a "one-tailed" (or "one-sided") test situation. For example, we might be testing a drug hypothesized to cause weight reduction in humans. The investigator is interested only in whether a weight *loss* occurs after the drug is taken. (In Example 7.2, using a two-sided test, we were interested in determining whether either weight loss or weight gain had occurred.) It is important to appreciate that the decision whether to test one-tailed or two-tailed hypotheses must be based on the scientific question being addressed, *before* data are collected.

In the present example, if there is either weight gain or no weight change, the drug will be considered a failure. Therefore, for this one-sided test, we should state $H_0: \mu \geq 0$ and $H_A: \mu < 0$. Here, the null hypothesis states that there is no mean weight loss (i.e., the mean weight change is greater than or equal to zero), and the alternate hypothesis states that there is a mean weight loss (i.e., the mean weight change is less than zero). By examining the alternate hypothesis, H_A, we see that H_0 will be rejected if t is in the left-hand critical region of the t distribution. In general,

$$\text{for } H_A: \mu < \mu_0,$$
$$\text{if } t \leq -t_{\alpha(1),\nu}, \text{ then reject } H_0.^{\dagger}$$

Example 7.3 summarizes such a set of 12 weight change data tested against this pair of hypotheses. From Appendix Table B.3 we find that $t_{0.05(1),11} = 1.796$, and the critical region for this test is shown in Figure 7.3. From this figure, and by examining Appendix Table B.3, we see that $t_{\alpha(1),\nu} = t_{2\alpha(2),\nu}$ or $t_{\alpha(2),\nu} = t_{\alpha/2(1),\nu}$; that is, for example, the critical value of t for a one-sided test at $\alpha = 0.05$ is the same as the critical value of t for a two-sided test at $\alpha = 0.10$.

*The term *replicate*, in the context of experimental design, was introduced by R. A. Fisher in 1926 (Miller, 2004a).

†For one-tailed testing of this H_0, probabilities of t up to 0.25 are indicated in Appendix Table B.3. If $t = 0$, then $P = 0.50$; so if $-t_{0.25(1),\nu} < t < 0$, then $0.25 < P < 0.50$; and if $t > 0$, then $P > 0.50$.

EXAMPLE 7.3 A One-Tailed t Test for the Hypotheses H_0: $\mu \geq 0$ and H_A: $\mu < 0$

The data are weight changes of humans, tabulated after administration of a drug proposed to result in weight loss. Each weight change (in kg) is the weight after minus the weight before drug administration.

0.2
−0.5
−1.3
−1.6
−0.7
0.4
−0.1
0.0
−0.6
−1.1
−1.2
−0.8

$n = 12$

$\overline{X} = -0.61$ kg

$s^2 = 0.4008\ \text{kg}^2$

$$s_{\overline{X}} = \sqrt{\frac{0.4008\ \text{kg}^2}{12}} = 0.18\ \text{kg}$$

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}} = \frac{-0.61\ \text{kg}}{0.18\ \text{kg}} = -3.389$$

$\nu = n - 1 = 11$

$t_{0.05(1),11} = 1.796.$

If $t \leq -t_{0.05(1),11}$, reject H_0.

Conclusion: reject H_0.

$$0.0025 < P(t \leq -3.389) < 0.005\ [P = 0.0030]$$

We conclude that the drug does cause weight loss.

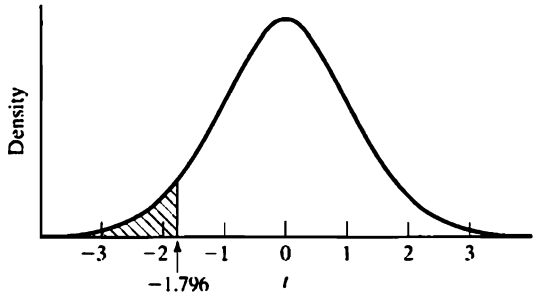


FIGURE 7.3: The distribution of t for $\nu = 11$, showing the critical region (shaded area) for a one-tailed test using $\alpha = 0.05$. (The critical value of t is −1.796.)

If we are interested in whether $\overline{X}$ is significantly *greater* than some value, μ_0, the hypotheses for the one-tailed test are H_0: $\mu \leq \mu_0$ and H_A: $\mu > \mu_0$. For example, a drug manufacturer might advertise that a product dissolves completely in gastric juice within 45 sec. The hypotheses appropriate for testing this claim are H_0: $\mu \leq 45$ sec and H_A: $\mu > 45$ sec, because we are not particularly interested in the possibility that the product dissolves faster than is claimed, but we wish to determine whether its dissolving time is longer than advertised. Thus, the rejection region would be in the right-hand tail, rather than in the left-hand tail (the latter being the case in Example 7.3). The details of such a test are shown in Example 7.4. In general,

for H_A: $\mu > \mu_0$,
if $t \geq t_{\alpha(1),\nu}$, then reject H_0.*

*For this H_0, if $t = 0$, then $P = 0.50$; therefore, if $0 < t < t_{0.25(1),\nu}$, then $0.25 < P < 0.50$, and if $t < 0$, then $P > 0.50$.

EXAMPLE 7.4 The One-Tailed t Test for the Hypotheses H_0: $\mu \leq 45$ sec and H_A: $\mu > 45$ sec

Dissolving times (in sec) of a drug in gastric juice: 42.7, 43.4, 44.6, 45.1, 45.6, 45.9, 46.8, 47.6.

H_0: $\mu \leq 45$ sec
H_A: $\mu > 45$ sec

$\alpha = 0.05$
$n = 8$
$\overline{X} = 45.21$ sec
SS $= 18.8288$ sec^2
$s^2 = 2.6898$ sec^2

$s_{\overline{X}} = 0.58$ sec

$$t = \frac{45.21 \text{ sec} - 45 \text{ sec}}{0.58 \text{ sec}} = 0.36$$

$\nu = 7$

$t_{0.05(1),7} = 1.895$

If $t \geq t_{0.05(1),7}$, reject H_0.

Conclusion: do not reject H_0.

$$P(t \geq 0.36) > 0.25 \; [P = 0.36]$$

We conclude that the mean dissolving time is not greater than 45 sec.

7.3 CONFIDENCE LIMITS FOR THE POPULATION MEAN

When Section 7.1 defined $t = (\overline{X} - \mu)/s_{\overline{X}}$, it was explained that 5% of all possible means from a normally distributed population with mean μ will yield t values that are either larger than $t_{0.05(2),\nu}$ or smaller than $-t_{0.05(2),\nu}$; that is, $|t| \geq t_{0.05(2),\nu}$ for 5% of the means. This connotes that 95% of all t values obtainable lie between the limits of $-t_{0.05(2),\nu}$ and $t_{0.05(2),\nu}$; this may be expressed as

$$P\left[-t_{0.05(2),\nu} \leq \frac{\overline{X} - \mu}{s_{\overline{X}}} \leq t_{0.05(2),\nu}\right] = 0.95. \tag{7.3}$$

It follows from this that

$$P[\overline{X} - t_{0.05(2),\nu}\, s_{\overline{X}} \leq \mu \leq \overline{X} + t_{0.05(2),\nu}\, s_{\overline{X}}] = 0.95. \tag{7.4}$$

The value of the population mean, μ, is not known, but we estimate it as $\overline{X}$, and if we apply Equation 7.4 to many samples from this population, for 95% of the samples the interval between $\overline{X} - t_{0.05(2),\nu}s_{\overline{X}}$ and $\overline{X} + t_{0.05(2),\nu}s_{\overline{X}}$ will include μ. As introduced in Section 6.4, this interval is called the *confidence interval* (abbreviated CI) for μ.

In general, the confidence interval for μ can be stated as

$$P[\overline{X} - t_{\alpha(2),\nu}\, s_{\overline{X}} \leq \mu \leq \overline{X} + t_{\alpha(2),\nu}\, s_{\overline{X}}] = 1 - \alpha. \tag{7.5}$$

As defined in Section 6.4, $\overline{X} - t_{\alpha(2),\nu}\, s_{\overline{X}}$ is called the *lower confidence limit* (abbreviated L_1); and $\overline{X} + t_{\alpha(2),\nu}\, s_{\overline{X}}$ is the *upper confidence limit* (abbreviated L_2); and the two confidence limits can be stated as

$$\overline{X} \pm t_{\alpha(2),\nu}\, s_{\overline{X}} \tag{7.6}$$

(reading "±" to be "plus or minus"). In expressing a confidence interval, we call the quantity $1 - \alpha$ (namely, $1 - 0.05 = 0.95$ in the present example)

the *confidence level* (or the *confidence coefficient*, or the *confidence probability*).*

Although $\overline{X}$ is the best estimate of μ, it is only an estimate (and not necessarily a very good one), and the calculation of the confidence interval for μ provides an expression of the precision of the estimate. Example 7.5, part (a), refers to the data of Example 7.1 and demonstrates the determination of the 95% confidence interval for the mean of the population from which the sample came. As the 95% confidence limits are computed to be $L_1 = 24.47°C$ and $L_2 = 25.59°C$, the 95% confidence interval may be expressed as $P(24.47°C \leq \mu \leq 25.59°C)$. The meaning of this kind of statement is commonly expressed in nonprobabilistic terms as having 95% *confidence* that the interval of 24.47°C to 25.59°C contains μ. This does *not* mean that there is a 95% *probability* that the interval constructed from this one sample contains the population mean, μ; but it does mean that 95% of the confidence limits computed for many independent random samples would bracket μ (or, this could be stated as the probability that the confidence interval from a future sample would contain μ). And, if the μ_0 in H_0 and H_A is within the confidence interval, then H_0 will not be rejected.

EXAMPLE 7.5 Computation of Confidence Intervals and Confidence Limits for the Mean, Using the Data of Example 7.1

(a) At the 95% confidence level:

$\overline{X} = 25.03°C$

$s_{\overline{X}} = 0.27°C$

$t_{0.05(2),24} = 2.064$

$\nu = 24$

$$\begin{aligned} 95\%\text{ confidence interval} &= \overline{X} \pm t_{0.05(2),24}\, s_{\overline{X}} \\ &= 25.03°C \pm (2.064)(0.27°C) \\ &= 25.03°C \pm 0.56°C \end{aligned}$$

$$\begin{aligned} 95\%\text{ confidence limits: } L_1 &= 25.03°C - 0.56°C = 24.47°C \\ L_2 &= 25.03°C + 0.56°C = 25.59°C \end{aligned}$$

(b) At the 99% confidence level:

$t_{0.01(2),24} = 2.797$

$$\begin{aligned} 99\%\text{ confidence interval} &= \overline{X} \pm t_{0.01(2),24}\, s_{\overline{X}} \\ &= 25.03° \pm (2.797)(0.27°C) \\ &= 25.03°C \pm 0.76°C \end{aligned}$$

$$\begin{aligned} 99\%\text{ confidence limits: } L_1 &= 25.03°C - 0.76°C = 24.27°C \\ L_2 &= 25.03°C + 0.76°C = 25.79°C \end{aligned}$$

In both parts (a) and (b), the hypothesized value, $\mu_0 = 24.3°C$ in Example 7.1, lies outside the confidence intervals. This indicates that H_0 would be rejected using either the 5% or the 1% level of significance.

*We owe the development of confidence intervals to Jerzy Neyman, between 1928 and 1933 (Wang, 2000), although the concept had been enunciated a hundred years before. Neyman introduced the terms *confidence interval* and *confidence coefficient* in 1934 (David, 1995). On rare occasion, the biologist may see reference to "fiducial intervals," a concept developed by R. A. Fisher beginning in 1930 and identical to confidence intervals in some, but not all, situations (Pfanzagl, 1978).

The smaller $s_{\overline{Y}}$ is, the smaller will be the confidence interval, meaning that μ is estimated more precisely when $s_{\overline{Y}}$ is small. Also, it can be observed from the calculation of $s_{\overline{Y}}$ (see Equation 6.8) that a large n will result in a small $s_{\overline{Y}}$ and, therefore, a narrower confidence interval. As introduced in Section 6.4, a parameter estimate from a large sample is generally more precise that an estimate of the same parameter from a small sample.

Setting confidence limits around μ has the same underlying assumptions as the testing of hypotheses about μ (Section 7.1a), and violating those assumptions can invalidate the stated level of confidence $(1 - \alpha)$.

If, instead of a 95% confidence interval, it is desired to state a higher level of confidence, say 99%, that L_1 and L_2 encompass the population mean, then $t_{0.01(1),24}$ rather than $t_{0.01(2),24}$ would be employed. From Appendix Table B.3 we find that $t_{0.01(1),24} = 2.797$, so the 99% confidence interval would be calculated as shown in Example 7.5, part (b), where it is determined that $P(24.27^\circ\text{C} \leq \mu \leq 25.79^\circ\text{C}) = 0.99$.

(a) One-Tailed Confidence Limits. As introduced in Section 6.4(a), one-tailed confidence intervals are appropriate in situations that warrant one-tailed hypothesis tests. Such a confidence interval employs a one-tailed critical value of t (i.e., $t_{\alpha(1),\nu}$) instead of a two-tailed critical value ($t_{\alpha(2),\nu}$). For H_0: $\mu \leq \mu_0$ and H_A: $\mu > \mu_0$, the confidence limits for μ are $L_1 = \overline{Y} - t_{\alpha(1),\nu} s_{\overline{Y}}$ and $L_2 = \infty$; and for H_0: $\mu \geq \mu_0$ and H_A: $\mu < \mu_0$, the confidence limits are $L_1 = -\infty$ and $L_2 = \overline{Y} + t_{\alpha(1),\nu} s_{\overline{Y}}$. For the situation in Example 7.4, in which H_0: $\mu \leq 45$ sec and H_A: $\mu > 45$ sec, L_1 would be 45.21 sec − (1.895)(0.58 sec) = 45.21 sec − 1.10 sec = 44.11 and $L_2 = \infty$. And the hypothesized μ_0 (45 sec) lies within the confidence interval, indicating that the null hypothesis is not rejected.

(b) Prediction Limits. While confidence limits express the precision with which a population characteristic is estimated, we can also indicate the precision with which future observations from this population can be predicted.

After calculating $\overline{Y}$ and s^2 from a random sample of n data from a population, we can ask what the mean would be from an additional random sample, of an additional m data, from the same population. The best estimate of the mean of those m additional data would be $\overline{Y}$, and the precision of that estimate may be expressed by this two-tailed *prediction interval* (abbreviated PI):

$$\overline{Y} \pm t_{\alpha(2),\nu} \sqrt{\frac{s^2}{m} + \frac{s^2}{n}}, \tag{7.7}$$

where $\nu = n - 1$ (Hahn and Meeker, 1991: 61–62). If the desire is to predict the value of one additional datum from that population (i.e., $m = 1$), then Equation 7.7 becomes

$$\overline{Y} \pm t_{\alpha(2),\nu} \sqrt{s^2 + \frac{s^2}{n}}. \tag{7.8}$$

The prediction interval will be wider than the confidence interval and will approach the confidence interval as m becomes very large. The use of Equations 7.7 and 7.8 is demonstrated in Example 7.6.

One-tailed prediction intervals are not commonly obtained but are presented in Hahn and Meeker (1991: 63), who also consider another kind of interval: the

tolerance interval, which may be calculated to contain at least a specified proportion (e.g., a specified percentile) of the sampled population; and in Patel (1989), who also discusses simultaneous prediction intervals of means from more than one future sample. The procedure is very much like that of Section 7.3a. If the desire is to obtain only a lower prediction limit, L_1 (while L_2 is considered to be ∞), then the first portion of Equation 7.7 (or 7.8) would be modified to be $\overline{X} - t_{\alpha(1),\nu}$ (i.e., the one-tailed t would be used); and if the intent is to express only an upper prediction limit, L_2 (while regarding L_1 to be $-\infty$), then we would use $\overline{X} + t_{\alpha(1),\nu}$. As an example, Example 7.6 might have asked what the highest mean body temperature is that would be predicted, with probability α, from an additional sample. This would involve calculating L_2 as indicated above, while $L_1 = -\infty$.

EXAMPLE 7.6 Prediction Limits for Additional Sampling from the Population Sampled in Example 7.1

From Example 7.1, which is a sample of 25 crab body temperatures,

$$n = 25,\ \overline{X} = 25.03°\text{C, and } s^2 = 1.80(°\text{C})^2.$$

(a) If we intend to collect 8 additional crab body temperatures from the same population from which the 25 data in Example 7.1 came, then (by Equation 7.7) we can be 95% confident that the mean of those 8 data will be within this prediction interval:

$$\begin{aligned} 25.03°\text{C} \pm t_{0.05(2),24}\sqrt{\frac{1.80(°\text{C})^2}{8} + \frac{1.80(°\text{C})^2}{2}} &= 25.03°\text{C} \pm 2.064(0.545°\text{C}) \\ &= 25.03°\text{C} \pm 1.12°\text{C}. \end{aligned}$$

Therefore, the 95% prediction limits for the predicted mean of these additional data are $L_1 = 23.91°\text{C}$ and $L_2 = 26.15°\text{C}$.

(b) If we intend to collect 1 additional crab body temperature from the same population from which the 25 data in Example 7.1 came, then (by Equation 7.8) we can be 95% confident that the additional datum will be within this prediction interval:

$$\begin{aligned} 25.03°\text{C} \pm t_{0.05(2),24}\sqrt{1.80(°\text{C})^2 + \frac{1.80(°\text{C})^2}{2}} &= 25.03°\text{C} \pm 2.064(1.368°\text{C}) \\ &= 25.03°\text{C} \pm 2.82°\text{C}. \end{aligned}$$

Therefore, the 95% prediction limits for this predicted datum are $L_1 = 22.21°\text{C}$ and $L_2 = 27.85°\text{C}$.

7.4 REPORTING VARIABILITY AROUND THE MEAN

It is very important to provide the reader of a research paper with information concerning the variability of the data reported. But authors of such papers are often unsure of appropriate ways of doing so, and not infrequently do so improperly.

If we wish to describe the population that has been sampled, then the sample mean ($\overline{X}$) and the standard deviation (s) may be reported. The range might also be reported, but in general it should not be stated without being accompanied by another measure of variability, such as s. Such statistics are frequently presented as in Table 7.1 or 7.2.

TABLE 7.1: Tail Lengths (in mm) of Field Mice from Different Localities

Location	n	$\overline{X} \pm$ SD (range in parentheses)
Bedford, Indiana	18	56.22 ± 1.33 (44.8 to 68.9)
Rochester, Minnesota	12	59.61 ± 0.82 (43.9 to 69.8)
Fairfield, Iowa	16	60.20 ± 0.92 (52.4 to 69.2)
Pratt, Kansas	16	53.93 ± 1.24 (46.1 to 63.6)
Mount Pleasant, Michigan	13	55.85 ± 0.90 (46.7 to 64.8)

TABLE 7.2: Evaporative Water Loss of a Small Mammal at Various Air Temperatures. Sample Statistics Are Mean ± Standard Deviation, with Range in Parentheses

	Air Temperature (°C)				
	16.2	24.8	30.7	36.8	40.9
Sample size	10	13	10	8	9
Evaporative water loss (mg/g/hr)	0.611 ± 0.164 (0.49 to 0.88)	0.643 ± 0.194 (0.38 to 1.13)	0.890 ± 0.212 (0.64 to 1.39)	1.981 ± 0.230 (1.50 to 2.36)	3.762 ± 0.641 (3.16 to 5.35)

If it is the author's intention to provide the reader with a statement about the precision of estimation of the population mean, the use of the standard error ($s_{\overline{X}}$) is appropriate. A typical presentation is shown in Table 7.3a. This table might instead be set up to show confidence intervals, rather than standard errors, as shown in Table 7.3b. The standard error is always smaller than the standard deviation. But this is not a reason to report the former in preference to the latter. The determination should be made on the basis of whether the desire is to describe variability within the population or precision of estimating the population mean.

There are three very important points to note about Tables 7.1, 7.2, 7.3a, and 7.3b. First, n should be stated somewhere in the table, either in the caption or in the body of the table. (Thus, the reader has the needed information to convert from SD to SE or from SE to SD, if so desired.) One should always state n when presenting sample statistics ($\overline{X}$, s, $s_{\overline{X}}$, range, etc.), and if a tabular presentation is prepared, it is very good practice to include n somewhere in the table, even if it is mentioned elsewhere in the paper.

Second, the measure of variability is clearly indicated. Not infrequently, an author will state something such as "the mean is 54.2 ± 2.7 g," with no explanation of what "± 2.7" denotes. This renders the statement worthless to the reader, because "± 2.7" will be assumed by some to indicate ± SD, by others to indicate ± SE, by others to

TABLE 7.3a: Enzyme Activities in the Muscle of Various Animals. Data Are $\bar{X} \pm$ SE, with n in Parentheses

	Enzyme Activity (μmole/min/g of tissue)	
Animal	Isomerase	Transketolase
Mouse	0.76 ± 0.09 (4)	0.39 ± 0.04 (4)
Frog	1.53 ± 0.08 (4)	0.18 ± 0.02 (4)
Trout	1.06 ± 0.12 (4)	0.24 ± 0.04 (4)
Crayfish	4.22 ± 0.30 (4)	0.26 ± 0.05 (4)

TABLE 7.3b: Enzyme Activities in the Muscle of Various Animals. Data Are $\bar{X} \pm$ 95% Confidence Limits

		Enzyme Activity (μmole/min/g of tissue)	
Animal	n	Isomerase	Transketolase
Mouse	4	0.76 ± 0.28	0.39 ± 0.13
Frog	4	1.53 ± 0.25	0.18 ± 0.05
Trout	4	1.06 ± 0.38	0.24 ± 0.11
Crayfish	4	4.22 ± 0.98	0.26 ± 0.15

indicate the 95% (or 99%, or other) confidence interval, and by others to indicate the range.* There is no widely accepted convention; one *must* state explicitly what quantity is meant by this type of statement. If such statements of '±" values appear in a table, then the explanation is best included somewhere in the table (either in the caption or in the body of the table), even if it is stated elsewhere in the paper.

Third, the units of measurement of the variable must be clear. There is little information conveyed by stating that the tail lengths of 24 birds have a mean of 8.42 and a standard error of 0.86 if the reader does not know whether the tail lengths were measured in centimeters, or inches, or some other unit. Whenever data appear in tables, the units of measurement should be stated somewhere in the table. Keep in mind that a table should be self-explanatory; one should not have to refer back and forth between the table and the text to determine what the tabled values represent.

Frequently, the types of information given in Tables 7.1, 7.2, 7.3a, and 7.3b are presented in graphs, rather than in tables. In such cases, the measurement scale is typically indicated on the vertical axis, and the mean is indicated in the body of the graph by a short horizontal line or some other symbol. The standard deviation, standard error, or a confidence interval for the mean is commonly indicated on such graphs via a vertical line or rectangle. Often the range is also included, and in such instances the SD or SE may be indicated by a vertical rectangle and the range by a vertical line. Some authors will indicate a confidence interval (generally 95%) in

*In older literature the ± symbol referred to yet another measure, known as the "probable error" (which fell into disuse in the early twentieth century). In a normal curve, the probable error (PE) is 0.6745 times the standard error, because $\bar{X} \pm$ PE includes 50% of the distribution. The term *probable error* was first used in 1815 by German astronomer Friedrich Wilhelm Bessel (1784–1846) (Walker, 1929: 24, 51, 186).

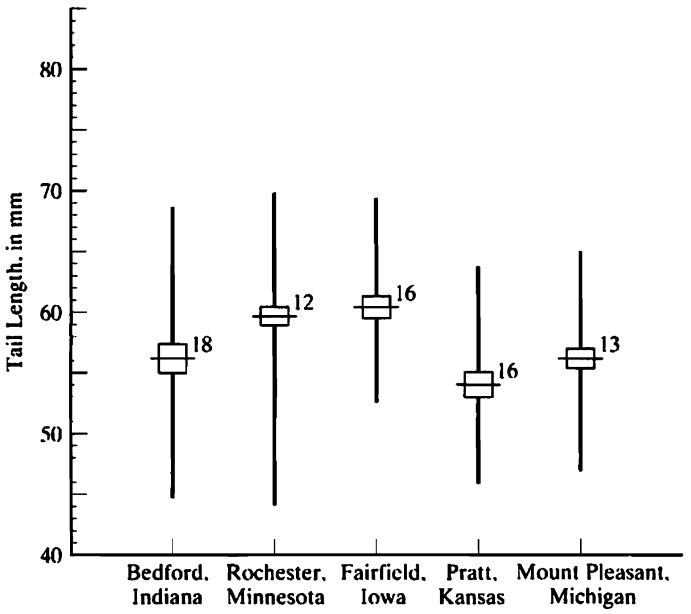


FIGURE 7.4: Tail lengths of male field mice from different localities, indicating the mean, the mean ± standard deviation (vertical rectangle), and the range (vertical line), with the sample size indicated for each location. The data are from Table 7.1.

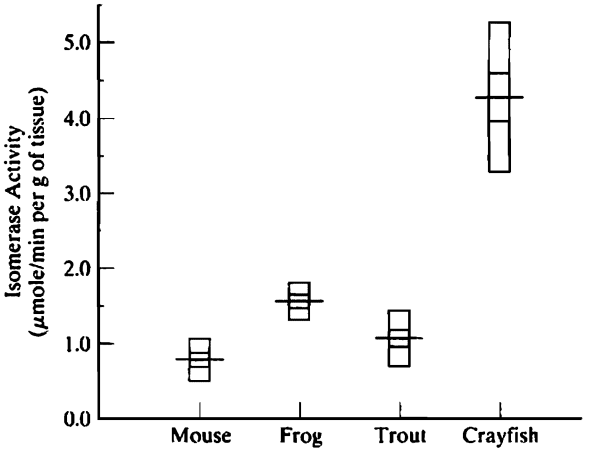


FIGURE 7.5: Levels of muscle isomerase in various animals. Shown is the mean ± standard error (shaded rectangle), and ± the 95% confidence interval (open rectangle). For each sample, $n = 4$. The data are from Tables 7.3a and 7.3b.

addition to the range and either SD or SE. Figures 7.4, 7.5, and 7.6 demonstrate how various combinations of these statistics may be presented graphically.

Instead of the mean and a measure of variability based on the variance, one may present tabular or graphical descriptions of samples using the median and quartiles (e.g., McGill, Tukey, and Larsen, 1978), or the median and its confidence interval. Thus, a graphical presentation such as in Figure 7.4 could have the range indicated by the vertical line, the median by the horizontal line, and the semiquartile range (Section 4.2) by the vertical rectangle. Such a graph is discussed in Section 7.5. Note that when the horizontal axis on the graph represents an interval or ratio scale

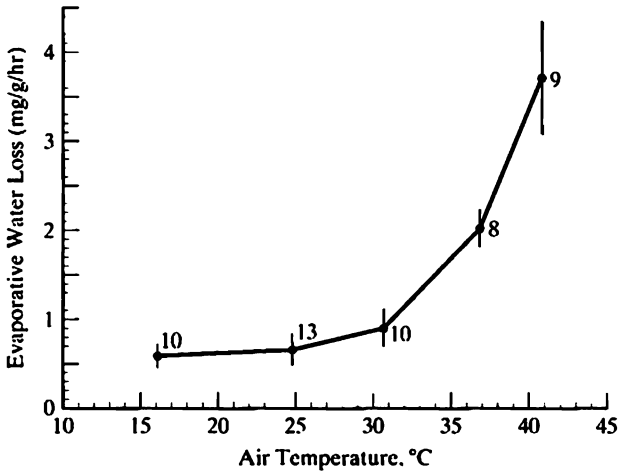


FIGURE 7.6: Evaporative water loss of a small mammal at various air temperatures. Shown at each temperature is the mean ± the standard deviation, and the sample size. The data are from Table 7.2.

variable (as in Figure 7.6), adjacent means may be connected by straight lines to aid in the recognition of trends.

In graphical presentation of data, as in tabular presentation, care must be taken to indicate clearly the following either on the graph or in the caption: The sample size (n), the units of measurement, and what measures of variability (if any) are indicated (e.g., SD, SE, range, 95% confidence interval).

Some authors present $\overline{X} \pm 2s_{\overline{X}}$ in their graphs. An examination of the t table (Appendix Table B.3) will show that, except for small samples, this expression will approximate the 95% confidence interval for the mean. But for small samples, the true confidence interval is, in fact, greater than $\overline{X} \pm 2s_{\overline{X}}$. Thus, the general use of this expression is not to be encouraged, and the calculation of the accurate confidence interval is the wiser practice.

A word of caution is in order for those who determine confidence limits, or SDs or SEs, for two or more means and, by observing whether or not the limits overlap, attempt to determine whether there are differences among the population means. Such a procedure is not generally valid (see Section 8.2); The proper methods for testing for differences between means are discussed in the next several chapters.

7.5 REPORTING VARIABILITY AROUND THE MEDIAN

The median and the lower and upper quartiles (Q_1 and Q_3) form the basis of a graphical presentation that conveys a rapid sense of the middle, the spread, and the symmetry of a set of data. As shown in Figure 7.7, a vertical box is drawn with its bottom at Q_1 and top at Q_3, meaning that the height of the box is the semi-quartile range (Q_3-Q_1). Then, the median is indicated by a horizontal line across the box. Next, a vertical line is extended from the bottom of the box to the smallest datum that is no farther from the box than 1.5 times the interquartile range; and a vertical line is drawn from the top of the box to the largest datum that is no farther from the box than 1.5 times the interquartile range. These two vertical lines, below and above the box, are termed "whiskers," so this graphical representation is called a *box plot* or *box-and-whiskers plot.** If any data are so deviant as to lie beyond the whiskers, they

*The term *box plot* was introduced by John W. Tukey in 1970 (David, 1995).

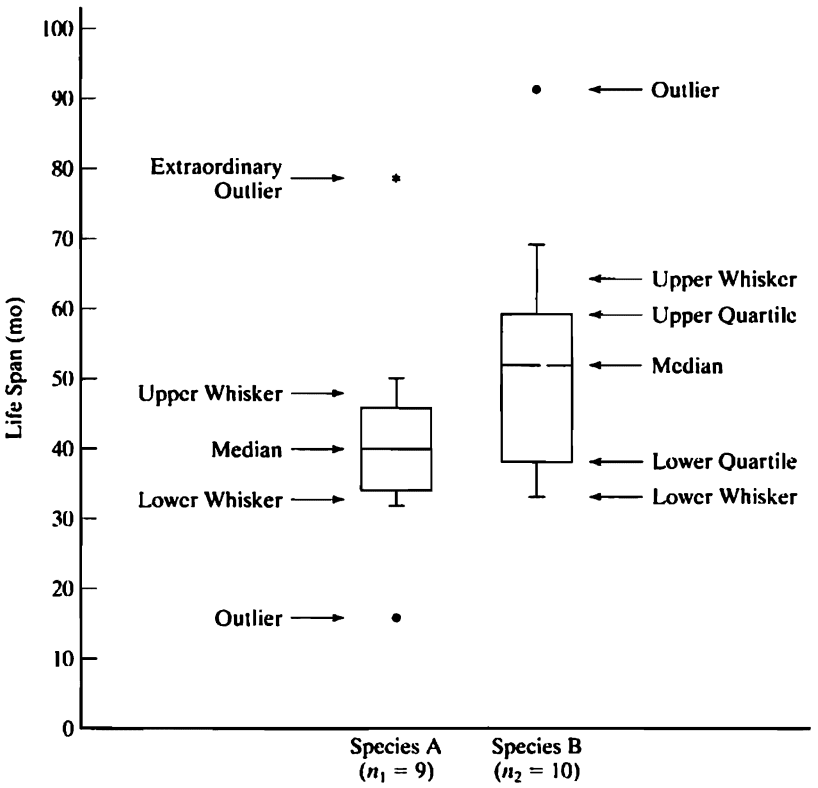


FIGURE 7.7: Box plots for the data of Example 3.3. (The wording with arrows is for instructional purposes and would not otherwise appear in such a graph.)

are termed *outliers* and are placed individually as small circles on the graph. If any are so aberrant as to lie at least 3 times the interquartile range from the box (let's call them "extraordinary outliers"), they may be placed on the graph with a distinctive symbol (such as "*") instead of a circle.* In addition, the size of the data set (n) should be indicated, either near the box plot itself or in the caption accompanying the plot.

Figure 7.7 presents a box plot for the two samples in Example 3.3. For species A, the median = 40 mo, $Q_1 = Q_{10/4} = X_{2.5} = 34.5$ mo, and $Q_3 = X_{10-2.5} = X_{7.5} = 46$ mo. The interquartile range is 46 mo − 34.5 mo = 11.5 mo, so 1.5 times the interquartile range is (1.5)(11.5 mo) = 17.25 mo, and 3 times the interquartile range is (3)(11.5) = 34.5 mo. Therefore, the upper whisker extends from the top of the box up to the largest datum that does not exceed 46 mo + 17.25 mo = 63.25 mo (and that datum is $X_8 = 50$ mo), and the lower whisker extends from the bottom of the box down to the smallest datum that is no smaller than 34.5 mo − 17.25 mo = 17.25 mo (namely, $X_2 = 32$ mo). Two of the data, $X_1 = 16$ mo and $X_9 = 79$ mo, lie farther from the box than the whiskers; thus they are outliers and, as X_9 lies more than 3 times the interquartile range from the box (i.e., is more than 34.5 mo greater than Q_3), it is an extraordinary outlier. Therefore, X_1 is indicated with a circle below the box and X_9 is denoted with a "*" above the box.

*The vertical distances above and below the box by an amount 1.5 times the interquartile range are sometimes called *inner fences*, with those using the factor of 3 being called *outer fences*. Also, the top (Q_3) and bottom (Q_1) of the box are sometimes called the *hinges* of the plot.

For species B, the median = 52 mo, $Q_1 = Q_{11/4} = Q_{2.75}$, which is rounded up to Q_1 = 38 mo, and $Q_3 = X_{11-3} = X_8$ = 59 mo. The interquartile range is 59 mo − 38 mo = 21 mo, so (1.5)(21 mo) = 31.5 mo and (3)(21 mo) = 63 mo. Thus, the upper whisker extends from the box up to the largest datum that does not exceed 59 mo + 21 mo = 80 mo (namely, X_9 = 69 mo), and the lower whisker extends from the box down to the smallest datum that is no smaller than 38 mo − 21 mo = 17 mo (namely, X_1 = 34 mo). As only X_{10} = 91 lies farther from the box than the whiskers, it is the only outlier in the sample of data for species B; it is not an extraordinary outlier because it is not (3)(21 mo), namely 63 mo, above the box.

Box plots are especially useful in visually comparing two or more sets of data. In Figure 7.7, we can quickly discern from the horizontal lines representing the medians that, compared to species A, species B has a greater median life span; and, as the box for species B is larger, that species' sample displays greater variability in life spans. Furthermore, it can be observed that species B has its median farther from the middle of the box, and an upper whisker much longer than the lower whisker, indicating that the distribution of life spans for this species is more skewed toward longer life than is the distribution for species A.

David (1995) attributes the 1970 introduction of box plots to J. W. Tukey, and the capability to produce such graphs appears in many computer software packages. Some authors and some statistical software have used multiplication factors other than 1.5 and 3 to define outliers, some have proposed modifications of box plots to provide additional information (e.g., making the width of each box proportional to the number of data, or to the square root of that number), and some employ quartile determination different from that in Section 4.2. Indeed, Frigge, Hoaglin, and Iglewicz (1989) report that, although common statistical software packages only rarely define the median (Q_2) differently than that presented in Section 3.2, they identified eight ways Q_1 and Q_3 are calculated in various packages. Unfortunately, the different presentations of box plots provide different impressions of the data, and some of the methods of expressing quartiles are not recommended by the latter authors.

7.6 SAMPLE SIZE AND ESTIMATION OF THE POPULATION MEAN

A commonly asked question is, "How large a sample must be taken to achieve a desired precision* in estimating the mean of a population?" The answer is related to the concept of a confidence interval, for a confidence interval expresses the precision of a sample statistic, and the precision increases (i.e., the confidence interval becomes narrower) as the sample size increases.

Let us write Equation 7.6 as $\overline{X} \pm d$, which is to say that $d = t_{\alpha(2),\nu} s_{\overline{X}}$. We shall refer to d as the half-width of the confidence interval, which means that μ is estimated to within $\pm d$. Now, the number of data we must collect to calculate a confidence interval of specified width depends upon: (1) the width desired (for a narrower confidence interval—i.e., more precision in estimating μ—requires a larger sample; (2) the variability in the population (which is estimated by s^2, and larger variability requires larger sample size); and (3) the confidence level specified (for greater confidence—e.g., 99% vs. 95%—requires a larger sample size).

*Recall from Section 2.4 that the precision of a sample statistic is the closeness with which it estimates the population parameter; it is not to be confused with the concept of the precision of a measurement (defined in Section 1.2), which is the nearness of repeated measurements to each

If we have a sample estimate (s^2) of the variance of a normal population, then we can estimate the required sample size for a future sample as

$$n = \frac{s^2 t^2_{\alpha(2),\nu}}{d^2}. \tag{7.9}$$

In this equation, s^2 is the sample variance, estimated with $\nu = n - 1$ degrees of freedom, d is the half-width of the desired confidence interval, and $1 - \alpha$ is the confidence level for the confidence interval. Two-tailed critical values of Student's t, with $\nu = n - 1$ degrees of freedom, are found in Appendix Table B.3.

There is a basic difficulty in solving Equation 7.9, however; the value of $t_{\alpha(2),(n-1)}$ depends upon n, the unknown sample size. The solution may be achieved by iteration—a process of trial and error with progressively more accurate approximations—as shown in Example 7.7. We begin the iterative process of estimation with an initial guess; the closer this initial guess is to the finally determined n, the faster we shall arrive at the final estimate. Fortunately, the procedure works well even if this initial guess is far from the final n (although the process is faster if it is a high, rather than a low, guess).

The reliability of this estimate of n depends upon the accuracy of s^2 as an estimate of the population variance, σ^2. As its accuracy improves with larger samples, one should use s^2 obtained from a sample with a size that is not a very small fraction of the n calculated from Equation 7.9.

EXAMPLE 7.7 Determination of Sample Size Needed to Achieve a Stated Precision in Estimating a Population Mean, Using the Data of Example 7.3

If we specify that we wish to estimate μ with a 95% confidence interval no wider than 0.5 kg, then $d = 0.25$ kg, $1 - \alpha = 0.95$, and $\alpha = 0.05$. From Example 7.3 we have an estimate of the population variance: $s^2 = 0.4008\ \text{kg}^2$.

Let us guess that a sample of 40 is necessary; then,

$$t_{0.05(2),39} = 2.023.$$

So we estimate (by Equation 7.7):

$$n = \frac{(0.4008)(2.023)^2}{(0.25)^2} = 26.2.$$

Next, we might estimate $n = 27$, for which $t_{0.05(2),26} = 2.056$, and we calculate

$$n = \frac{(0.4008)(2.056)^2}{(0.25)^2} = 27.1.$$

Therefore, we conclude that a sample size greater than 27 is required to achieve the specified confidence interval.

7.7 SAMPLE SIZE, DETECTABLE DIFFERENCE, AND POWER IN TESTS CONCERNING THE MEAN

(a) Sample Size Required. If we are to perform a one-sample test as described in Section 7.1 or 7.2, then it is desirable to know how many data should be collected

to detect a specified difference with a specified power. An estimate of the minimum sample size (n) required will depend upon σ^2, the population variance (which can be estimated by s_p^2 from previous similar studies). The beginning of Section 8.4 lists considerations (based upon Lenth, 2001) relevant to the determination of sample size.

We may specify that we wish to perform a t test with a probability of α of committing a Type I error and a probability of β of committing a Type II error; and we can state that we want to be able to detect a difference between μ and μ_0 as small as δ (where μ is the actual population mean and μ_0 is the mean specified in the null hypothesis).* To test at the α significance level with $1 - \beta$ power, the minimum sample size required to detect δ is

$$n = \frac{s^2}{\delta^2}(t_{\alpha,\nu} + t_{\beta(1),\nu})^2, \tag{7.10}$$

where α can be either $\alpha(1)$ or $\alpha(2)$, respectively, depending on whether a one-tailed or two-tailed test is to be used. However, ν depends on n, so n cannot be calculated directly but must be obtained by iteration† (i.e., by a series of estimations, each estimation coming closer to the answer than that preceding). This is demonstrated in Example 7.8.

Equation 7.10 provides better estimates of n when s^2 is a good estimate of the population variance, σ^2, and the latter estimate improves when s^2 is calculated from larger samples. Therefore, it is most desirable that s^2 be obtained from a sample with a size that is not a small fraction of the estimate of n; and it can then be estimated how large an n is needed to repeat the experiment and use the resulting data to test with the designated α, β, and δ.

EXAMPLE 7.8 Estimation of Required Sample Size to Test H_0: $\mu = \mu_0$

How large a sample is needed to reject the null hypothesis of Example 7.2 when sampling from the population in that example? We wish to test at the 0.05 level of significance with a 90% chance of detecting a population mean different from $\mu_0 = 0$ by as little as 1.0 g. In Example 7.2, $s^2 = 1.5682\ \text{g}^2$.

Let us guess that a sample size of 20 would be required. Then, $\nu = 19$, $t_{0.05(2),19} = 2.093$, $\beta = 1 - 0.90 = 0.10$, $t_{0.10(1),19} = 1.328$, and we use Equation 7.8 to calculate

$$n = \frac{1.5682}{(1.0)^2}(2.093 + 1.328)^2 = 18.4.$$

We now use $n = 19$ as an estimate, in which case $\nu = 18$, $t_{0.05(2),18} = 2.101$, $t_{0.10(1),18} = 1.330$, and

$$n = \frac{1.5682}{(1.0)^2}(2.101 + 1.330)^2 = 18.5.$$

Thus, we conclude that a new sample of at least 19 data may be taken from this population to test the above hypotheses with the specified α, β, and δ.

*δ is lowercase Greek delta.

†If the population variance, σ^2, were actually known (a most unlikely situation), rather than estimated by s^2, then Z_α would be substituted for t_α in this and the other computations in this section, and n would be determined in one step instead of iteratively.

(b) Minimum Detectable Difference. By rearranging Equation 7.10, we can ask how small a δ (the difference between μ and μ_0) can be detected by the t test with $1 - \beta$ power, at the α level of significance, using a sample of specified size n:

$$\delta = \sqrt{\frac{s^2}{n}}\left(t_{\alpha,\nu} + t_{\beta(1),\nu}\right), \tag{7.11}$$

where $t_{\alpha,\nu}$ can be either $t_{\alpha(1),\nu}$ or $t_{\alpha(2),\nu}$, depending on whether a one-tailed or two-tailed test is to be performed. The estimation of δ is demonstrated in Example 7.9. Some literature (e.g., Cohen, 1988: 811–814) refers to the "effect size," a concept similar to minimum detectable difference.

EXAMPLE 7.9 Estimation of Minimum Detectable Difference in a One-Sample t Test for H_0: $\mu = \mu_0$

In the two-tailed test of Example 7.2, what is the smallest difference (i.e., difference between μ and μ_0) that is detectable 90% of the time using a sample of 25 data and a significance level of 0.05?

Using Equation 7.9:

$$\begin{aligned}\delta &= \sqrt{\frac{1.5682}{25}}\left(t_{0.05(2),24} + t_{0.10(1),24}\right)\\ &= (0.25)(2.064 + 1.318)\\ &= 0.85 \text{ g.}\end{aligned}$$

(c) Power of One-Sample Testing. If our desire is to express the probability of correctly rejecting a false H_0 about μ, then we seek to estimate the power of a t test. Equation 7.10 can be rearranged to give

$$t_{\beta(1),\nu} = \frac{\delta}{\sqrt{\frac{s^2}{n}}} - t_{\alpha,\nu}, \tag{7.12}$$

where α refers to either $\alpha(2)$ or $\alpha(1)$, depending upon whether the null hypothesis to be tested is two-tailed or one-tailed, respectively. As shown in Example 7.10, for a stipulated δ, α, s^2, and sample size, we can express $t_{\beta(1),\nu}$. Consulting Appendix Table B.3 allows us to convert $t_{\beta(1),\nu}$ to β, but only roughly (e.g., $\beta > 0.25$ in Example 7.10). However, $t_{\beta(1),\nu}$ may be considered to be approximated by $Z_{\beta(1)}$, so Appendix Table B.2 may be used to determine β.* Then, the power of the test is expected to be $1 - \beta$, as shown in Example 7.10. Note that this is the estimated power of a test to be run on a new sample of data from this population, *not* the power of the test performed in Example 7.2.

*Some calculators and computer programs yield β given $t_{\beta,\nu}$. Approximating $t_{\beta(1),\nu}$ by $Z_{\beta(1)}$ apparently yields a β that is an underestimate (and a power that is an overestimate) of no more than 0.01 for ν of at least 11 and no more than 0.02 for ν of at least 7.

EXAMPLE 7.10 Estimation of the Power of a One-Sample *t* Test for H_0: $\mu = \mu_0$

What is the probability of detecting a true difference (i.e., a difference between μ and μ_0) of at least 1.0 g, using $\alpha = 0.05$ for the hypotheses of Example 7.2, if we run the experiment again using a sample of 15 from the same population?

For $n = 15$, $\nu = 14$; $\alpha = 0.05$, $t_{0.05(2),14} = 2.145$, $s^2 = 1.5682\ \text{g}^2$, and $\delta = 1.0$ g; and we use Equation 7.12 to find

$$t_{\beta(1),14} = \frac{1.0}{\sqrt{\frac{1.5682\ \text{g}^2}{15}}} - 2.145$$
$$= 0.948.$$

Consulting Appendix Table B.3 tells us that, for $t_{\beta(1),14} = 0.948$, $0.10 < \beta < 0.25$, so we can say that the power would be $0.75 < 1 - \beta < 0.90$. Alternatively, by considering 0.948 to be a normal deviate and consulting Appendix Table B.2, we conclude that $\beta = 0.17$ and that the power of the test is $1 - \beta = 0.83$. (The exact probabilities, by computer, are $\beta = 0.18$ and power = 0.82.)

When the concept of power was introduced in the discussion "Statistical Errors in Hypothesis Testing" in Section 6.4, it was stated that, for a given sample size (n), α is inversely related to β; that is, the lower the probability of committing a Type I error, the greater the probability of committing a Type II error. It was also noted that α and β can be lowered simultaneously by increasing n. Power is also greater for one-tailed than for two-tailed tests, but recall (from the end of Section 6.4 and from Section 7.2) that power is *not* the criterion for performing a one-tailed instead of a two-tailed test. These relationships are shown in Table 7.4. Table 7.5 shows how power is related to n, s^2, and δ. It can be seen that, for a given s^2 and δ, an increased sample size (n) results in an increase in power. Also, for a given n and δ, power increases as s^2 decreases, so a smaller variability among the data yields greater power. And for a given n and s^2, power increases as δ increases, meaning there is greater power in detecting large differences than there is in detecting small differences.

Often a smaller s^2 is obtained by narrowing the definition of the population of interest. For example, the data of Example 7.2 may vary as much as they do because the sample contains animals of different ages, or of different strains, or of both sexes. It may be wiser to limit the hypothesis, and the sampling, to animals of the same sex and strain and of a narrow range of ages. And power can be increased by obtaining more precise measurements; also, greater power is associated with narrower confidence intervals. A common goal is to test with a power between 0.75 and 0.90.

7.8 SAMPLING FINITE POPULATIONS

In general we assume that a sample from a population is a very small portion of the totality of data in that population. Essentially, we consider that the population is infinite in size, so that the removal of a relatively small number of data from the population does not noticeably affect the probability of selecting further data.

However, if the sample size, n, is an appreciable portion of the population size (a very unusual circumstance), N (say, at least 5%), then we are said to be sampling

TABLE 7.4: Relationship between α, β, Power $(1 - \beta)$, and n, for the Data of Example 7.9
Sample Variance ($s^2 = 1.5682$ g^2) and True Difference ($\delta = 1.0$ g) of Example 7.10, Using Equation 7.12

Two-Tailed Test				One-Tailed Test			
n	α	β	$1 - \beta$	n	α	β	$1 - \beta$
10	0.10	0.25	0.75	10	0.10	0.14	0.86
10	0.05	0.40	0.60	10	0.05	0.25	0.75
10	0.01	0.76	0.24	10	0.01	0.61	0.39
12	0.10	0.18	0.82	12	0.10	0.09	0.91
12	0.05	0.29	0.71	12	0.05	0.18	0.82
12	0.01	0.73	0.27	12	0.01	0.48	0.52
15	0.10	0.10	0.90	15	0.10	0.05	0.95
15	0.05	0.18	0.82	15	0.05	0.10	0.90
15	0.01	0.45	0.55	15	0.01	0.32	0.68
20	0.10	0.04	0.96	20	0.10	0.02	0.98
20	0.05	0.08	0.92	20	0.05	0.04	0.96
20	0.01	0.24	0.76	20	0.01	0.16	0.84

TABLE 7.5: Relationship between n, s^2, δ, and Power (for Testing at $\alpha = 0.05$) for the Hypothesis of Example 7.9, Using Equation 7.10

n	s^2	δ	Power of Two-Tailed Test	Power of One-Tailed Test
Effect of n				
10	1.5682	1.0	0.60	0.75
12	1.5682	1.0	0.71	0.82
15	1.5682	1.0	0.82	0.90
20	1.5682	1.0	0.92	0.96
Effect of s^2				
12	2.0000	1.0	0.60	0.74
12	1.5682	1.0	0.71	0.82
12	1.0000	1.0	0.88	0.94
Effect of δ				
12	1.5682	1.0	0.71	0.82
12	1.5682	1.2	0.86	0.92
12	1.5682	1.4	0.96	0.97

a *finite population*. In such a case, $\overline{X}$ is a substantially better estimate of μ the closer n is to N; specifically,

$$s_{\overline{X}} = \sqrt{\frac{s^2}{n}}\sqrt{1 - \frac{n}{N}} = \sqrt{\left(\frac{s^2}{n}\right)\left(1 - \frac{n}{N}\right)}, \tag{7.13}$$

where n/N is the *sampling fraction* and $1 - n/N$ is referred to as the *finite population correction*.*

Obviously, from Equation 7.13, when n is very small compared to N, then the sampling fraction is almost zero, the finite population correction will be nearly one,

*One may also calculate $1 - n/N$ as $(N - n)/N$.

and $s_{\overline{X}}$ will be nearly $\sqrt{s^2/n}$, just as we have used (Equation 6.8) when assuming the population size, N, to be infinite. As n becomes closer to N, the correction becomes smaller, and $s_{\overline{X}}$ becomes smaller, which makes sense intuitively. If $n = N$, then $1 - n/N = 0$ and $s_{\overline{X}} = 0$, meaning there is no error at all in estimating μ if the sample consists of the entire population; that is, $\overline{X} = \mu$ if $n = N$. In computing confidence intervals when sampling finite populations (i.e., when n is not a negligibly small fraction of N), Equation 7.13 should be used instead of Equation 6.8.

If we are determining the sample size required to estimate the population mean with a stated precision (Section 7.6), and the sample size is an appreciable fraction of the population size, then the required sample size is calculated as

$$m = \frac{n}{1 + (n - 1)/N} \tag{7.14}$$

(Cochran, 1977: 77–78), where n is from Equation 7.9.

7.9 HYPOTHESES CONCERNING THE MEDIAN

In Example 7.2 we examined a sample of weight change data in order to ask whether the mean change in the sampled population was different from zero. Analogously, we may test hypotheses about the population median, M, such as testing H_0: $M = M_0$ against H_A: $M \neq M_0$, where M_0 can be zero or any other hypothesized population median.*

A simple method for testing this two-tailed hypothesis is to determine the confidence limits for the population median, as discussed in Section 23.9, and reject H_0 (with probability $\leq \alpha$ of a Type I error) if $M_0 \leq L_1$ or $M_0 \geq L_2$. This is essentially a binomial test (Section 23.6), where we consider the number of data $< M_0$ as being in one category and the number of data $> M_0$ being in the second category. If either of these two numbers is less than or equal to the critical value in Appendix Table B.27, then H_0 is rejected. (Data equal to M_0 are ignored in this test.)

For one-tailed hypotheses about the median, the binomial test may also be employed. For H_0: $M \geq M_0$ versus H_A: $M < M_0$, H_0 is rejected if the number of data less than M_0 is $\leq$ the one-tailed critical value, $C_{\alpha(1),n}$. For H_0: $M \leq M_0$ versus H_A: $M > M_0$, H_0 is rejected if the number of data greater than M_0 is $\geq n - C_{\alpha(1),n}$.

As an alternative to the binomial test, for either two-tailed or one-tailed hypotheses, we may use the more powerful Wilcoxon signed-rank test. The Wilcoxon procedure is applied as a one-sample median test by ranking the data as described in Section 9.5 and assigning a minus sign to each rank associated with a datum $< M_0$ and a plus sign to each associated with a datum $> M_0$. Any rank equal to M_0 is ignored in this procedure. The sum of the ranks with a plus sign is called T_+ and the sum of the ranks with a minus sign is T_-, with the test then proceeding as described in Section 9.5. The Wilcoxon test assumes that the sampled population is symmetric (in which case the median and mean are identical and this procedure becomes a hypothesis test about the mean as well as about the median, but the one-sample t test is typically a more powerful test about the mean). Section 9.5 discusses this test further.

7.10 CONFIDENCE LIMITS FOR THE POPULATION MEDIAN

The sample median (Section 3.2) is used as the best estimate of M, the population median. Confidence limits for M may be determined by considering the binomial distribution, as discussed in Section 24.9.

*Here M represents the Greek capital letter mu.

HYPOTHESES CONCERNING THE VARIANCE

The sampling distribution of means is a symmetrical distribution, approaching the normal distribution as n increases. But the sampling distribution of variances is not symmetrical, and neither the normal nor the t distribution may be employed to test hypotheses about σ^2 or to set confidence limits around σ^2. However, theory states that

$$\chi^2 = \frac{\nu s^2}{\sigma^2} \tag{7.15}$$

(if the sample came from a population with a normal distribution), where χ^2 represents a statistical distribution* that, like t, varies with the degrees of freedom, ν, where $\nu = n - 1$. Critical values of $\chi^2_{\alpha,\nu}$ are found in Appendix Table B.1.

Consider the pair of two-tailed hypotheses, H_0: $\sigma^2 = \sigma_0^2$ and H_A: $\sigma^2 \neq \sigma_0^2$, where σ_0^2 may be any hypothesized population variance. Then, simply calculate

$$\chi^2 = \frac{\nu s^2}{\sigma_0^2} \quad \text{or, equivalently,} \quad \chi^2 = \frac{SS}{\sigma_0^2}, \tag{7.16}$$

and if the calculated χ^2 is $\geq \chi^2_{\alpha/2,\nu}$ or $\leq \chi^2_{(1-\alpha/2),\nu}$, then H_0 is rejected at the α level of significance. For example, if we wished to test H_0: $\sigma^2 = 1.0(^\circ C)^2$ and H_A: $\sigma^2 \neq 1.0(^\circ C)^2$ for the data of Example 7.1, with $\alpha = 0.05$, we would first calculate $\chi^2 = SS/\sigma_0^2$. In this example, $\nu = 24$ and $s^2 = 1.80(^\circ C)^2$, so $SS = \nu s^2 = 43.20(^\circ C)^2$. Also, as σ^2 is hypothesized to be $1.0(^\circ C)^2$, $\chi^2 = SS/\sigma_0^2 = 43.20(^\circ C)^2/1.0(^\circ C)^2 = 43.20$. Two critical values are to be obtained from the chi-square table (Appendix Table B.1): $\chi^2_{(0.05/2),24} = \chi^2_{0.025,24} = 39.364$ and $\chi^2_{(1-0.05/2),24} = \chi^2_{0.975,24} = 12.401$. As the calculated χ^2 is more extreme than one of these critical values (i.e., the calculated χ^2 is > 39.364), H_0 is rejected, and we conclude that the sample of data was obtained from a population having a variance different from $1.0(^\circ C)^2$.

It is more common to consider one-tailed hypotheses concerning variances. For the hypotheses H_0: $\sigma^2 \leq \sigma_0^2$ and H_A: $\sigma^2 > \sigma_0^2$, H_0 is rejected if the χ^2 calculated from Equation 7.16 is $\geq \chi^2_{\alpha,\nu}$. For H_0: $\sigma^2 \geq \sigma_0^2$ and H_A: $\sigma^2 < \sigma_0^2$, a calculated χ^2 that is $\leq \chi^2_{(1-\alpha),\nu}$ is grounds for rejecting H_0. For the data of Example 7.4, a manufacturer might be interested in whether the variability in the dissolving times of the drug is greater than a certain value—say, 1.5 sec. Thus, H_0: $\sigma^2 \leq 1.5$ sec^2 and H_A: $\sigma^2 > 1.5$ sec^2 might be tested, as shown in Example 7.11.

EXAMPLE 7.11 A One-Tailed Test for the Hypotheses H_0: $\sigma^2 \leq 1.5$ sec^2 and H_A: $\sigma^2 > 1.5$ sec^2, Using the Data of Example 7.4

$SS = 18.8288$ sec^2

$\nu = 7$

$s^2 = 2.6898$ sec^2

$$\chi^2 = \frac{SS}{\sigma_0^2} = \frac{18.8288 \text{ sec}^2}{1.5 \text{ sec}^2} = 12.553$$

$\chi^2_{0.05,7} = 14.067$

*The Greek letter "chi" (which in lowercase is χ) is pronounced as the "ky" in "sky."

Since $12.553 < 14.067$, H_0 is not rejected.

$$0.05 < P < 0.10 \quad [P = 0.084]$$

We conclude that the variance of dissolving times is no more than 1.5 sec^2.

As is the case for testing hypotheses about a population mean, μ (Sections 7.1 and 7.2), the aforementioned testing of hypotheses about a population variance, σ, depends upon the sample's having come from a population of normally distributed data. However, the F test for variances is not as robust as the t test for means; that is, it is not as resistant to violations of this underlying assumption of normality. The probability of a Type I error will be very different from the specified α if the sampled population is nonnormal, even if it is symmetrical. And, likewise, α will be distorted if there is substantial asymmetry (say, $|\sqrt{b_1}| > 0.6$), even if the distribution is normal (Pearson and Please, 1975).

7.12 CONFIDENCE LIMITS FOR THE POPULATION VARIANCE

Confidence intervals may be determined for many parameters other than the population mean, in order to express the precision of estimates of those parameters.

By employing the χ^2 distribution, we can define an interval within which there is a $1 - \alpha$ chance of including σ^2 in repeated sampling. Appendix Table B.1 tells us the probability of a calculated χ^2 being greater than that in the Table. If we desire to know the two χ^2 values that enclose $1 - \alpha$ of the chi-square curve, we want the portion of the curve between $\chi^2_{(1-\alpha/2),\nu}$ and $\chi^2_{\alpha/2,\nu}$ (for a 95% confidence interval, this would mean the area between $\chi^2_{0.975,\nu}$ and $\chi^2_{0.025,\nu}$). It follows from Equation 7.13 that

$$\chi^2_{(1-\alpha/2),\nu} \leq \frac{\nu s^2}{\sigma^2} \leq \chi^2_{\alpha/2,\nu}, \tag{7.17}$$

and

$$\frac{\nu s^2}{\chi^2_{\alpha/2,\nu}} \leq \sigma^2 \leq \frac{\nu s^2}{\chi^2_{(1-\alpha/2),\nu}}. \tag{7.18}$$

Since $\nu s^2 = SS$, we can also write Equation 7.16 as

$$\frac{SS}{\chi^2_{\alpha/2,\nu}} \leq \sigma^2 \leq \frac{SS}{\chi^2_{(1-\alpha/2),\nu}}. \tag{7.19}$$

Referring back to the data of Example 7.1, we would calculate the 95% confidence interval for σ^2 as follows. As $\nu = 24$ and $s^2 = 1.80(^\circ C)^2$, $SS = \nu s^2 = 43.20(^\circ C)^2$. From Appendix Table B.1, we find $\chi^2_{0.025,24} = 39.364$ and $\chi^2_{0.975,24} = 12.401$. Therefore, $L_1 = SS/\chi^2_{\alpha/2,\nu} = 43.20(^\circ C)^2/39.364 = 1.10(^\circ C)^2$, and $L_2 = SS/\chi^2_{(1-\alpha),\nu} = 43.20(^\circ C)/12.401 = 3.48(^\circ C)^2$. If the null hypothesis H_0: $\sigma^2 = \sigma_0^2$ would have been tested and rejected for some specified variance, σ_0, then σ_0 would be outside of the confidence interval (i.e., σ_0^2 would be either less than L_1 or greater than L_2). Note that the confidence limits, $1.10(^\circ C)^2$ and $3.48(^\circ C)^2$, are not symmetrical around s^2; that is, the distance from L_1 to s^2 is not the same as the distance from s^2 to L_2.

To obtain the $1 - \alpha$ confidence interval for the population standard deviation, simply use the square roots of the confidence limits for σ^2, so that

$$\sqrt{\frac{SS}{\chi^2_{\alpha/2,\nu}}} \le \sigma \le \sqrt{\frac{SS}{\chi^2_{(1-\alpha/2),\nu}}}. \tag{7.20}$$

For the preceding example, the 95% confidence interval for σ would be $\sqrt{1.10(^\circ C)^2} \le \sigma \le \sqrt{3.48(^\circ C)^2}$, or $1.0^\circ C \le \sigma \le 1.9^\circ C$.

The end of Section 7.11 cautioned that testing hypotheses about σ^2 is adversely affected if the sampled population is nonnormal (even if it is symmetrical) or if the population is not symmetrical (even if it is normal). Determination of confidence limits also suffers from this unfavorable effect.

(a) One-Tailed Confidence Limits. In a fashion analogous to estimating a population mean via a one-tailed confidence interval, a one-tailed interval for a population variance is applicable in situations where a one-tailed hypothesis test for the variance is appropriate. For H_0: $\sigma^2 \le \sigma_0^2$ and H_A: $\sigma^2 > \sigma_0^2$, the one-tailed confidence limits for σ^2 are $L_1 = SS/\chi_{\alpha,\nu}$ and $L_2 = \infty$; and for H_0: $\sigma^2 \ge \sigma_0^2$ and H_A: $\sigma^2 < \sigma_0^2$, the confidence limits are $L_1 = 0$ and $L_2 = SS/\chi^2_{1-\alpha,\nu}$. Considering the data in Example 7.4, in which H_0: $\sigma^2 \le 45$ sec^2 and H_A: $\sigma > 45$ sec^2, for 95% confidence, L_1 would be $SS/\chi^2_{0.05,7} = 18.8188 \text{ sec}^2/14.067 = 1.34 \text{ sec}^2$ and $L_2 = \infty$. The hypothesized σ_0^2 (45 sec^2) lies within the confidence interval, indicating that the null hypothesis would not be rejected.

If the desire is to estimate a population's standard deviation (σ) instead of the population variance (σ^2), then simply substitute σ for σ^2 and σ_0 for σ_0^2 above and use the square root of L_1 and L_2 (bearing in mind that $\sqrt{\infty} = \infty$).

(b) Prediction Limits. We can also estimate the variance that would be obtained from an additional random sample of m data from the same population. To do so, the following two-tailed $1 - \alpha$ prediction limits may be determined:

$$L_1 = \frac{s^2}{F_{\alpha(2),n-1,m-1}} \tag{7.21}$$

$$L_2 = s^2 F_{\alpha(2),m-1,n-1} \tag{7.22}$$

(Hahn, 1972; Hahn and Meeker, 1991: 64, who also mention one-tailed prediction intervals; Patel, 1989). A prediction interval for s would be obtained by taking the square roots of the prediction limits for s^2.

The critical values of F, which will be employed many times later in this book, are given in Appendix Table B.4. These will be written in the form F_{α,ν_1,ν_2}, where ν_1 and ν_2 are termed the "numerator degrees of freedom" and "denominator degrees of freedom," respectively (for a reason that will be apparent in Section 8.5). So, if we wished to make a prediction about the variance (or standard deviation) that would be obtained from an additional random sample of 10 data from the population from which the sample in Example 7.1 came, $n = 25$, $n - 1 = 24$, $m = 10$, and $m - 1 = 9$; and to compute the 95% two-tailed prediction interval, we would consult Table B.4 and obtain $F_{\alpha(2),n-1,m-1} = F_{0.05(2),24,9} = 3.61$ and $F_{\alpha(2),m-1,n-1} = F_{0.05(2),9,24} = 2.79$. Thus, the prediction limits would be $L_1 = 1.80(^\circ C)^2/3.61 = 0.50(^\circ C)^2$ and $L_2 = [1.80(^\circ C)^2][2.79] = 5.02(^\circ C)^2$.

7.13 POWER AND SAMPLE SIZE IN TESTS CONCERNING THE VARIANCE

(a) Sample Size Required. We may ask how large a sample would be required to perform the hypothesis tests of Section 7.12 at a specified power. For the hypotheses H_0: $\sigma^2 \leq \sigma_0^2$ versus H_A: $\sigma^2 > \sigma_0^2$, the minimum sample size is that for which

$$\frac{\chi^2_{1-\beta,\nu}}{\chi^2_{\alpha,\nu}} = \frac{\sigma_0^2}{s^2}, \tag{7.23}$$

and this sample size, n, may be found by iteration (i.e., by a directed trial and error), as shown in Example 7.12. The ratio on the left side of Equation 7.23 increases in magnitude as n increases.

EXAMPLE 7.12 Estimation of Required Sample Size to Test H_0: $\sigma^2 \leq \sigma_0^2$ versus H_A: $\sigma^2 > \sigma_0^2$

How large a sample is needed to reject H_0 : $\sigma^2 \leq 1.50$ sec^2, using the data of Example 7.11, if we test at the 0.05 level of significance and with a power of 0.90? (Therefore, $\alpha = 0.05$ and $\beta = 0.10$.)

From Example 7.11, $s^2 = 2.6898$ sec^2. As we have specified $\sigma_0^2 = 1.75$ sec^2, $\sigma_0^2/s^2 = 0.558$.

To begin the iterative process of estimating n, let us guess that a sample size of 30 would be required. Then,

$$\frac{\chi^2_{0.90,29}}{\chi^2_{0.05,29}} = \frac{19.768}{42.557} = 0.465.$$

Because $0.465 < 0.558$, our estimate of n is too low. So we might guess that $n = 50$ is required:

$$\frac{\chi^2_{0.90,49}}{\chi^2_{0.05,49}} = \frac{36.818}{66.339} = 0.555.$$

Because 0.555 is a little less than 0.558, $n = 50$ is a little too low and we might guess $n = 55$, for which $\chi^2_{0.90,54}/\chi^2_{0.05,54} = 41.183/70.153 = 0.571$.

Because 0.571 is greater than 0.558, our estimate of n is high, so we could try $n = 51$, for which $\chi^2_{0.90,50}/\chi^2_{0.05,50} = 37.689/67.505 = 0.558$.

Therefore, we estimate that a sample size of at least 51 is required to perform the hypothesis test with the specified characteristics.

For the hypotheses H_0: $\sigma^2 \geq \sigma_0^2$ versus H_A: $\sigma^2 < \sigma_0^2$, the minimum sample size is that for which

$$\frac{\chi^2_{\beta,\nu}}{\chi^2_{1-\alpha,\nu}} = \frac{\sigma_0^2}{s^2}. \tag{7.24}$$

(b) Power of the Test. If we plan to test the one-tailed hypotheses H_0: $\sigma^2 \leq \sigma_0^2$ versus H_A: $\sigma^2 > \sigma_0^2$, using the α level of significance and a sample size of n, then the power of the test would be

$$1 - \beta = P(\chi^2 \geq \chi^2_{\alpha,\nu}\sigma_0^2/s^2). \tag{7.25}$$

Thus, if the experiment of Example 7.11 were to be repeated with the same sample size, then $n = 8$, $\nu = 7$, $\alpha = 0.05$, $\chi^2_{0.05,9} = 14.067$, $s^2 = 2.6898\ \text{sec}^2$, $\sigma_0^2 = 1.5\ \text{sec}^2$, and the predicted power of the test would be

$$1 - \beta = P[\chi^2 \geq (14.067)(1.5)/2.6898] = P(\chi^2 \geq 7.845).$$

From Appendix Table B.1 we see that, for $\chi^2 \geq 7.845$ with $\nu = 7$, P lies between 0.25 and 0.50 (that is, $0.25 < P < 0.50$). By linear interpolation between $\chi^2_{0.25,7}$ and $\chi^2_{0.50,7}$, we estimate $P(\chi^2 \geq 7.845)$, which is the predicted power of the test, to be 0.38.* If greater power is preferred for this test, we can determine what power would be expected if the experiment were performed with a larger sample size, say $n = 40$. In that case, $\nu = 39$, $\chi^2_{0.05,39} = 54.572$, and the estimate of the power of the test would be

$$1 - \beta = P[\chi^2 \geq (54.572)(1.5)/2.6898] = P(\chi^2 \geq 30.433).$$

Consulting Table B1 for $\nu = 39$, we see that $0.75 < P < 0.90$. By linear interpolation between $\chi^2_{0.75,39}$ and $\chi^2_{0.90,39}$, we estimate $P(\chi^2 \geq 54.572)$, the power of the test, to be 0.82.†

One-tailed testing of H_0: $\sigma^2 \geq \sigma_0^2$ versus H_A: $\sigma^2 \leq \sigma_0^2$ would also employ Equation 7.25. For two-tailed testing of H_0: $\sigma^2 = \sigma_0^2$ versus H_A: $\sigma^2 \neq \sigma_0^2$, substitute $\chi^2_{\alpha/2,\nu}$ for $\chi^2_{\alpha,\nu}$ in Equation 7.25.

7.14 HYPOTHESES CONCERNING THE COEFFICIENT OF VARIATION

Although rarely done, it is possible to ask whether a sample of data is likely to have come from a population with a specified coefficient of variation, call it $(\sigma/\mu)_0$. This amounts to testing of the following pair of two-tailed hypotheses: H_0: $\sigma/\mu = (\sigma/\mu)_0$ and H_A: $\sigma/\mu \neq (\sigma/\mu)_0$. Among the testing procedures proposed, that presented by Miller (1991) works well for a sample size of at least 10 if the sampled population is normal with a mean > 0, with a variance > 0, and with a coefficient of variation, σ/μ, no greater than 0.33. For one-tailed testing (i.e., H_0: $\sigma/\mu \leq (\sigma/\mu)_0$ vs. H_A: $\sigma/\mu > (\sigma/\mu)_0$, or H_0: $\sigma/\mu \geq (\sigma/\mu)_0$ vs. H_A: $\sigma/\mu < (\sigma/\mu)_0$), the test statistic is

$$Z = \frac{\sqrt{n-1}\,[V - (\sigma/\mu)_0]}{(\mu/\sigma)_0\sqrt{0.5 + (\sigma/\mu)_0^2}}, \tag{7.26}$$

the probability of which may be obtained from Appendix Table B.2; or Z may be compared to the critical values of Z_α, read from the last line of Appendix Table B.3. Miller also showed this procedure to yield results very similar to those from a χ^2 approximation by McKay (1932) that, although applicable for n as small as 5, lacks power at such small sample sizes.

Miller and Feltz (1997) present an estimate of the power of this test.

*See the beginning of Appendix B for a discussion of interpolation. In this example, linear interpolation yields $P = 0.38$, harmonic interpolation concludes $P = 0.34$, and the true probability (from appropriate computer software) is $P = 0.35$. So interpolation gave very good approximations.

†The actual probability (via computer) is 0.84, while linear and harmonic interpolations each produced a probability of 0.82, an excellent approximation.

7.15 CONFIDENCE LIMITS FOR THE POPULATION COEFFICIENT OF VARIATION

The $1 - \alpha$ confidence limits for the population coefficient of variation may be estimated as

$$V \pm \frac{V\sqrt{0.5 + V^2}Z_{\alpha/2}}{\sqrt{\nu}}; \tag{7.27}$$

see Miller and Feltz (1997).

7.16 HYPOTHESES CONCERNING SYMMETRY AND KURTOSIS

Section 6.5 introduced the assessment of a population's departure from a normal distribution, including a consideration of a population parameter, $\sqrt{\beta_1}$, for symmetry around the mean and a parameter, β_2, for kurtosis; and their respective sample statistics are $\sqrt{b_1}$ and b_2. Methods will now be discussed for testing hypotheses about a population's symmetry and kurtosis. Such hypotheses are not often employed, but they are sometimes called upon to conclude whether a sampled population follows a normal distribution, and they do appear in some statistical computer packages.

(a) Testing Symmetry around the Mean. The two-tailed hypotheses H_0: $\sqrt{\beta_1} = 0$ versus H_0: $\sqrt{\beta_1} \neq 0$ address the question of whether a sampled population's distribution is symmetrical around its mean. The sample symmetry measure, $\sqrt{b_1}$, is an estimate of $\sqrt{\beta_1}$ and may be calculated by Equation 6.16. Its absolute value may then be compared to critical values, $(\sqrt{b_1})_{\alpha(2),n}$, in Appendix Table B.22.

As an illustration of this, let us say that the data of Example 6.7 yield $\sqrt{b_1} = 0.351$. To test the above H_0 at the 5% level of significance, the critical value from Table B.22 is $(\sqrt{b_1})_{0.05(2),70} = 0.556$. So, H_0 is not rejected and the table indicates that $P(|\sqrt{b_1}| > 0.10)$.

One-tailed testing could be employed if the interest were solely in whether the distribution is skewed to the right (H_0: $\sqrt{\beta_1} \leq 0$ vs. H_0: $\sqrt{\beta_1} > 0$), in which case H_0 would be rejected if $\sqrt{b_1} \geq (\sqrt{b_1})_{\alpha(1),n}$. Or, a one-tailed test of H_0: $\sqrt{\beta_1} \geq 0$ versus H_0: $\sqrt{\beta_1} < 0$ could be used to test specifically whether the distribution is skewed to the left; and H_0 would be rejected if $\sqrt{b_1} \leq -(\sqrt{b_1})_{\alpha(1),n}$.

If the sample size, n, does not appear in Table B.22, a conservative approach (i.e., one with lowered power) would be to use the largest tabled n that is less than the n of our sample; for example, if n were 85, we would use critical values for $n = 80$. Alternatively, a critical value could be estimated, from the table's critical values for n's immediately above and below n under consideration, using linear or harmonic interpolation (see the introduction to Appendix B), with harmonic interpolation appearing to be a little more accurate. There is also a method (D'Agostino, 1970, 1986; D'Agostino, Belanger, and D'Agostino, 1990) by which to approximate the exact probability of H_0.

(b) Testing Kurtosis. Our estimate of a population's kurtosis (β_2) is b_2, given by Equation 6.17. We can ask whether the population is not mesokurtic by the two-tailed hypotheses H_0: $\beta_2 = 3$ versus H_0: $\beta_1 \neq 3$. Critical values for this test are presented in Table B.23, and H_0 is rejected if b_2 is *either* less than the lower-tail critical value for $(b_2)_{\alpha(2),n}$ *or* greater than the upper-tail critical value for $(b_2)_{\alpha(2),n}$.

For the data of Example 6.7, $b_2 = 2.25$. To test the above H_0 at the 5% level of significance, we find that critical values for $n = 70$ do not appear in Table B.23. A conservative procedure (i.e., one with lowered power) is to employ the critical values for the

tabled critical values for the largest n that is less than our sample's n. In our example, this is $n = 50$, and H_0 is rejected if b_2 is *either* less than the lower-tail $(b_2)_{0.05(2),50} = 2.06$ *or* greater than the upper-tail $(b_2)_{0.05(2),50} = 4.36$. In the present example, $b_2 = 2.25$ is neither less than 2.06 nor greater than 4.36, so H_0 is not rejected. And, from Table B.23, we see that $0.05 < P < 0.10$. Rather than using the nearest lower n in Table B.23, we could engage in linear or harmonic interpolation between tabled critical values (see introduction to Appendix B), with harmonic interpolation apparently a little more accurate. There is also a method (D'Agostino, 1970, 1986; D'Agostino, Belanger, and D'Agostino, 1990) to approximate the exact probability of H_0.

One-tailed testing could be employed if the interest is solely in whether the population's distribution is leptokurtic, for which H_0: $\beta_2 \leq 3$ versus H_0: $\beta_2 > 3$ would apply; and H_0 would be rejected if $b_2 \geq$ the upper-tail $(b_2)_{\alpha(1),n}$. Or, if testing specifically whether the distribution is platykurtic, a one-tailed test of H_0: $\beta_2 \geq 3$ versus H_0: $\beta_2 < 3$ would be applicable; and H_0 would be rejected if $b_2 \leq$ the lower-tail $(b_2)_{\alpha(1),n}$.

EXAMPLE 7.13 Two-Tailed Nonparametric Testing of Symmetry Around the Median, Using the Data of Example 6.7 and the Wilcoxon Test of Section 9.5

H_0: The population of data from which this sample came is distributed symmetrically around its median.

H_A: The population is not distributed symmetrically around its median.

$n = 70$; median $= X_{(70+1)/2} = X_{35.5} = 70.5$ in.

X (in.)	d (in.)	f	$\lvert d \rvert$ (in.)	Rank of $\lvert d \rvert$	Signed rank of $\lvert d \rvert$	(f)(Signed rank)
63	−7.5	2	7.5	69.5	−69.5	−139
64	−6.5	2	6.5	67.5	−67.5	−135
65	−5.5	3	5.5	64	−64	−192
66	−4.5	5	4.5	57.5	−57.5	−287.5
67	−3.5	4	3.5	48.5	−48.5	−194
68	−2.5	6	2.5	35.5	−35.5	−213
69	−1.5	5	1.5	21.5	−21.5	−107.5
70	−0.5	8	0.5	8	−8	−64
71	0.5	7	0.5	8	8	56
72	1.5	7	1.5	21.5	21.5	160.5
73	2.5	10	2.5	35.5	35.5	355
74	3.5	6	3.5	48.5	48.5	291
75	4.5	3	4.5	57.5	57.5	172.5
76	5.5	2	5.5	64	64	128
		70				

$T_- = 1332$

$T_+ = 1163$

$T_{0.05(2),70} = 907$ (from Appendix Table B.12)

As neither T_- nor $T_+ < T_{0.05(2),70}$, do not reject H_0. [$P > 0.50$]

(c) Testing Symmetry around the Median. Symmetry of dispersion around the median instead of the mean may be tested nonparametrically by using the Wilcoxon paired-sample test of Section 9.5 (also known as the Wilcoxon signed-rank test). For each datum (X_i) we compute the deviation from the median ($d_i = X_i -$ median) and then analyze the d_i's as in Section 9.5. For the two-tailed test (considering both T_- and T_+ in the Wilcoxon test), the null hypothesis is H_0: The underlying distribution is symmetrical around (i.e., is not skewed from) the median. For a one-tailed test, T_- is the critical value for H_0: The underlying distribution is not skewed to the right of the median; and T_+ is the critical value for H_0: The underlying distribution is not skewed to the left of the median. This test is demonstrated in Example 7.13.

EXERCISES

7.1. The following data are the lengths of the menstrual cycle in a random sample of 15 women. Test the hypothesis that the mean length of human menstrual cycle is equal to a lunar month (a lunar month is 29.5 days).

The data are 26, 24, 29, 33, 25, 26, 23, 30, 31, 30, 28, 27, 29, 26, and 28 days.

7.2. A species of marine arthropod lives in seawater that contains calcium in a concentration of 32 mmole/kg of water. Thirteen of the animals are collected and the calcium concentrations in their coelomic fluid are found to be: 28, 27, 29, 29, 30, 30, 31, 30, 33, 27, 30, 32, and 31 mmole/kg. Test the appropriate hypothesis to conclude whether members of this species maintain a coelomic calcium concentration less than that of their environment.

7.3. Present the following data in a graph that shows the mean, standard error, 95% confidence interval, range, and number of observations for each month.

Table of Caloric Intake (kcal/g of Body Weight) **of Squirrels**

Month	*Number of Data*	*Mean*	*Standard Error*	*Range*
January	13	0.458	0.026	0.289–0.612
February	12	0.413	0.027	0.279–0.598
March	17	0.327	0.018	0.194–0.461

7.4. A sample of size 18 has a mean of 13.55 cm and a variance of 6.4512 cm^2.

(a) Calculate the 95% confidence interval for the population mean.

(b) How large a sample would have to be taken from this population to estimate μ to within 1.00 cm, with 95% confidence?

(c) to within 2.00 cm with 95% confidence?

(d) to within 2.00 cm with 99% confidence?

(e) For the data of Exercise 7.4, calculate the 95% prediction interval for what the mean would be of an additional sample of 10 data from the same population.

7.5. We want to sample a population of lengths and to perform a test of H_0: $\mu = \mu_0$ versus H_A: $\mu \neq \mu_0$, at the 5% significance level, with a 95% probability of rejecting H_0 when $|\mu - \mu_0|$ is at least 2.0 cm. The estimate of the population variance, σ^2, is $s^2 = 8.44$ cm^2.

(a) What minimum sample size should be used?

(b) What minimum sample size would be required if α were 0.01?

(c) What minimum sample size would be required if $\alpha = 0.05$ and power = 0.99?

(d) If $n = 25$ and $\alpha = 0.05$, what is the smallest difference, $|\mu - \mu_0|$, that can be detected with 95% probability?

(e) If $n = 25$ and $\alpha = 0.05$, what is the probability of detecting a difference, $|\mu - \mu_0|$, as small as 2.0 cm?

7.6. There are 200 members of a state legislature. The ages of a random sample of 50 of them are obtained, and it is found that $\bar{X} = 53.87$ yr and $s = 9.89$ yr.

(a) Calculate the 95% confidence interval for the mean age of all members of the legislature.

(b) If the above $\bar{X}$ and s had been obtained from a random sample of 100 from this population, what would the 95% confidence interval for the population mean have been?

7.7. For the data of Exercise 7.4:

(a) Calculate the 95% confidence interval for the population variance.

(b) Calculate the 95% confidence interval for the population standard deviation.

(c) Using the 5% level of significance, test H_0: $\sigma^2 \leq 4.4000$ cm^2 versus H_A: $\sigma^2 > 4.4000$ cm^2.

(d) Using the 5% level of significance, test H_0: $\sigma \geq 3.00$ cm versus H_A : $\sigma < 3.00$ cm.

(e) How large a sample is needed to test H_0: $\sigma^2 \leq 5.0000$ cm^2 if it is desired to test

at the 0.05 level of significance with 75% power?

For the data of Exercise 7.4, calculate the 95% prediction interval for what the variance and standard deviation would be of an additional sample of 20 data from the same population.

7.8. A sample of 100 body weights has $\sqrt{b_1} = 0.375$ and $b_2 = 4.20$.

(a) Test H_0: $\sqrt{\beta_1} = 0$ and H_A: $\sqrt{\beta_1} \neq 0$, at the 5% significance level.

(b) Test H_0: $\beta_2 = 3$ and H_A: $\beta_2 \neq 3$, at the 5% significance level.

CHAPTER 8

Two-Sample Hypotheses

Among the most commonly employed biostatistical procedures is the comparison of two samples to infer whether differences exist between the two populations sampled. This chapter will consider hypotheses comparing two population means, medians, variances (or standard deviations), coefficients of variation, and indices of diversity. In doing so, we introduce another very important sampling distribution, the F distribution—named for its discoverer, R. A. Fisher—and will demonstrate further use of Student's t distribution.

The objective of many two-sample hypotheses is to make inferences about population parameters by examining sample statistics. Other hypothesis-testing procedures, however, draw inferences about populations without referring to parameters. Such procedures are called *nonparametric* methods, and several will be discussed in this and following chapters.

8.1 TESTING FOR DIFFERENCE BETWEEN TWO MEANS

A very common situation for statistical testing is where a researcher desires to infer whether two population means are the same. This can be done by analyzing the difference between the means of samples taken at random from those populations.

Example 8.1 presents the results of an experiment in which adult male rabbits were divided at random into two groups, one group of six and one group of seven.* The members of the first group were given one kind of drug (called "B"), and the

*Sir Ronald Aylmer Fisher (1890–1962) is credited with the first explicit recommendation of the important concept of assigning subjects *at random* to groups for different experimental treatments (Bartlett, 1965; Fisher, 1925b; Rubin, 1990).

members of the second group were given another kind of drug (called "G"). Blood is to be taken from each rabbit and the time it takes the blood to clot is to be recorded.

EXAMPLE 8.1 A Two-Sample *t* Test for the Two-Tailed Hypotheses, H_0: $\mu_1 = \mu_2$ and H_A: $\mu_1 \neq \mu_2$ (Which Could Also Be Stated as H_0: $\mu_1 - \mu_2 = 0$ and H_A: $\mu_1 - \mu_2 \neq 0$). The Data Are Blood-Clotting Times (in Minutes) of Male Adult Rabbits Given One of Two Different Drugs

H_0: $\mu_1 = \mu_2$
H_A: $\mu_1 \neq \mu_2$

Given drug B	*Given drug G*
8.8	9.9
8.4	9.0
7.9	11.1
8.7	9.6
9.1	8.7
9.6	10.4
	9.5

$n_1 = 6$ $n_2 = 7$
$\nu_1 = 5$ $\nu_2 = 6$
$\overline{X}_1 = 8.75$ min $\overline{X}_2 = 9.74$ min
$SS_1 = 1.6950$ min^2 $SS_2 = 4.0171$ min^2

$$s_p^2 = \frac{SS_1 + SS_2}{\nu_1 + \nu_2} = \frac{1.6950 + 4.0171}{5 + 6} = \frac{5.7121}{11} = 0.5193 \text{ min}^2$$

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{\frac{0.5193}{6} + \frac{0.5193}{7}} = \sqrt{0.0866 + 0.0742}$$
$$= \sqrt{0.1608} = 0.40 \text{ min}$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} = \frac{8.75 - 9.74}{0.40} = \frac{-0.99}{0.40} = -2.475$$

$t_{0.05(2),\nu} = t_{0.05(2),11} = 2.201$

Therefore, reject H_0.

$$0.02 < P(|t| \geq 2.475) < 0.05 \quad [P = 0.031]$$

We conclude that mean blood-clotting time is not the same for subjects receiving drug B as it is for subjects receiving drug G.

We can ask whether the mean of the population of blood-clotting times of all adult male rabbits who might have been administered drug B (let's call that mean μ_1) is the same as the population mean for blood-clotting times of all adult male rabbits who might have been given drug G (call it μ_2). This would involve the two-tailed hypotheses H_0: $\mu_1 - \mu_2 = 0$ and H_A: $\mu_1 - \mu_2 \neq 0$; and these hypotheses are commonly expressed in their equivalent forms: H_0: $\mu_1 = \mu_2$ and H_A: $\mu_1 \neq \mu_2$. The data from this experiment are presented in Example 8.1.

In this example, a total of 13 members of a biological population (adult male rabbits) were divided at random into two experimental groups, each group to receive treatment with one of the drugs. Another kind of testing situation with two independent samples is where the two groups are predetermined. For example, instead of desiring to test the effect of two drugs on blood-clotting time, a researcher might want to compare the mean blood-clotting time of adult male rabbits to that of adult female rabbits, in which case one of the two samples would be composed of randomly chosen males and the other sample would comprise randomly selected females. In that situation, the researcher would not specify which rabbits will be designated as male and which as female; the sex of each animal (and, therefore, the experimental group to which each is assigned) is determined before the experiment is begun. Similarly, it might have been asked whether the mean blood-clotting time is the same in two strains (or two ages, or two colors) of rabbits. Thus, in Example 8.1 there is random allocation of animals to the two groups to be compared, while in the other examples in this paragraph, there is random sampling of animals within each of two groups that are already established. The statistical hypotheses and the statistical testing procedure are the same in both circumstances.

If the two samples came from two normally distributed populations, and if the two populations have equal variances, then a t value to test such hypotheses may be calculated in a manner analogous to its computation for the one-sample t test introduced in Section 7.1. The t for testing the preceding hypotheses concerning the difference between two population means is

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}. \tag{8.1}$$

The quantity $\overline{X}_1 - \overline{X}_2$ is the difference between the two sample means; and $s_{\overline{X}_1 - \overline{X}_2}$ is the standard error of the difference between the sample means (explained further below), which is a measure of the variability of the data within the two samples. Therefore, Equation 8.1 compares the differences between two means to the differences among all the data (a concept to be enlarged upon when comparing more than two means—in Chapter 10 and beyond).

The quantity $s_{\overline{X}_1 - \overline{X}_2}$, along with $s^2_{\overline{X}_1 - \overline{X}_2}$, the variance of the difference between the means, needs to be considered further. Both $s^2_{\overline{X}_1 - \overline{X}_2}$ and $s_{\overline{X}_1 - \overline{X}_2}$ are statistics that can be calculated from the sample data and are estimates of the population parameters, $\sigma^2_{\overline{X}_1 - \overline{X}_2}$ and $\sigma_{\overline{X}_1 - \overline{X}_2}$, respectively. It can be shown mathematically that the variance of the difference between two independent variables is equal to the sum of the variances of the two variables, so that $\sigma^2_{\overline{X}_1 - \overline{X}_2} = \sigma^2_{\overline{X}_1} + \sigma^2_{\overline{X}_2}$. Independence means that there is no association correlation between the data in the two populations.* As $\sigma^2_{\overline{X}} = \sigma^2/n$, we can write

$$\sigma^2_{\overline{X}_1 - \overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \tag{8.2}$$

Because the two-sample t test requires that we assume $\sigma_1^2 = \sigma_2^2$, we can write

$$\sigma^2_{\overline{X}_1 - \overline{X}_2} = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}. \tag{8.3}$$

*If there is a unique relationship between each datum in one sample and a specific datum in another sample, then the data are considered *paired* and the considerations of Chapter 9 apply instead of the methods of the present chapter.

Thus, to calculate the estimate of $\sigma^2_{\overline{X}_1-\overline{X}_2}$, we must have an estimate of σ^2. Since both s_1^2 and s_2^2 are assumed to estimate σ^2, we compute the *pooled variance*, s_p^2, which is then used as the best estimate of σ^2:

$$s_p^2 = \frac{SS_1 + SS_2}{\nu_1 + \nu_2}, \tag{8.4}$$

and

$$s^2_{\overline{X}_1-\overline{X}_2} = \frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}. \tag{8.5}$$

Thus,*

$$s_{\overline{X}_1-\overline{X}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}, \tag{8.6}$$

and Equation 8.1 becomes

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}, \tag{8.7a}$$

which for equal sample sizes (i.e., $n_1 = n_2$, so each sample size may be referred to as n),

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{2s_p^2}{n}}}. \tag{8.7b}$$

Example 8.1 summarizes the procedure for testing the hypotheses under consideration. The critical value to be obtained from Appendix Table B.3 is $t_{\alpha(2),(\nu_1+\nu_2)}$, the two-tailed t value for the α significance level, with $\nu_1 + \nu_2$ degrees of freedom. We shall also write this as $t_{\alpha(2),\nu}$, defining the pooled degrees of freedom to be

$$\nu = \nu_1 + \nu_2 \quad \text{or, equivalently,} \quad \nu = n_1 + n_2 - 2. \tag{8.8}$$

In the two-tailed test, H_0 will be rejected if either $t \geq t_{\alpha(2),\nu}$ or $t \leq -t_{\alpha(2),\nu}$. Another way of stating this is that H_0 will be rejected if $|t| \geq t_{\alpha(2),\nu}$.

This statistical test asks what the probability is of obtaining two independent samples with means ($\overline{X}_1$ and $\overline{X}_2$) at least this different by random sampling from populations whose means (μ_1 and μ_2) are equal. And, if that probability is α or less, then H_0: $\mu_1 = \mu_2$ is rejected and it is declared that there is good evidence that the two population means are different.†

H_0: $\mu_1 = \mu_2$ may be written H_0: $\mu_1 - \mu_2 = 0$ and H_A: $\mu_1 \neq \mu_2$ as H_A: $\mu_1 - \mu_2 \neq 0$; the generalized two-tailed hypotheses are H_0: $\mu_1 - \mu_2 = \mu_0$ and H_A: $\mu_1 - \mu_2 \neq \mu_0$, tested as

$$t = \frac{|\overline{X}_1 - \overline{X}_2| - \mu_0}{s_{\overline{X}_1-\overline{X}_2}}, \tag{8.9}$$

where μ_0 may be any hypothesized difference between population means.

*The standard error of the difference between means may also be calculated as $s_{\overline{X}_1-\overline{X}_2} = \sqrt{Ns_p^2/(n_1n_2)}$, where $N = n_1 + n_2$.

†Instead of testing this hypotheses, a hypothesis of "correlation" (Section 19.11b) could be tested, which would ask whether there is a significant linear relationship between the magnitude of X and the group from which it came. This is not commonly done.

By the procedure of Section 8.9, one can test whether the measurements in one population are a specified amount as large as those in a second population.

(a) One-Tailed Hypotheses about the Difference between Means. One-tailed hypotheses can be tested in situations where the investigator is interested in detecting a difference in only one direction. For example, a gardener may use a particular fertilizer for a particular kind of plant, and a new fertilizer is advertised as being an improvement. Let us say that plant height at maturity is an important characteristic of this kind of plant, with taller plants being preferable. An experiment was run, raising ten plants on the present fertilizer and eight on the new one, with the resultant eighteen plant heights shown in Example 8.2. If the new fertilizer produces plants that are shorter than, or the same height as, plants grown with the present fertilizer, then we shall decide that the advertising claims are unfounded; therefore, the statements of $\mu_1 > \mu_2$ and $\mu_1 = \mu_2$ belong in the same hypothesis, namely the null hypothesis, H_0. If, however, mean plant height is indeed greater with the newer fertilizer, then it shall be declared to be distinctly better, with the alternate hypothesis (H_A: $\mu_1 < \mu_2$) concluded to be the true statement. The t statistic is calculated by Equation 8.1, just as for the two-tailed test. But this calculated t is then compared with the critical value $t_{\alpha(1),\nu}$, rather than with $t_{\alpha(2),\nu}$.

EXAMPLE 8.2 A Two-Sample t Test for the One-Tailed Hypotheses, H_0: $\mu_1 \geq \mu_2$ and H_A: $\mu_1 < \mu_2$ (Which Could Also Be Stated as H_0: $\mu_1 - \mu_2 \geq 0$ and H_A: $\mu_1 - \mu_2 < 0$). The Data Are Heights of Plants, Each Grown with One of Two Different Fertilizers

H_0: $\mu_1 \geq \mu_2$
H_A: $\mu_1 < \mu_2$

Present fertilizer	*Newer fertilizer*
48.2 cm	52.3 cm
54.6	57.4
58.3	55.6
47.8	53.2
51.4	61.3
52.0	58.0
55.2	59.8
49.1	54.8
49.9	
52.6	
$n_1 = 10$	$n_2 = 8$
$\nu_1 = 9$	$\nu_2 = 7$
$\overline{X}_1 = 51.91$ cm	$\overline{X}_2 = 56.55$ cm
$SS_1 = 102.23$ cm^2	$SS_2 = 69.20$ cm^2

$$s_p^2 = \frac{102.23 + 69.20}{9 + 7} = \frac{171.43}{16} = 10.71 \text{ cm}^2$$

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{10.71}{10} + \frac{10.71}{8}} = \sqrt{2.41} = 1.55 \text{ cm}$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} = \frac{51.91 - 56.55}{1.55} = \frac{-4.64}{1.55} = -2.99$$

$t_{0.05(1),16} = 1.746$

As t of -2.99 is less than -1.746, H_0 is rejected.

$$0.0025 < P < 0.005 \quad [P = 0.0043]$$

The mean plant height is greater with the newer fertilizer.

In other cases, the one-tailed hypotheses, H_0: $\mu_1 \leq \mu_2$ and H_A: $\mu_1 > \mu_2$, may be appropriate. Just as introduced in the one-sample testing of Sections 7.1 and 7.2, the following summary of procedures applies to two-sample t testing:

For H_A: $\mu_1 \neq \mu_2$, if $|t| \geq t_{\alpha(2),\nu}$, then reject H_0.

For H_A: $\mu_1 < \mu_2$, if $t \leq -t_{\alpha(1),\nu}$, then reject H_0.*

For H_A: $\mu_1 > \mu_2$, if $t \geq t_{\alpha(1),\nu}$, then reject H_0.†

As indicated in Section 6.3, the null and alternate hypotheses are to be decided upon *before* the data are collected.

Also, H_0: $\mu_1 \leq \mu_2$ and H_A: $\mu_1 > \mu_2$ may be written as H_0: $\mu_1 - \mu_2 \leq 0$ and H_A: $\mu_1 - \mu_2 > 0$, respectively. The generalized hypotheses for this type of one-tailed test are H_0: $\mu_1 - \mu_2 \leq \mu_0$ and H_A: $\mu_1 - \mu_2 > \mu_0$, for which the t is

$$t = \frac{\overline{X}_1 - \overline{X}_2 - \mu_0}{s_{\overline{X}_1 - \overline{X}_2}}, \tag{8.10}$$

and μ_0 may be any specified value of $\mu_1 - \mu_2$.

Lastly, H_0: $\mu_1 \geq \mu_2$ and H_A: $\mu_1 < \mu_2$ may be written as H_0: $\mu_1 - \mu_2 \geq 0$ and H_A: $\mu_1 - \mu_2 < 0$, and the generalized one-tailed hypotheses of this type are H_0: $\mu_1 - \mu_2 \geq \mu_0$ and H_A: $\mu_1 - \mu_2 < \mu_0$, with the appropriate t statistic being that of Equation 8.10. For example, the gardener collecting the data of Example 8.2 may have decided, because the newer fertilizer is more expensive than the other, that it should be used only if the plants grown with it averaged at least 5.0 cm taller than plants grown with the present fertilizer. Then, $\mu_0 = \mu_1 - \mu_2 = -5.0$ cm and, by Equation 8.10, we would calculate $t = (51.91 - 56.55 + 5.0)/1.55 = 0.36/1.55 = 0.232$, which is not $\geq$ the critical value shown in Example 8.2; so H_0: $\mu_1 - \mu_2 \geq -5.0$ cm is not rejected. The following summary of procedures applies to these general hypotheses:

For H_A: $\mu_1 - \mu_2 \neq \mu_0$, if $|t| \geq t_{\alpha(2),\nu}$, then reject H_0.

For H_A: $\mu_1 - \mu_2 < \mu_0$, if $t \leq -t_{\alpha(1),\nu}$, then reject H_0.

For H_A: $\mu_1 - \mu_2 > \mu_0$, if $t \geq t_{\alpha(1),\nu}$, then reject H_0.

*For this one-tailed hypothesis test, probabilities of t up to 0.25 are indicated in Appendix Table B.3. If $t = 0$, then $P = 0.50$; so if $-t_{0.25(1),\nu} < t < 0$, then $0.25 < P < 0.50$; and if $t > 0$ then $P > 0.50$.

†For this one-tailed hypothesis test, $t = 0$ indicates $P = 0.50$; therefore, if $0 < t < t_{0.25(1),\nu}$, then $0.25 < P < 0.50$; and if $t < 0$, then $P > 0.50$.

(b) Violations of the Two-Sample *t*-test Assumptions. The validity of two-sample t testing depends upon two basic assumptions: that the two samples came at random from normal populations and that the two populations had the same variance. Populations of biological data will not have distributions that are exactly normal or variances that are exactly the same. Therefore, it is fortunate that numerous studies, over 70 years, have shown that this t test is robust enough to withstand considerable nonnormality and some inequality of variances. This is especially so if the two sample sizes are equal or nearly equal, particularly when two-tailed hypotheses are tested (e.g., Boneau, 1960; Box, 1953; Cochran, 1947; Havlicek and Peterson, 1974; Posten, Yen, and Owen, 1982; Srivastava, 1958; Stonehouse and Forrester, 1998; Tan, 1982; Welch, 1938) but also in one-tailed testing (Posten, 1992).

In general, the larger and the more equal in size the samples are, the more robust the test will be; and sample sizes of at least 30 provide considerable resistance effects of violating the t-test assumptions when testing at $\alpha = 5\%$ (i.e., the 0.05 level of signficance), regardless of the disparity between σ_1^2 and σ_2^2 (Donaldson, 1968; Ramsey, 1980; Stonehouse and Forrester, 1998); larger sample sizes are needed for smaller α's, smaller n's will suffice for larger significance levels, and larger samples are required for larger differences between σ_1 and σ_2.

Hsu (1938) reported remarkable robustness, even in the presence of very unequal variances and very small samples, if $n_1 = n_2 + 1$ and $\sigma_1^2 > \sigma_2^2$. So, if it is believed (by inspecting s_1^2 and s_2^2) that the population variances (σ_1^2 and σ_2^2) are dissimilar, one might plan experiments that have samples that are unequal in size by 1, where the larger sample comes from the population with the larger variance. But the procedure of Section 8.1c, below, has received a far greater amount of study and is much more commonly employed.

The two-sample t test is very robust to nonnormality if the population variances are the same (Kohr and Games, 1974; Posten, 1992; Posten, Yeh, and Owen, 1982; Ramsey, 1980; Stonehouse and Forrester, 1998; Tomarkin and Serlin, 1986). If the two populations have the same variance and the same shape, the test works well even if that shape is extremely nonnormal (Stonehouse and Forrester, 1998; Tan, 1982). Havlicek and Peterson (1974) specifically discuss the effect of skewness and leptokurtosis.

If the population variances are unequal but the sample sizes are the same, then the probability of a Type I error will tend to be greater than the stated α (Havlicek and Peterson, 1974; Ramsey, 1980), and the test is said to be *liberal*. As seen in Table 8.1a, this departure from α will be less for smaller differences between σ_1^2 and σ_2^2 and for larger sample sizes. (The situation with the most heterogeneous variances is where σ_1^2/σ_2^2 is zero (0) or infinity (∞).)

If the two variances are not equal *and* the two sample sizes are not equal, then the probability of a Type I error will differ from the stated α. If the larger σ^2 is associated with the larger sample, this probability will be less than the stated α (and the test is called *conservative*) and this probability will be greater than the stated α (and the test is called *liberal*) if the smaller sample came from the population with the larger variance (Havlicek and Peterson, 1974; Ramsey, 1980; Stonehouse and Forrester, 1998; Zimmerman, 1987).* The greater the difference between variances, the greater will be the disparity between the probability of a Type I error and the specified α, larger differences will also result in greater departure from α. Table 8.1b

*The reason for this can be seen from Equations 8.4–8.7a: If the larger s_i^2 is coupled with the larger n_i, then the numerator of s_p^2 (which is $\nu_1 s_1^2 + \nu_2 s_2^2$) is greater than if the larger variance is associated with the smaller n. This makes s_p^2 larger, which translates into a larger $s_{\bar{X}_1 - \bar{X}_2}^2$, which produces a smaller t, resulting in a probability of a Type I error lower than the stipulated α.

TABLE 8.1a: Maximum Probabilities of Type I Error when Applying the Two-Tailed *(or, One-Tailed)* t Test to Two Samples of Various Equal Sizes ($n_1 = n_2 = n$), Taken from Normal Populations Having Various Variance Ratios, σ_1^2/σ_2^2

σ_1^2/σ_2^2	n:	3	5	10	15 [16]	20	30	∞
				For $\alpha = 0.05$				
3.33 or 0.300		0.059	0.056	0.054	0.052	0.052	0.051	0.050
5.00 or 0.200		0.064	0.061	0.056	0.054	0.053	0.052	0.050
10.00 or 0.100			0.068	0.059	0.056	0.055	0.053	0.050
∞ or 0		0.109	0.082	0.065	0.060	0.057	0.055	0.050
∞ or 0		*(0.083)*	*(0.068)*	*(0.058)*	*(0.055)*	*(0.054)*	*(0.053)*	*(0.050)*
				For $\alpha = 0.01$				
3.33 or 0.300		0.013	0.013	0.012	[0.011]	0.011	0.011	0.010
5.00 or 0.200		0.015	0.015	0.013	[0.012]	0.011	0.011	0.010
10.00 or 0.100		0.020	0.019	0.015	[0.013]	0.012	0.012	0.010
∞ or 0		0.044	0.028	0.018	[0.015]	0.014	0.013	0.010
∞ or 0		*(0.032)*	*(0.022)*	*(0.015)*	*(0.014)*	*(0.013)*	*(0.012)*	*(0.010)*

These probabilities are gleaned from the extensive analysis of Ramsey (1980), and from Table 1 of Posten, Yeh, and Owen (1982).

shows this for various sample sizes. For example, Table 8.1a indicates that if 20 data are distributed as $n_1 = n_2 = 10$ and the two-tailed t test is performed at the 0.05 significance level, the probability of a Type I error approaches 0.065 for greatly divergent population variances. But in Table 8.1b we see that if $\alpha = 0.05$ is used and 20 data are distributed as $n_1 = 9$ and $n_2 = 11$, then the probability of a Type I error can be as small as 0.042 (if the sample of 11 came from the population with the larger variance) or as large as 0.096 (if the sample of 9 came from the population with the smaller variance).

Section 6.3b explained that a decrease in the probability of the Type I error (α) is associated with an increase in the probability of a Type II error (β); and, because power is $1 - \beta$, an increase in β means a decrease in the power of the test ($1 - \beta$). Therefore, for situations described above as conservative—that is, P(Type I error) $< \alpha$—there will generally be less power than if the population variances were all equal; and when the test is liberal—that is, P(Type I error) $> \alpha$—there will generally be more power than if the variances were equal. (See also Zimmerman and Zumbo, 1993.)

The power of the two-tailed t test is affected very little by small or moderate skewness in the sampled populations, especially if the sample sizes are equal, but there can be a serious effect on one-tailed tests. As for kurtosis, the actual power of the test is less than that discussed in Section 8.4 when the populations are platykurtic and greater when they are leptokurtic, especially for small sample sizes (Boneau, 1960; Glass, Peckham, and Sanders, 1972). The adverse effect of nonnormality is less with large sample sizes (Srivastava, 1958).

(c) The Two-sample t Test with Unequal Variances. As indicated above, the t test for difference between two means is robust to some departure from its underlying assumptions; but it is not dependable when the two population variances are very different. The latter situation is known as the Behrens-Fisher problem, referring to the early work on it by Behrens (1929) and Fisher (e.g., 1939b), and numerous

TABLE 8.1b: Maximum Probabilities of Type I Error when Applying the Two-Tailed *(or One-Tailed)* t Test to Two Samples of Various Unequal Sizes, Taken from Normal Populations Having the Largest Possible Difference between Their Variances

		For $\alpha = 0.05$		For $\alpha = 0.01$	
n_1	n_2	σ_1^2 large	σ_1^2 small	σ_1^2 large	σ_1^2 small
11	9	0.042	0.096	0.0095	0.032
		(0.041)	*(0.079)*	*(0.0088)*	*(0.026)*
22	18	0.036	0.086	0.0068	0.026
		(0.038)	*(0.073)*	*(0.0068)*	*(0.021)*
33	27	0.034	0.082	0.0059	0.024
		(0.037)	*(0.072)*	*(0.0062)*	*(0.020)*
55	45	0.032	0.080	0.0053	0.022
		(0.036)	*(0.070)*	*(0.0057)*	*(0.019)*
12	8	0.025	0.13	0.0045	0.054
		(0.028)	*(0.098)*	*(0.0046)*	*(0.040)*
24	16	0.020	0.12	0.0029	0.044
		(0.025)	*(0.096)*	*(0.0033)*	*(0.034)*
36	24	0.019	0.12	0.0024	0.041
		(0.023)	*(0.094)*	*(0.0029)*	*(0.032)*
60	40	0.018	0.11	0.0021	0.039
		(0.023)	*(0.092)*	*(0.0026)*	*(0.031)*

From Posten, Yeh, and Owen (1982) and Posten (1992).

other studies of this problem have ensued (e.g., Best and Raynor, 1987; Dixon and Massey, 1969: 119; Fisher and Yates, 1963: 60–61;* Gill, 1971; Kim and Cohen, 1998; Lee and Fineberg, 1991; Lee and Gurland, 1975; Satterthwaite, 1946; Scheffé, 1970; Zimmerman and Zumbo, 1993). Several solutions have been proffered, and they give very similar results except for very small samples. One of the easiest, yet reliable, of available procedures is that attributed to Smith (1936) and is often known as the "Welch approximate t"† (Davenport and Webster, 1975; Mehta and Srinivasan, 1971; Wang, 1971; Welch, 1936, 1938, 1947). It has been shown to perform well with respect to Type I error, and it requires no special tables.

The test statistic is that of Equation 8.1 or 8.9, but with $s_{\overline{X}_1 - \overline{X}_2}$ (the standard error of the difference between the means) calculated with the two separate variances instead of with a pooled variance; that is,

$$s'_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{8.11a}$$

instead of Equation 8.6. And, because $s_{\overline{X}_i} = s_i^2/n_i$ (Equation 6.7), this can be written equivalently as

$$s'_{\overline{X}_1 - \overline{X}_2} = \sqrt{s_{\overline{X}_1}^2 + s_{\overline{X}_2}^2}. \tag{8.11b}$$

*In Fisher and Yates (1963), s refers to the standard error, not the standard deviation.

†Bernard Lewis Welch (1911–1989), English statistician. (See Mardia, 1990.)

Therefore, Equation 8.1 becomes

$$t' = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s^2_{\overline{X}_1} + s^2_{\overline{X}_2}}}, \tag{8.11c}$$

Equation 8.11 becomes

$$t' = \frac{\overline{X}_1 - \overline{X}_2 - \mu_0}{\sqrt{s^2_{\overline{X}_1} + s^2_{\overline{X}_2}}}, \tag{8.11d}$$

and two-tailed and one-tailed hypotheses are tested as described earlier for t.

Tables of critical values of t' have been published, but they are not extensive. Satterthwaite (1946) and Scheffé (1970) approximated the distribution of t' well by using t with degrees of freedom of

$$\nu' = \frac{\left(s^2_{\overline{X}_1} + s^2_{\overline{X}_2}\right)^2}{\frac{\left(s^2_{\overline{X}_1}\right)^2}{n_1 - 1} + \frac{\left(s^2_{\overline{X}_2}\right)^2}{n_2 - 1}}. \tag{8.12}$$

These degrees of freedom can be as small as $n_1 - 1$ or $n_2 - 1$, whichever is smaller, and as large as $n_1 + n_2 - 2$. However, ν' is typically not an integer, so the critical value of t' often will not be found in Appendix Table B.3. If ν' is not an integer, the needed critical value, $t_{\alpha,\nu}$, can be obtained via some computer software; or these values can be interpolated from the t's in Table B.3 (the beginning of Appendix B explains interpolation, and at the end of Table B.3 there is an indication of the accuracy of interpolation for t); or, less accurately, the closest integer to ν' (or, to be conservative, the nearest integer less than ν') can be used as the degrees of freedom in Table B.3. The Behrens-Fisher test is demonstrated in Example 8.2a.*

*In the highly unlikely situation where the variances (σ_1^2 and σ_2^2) of the two sampled populations are known, the test for difference between means could be effected with

$$Z = \frac{|\overline{X}_1 - \overline{X}_2| - \mu_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}; \tag{8.12a}$$

and it can be recalled that $Z_\alpha = t_{\alpha,\infty}$. If the variance ($\sigma_1^2$) of one of the two populations is known, this test statistic and degrees of freedom may be employed (Maity and Sherman, 2006):

$$t = \frac{|\overline{X}_1 - \overline{X}_2| - \mu_0}{\sqrt{\sigma_1^2/n_1 + s_2^2/n_2}}; \tag{8.12b}$$

$$\nu' = \frac{\left(\sigma_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{s_2^2/n_2}{n_2 - 1}}. \tag{8.12c}$$

EXAMPLE 8.2a The Behrens-Fisher Test for the Two-Tailed Hypotheses, H_0: $\mu_1 = \mu_2$ and H_A: $\mu_1 \neq \mu_2$

The data are the times for seven cockroach eggs to hatch at one laboratory temperature and for eight eggs to hatch at another temperature.

H_0: $\mu_1 = \mu_2$
H_A: $\mu_1 \neq \mu_2$

At 30°C	At 10°C
40 days	36 days
38	45
32	32
37	52
39	59
41	41
35	48
	55

$n_1 = 7$	$n_2 = 8$
$\nu_1 = 6$	$\nu_2 = 7$
$\overline{X}_1 = 37.4$ days	$\overline{X}_2 = 46.0$ days
$SS_1 = 57.71$ days^2	$SS_2 = 612.00$ days^2
$s_1^2 = 9.62$ days^2	$s_2^2 = 87.43$ days^2
$s_{\overline{X}_1}^2 = 1.37$ days^2	$s_{\overline{X}_2}^2 = 10.93$ days^2

$$s'_{\overline{X}_1 - \overline{X}_2} = \sqrt{s_{\overline{X}_1}^2 + s_{\overline{X}_2}^2} = \sqrt{1.37 + 10.83} = 3.51 \text{ days}$$

$$t' = \frac{\overline{X}_1 - \overline{X}_2}{s'_{\overline{X}_1 - \overline{X}_2}} = \frac{37.4 - 46.00}{3.51} = -2.450$$

$$\nu' = \frac{\left(s_{\overline{X}_1}^2 + s_{\overline{X}_2}^2\right)^2}{\frac{\left(s_{\overline{X}_1}^2\right)^2}{\nu_1} + \frac{\left(s_{\overline{X}_2}^2\right)^2}{\nu_2}} = \frac{(1.37 + 10.93)^2}{\frac{(1.37)^2}{6} + \frac{(10.93)^2}{7}} = 8.7$$

$t_{0.05(2),8.7} = 2.274$*

Therefore, reject H_0.

[$P = 0.038$.]*

*These values were obtained by computer.

As with t, the robustness of t' is greater with large and with equal sample sizes. If $\sigma_1^2 = \sigma_2^2$, then either t or t' can be used, but t will be the more powerful procedure (Ramsey, 1980), but generally with only a very slight advantage over t' (Best and Rayner, 1987). If $n_1 = n_2$ and $s_1^2 = s_2^2$, then $t' = t$ and $\nu' = \nu$; but t' is not as powerful and not as robust to nonnormality as t is (Stonehouse and Forrester, 1998; Zimmerman and Zumbo, 1993). However, Best and Rayner (1987) found t' to be much better when the variances and the sample sizes are unequal. They, and Davenport and Webster (1975), reported that the probabilty of a Type I error in the t' test is related to the ratio $(n_2\sigma_1^2)/(n_1\sigma_2^2)$ (let's call this ratio r for the present): When $r > 1$ and $n_1 > n_2$, then this error is near the α specified for the significance test; when $r > 1$ and $n_1 < n_2$, then the error diverges from that α to an extent reflecting the magnitude of r and the difference betwen n_1 and n_2. And, if $r < 1$, then the error is close to the stated α if $n_1 < n_2$, and it departs from that α if $n_1 > n_2$ (differing to a greater extent as the difference between the sample sizes is larger and the size of r is greater). But, larger sample sizes result in less departure from the α used in the hypothesis test.

The effect of heterogeneous variances on the t test can be profound. For example, Best and Rayner (1987) estimated that a t test with $n_1 = 5$ and $n_2 = 15$, and $\sigma_1/\sigma_2 = 4$, has a probabilty of a Type I error using t of about 0.16; and, for those sample sizes when $\sigma_1/\sigma_2 = 0.25$, P(Type I error) is about 0.01; but the probability of that error in those cases is near 0.05 if t' is employed. When the two variances are unequal, the Brown-Forsythe test mentioned in Section 10.1g could also be employed and would be expected to perform similarly to the Behrens-Fisher test, though generally not as well.

If the Behrens-Fisher test concludes difference between the means, a confidence interval for that difference may be obtained in a manner analogous to that in Section 8.2: The procedure is to substitute $s'_{\overline{X}_1 - \overline{X}_2}$ for $s_{\overline{X}_1 - \overline{X}_2}$ and to use ν' instead of ν in Equation 8.14.

Because the t test is adversely affected by heterogeneity of variances, some authors have recommended a two-step testing process: (1) The two sample variances are compared, and (2) only if the two population variances are concluded to be similar should the t test be employed. The similarity of variances may be tested by the procedures of Section 8.5. However, considering that the Behrens-Fisher t' test is so robust to variance inequality (and that the most common variance-comparison test performs very poorly when the distributions are nonnormal or asymmetrical), the routine test of variances is not recommended as a precursor to the testing of means by either t or t' (even though some statistical software packages perform such a test). Gans (1991) and Markowski and Markowski (1990) enlarge upon this conclusion; Moser and Stevens (1992) explain that there is no circumstance when the testing of means using either t or t' is improved by preliminary testing of variances; and Sawilowski (2002) and Wehrhahn and Ogawa (1978) state that the t test's probability of a Type I error may differ greatly from the stated α if such two-step testing is employed.

(d) Which Two-Sample Test to Use. It is very important to inform the reader of a research report specifically what statistical procedures were used in the presentation and analysis of data. It is also generally advisable to report the size (n), the mean ($\overline{X}$), and the variability (variance, standard deviation, or standard error) of each group of data; and confidence limits for each mean and for the difference between the means (Section 8.2) may be expressed if the mean came from a normally distributed population. Visualization of the relative magnitudes of means and measures of variability may be aided by tables or graphs such as described in Section 7.4.

Major choices of statistical methods for comparing two samples are as follows:

- If the two sampled populations are normally distributed and have identical variances (or if they are only slightly to moderately nonnormal and have similar variances): The t test for difference between means is appropriate and preferable. (However, as samples nearly always come from distributions that are not exactly normal with exactly the same variances, conclusions to reject or not reject a null hypothesis should not be considered definitive when the probability associated with t is very near the specified α. For example, if testing at the 5% level of significance, it should not be emphatically declared that H_0 is false if the probability of the calculated t is 0.048. The conclusion should be expressed with caution and, if feasible, the experiment should be repeated—perhaps with more data.)
- If the two sampled populations are distributed normally (or are only slightly to moderately nonnormal), but they have very dissimilar variances: The Behrens-Fisher test of Section 8.1c is appropriate and preferable to compare the two means.
- If the two sampled populations are very different from normally distributed, but they have similar distribution shapes and variances: The Mann-Whitney test of Section 8.11 is appropriate and preferable.
- If the two sampled populations have distributions greatly different from normal and do not have similar distributions and variances: (1) Consider the procedures of Chapter 13 for data that do not exhibit normality and variance equality but that can be transformed into data that are normal and homogeneous of variance; or (2) refer to the procedure mentioned at the end of Section 8.11, which modifies the Mann-Whitney test for Behrens-Fisher situations; or (3) report the mean and variability for each of the samples, perhaps also presenting them in tables and/or graphs (as in Section 7.4), and do not perform hypothesis testing.*

(e) Replication of Data. It is important to use data that are true replicates of the variable to be tested (and recall that a replicate is the smallest experimental unit to which a treatment is independently applied). In Example 8.1 the purpose of the experiment was to ask whether there is a difference in blood-clotting times between persons administered two different drugs. This necessitates obtaining a blood measurement on each of n_1 individuals in the first sample (receiving one of the drugs) and n_2 individuals in the second sample (receiving the other drug). It would not be valid to use n_1 measurements from a single person and n_2 measurements from another person, and to do so would be engaging in what Hurlbert (1984), and subsequently many others, discuss as *pseudoreplication*.

8.2 CONFIDENCE LIMITS FOR POPULATION MEANS

In Section 7.3, we defined the confidence interval for a population mean as $\overline{X} \pm t_{\alpha(2),\nu} s_{\overline{X}}$, where $s_{\overline{X}}$ is the best estimate of $\sigma_{\overline{X}}$ and is calculated as $\sqrt{s^2/n}$. For the

*Another procedure, seldom encountered but highly recommended by Yuen (1974), is to perform the Behrens-Fisher test on *trimmed means* (also known as "truncated means"). A trimmed mean is a sample mean calculated after deleting data from the extremes of the tails of the data distribution. There is no stipulated number of data to be deleted, but it is generally the same number for each tail. The degrees of freedom are those pertaining to the number of data remaining after the deletion.

two-sample situation where we assume that $\sigma_1^2 = \sigma_2^2$, the confidence interval for either μ_1 or μ_2 is calculated using s_p^2 (rather than either s_1^2 or s_2^2) as the best estimate of σ^2, and we use the two-tailed tabled t value with $\nu = \nu_1 + \nu_2$ degrees of freedom. Thus, for μ_i (where i is either 1 or 2, referring to either of the two samples), the $1 - \alpha$ confidence interval is

$$\overline{X}_i \pm t_{\alpha(2),\nu}\sqrt{\frac{s_p^2}{n_i}}. \tag{8.13}$$

For the data of Example 8.1, $\sqrt{s_p^2/n_2} = \sqrt{0.5193\ \text{min}^2/7} = 0.27$ min. Thus, the 95% confidence interval for μ_2 would be 9.74 min $\pm$ (2.201)(0.27 min) = 9.74 min $\pm$ 0.59 min, so that L_1 (the lower confidence limit) = 9.15 min and L_2 (the upper confidence limit) = 10.33 min, and we can declare with 95% confidence that, for the population of blood-clotting times after treatment with drug G, the population mean, μ_2, is no smaller than 9.15 min and no larger than 10.33 min. This may be written as $P(9.15\ \text{min} \leq \mu_2 \leq 10.33\ \text{min}) = 0.95$. The confidence interval for the population mean of data after treatment with drug B would be 8.75 min $\pm(2.201)\sqrt{0.5193\ \text{min}^2/6} = 8.75\ \text{min} \pm 0.64\ \text{min}$; so $L_1 = 8.11$ min and $L_2 = 9.39$ min. Further interpretation of the meaning of the confidence interval for each of these two population means is in Section 7.3.

Confidence limits for the difference between the two population means can also be computed. The $1 - \alpha$ confidence interval for $\mu_1 - \mu_2$ is

$$\overline{X}_1 - \overline{X}_2 \pm t_{\alpha(2),\nu}s_{\overline{X}_1 - \overline{X}_2}. \tag{8.14}$$

Thus, for Example 8.1, the 95% confidence interval for $\mu_1 - \mu_2$ is (8.75 min $-$ 9.74 min) $\pm$ (2.201)(0.40 min) = -0.99 min $\pm$ 0.88 min. Thus, $L_1 = -1.87$ min and $L_2 = -0.11$ min, and we can write $P(-1.87\ \text{min} \leq \mu_1 - \mu_2 \leq -0.11\ \text{min}) = 0.95$.

If H_0: $\mu_1 = \mu_2$ is not rejected, then both samples are concluded to have come from populations having identical means, the common mean being denoted as μ. The best estimate of μ is the "pooled" or "weighted" mean:

$$\overline{X}_p = \frac{n_1\overline{X}_1 + n_2\overline{X}_2}{n_1 + n_2}, \tag{8.15}$$

which is the mean of the combined data from the two samples. Then the $1 - \alpha$ confidence interval for μ is

$$\overline{X}_p \pm t_{\alpha(2),\nu}\sqrt{\frac{s_p^2}{n_1 + n_2}}. \tag{8.16}$$

If H_0 is not rejected, it is the confidence interval of Equation 8.16, rather than those of Equations 8.13 and 8.14, that one would calculate.

As is the case with the t test, these confidence intervals are computed with the assumption that the two samples came from normal populations with the same variance. If the sampled distributions are far from meeting these conditions, then confidence intervals should be eschewed or, if they are reported, they should be presented with the caveat that they are only approximate.

If a separate $1 - \alpha$ confidence interval is calculated for μ_1 and for μ_2, it may be tempting to draw a conclusion about H_0: $\mu_1 = \mu_2$ by observing whether the two confidence intervals overlap. Overlap is the situation where L_1 for the larger mean is less than L_2 for the smaller mean, and such conclusions are made visually

enticing if the confidence intervals are presented in a graph (such as in Figure 7.5) or in a table (e.g., as in Table 7.3b). However, this is *not* a valid procedure for hypothesis testing (e.g., Barr, 1969; Browne, 1979; Ryan and Leadbetter, 2002; Schenker and Gentleman, 2001). If there is no overlap and the population means are consequently concluded to be different, this inference will be associated with a Type I error probability less than the specified α (very much less if the two standard errors are similar); and if there is overlap, resulting in failure to reject H_0, this conclusion will be associated with a probability of a Type II error greater than (i.e., a power less than) if the appropriate testing method were used. As an illustration if this, the data of Example 8.1 yield $L_1 = 8.11$ min and $L_2 = 9.39$ min for the mean of group B and $L_1 = 9.15$ min and $L_2 = 10.33$ min for the mean of group G; and the two confidence intervals overlap even though the null hypothesis is rejected.

(a) One-Tailed Confidence Limits for Difference between Means. If the two-sample t test is performed to assess one-tailed hypotheses (Section 7.2), then it is appropriate to determine a one-tailed confidence interval (as was done in Section 7.3a following a one-tailed one-sample t test). Using one-tailed critical values of t, the following confidence limits apply:

For H_0: $\mu_1 \leq \mu_2$ versus $\mu_1 > \mu_2$, or $\mu_1 - \mu_2 \leq \mu_0$ versus H_0: $\mu_1 - \mu_2 > \mu_0$:

$$L_1 = \overline{X} - \left(t_{\alpha(1),\nu}\right)\left(s_{\overline{X}_1 - \overline{X}_2}\right) \text{ and } L_2 = \infty.$$

For H_0: $\mu_1 \geq \mu_2$ versus $\mu_1 < \mu_2$, or $\mu_1 - \mu_2 \geq \mu_0$ versus $\mu_1 - \mu_2 < \mu_0$:

$$L_1 = -\infty \text{ and } L_2 = \overline{X} + \left(t_{\alpha(1),\nu}\right)\left(s_{\overline{X}_1 - \overline{X}_2}\right).$$

In Example 8.2, one-tailed confidence limits would be $L_1 = -\infty$ and $L_2 = (1.746)(1.55) = 2.71$ cm.

(b) Confidence Limits for Means when Variances Are Unequal. If the population variances are judged to be different enough to warrant using the Behrens-Fisher test (Section 8.1c) for H_0: $\mu_1 = \mu_2$, then the computation of confidence limits is altered from that shown above. If this test rejects the null hypothesis, a confidence interval for each of the two population means (μ_1 and μ_2) and a CI for the difference between the means ($\mu_1 - \mu_2$) should be determined. The $1 - \alpha$ confidence interval for μ_i is obtained as

$$\overline{X}_i \pm t_{\alpha(2),\nu'}\sqrt{\frac{s_i^2}{n_i}}, \text{ which is } \overline{X}_i \pm t_{\alpha(2),\nu'}\sqrt{s_{\overline{X}_i}^2}, \tag{8.17}$$

rather than by Equation 8.13, where ν' is from Equation 8.12. The confidence interval for the difference between the two population means is computed to be

$$\overline{X}_1 - \overline{X}_2 \pm \left(t_{\alpha(2),\nu'}\right)\left(s'_{\overline{X}_1 - \overline{X}_2}\right) \tag{8.18}$$

rather than by Equation 8.14, where $s'_{\overline{X}_1 - \overline{X}_2}$ is from Equation 8.11a or 8.11b. One-tailed confidence intervals are obtained as shown in Section 8.2a above, but using $s'_{\overline{X}_1 - \overline{X}_2}$ instead of $s_{\overline{X}_1 - \overline{X}_2}$. A confidence interval (two-tailed or one-tailed) for $\mu_1 - \mu_2$ includes zero when the associated H_0 is not rejected.

In Example 8.2a, H_0 is rejected, so it is appropriate to determine a 95% CI for μ_1, which is $37.4 \pm 2.274\sqrt{1.37} = 37.4$ days $\pm$ 2.7 days; for μ_2, which is $46.0 \pm 2.274\sqrt{10.93} = 46.0$ days $\pm$ 7.5 days; and for $\mu_1 - \mu_2$, which is $37.4 - 46.0 \pm (2.274)(3.51) = -8.6$ days $\pm$ 7.98 days.

If H_0: $\mu_1 = \mu_2$ is not rejected, then a confidence interval for the common mean, $\overline{X}_p$ (Equation 8.15), may be obtained by using the variance of the combined data from the two samples (call it s_t^2) and the degrees of freedom for those combined data ($\nu_t = n_1 + n_2 - 1$):

$$\overline{X}_p \pm t_{\alpha(2),\nu_t}\sqrt{\frac{s_t^2}{n_1 + n_2}}. \tag{8.19}$$

(c) Prediction Limits. As introduced in Section 7.3b, we can predict statistical characteristics of future sampling from populations from which samples have previously been analyzed. Such a desire might arise with data from an experiment such as in Example 8.1. Data were obtained from six animals treated with one drug and from seven animals treated with a second drug; and the mean blood-clotting times were concluded to be different under these two treatments. Equations 8.14 and 8.18 showed how confidence intervals can be obtained for the difference between means of two samples. It could also be asked what the difference between the means would be of an additional sample of m_1 animals treated with the first drug and a sample of an additional m_2 animals treated with the second.

For those two additional samples the best prediction of the difference between the two sample means would be $\overline{X}_1 - \overline{X}_2$, which in Example 8.1 is 8.75 min $-$ 9.74 min $= -0.99$ min; and there would be a $1 - \alpha$ probability that the difference between the two means would be contained in this prediction interval:

$$\overline{X}_1 - \overline{X}_2 \pm t_{\alpha(2),\nu}\sqrt{s_c^2}, \tag{8.19a}$$

where

$$s_c^2 = \frac{s_p^2}{m_1} + \frac{s_p^2}{n_1} + \frac{s_p^2}{m_2} + \frac{s_p^2}{n_2} \tag{8.19b}$$

(Hahn, 1977). For example, if an additional sample of 10 data were to be obtained for treatment with the first drug and an additional sample of 12 were to be acquired for treatment with the second drug, the 95% prediction limits for the difference between means would employ $s_p^2 = 0.5193\ \text{min}^2$, $n_1 = 6$, $n_2 = 7$, $m_1 = 10$, $m_2 = 12$, $t_{0.05(2),\nu=2.201}$, and $\nu = 11$; and $s_c^2 = 0.51$ min, so the 95% prediction limits would be $L_1 = -2.11$ min and $L_2 = 0.13$ min.

As the above procedure uses the pooled variance, s_p^2, it assumes that the two sampled populations have equal variances. If the two variances are thought to be quite different (the Behrens-Fisher situation discussed in Section 8.1c), then it is preferable to calculate the prediction interval as

$$\overline{X}_1 - \overline{X}_2 \pm t_{\alpha(2),\nu'}\sqrt{s_c^2}, \tag{8.19c}$$

where

$$s_c^2 = \frac{s_1^2}{m_1} + \frac{s_1^2}{n_1} + \frac{s_2^2}{m_2} + \frac{s_2^2}{n_2} \tag{8.19d}$$

2. Sample sizes not large enough to result in detection of a difference of biological importance can expend resources without yielding useful results, *and* sample sizes larger than needed to detect a difference of biological importance can result in unnecessary expenditure of resources.
3. Sample sizes not large enough to detect a difference of biological importance can expose subjects in the study to potentially harmful factors without advancing knowledge, *and* sample sizes larger than needed to detect a difference of biological importance can expose more subjects than necessary to potentially harmful factors or deny them exposure to potentially beneficial ones.

Assuming each sample comes from a normal population and the population variances are similar, we can estimate the minimum sample size to use to achieve desired test characteristics:

$$n \geq \frac{2s_p^2}{\delta^2}(t_{\alpha,\nu} + t_{\beta(1),\nu})^2 \tag{8.22}$$

(Cochran and Cox, 1957: 19–21).* Here, δ is the smallest population difference we wish to detect: $\delta = \mu_1 - \mu_2$ for the hypothesis test for which Equation 8.1 is used; $\delta = |\mu_1 - \mu_2| - \mu_0$ when Equation 8.9 is appropriate; $\delta = \mu_1 - \mu_2 - \mu_0$ when performing a test using Equation 8.10. In Equation 8.22, $t_{\alpha,\nu}$ may be either $t_{\alpha(1),\nu}$ or $t_{\alpha(2),\nu}$, depending, respectively, on whether a one-tailed or two-tailed test is to be performed.

Note that the required sample size depends on the following four quantities:

- δ, the minimum detectable difference between population means.† If we desire to detect a very small difference between means, then we shall need a larger sample than if we wished to detect only large differences.
- σ^2, the population variance. If the variability within samples is great, then a larger sample size is required to achieve a given ability of the test to detect differences between means. We need to know the variability to expect among the data; assuming the variance is the same in each of the two populations sampled, σ^2 is estimated by the pooled variance, s_p^2, obtained from similar studies.
- The significance level, α. If we perform the t test at a low α, then the critical value, $t_{\alpha,\nu}$, will be large and a large n is required to achieve a given ability to detect differences between means. That is, if we desire a low probability of committing a Type I error (i.e., falsely rejecting H_0), then we need large sample sizes.
- The power of the test, $1 - \beta$. If we desire a test with a high probability of detecting a difference between population means (i.e., a low probability of committing a Type II error), then $\beta(1)$ will be small, $t_{\beta(1)}$ will be large, and large sample sizes are required.

Example 8.4 shows how the needed sample size may be estimated. As $t_{\alpha(2),\nu}$ and $t_{\beta(1),\nu}$ depend on n, which is not yet known, Equation 8.22 must be solved iteratively, as we did with Equation 7.10. It matters little if the initial guess for n is inaccurate. Each iterative step will bring the estimate of n closer to the final result (which is

*The method of Section 10.3 may also be used for estimation of sample size, but it offers no substantial advantage over the present procedure.

†δ is lowercase Greek delta. If μ_0 in the statistical hypotheses is not zero (see discussion surrounding Equations 8.9 and 8.10), then δ is the amount by which the absolute value of the difference between the population means differs from μ_0.

declared when two successive iterations fail to change the value of n rounded to the next highest integer). In general, however, fewer iterations are required (i.e., the process is quicker) if one guesses high instead of low.

EXAMPLE 8.4 Estimation of Required Sample Size for a Two-Sample t Test

We desire to test for significant difference between the mean blood-clotting times of persons using two different drugs. We wish to test at the 0.05 level of significance, with a 90% chance of detecting a true difference between population means as small as 0.5 min. The within-population variability, based on a previous study of this type (Example 8.1), is estimated to be 0.52 min^2.

Let us guess that sample sizes of 100 will be required. Then, $\nu = 2(n - 1) = 2(100 - 1) = 198, t_{0.05(2),198} \approx 1.972, \beta = 1 - 0.90 = 0.10, t_{0.10(1),198} = 1.286$, and we calculate (by Equation 8.22):

$$n \geq \frac{2(0.52)}{(0.5)^2}(1.972 + 1.286)^2 = 44.2.$$

Let us now use $n = 45$ to determine $\nu = 2(n - 1) = 88$, $t_{0.05(2),88} = 1.987$, $t_{0.10(1),88} = 1.291$, and

$$n \geq \frac{2(0.52)}{(0.5)^2}(1.987 + 1.291)^2 = 44.7.$$

Therefore, we conclude that each of the two samples should contain at least 45 data.

If n_1 were constrained to be 30, then, using Equation 8.21, the required n_2 would be

$$n_2 = \frac{(44.7)(30)}{2(30) - 44.7} = 88.$$

For a given total number of data ($n_1 + n_2$), maximum test power and robustness occur when $n_1 = n_2$ (i.e., the sample sizes are equal). There are occasions, however, when equal sample sizes are impossible or impractical. If, for example, n_1 were fixed, then we would first determine n by Equation 8.22 and then find the required size of the second sample by Equation 8.21, as shown in Example 8.4. Note, from this example, that a total of $45 + 45 = 90$ data are required in the two equal-sized samples to achieve the desired power, whereas a total of $30 + 88 = 118$ data are needed if the two samples are as unequal as in this example. If $2n - 1 \leq 0$, then see the discussion following Equation 8.21.

(b) Minimum Detectable Difference. Equation 8.22 can be rearranged to estimate how small a population difference (δ, defined above) would be detectable with a given sample size:

$$\delta \geq \sqrt{\frac{2s_p^2}{n}}(t_{\alpha,\nu} + t_{\beta(1),\nu}). \tag{8.23}$$

The estimation of δ is demonstrated in Example 8.5.

EXAMPLE 8.5 Estimation of Minimum Detectable Difference in a Two-Sample *t* Test

In two-tailed testing for significant difference between mean blood-clotting times of persons using two different drugs, we desire to use the 0.05 level of significance and sample sizes of 20. What size difference between means do we have a 90% chance of detecting?

Using Equation 8.23 and the sample variance of Example 8.1, we calculate:

$$\delta = \sqrt{\frac{2(0.5193)}{20}}(t_{0.05(2),38} + t_{0.10(1),38})$$
$$= (0.2279)(2.024 + 1.304) = 0.76 \text{ min.}$$

In a Behrens-Fisher situation (i.e., if we don't assume that $\sigma_1^2 = \sigma_2^2$), Equation 8.23 would employ $\sqrt{s_1^2/n + s_2^2/n}$ instead of $\sqrt{2s_p^2/n}$.

(c) Power of the Test. Further rearrangement of Equation 8.22 results in

$$t_{\beta(1),\nu} \leq \frac{\delta}{\sqrt{\frac{2s_p^2}{n}}} - t_{\alpha,\nu}, \tag{8.24}$$

which is analogous to Equation 7.12 in Section 7.7. On computing $t_{\beta(1),\nu}$, one can consult Appendix Table B.3 to determine $\beta(1)$, whereupon $1 - \beta(1)$ is the power. But this generally will only result in declaring a range of power (e.g., $0.75 <$ power < 0.90). Some computer programs can provide the exact probability of $\beta(1)$, or we may, with only slight overestimation of power (as noted in the footnote in Section 7.7) consider $t_{\beta(1)}$ to be approximated by a normal deviate and may thus employ Appendix Table B.2.

If the two population variances are not assumed to be the same, then $\sqrt{s_1^2/n + s_2^2/n}$ would be used in place of $\sqrt{2s_p^2/n}$ in Equation 8.24.

The above procedure for estimating power is demonstrated in Example 8.6, along with the following method (which will be expanded on in the chapters on analysis of variance). We calculate

$$\phi = \sqrt{\frac{n\delta^2}{4s_p^2}} \tag{8.25}$$

(derived from Kirk, 1995: 182) and ϕ (lowercase Greek phi) is then located in Appendix Figure B.1a, along the lower axis (taking care to distinguish between ϕ's for $\alpha = 0.01$ and $\alpha = 0.05$). Along the top margin of the graph are indicated pooled degrees of freedom, ν, for α of either 0.01 or 0.05 (although the symbol ν_2 is used on the graph for a reason that will be apparent in later chapters). By noting where ϕ vertically intersects the curve for the appropriate ν, one can read across to either the left or right axis to find the estimate of power. As noted in Section 7.7c, the calculated power is an estimate of the probability of rejecting a false null hypothesis in future statistical tests; it is not the probability of rejecting H_0 in tests performed on the present set of data.

EXAMPLE 8.6 Estimation of the Power of a Two-Sample *t* Test

What would be the probability of detecting a true difference of 1.0 min between mean blood-clotting times of persons using the two drugs of Example 8.1, if $n_1 = n_2 = 15$, and $\alpha(2) = 0.05$?

For $n = 15$, $\nu = 2(n - 1) = 28$ and $t_{0.05(2),28} = 2.048$. Using Equation 8.24:

$$t_{\beta(1),28} \leq \frac{1.0}{\sqrt{\frac{2(0.5193)}{15}}} - 2.048 = 1.752.$$

Consulting Appendix Table B.3, we see that, for one-tailed probabilities and $\nu = 28$: $0.025 < P(t \geq 1.752) < 0.05$, so $0.025 < \beta < 0.05$.

$$\text{Power} = 1 - \beta, \text{ so } 0.95 < \text{power} < 0.975.$$

Or, by the normal approximation, we can estimate β by $P(Z \geq 1.752) = 0.04$. So power = 0.96. [The exact figures are $\beta = 0.045$ and power = 0.955.]

To use Appendix Figure B.1, we calculate

$$\phi = \sqrt{\frac{n\delta^2}{4s_p^2}} = \sqrt{\frac{(15)(1.0)}{4(0.5193)}} = 2.69.$$

In the first page of Appendix Figure B.1, we find that $\phi = 2.69$ and $\nu(= \nu_2) = 28$ are associated with a power of about 0.96.

(d) Unequal Sample Sizes. For a given total number of data, $n_1 + n_2$, the two-sample t test has maximum power and robustness when $n_1 = n_2$. However, if $n_1 \neq n_2$, the above procedure for determining minimum detectable difference (Equation 8.23) and power (Equations 8.24 and 8.25) can be performed using the harmonic mean of the two sample sizes (Cohen, 1988: 42):

$$n = \frac{2n_1n_2}{n_1 + n_2}. \tag{8.26}$$

Thus, for example, if $n_1 = 6$ and $n_2 = 7$, then

$$n = \frac{2(6)(7)}{6 + 7} = 6.46.$$

8.5 TESTING FOR DIFFERENCE BETWEEN TWO VARIANCES

If we have two samples of measurements, each sample taken at random from a normal population, we might ask if the variances of the two populations are equal. Consider the data of Example 8.7, where s_1^2, the estimate of σ_1^2, is 21.87 moths^2, and s_2^2, the estimate of σ_2^2, is 12.90 moths^2. The two-tailed hypotheses can be stated as H_0: $\sigma_1^2 = \sigma_2^2$ and H_A: $\sigma_1^2 \neq \sigma_2^2$, and we can ask, What is the probability of taking two samples from two populations having identical variances and having the two sample variances be as different as are s_1^2 and s_2^2? If this probability is rather low (say ≤ 0.05, as in previous chapters), then we reject the veracity of H_0 and conclude that the two samples came from populations having unequal variances. If the probability is greater

than α, we conclude that there is insufficient evidence to conclude that the variances of the two populations are not the same.

(a) Variance-Ratio Test. The hypotheses may be submitted to the two-sample *variance-ratio test*, for which one calculates

$$F = \frac{s_1^2}{s_2^2} \quad \text{or} \quad F = \frac{s_2^2}{s_1^2}, \text{ whichever is larger.*} \tag{8.27}$$

That is, the larger variance is placed in the numerator and the smaller in the denominator. We then ask whether the calculated ratio of sample variances (i.e., F) deviates so far from 1.0 as to enable us to reject H_0 at the α level of significance. For the data in Example 8.7, the calculated F is 1.70. The critical value, $F_{0.05(2),10,9}$, is obtained from Appendix Table B.4 and is found to be 3.59. As $1.70 < 3.59$, we do not reject H_0†.

Note that we consider degrees of freedom associated with the variances in both the numerator and denominator of the variance ratio. Furthermore, it is important to realize that F_{α,ν_1,ν_2} and F_{α,ν_2,ν_1} are not the same (unless, of course, $\nu_1 = \nu_2$), so the numerator and denominator degrees of freedom must be referred to in the correct order.

If $H_0: \sigma_1^2 = \sigma_2^2$ is not rejected, then s_1^2 and s_2^2 are assumed to be estimates of the same population variance, σ^2. The best estimate of this σ^2 that underlies both samples is called the *pooled variance* (introduced as Equation 8.4):

$$s_p^2 = \frac{SS_1 + SS_2}{\nu_1 + \nu_2} = \frac{\nu_1 s_1^2 + \nu_2 s_2^2}{\nu_1 + \nu_2}. \tag{8.28}$$

One-tailed hypotheses may also be submitted to the variance ratio test. For $H_0: \sigma_1^2 \geq \sigma_2^2$ and $H_A: \sigma_1^2 < \sigma_2^2$, s_2^2 is always used as the numerator of the variance ratio; for $H_0: \sigma_1^2 \leq \sigma_2^2$ and $H_A: \sigma_1^2 > \sigma_2^2$, s_1^2 is always used as the numerator. (A look at the alternate hypothesis tells us which variance belongs in the numerator of F in order to make $F > 1$.)

The critical value for a one-tailed test is $F_{\alpha(1),\nu_1,\nu_2}$ from Appendix Table B.4, where ν_1 is the degrees of freedom associated with the numerator of F and ν_2 is the degrees of freedom associated with the denominator. Example 8.8 presents the data submitted to the hypothesis test for whether seeds planted in a greenhouse have less variability in germination time than seeds planted outside.

The variance-ratio test is not a robust test, being severely and adversely affected by sampling nonnormal populations (e.g., Box, 1953; Church and Wike, 1976; Markowski and Markowski, 1990; Pearson, 1932; Tan, 1982), with deviations from mesokurtosis somewhat more important than asymmetry; and in cases of nonnormality the probability of a Type I error can be very much greater than α.

*What we know as the F statistic is a ratio of the variances of two normal distributions and was first described by R. A. Fisher in 1924 (and published in 1928) (Lehmann, 1999); the statistic was named in his honor by G. W. Snedecor (1934: 15).

†Some calculators and many computer programs have the capability of determining the probability of a given F. For the present example, we would thereby find that $P(F \geq 1.70) = 0.44$.

EXAMPLE 8.7 The Two-Tailed Variance Ratio Test for the Hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ and H_A: $\sigma_1^2 \neq \sigma_2^2$. The Data Are the Numbers of Moths Caught During the Night by 11 Traps of One Style and 10 Traps of a Second Style

H_0: $\sigma_1^2 = \sigma_2^2$

H_A: $\sigma_1^2 \neq \sigma_2^2$

$\alpha = 0.05$

Trap type 1	*Trap type 2*
41	52
35	57
33	62
36	55
40	64
46	57
31	56
37	55
34	60
30	59
38	

$n_1 = 11$	$n_2 = 10$
$\nu_1 = 10$	$\nu_2 = 9$
$SS_1 = 218.73$ moths^2	$SS_2 = 116.10$ moths^2
$s_1^2 = 21.87$ moths^2	$s_2^2 = 12.90$ moths^2

$$F = \frac{s_1^2}{s_2^2} = \frac{21.87}{12.90} = 1.70$$

$F_{0.05(2),10,9} = 3.96$

Therefore, do not reject H_0.

$$P(0.20 < F < 0.50)[P = 0.44]$$

$$s_p^2 = \frac{218.73 \text{ moths}^2 + 116.10 \text{ moths}^2}{10 + 9} = 17.62 \text{ moths}^2$$

The conclusion is that the variance of numbers of moths caught is the same for the two kinds of traps.

(b) Other Two-Sample Tests for Variances. A large number of statistical procedures to test differences between variances have been proposed and evaluated (e.g., Brown and Forsythe, 1974c; Church and Wike, 1976; Draper and Hunter, 1969; Levene, 1960; Miller, 1972; O'Neill and Mathews, 2000), often with the goal of avoiding the

EXAMPLE 8.8 A One-Tailed Variance-Ratio Test for the Hypothesis That the Germination Time for Pine Seeds Planted in a Greenhouse Is Less Variable Than for Pine Seeds Planted Outside

H_0: $\sigma_1^2 \geq \sigma_2^2$

H_A: $\sigma_1^2 < \sigma_2^2$

$\alpha = 0.05$

Germination Time (in Days) of Pine Seeds

Greenhouse	*Outside*
69.3	69.5
75.5	64.6
81.0	74.0
74.7	84.8
72.3	76.0
78.7	93.9
76.4	81.2
	73.4
	88.0

$n_1 = 7$	$n_2 = 9$
$\nu_1 = 6$	$\nu_2 = 8$
$SS_1 = 90.57$ days^2	$SS_2 = 700.98$ days^2
$s_1^2 = 15.10$ days^2	$s_2^2 = 87.62$ days^2

$$F = \frac{87.62}{15.10} = 5.80$$

$F_{0.05(1),8,6} = 4.15$

Therefore, reject H_0.

$$0.01 < P(F \geq 5.80) < 0.025 \quad [P = 0.023]$$

The conclusion is that the variance in germination time is less in plants grown in the greenhouse than in those grown outside.

lack of robustness of the variance-ratio test when samples come from nonnormal populations of data. A commonly encountered one is Levene's test, and its various modifications, which is typically less affected by nonnormal distributions than the variance-ratio test is.

The concept is to perform a two-sample t test (two-tailed or one-tailed, as the situation warrants; see Section 8.1), not on the values of X in the two samples but on values of the data after conversion to other quantities. A common conversion is to employ the deviations of each X from its group mean or median; that is, the two-sample t test is performed on $|X_{ij} - \overline{X}_i|$ or on $|X_{ij} -$ median of group $i|$. Other

data conversions, such as the square root or the logarithm of $|X_{ij} - \overline{X}_i|$, have also been examined (Brown and Forsythe, 1974c).

Levene's test is demonstrated in Example 8.9 for two-tailed hypotheses, and X' is used to denote $|X_i - \overline{X}|$. This procedure may also be employed to test one-tailed hypotheses about variances, either H_0: $\sigma_1^2 \geq \sigma_2^2$ vs. H_A: $\sigma_1^2 < \sigma_2^2$, or H_0: $\sigma_1^2 \leq \sigma_2^2$ vs. H_A: $\sigma_1^2 > \sigma_2^2$. This would be done by the one-tailed t-testing described in Section 8.1, using σ^2 in place of μ in the hypothesis statements and using $|X_i - \overline{X}|$ instead of X_i in the computations.

EXAMPLE 8.9 The Two-Sample Levene Test for H_0: $\sigma_1^2 = \sigma_2^2$ and H_A: $\sigma_1^2 \neq \sigma_2^2$. The Data Are Those of Example 8.7

H_0: $\sigma_1^2 = \sigma_2^2$

H_A: $\sigma_1^2 \neq \sigma_2^2$

$\alpha = 0.05$

For group 1: $\Sigma X = 401$ moths, n = 11, $\nu = 10$, $\overline{X} = 36.45$ moths.

For group 2: $\Sigma X = 577$ moths, n = 10, $\nu = 9$, $\overline{X} = 57.70$ moths.

Trap Type 1		*Trap Type 2*	
X_i	$X' = \lvert X_i - \overline{X}\rvert$	X_i	$X' = \lvert X_i - \overline{X}\rvert$
41	4.55	52	5.70
35	1.45	57	0.70
33	3.45	62	4.30
36	0.45	55	2.70
40	3.55	64	6.30
46	9.55	57	0.70
31	5.45	56	1.70
37	0.55	55	2.70
34	2.45	60	2.30
30	6.45	59	1.30
38	1.55		
ΣX_i = 401 moths	$\Sigma X_i' = \Sigma\lvert X_i - \overline{X}\rvert$ = 39.45 moths	ΣX_i = 577 moths	$\Sigma X_i' = \Sigma\lvert X_i - \overline{X}\rvert$ = 28.40 moths

For the absolute values of the deviations from the mean:

$X_1' = 39.45 \text{ moths}/11 = 3.59 \text{ moths}$ $\quad SS_1' = 77.25 \text{ moths}^2$

$X_2' = 28.40 \text{ moths}/10 = 2.84 \text{ moths}$ $\quad SS_2' = 35.44 \text{ moths}^2$

the calculated confidence intervals are only approximations, with the approximation poorer the further from normality the populations are.

Meeker and Hahn (1980) discuss calculation of prediction limits for the variance ratio and provide special tables for that purpose.

8.7 SAMPLE SIZE AND POWER IN TESTS FOR DIFFERENCE BETWEEN TWO VARIANCES

(a) Sample Size Required. In considering the variance-ratio test of Section 8.5, we may ask what minimum sample sizes are required to achieve specified test characteristics. Using the normal approximation recommended by Desu and Raghavarao (1990: 35), the following number of data is needed in each sample to test at the α level of significance with power of $1 - \beta$:

$$n = \left[\frac{Z_\alpha + Z_{\beta(1)}}{\ln\left(\frac{s_1^2}{s_2^2}\right)} \right]^2 + 2. \tag{8.32}$$

For analysts who prefer performing calculations with "common logarithms" (those employing base 10) to using "natural logarithms" (those in base e),* Equation 8.32 may be written equivalently as

$$n = \left[\frac{Z_\alpha + Z_{\beta(1)}}{(2.30259)\log\left(\frac{s_1^2}{s_2^2}\right)} \right]^2 + 2. \tag{8.33}$$

This sample-size estimate assumes that the samples are to be equal in size, which is generally preferable. If, however, it is desired to have unequal sample sizes (which will typically require more total data to achieve a particular power), one may specify that ν_1 is to be m times the size of ν_2; then (after Desu and Raghavarao, 1990: 35):

$$m = \frac{n_1 - 1}{n_2 - 1}, \tag{8.34}$$

$$n_2 = \frac{(m + 1)(n - 2)}{2m} + 2, \tag{8.35}$$

and

$$n_1 = m(n_2 - 1) + 1. \tag{8.36}$$

*In this book, *ln* will denote the natural, or Naperian, logarithm, and *log* will denote the common, or Briggsian, logarithm. These are named for the Scottish mathematician John Napier (1550–1617), who devised and named logarithms, and the English mathematician Henry Briggs (1561–1630), who adapted this computational method to base 10; the German astronomer Johann Kepler (1550–1617) was the first to use the abbreviation "Log," in 1624, and Italian mathematician Bonaventura Cavalieri (1598–1647) was the first to use "log" in 1632 (Cajori, 1928/9, Vol. II: 105–106; Gullber, 1997: 152). Sometimes $\log_e$ and $\log_{10}$ will be seen instead of ln and log, respectively.

As in Section 8.5, determination of whether s_1^2 or s_2^2 is placed in the numerator of the variance ratio in Equation 8.32 depends upon the hypothesis test, and Z_α is either a one-tailed or two-tailed normal deviate depending upon the hypothesis to be tested; n_1 and n_2 correspond to s_1^2 and s_2^2, respectively. This procedure is applicable if the variance ratio is >1.

(b) Power of the Test. We may also estimate what the power of the variance ratio test would be if specified sample sizes were used. If the two sample sizes are the same (i.e., $n = n_1 = n_2$), then Equations 8.32 and 8.33 may be rearranged, respectively, as follows:

$$Z_{\beta(1)} = \sqrt{n - 2}\ln\left(\frac{s_1^2}{s_2^2}\right) - Z_\alpha; \tag{8.37}$$

$$Z_{\beta(1)} = \sqrt{n - 2}(2.30259)\log\left(\frac{s_1^2}{s_2^2}\right) - Z_\alpha. \tag{8.38}$$

After $Z_{\beta(1)}$ is calculated, $\beta(1)$ is determined from the last line of Appendix Table B.3, or from Appendix Table B.2, or from a calculator or computer that gives probability of a normal deviate; and power $= 1 - \beta(1)$. If the two sample sizes are not the same, then the estimation of power may employ

$$Z_{\beta(1)} = \sqrt{\frac{2m(n_2 - 2)}{m + 1}}\ln\left(\frac{s_1^2}{s_2^2}\right) - Z_\alpha \tag{8.39}$$

or

$$Z_{\beta(1)} = \sqrt{\frac{2m(n_2 - 2)}{m + 1}}(2.30259)\log\left(\frac{s_1^2}{s_2^2}\right) - Z_\alpha, \tag{8.40}$$

where m is as in Equation 8.34.

8.8 TESTING FOR DIFFERENCE BETWEEN TWO COEFFICIENTS OF VARIATION

A very useful property of coefficients of variation is that they have no units of measurement. Thus, V's may be compared even if they are calculated from data having different units, as is the case in Example 8.10. And it may be desired to test the null hypothesis that two samples came from populations with the same coefficients of variation.

EXAMPLE 8.10 A Two-Tailed Test for Difference Between Two Coefficients of Variation

H_0: The intrinsic variability of male weights is the same as the intrinsic variability of male heights (i.e., the population coefficients of variation of weight and height are the same, namely H_0: $\sigma_1/\mu_1 = \sigma_2/\mu_2$).

H_0: The intrinsic variability of male weight is not the same as the intrinsic variability of male heights (i.e., the population coefficients of variation of weight and height are not the same, namely H_0: $\sigma_1/\mu_1 \neq \sigma_2/\mu_2$).

(a) The variance-ratio test.

Weight (kg)	*Log of weight*	*Height* (cm)	*Log of height*
72.5	1.86034	183.0	2.26245
71.7	1.85552	172.3	2.23629
60.8	1.78390	180.1	2.25551
63.2	1.80072	190.2	2.27921
71.4	1.85370	191.4	2.28194
73.1	1.86392	169.6	2.22943
77.9	1.89154	166.4	2.22115
75.7	1.87910	177.6	2.24944
72.0	1.85733	184.7	2.26647
69.0	1.83885	187.5	2.27300
		179.8	2.25479

$n_1 = 10$	$n_2 = 11$
$\nu_1 = 9$	$\nu_2 = 10$
$\overline{X}_1 = 70.73$ kg	$\overline{X}_2 = 180.24$ cm
$SS_1 = 246.1610$ kg^2	$SS_2 = 678.9455$ cm^2
$s_1^2 = 27.3512$ kg^2	$s_2^2 = 67.8946$ cm^2
$s_1 = 5.23$ kg	$s_2 = 8.24$ cm
$V_1 = 0.0739$	$V_2 = 0.0457$
$(SS_{log})_1 = 0.00987026$	$(SS_{log})_2 = 0.00400188$
$(s_{log}^2)_1 = 0.0010967$	$(s_{log}^2)_2 = 0.00040019$

$$F = \frac{0.0010967}{0.00040019} = 2.74$$

$F_{0.05(2),9,10} = 3.78$

Therefore, do not reject H_0.

$$0.10 < P < 0.20 \quad [P = 0.13]$$

It is concluded that the coefficient of variation is the same for the population of weights as it is for the population of heights.

(b) The Z test.

$$V_p = \frac{\nu_1 V_1 + \nu_2 V_2}{\nu_1 + \nu_2} = \frac{9(0.0739) + 10(0.0457)}{9 + 10} = \frac{1.1221}{19} = 0.0591$$

$V_p^2 = 0.003493$

$$Z = \frac{V_1 - V_2}{\sqrt{\left(\frac{V_p^2}{\nu_1} + \frac{V_p^2}{\nu_2}\right)(0.5 + V_p^2)}}$$

$$= \frac{0.0739 - 0.0457}{\sqrt{\left(\frac{0.003493}{9} + \frac{0.003493}{10}\right)(0.5 + 0.003493)}}$$

$$= \frac{0.0282}{0.0193} = 1.46$$

$Z_{0.05(2)} = t_{0.05(2),\infty} = 1.960$

Do not reject H_0.

$$0.10 < P < 0.20 \quad [P = 0.14]$$

It is concluded that the coefficient of variation is the same for the population of weights as it is for the population of heights.

Lewontin (1966) showed that

$$F = \frac{\left(s^2_{\log}\right)_1}{\left(s^2_{\log}\right)_2} \text{ or } F = \frac{\left(s^2_{\log}\right)_2}{\left(s^2_{\log}\right)_1} \tag{8.41}$$

may be used for a variance-ratio test, analogously to Equation 8.27. In Equation 8.41, $\left(s^2_{\log}\right)_i$ refers to the variance of the logarithms of the data in Sample i, where logarithms to any base may be employed. This procedure is applicable only if all of the data are positive (i.e., > 0), and it is demonstrated in Example 8.10a. Either two-tailed or one-tailed hypotheses may be tested, as shown in Section 8.5.

This variance-ratio test requires that the logarithms of the data in each sample come from a normal distribution. A procedure advanced by Miller (1991) allows testing when the data, not their logarithms, are from normal distributions (that have positive means and variances). The test statistic, as demonstrated in Example 8.10b, is

$$Z = \frac{V_1 - V_2}{\sqrt{\left(\frac{V_p^2}{\nu_1} + \frac{V_p^2}{\nu_2}\right)(0.5 + V_p^2)}}, \tag{8.42}$$

where

$$V_p = \frac{\nu_1 V_1 + \nu_2 V_2}{\nu_1 + \nu_2} \tag{8.43}$$

is referred to as the "pooled coefficient of variation," which is the best estimate of the population coefficient of variation, σ/μ, that is common to both populations if the null hypothesis of no difference is true.

This procedure is shown, as a two-tailed test, in Example 8.10b. Recall that critical values of Z may be read from the last line of the table of critical values of t (Appendix Table B.3), so $Z_{\alpha(2)} = t_{\alpha(2),\infty}$. One-tailed testing is also possible, in which case the alternate hypothesis would declare a specific direction of difference and one-tailed critical values ($t_{\infty(1),\alpha}$) would be consulted. This test works best if there are at least 10 data in each sample and each population's coefficient of variation is no larger than 0.33. An estimate of the power of the test is given by Miller and Feltz (1997).

8.9 CONFIDENCE LIMITS FOR THE DIFFERENCE BETWEEN TWO COEFFICIENTS OF VARIATION

Miller and Feltz (1997) have provided this $1 - \alpha$ confidence interval for $\sigma_1/\mu_1 - \sigma_2/\mu_2$, where the two sampled populations are normally distributed:

$$V_1 - V_2 \pm Z_{\alpha(2)}\sqrt{\frac{V_1^2}{\nu_1}(0.5 + V_1^2) + \frac{V_2^2}{\nu_1}(0.5 + V_2^2)}. \tag{8.44}$$

8.10 NONPARAMETRIC STATISTICAL METHODS

There is a large body of statistical methods that do not require the estimation of population parameters (such as μ and σ) and that test hypotheses that are not statements about population parameters. These statistical procedures are termed *nonparametric tests.** These are in contrast to procedures such as t tests, which are called *parametric tests* and which do rely upon estimates of population parameters and upon the statement of parameters in the statistical hypotheses. Although they may assume that the sampled populations have the same dispersion or shape, nonparametric methods typically do not make assumptions about the nature of the populations' distributions (e.g., there is no assumption of normality); thus they are sometimes referred to as *distribution-free tests.*† Both parametric and nonparametric tests require that the data have come at random from the sampled populations.

Nonparametric tests (such as the two-sample testing procedure described in Section 8.11) generally may be applied to any situation where we would be justified in employing a parametric test (such as the two-sample t test), as well as in some instances where the assumptions of the latter are untenable. If either the parametric or nonparametric approach is applicable, then the former will generally be more powerful than the latter (i.e., the parametric method will typically have a lower probability of committing a Type II error). However, often the difference in power is not great and can be compensated by a small increase in sample size for the nonparametric test. When the underlying assumptions of a parametric test are seriously violated, then the nonparametric counterpart may be decidedly more powerful.

Most nonparametric statistical techniques convert observed data to the ranks of the data (i.e., their numerical order). For example, measurements of 2.1, 2.3, 2.9, 3.6, and 4.0 kg would be analyzed via their ranks of 1, 2, 3, 4, and 5. A possible disadvantage of this rank transformation of data is that some information is lost (for example, the same ranks would result from measurements of 1.1, 1.3, 2.9, 4.6, and 5.0 kg). A possible advantage is that outliers (see Section 2.5) will have much less influence (for example, the same ranks would result from measurements of 2.1, 2.3, 2.9, 3.6, and 25.0 kg).

It is sometimes counseled that only nonparametric testing may be employed when dealing with ordinal-scale data, but such advice is based upon what Gaito (1980) calls "an old misconception"; this issue is also discussed by Anderson (1961), Gaito (1960), Savage (1957), and Stevens (1968). Interval-scale or ratio-scale measurements are not intrinsically required for the application of parametric testing procedures. Thus parametric techniques may be considered for ordinal-scale data *if* the assumptions of such methods are met—typically, random sampling from normally distributed populations with homogeneity of variances. But ordinal data often come

*The term *nonparametric* was first used by Jacob Wolfowitz in 1942 (David, 1995; Noether, 1984).

†The terms *nonparametric* and *distribution-free* are commonly used interchangeably, but they do not both define exactly the same set of statistical techniques (Noether, 1984).

from nonnormal populations, in which case properly subjecting them to parametric analysis depends upon the robustness of the test to the extent of nonnormality present.

8.11 TWO-SAMPLE RANK TESTING

Several nonparametric procedures, with various characteristics and assumptions, have been proposed for testing differences between the dispersions, or variabilities, of two populations (e.g., see Hettmansperger and McKean, 1998: 118–127; Hollander and Wolfe, 1999: 141–188; Sprent and Smeeton, 2001: 175–185). A far more common desire for nonparametric testing is to compare two populations' central tendencies (i.e., locations on the measurement scale) when underlying assumptions of the t test are not met. The most frequently employed such test is that originally proposed, for equal sample sizes, by Wilcoxon (1945)* and independently presented by Mann and Whitney (1947), for equal or unequal n's. It is called the Wilcoxon-Mann-Whitney test or, more commonly, the Mann-Whitney test.

(a) The Mann-Whitney Test. For this test, as for many other nonparametric procedures, the actual measurements are not employed, but we use instead the ranks of the measurements. The data may be ranked either from the highest to lowest or from the lowest to the highest values. Example 8.11 ranks the measurements from highest to lowest: The greatest height in either of the two groups is given rank 1, the second greatest height is assigned rank 2, and so on, with the shortest height being assigned rank N, where

$$N = n_1 + n_2. \tag{8.45}$$

A Mann-Whitney statistic is then calculated as

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1, \tag{8.46}$$

where n_a and n_2 are the number of observations in samples 1 and 2, respectively, and R_1 is the sum of the ranks in sample 1. The Mann-Whitney statistic can also be calculated as

$$U' = n_2 n_1 + \frac{n_2(n_2 + 1)}{2} - R_2 \tag{8.47}$$

(where R_2 is the sum of the ranks of the observations in sample 2), because the labeling of the two samples as 1 and 2 is arbitrary.† If Equation 8.46 has been used to calculate U, then U' can be obtained quickly as

$$U' = n_1 n_2 - U; \tag{8.48}$$

*Wilcoxon may have proposed this test primarily to avoid the drudgery of performing numerous t tests in a time before ubiquitous computer availability (Noether, 1984). Kruskal (1957) gives additional history, including identification of seven independent developments of the procedure Wilcoxon introduced, two of them prior to Wilcoxon, the earliest being by the German psychologist Gustav Deuchler in 1914.

†The Wilcoxon two-sample test (sometimes referred to as the Wilcoxon rank-sum test) uses a test statistic commonly called W, which is R_1 or R_2; the test is equivalent to the Mann-Whitney test, for $U = R_2 - n_2(n_2 + 1)/2$ and $U' = R_1 - n_1(n_1 + 1)/2$. U (or U') is also equal to the number of data in one sample that are exceeded by each datum in the other sample. Note in Example 8.11: For females, ranks 7 and 8 each exceed 6 male ranks and ranks 10, 11, and 12 each exceed all 7 males ranks, for a total of $6 + 6 + 7 + 7 + 7 = 33 = U$; for males, rank 9 exceeds 2 female ranks for a total of $2 = U'$.

EXAMPLE 8.11 The Mann-Whitney Test for Nonparametric Testing of the Two-Tailed Null Hypothesis That There Is No Difference Between the Heights of Male and Female Students

H_0: Male and female students are the same height.

H_A: Male and female students are not the same height.

$\alpha = 0.05$

Heights of males	*Heights of females*	*Ranks of male heights*	*Ranks of female heights*
193 cm	178 cm	1	6
188	173	2	8
185	168	3	10
183	165	4	11
180	163	5	12
175		7	
170		9	
$n_1 = 7$	$n_2 = 5$	$R_1 = 31$	$R_2 = 47$

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$
$$= (7)(5) + \frac{(7)(8)}{2} - 31$$
$$= 35 + 28 - 31$$
$$= 32$$
$$U' = n_1 n_2 - U$$
$$= (7)(5) - 32$$
$$= 3$$

$U_{0.05(2),7,5} = U_{0.05(2),5,7} = 30$

As $32 > 30$, H_0 is rejected.

$0.01 < P(U \geq 32 \text{ or } U' \leq 3) < 0.02 \quad [P = 0.018]^*$

Therefore, we conclude that height is different for male and female students.

and if Equation 8.47 has been used to compute U', then U can be ascertained as

$$U = n_1 n_2 - U'. \tag{8.49}$$

*In many of the examples in this book, the exact probability of a statistic from a nonparametric test (such as U) will be given within brackets. In some cases, this probability is obtainable from published sources (e.g., Owen, 1962). It may also be given by computer software, in which case there are two cautions: The computer result may not be accurate to the number of decimal places given, and the computer may have used an approximation (such as the normal approximation in the case of U; see Section 8.11d), which may result in a probability departing substantially from the exact probability, especially if the sample sizes are small.

For the two-tailed hypotheses, H_0: male and female students are the same height and H_A: male and female students are not the same height, the calculated U or U'—whichever is larger—is compared with the two-tailed value of $U_{\alpha(2),n_1,n_2}$ found in Appendix Table B.11. This table is set up assuming $n_1 \leq n_2$, so if $n_1 > n_2$, simply use $U_{\alpha(2),n_2,n_1}$ as the critical value. If either U or U' is as great as or greater than the critical value, H_0 is rejected at the α level of significance. A large U or U' will result when a preponderance of the large ranks occurs in one of the samples. As shown in Example 8.11, neither parameters nor parameter estimates are employed in the statistical hypotheses or in the calculations of U and U'.

The values of U in the table are those for probabilities less than or equal to the column headings. Therefore, the U of 32 in Example 8.11 is seen to have a probability of $0.01 < P \leq 0.02$. If the calculated U would have been 31, its probability would have been expressed as $0.02 < P < 0.05$.

We may assign ranks either from large to small data (as in Example 8.11), or from small to large, calling the smallest datum rank 1, the next largest rank 2, and so on. The value of U obtained using one ranking procedure will be the same as the value of U' using the other procedure. In a two-tailed test both U and U' are employed, so it makes no difference from which direction the ranks are assigned.

In summary, we note that after ranking the combined data of the two samples, we calculate U and U' using either Equations 8.46 and 8.48, which requires the determination of R_1, or Equations 8.47 and 8.49, which requires R_2. That is, the sum of the ranks for only one of the samples is needed. However, we may wish to compute both R_1 and R_2 in order to perform the following check on the assignment of ranks (which is especially desirable in the somewhat more complex case of assigning ranks to tied data, as will be shown below):

$$R_1 + R_2 = \frac{N(N + 1)}{2}. \tag{8.50}$$

Thus, in Example 8.11,

$$R_1 + R_2 = 30 + 48 = 78$$

should equal

$$\frac{N(N + 1)}{2} = \frac{12(12 + 1)}{2} = 78.$$

This provides a check on (although it does not guarantee the accuracy of) the assignment of ranks.

Note that hypotheses for the Mann-Whitney test are not statements about parameters (e.g., means or medians) of the two populations. Instead, they address the more general, less specific question of whether the two population distributions of data are the same. Basically, the question asked is whether it is likely that the two samples came at random from the two populations described in the null hypothesis. If samples at least that different would occur with a probability that is small (i.e., less than the significance level, such as 0.05), then H_0 is rejected.

The Mann-Whitney procedure serves to test for difference between medians under certain circumstances (such as when the two sampled populations have symmetrical distributions), but in general it addresses the less specific hypothesis of similarity between the two populations' distributions. The Watson test of Section 26.6 may also be employed when the Mann-Whitney test is applicable, but the latter is easier to perform and is more often found in statistical software.

(b) The Mann-Whitney Test with Tied Ranks. Example 8.12 demonstrates an important consideration encountered in tests requiring the ranking of observations. When two or more observations have exactly the same value, they are said to be *tied*. The rank assigned to each of the tied ranks is the mean of the ranks that would have been assigned to these ranks had they not been tied.* For example, in the present set of data, which are ranked from low to high, the third and fourth lowest values are tied at 32 words per minute, so they are each assigned the rank of $(3 + 4)/2 = 3.5$. The eighth, ninth, and tenth observations are tied at 44 words per minute, so each of them receives the rank of $(8 + 9 + 10)/3 = 9$. Once the ranks have been assigned by this procedure, U and U' are calculated as previously described.

(c) The One-Tailed Mann-Whitney Test. For one-tailed hypotheses we need to declare which tail of the Mann-Whitney distribution is of interest, as this will determine whether U or U' is the appropriate test statistic. This consideration is presented in Table 8.2. In Example 8.12 we have data that were ranked from lowest to highest and the alternate hypothesis states that the data in group 1 are greater in magnitude than those in group 2. Therefore, we need to compute U' and compare it to the one-tailed critical value, $U_{\alpha(1),n_1,n_2}$, from Appendix Table B.11.

TABLE 8.2: The Appropriate Test Statistic for the One-Tailed Mann-Whitney Test

	H_0: Group 1 $\geq$ Group 2 H_A: Group 1 $<$ Group 2	H_0: Group 1 $\leq$ Group 2 H_A: Group 1 $>$ Group 2
Ranking done from low to high	U	U'
Ranking done from high to low	U'	U

(d) The Normal Approximation to the Mann-Whitney Test. Note that Appendix Table B.11 can be used only if the size of the smaller sample does not exceed twenty and the size of the larger sample does not exceed forty. Fortunately, the distribution of U approaches the normal distribution for larger samples. For large n_1 and n_2 we use the fact that the U distribution has a mean of

$$\mu_U = \frac{n_1 n_2}{2}, \tag{8.51}$$

which may be calculated, equivalently, as

$$\mu_U = \frac{U + U'}{2}, \tag{8.51a}$$

and a standard error of

$$\sigma_U = \sqrt{\frac{n_1 n_2 (N + 1)}{12}}, \tag{8.52}$$

*Although other procedures have been proposed to deal with ties, assigning the rank mean has predominated for a long time (e.g., Kendall, 1945).

EXAMPLE 8.12 The One-Tailed Mann-Whitney Test Used to Determine the Effectiveness of High School Training on the Typing Speed of College Students. This Example Also Demonstrates the Assignment of Ranks to Tied Data

H_0: Typing speed is not greater in college students having had high school typing training.

H_A: Typing speed is greater in college students having had high school typing training.

$\alpha = 0.05$

Typing Speed (words per minute)	
With training (rank in parentheses)	*Without training (rank in parentheses)*
44 (9)	32 (3.5)
48 (12)	40 (7)
36 (6)	44 (9)
32 (3.5)	44 (9)
51 (13)	34 (5)
45 (11)	30 (2)
54 (14)	26 (1)
56 (15)	
$n_1 = 8$	$n_2 = 7$
$R_1 = 83.5$	$R_2 = 36.5$

Because ranking was done from low to high and the alternate hypothesis states that the data of group one are larger than the data of group two, use U' as the test statistic (as indicated in Table 8.2).

$$U' = n_2 n_1 + \frac{n_2(n_2 + 1)}{2} - R_2$$

$$= (7)(8) + \frac{(7)(8)}{2} - 36.5$$

$$= 56 + 28 - 36.5$$

$$= 47.5$$

$U_{0.05(1),8,7} = U_{0.05(1),7,8} = 43$

As $47.5 > 43$, reject H_0.

$$0.01 < P < 0.025 \quad [P = 0.012]$$

Consequently, it is concluded that college-student typing speed is greater for students who had typing training in high school.

where $N = n_1 + n_2$, as used earlier. Thus, if a U, or a U', is calculated from data where either n_1 or n_2 is greater than that in Appendix Table B.11, its significance can be determined by computing

$$Z = \frac{U - \mu_U}{\sigma_U} \tag{8.53}$$

or, using a correction for continuity, by

$$Z_c = \frac{|U - \mu_U| - 0.5}{\sigma_U}. \tag{8.54}$$

The continuity correction is included to account for the fact that Z is a continuous distribution, but U is a discrete distribution. However, it appears to be advisable only if the two-tailed P is about 0.05 or greater (as seen from an expansion of the presentation of Lehmann, 1975: 17).

Recalling that the t distribution with $\nu = \infty$ is identical to the normal distribution, the critical value, Z_α, is equal to the critical value, $t_{\alpha,\infty}$. The normal approximation is demonstrated in Example 8.13. When using the normal approximation for two-tailed testing, only U *or* U' (not both) need be calculated. If U' is computed instead of U, then U' is simply substituted for U in Equation 8.53 or 8.54, the rest of the testing procedure remaining the same.

EXAMPLE 8.13 The Normal Approximation to a One-Tailed Mann-Whitney Test to Determine Whether Animals Raised on a Dietary Supplement Reach a Greater Body Weight Than Those Raised on an Unsupplemented Diet

In the experiment, 22 animals (group 1) were raised on the supplemented diet, and 46 were raised on the unsupplemented diet (group 2). The body weights were ranked from 1 (for the smallest weight) to 68 (for the largest weight), and U was calculated to be 282.

H_0: Body weight of animals on the supplemented diet are not greater than those on the unsupplemented diet.

H_A: Body weight of animals on the supplemented diet are greater than those on the unsupplemented diet.

$n_1 = 22, n_2 = 46, N = 68$

$U = 282$

$$U' = n_1 n_2 - U = (22)(46) - 282 = 1012 - 282 = 730$$

$$\mu_U = \frac{n_1 n_2}{2} = \frac{(22)(46)}{2} = 506$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (N + 1)}{12}} = \sqrt{\frac{(22)(46)(68 + 1)}{12}} = 76.28$$

$$Z = \frac{U' - \mu_U}{\sigma_U} = \frac{224}{76.28} = 2.94$$

For a one-tailed test at $\alpha = 0.05, t_{0.05(1),\infty} = Z_{0.05(1)} = 1.6449$.

As $Z = 2.94 > 1.6449$, reject H_0. $[P = 0.0016]$

So we conclude that the supplemental diet results in greater body weight.

• One-tailed testing may also be performed using the normal approximation. Here one computes either U or U', in accordance with Table 8.2, and uses it in either Equation 8.55 or 8.56, respectively, inserting the correction term (-0.5) if P is about

0.025 or greater:

$$Z_c = \frac{U - \mu_U - 0.5}{\sigma_U}, \text{ if } U \text{ is used, or} \tag{8.55}$$

$$Z_c = \frac{U' - \mu_U - 0.5}{\sigma_U}, \text{ if } U' \text{ is used.} \tag{8.56}$$

The resultant Z_c is then compared to the one-tailed critical value, $Z_{\alpha(1)}$, or, equivalently, $t_{\alpha(1),\infty}$; and if $Z \geq$ the critical value, then H_0 is rejected.*

If tied ranks exist and the normal approximation is utilized, the computations are slightly modified as follows. One should calculate the quantity

$$\sum t = \sum (t_i^3 - t_i), \tag{8.57}$$

where t_i is the number of ties in a group of tied values, and the summation is performed over all groups of ties. Then,

$$\sigma_U = \sqrt{\frac{n_1 n_2}{N^2 - N} \cdot \frac{N^3 - N - \sum t}{12}}, \tag{8.58}$$

and this value is used in place of that from Equation 8.52. (The computation of $\sum t$ is demonstrated, in a similar context, in Example 10.11.)

The normal approximation is best for $\alpha(2) = 0.10$ or 0.05 [or for $\alpha(1) = 0.05$ or 0.025] and is also good for $\alpha(2) = 0.20$ or 0.02 [or for $\alpha(1) = 0.10$ or 0.01], with the approximation improving as sample sizes increase; for more extreme significance levels it is not as reliable, especially if n_1 and n_2 are dissimilar. Fahoome (2002) determined that the normal approximation (Equation 8.53) performed well at the two-tailed 0.05 level of significance (i.e., the probability of a Type I error was between 0.045 and 0.055) for sample sizes as small as 15, and at $\alpha(2) = 0.01$ (for P(Type I error) between 0.009 and 0.011) for n_1 and n_2 of at least 29. Indeed, in many cases with even smaller sample sizes, the normal approximation also yields Type I error probabilities very close to the exact probabilities of U obtained from specialized computer software (especially if there are few or no ties).† Further observations on the accuracy of this approximation are given at the end of Appendix Table B.11.

Buckle, Kraft, and van Eeden (1969) propose another distribution, which they refer to as the "uniform approximation." They show it to be more accurate for $n_1 \neq n_2$, especially when the difference between n_1 and n_2 is great, and especially for small α.

Fix and Hodges (1955) describe an approximation to the Mann-Whitney distribution that is much more accurate than the normal approximation but requires very involved computation. Hodges, Ramsey, and Wechsler (1990) presented a simpler method for a modified normal approximation that provides very good results for probabilities of about 0.001 or greater. Also, the two-sample t test may be applied to the ranks of the data (what is known as using the *rank transformation* of the data), with the probability of the resultant t approaching the exact probability for very large n. But these procedures do not appear to be generally preferable to the normal approximation described above, at least for the probabilities most often of interest.

*By this procedure, Z must be positive in order to reject H_0. If it is negative, then the probability of H_0 being true is $P > 0.50$.

†As a demonstration of this, in Example 8.11 the exact probability is 0.018 and the probability by the normal approximation is 0.019; and for Example 8.12, the exact probability and the normal approximation are both 0.012. In Exercise 8.12, P for U is 0.53 and P for Z is 0.52; and in Exercise 8.13, P for U is 0.41 and P for Z_c is 0.41.

(e) The Mann-Whitney Test with Ordinal Data. The Mann-Whitney test may also be used for ordinal data. Example 8.14 demonstrates this procedure. In this example, 25 undergraduate students were enrolled in an invertebrate zoology course. Each student was guided through the course by one of two teaching assistants, but the same examinations and grading criteria were applied to all students. On the basis of the students' final grades in the course, we wish to test the null hypothesis that students (students in general, not just these 25) perform equally well under both teaching assistants. The variable measured (i.e., the final grade) results in ordinal data, and the hypothesis is amenable to examination by the Mann-Whitney test.

(f) Mann-Whitney Hypotheses Employing a Specified Difference Other Than Zero. Using the two-sample t test, one can examine hypotheses such as H_0: $\mu_1 - \mu_2 = \mu_0$, where μ_0 is not zero. Similarly, the Mann-Whitney test can be applied to hypotheses such as H_0: males are at least 5 cm taller than females (a one-tailed hypothesis with data such as those in Example 8.11) or H_0: the letter grades of students in one course are at least one grade higher that those of students in a second course (a one-tailed hypothesis with data such as those in Example 8.14). In the first hypothesis, one would list all the male heights but list all the female heights after increasing each of them by 5 cm. Then these listed heights would be ranked and the Mann-Whitney analysis would proceed as usual. For testing the second hypothesis, the letter grades for the students in the first course would be listed unchanged, with the grades for the second course increased by one letter grade before listing. Then all the listed grades would be ranked and subjected to the Mann-Whitney test.*

When dealing with ratio- or interval-scale data, it is also possible to propose hypotheses employing a multiplication, rather than an addition, constant. Consider the two-tailed hypothesis H_0: the wings of one species of insect are two times the length of the wings of a second species. We could test this by listing the wing lengths of the first species, listing the wing lengths of the second species after multiplying each length by two, and then ranking the members of the combined two lists and subjecting the ranks to the Mann-Whitney test. The parametric t testing procedure, which assumes equal population variance, ordinarily would be inapplicable for such a hypothesis, because multiplying the data by a constant changes the variance of the data by the square of the constant.

(g) Violations of the Mann-Whitney Test Assumptions. If the underlying assumptions of the parametric analog of a nonparametric test are met, then either procedure may be employed but the parametric test will be the more powerful. The Mann-Whitney test is one of the most powerful of nonparametric tests. When the t-test assumptions are met, the power of the Mann-Whitney test approaches 95.5% (i.e., $3/\pi$) of the power of the t test as sample size increases (Mood, 1954).† And

*To increase these grades by one letter each, a grade of "B" would be changed to an "A," a "C" changed to a "B," and so on; a grade of "A" would have to be increased to a grade not on the original scale (e.g., call it a "Z") and, when ranking, we simply have to keep in mind that this new grade is higher than an "A."

†Mood (1954) credits an earlier statement of this to 1948 lecture notes of E. J. G. Pitman and to a 1950 Dutch publication by H. R. Van der Vaart. The statement that statistical test A is 0.955 as powerful as test B means that the power of test A with sample size of n tends (as n increases) toward having the same power as test B with sample size of $0.955n$; and this is referred to as the *asymptotic relative efficiency* (ARE) of test A compared to test B. Because of its development by Australian statistician Edwin James George Pitman (1897–1993), ARE is often called *Pitman efficiency*, which distinguishes it from a less commonly encountered definition of asymptotic relative

EXAMPLE 8.14 The Mann-Whitney Test for Ordinal Data

H_0: The performance of students is the same under the two teaching assistants.
H_A: Students do not perform equally well under the two teaching assistants.

$\alpha = 0.05$

Teaching Assistant A		Teaching Assistant B	
Grade	*Rank of grade*	*Grade*	*Rank of grade*
A	3	A	3
A	3	A	3
A	3	B+	7.5
A−	6	B+	7.5
B	10	B	10
B	10	B−	12
C+	13.5	C	16.5
C+	13.5	C	16.5
C	16.5	C−	19.5
C	16.5	D	22.5
C−	19.5	D	22.5
		D	22.5
		D	22.5
		D−	25
$n_1 = 11$		$n_2 = 14$	
$R_1 = 114.5$		$R_2 = 210.5$	

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$
$$= (11)(14) + \frac{(11)(12)}{2} - 114.5$$
$$= 154 + 66 - 114.5$$
$$= 105.5$$

$$U' = n_1 n_2 - U$$
$$= (11)(14) - 105.5$$
$$= 48.5$$

$U_{0.05(2),11,14} = 114$

As $105.5 < 114$, do not reject H_0.

$$0.10 < P(U \geq 105.5 \text{ or } U \leq 48.5) < 0.20$$

Thus, the conclusion is that student performance is the same under both teaching assistants.

efficiency by Bahadur (1967; Blair and Higgins, 1985). Although Pitman efficiency is defined in terms of very large n, it is generally a good expression of relative efficiency of two tests even with small n (Conover, 1999: 112).

for some extremely nonnormal distributions, the Mann-Whitney test is immensely more powerful (Blair and Higgins, 1980a, 1980b; Blair, Higgins, and Smitley, 1980; Hodges and Lehman, 1956). The power of the Mann-Whitney test will never be less than 86.4% of the power of the t test (Conover, 1997: 297; Hodges and Lehman, 1956).

The Mann-Whitney test does not assume normality of the sampled populations as the t test does, but the calculated U is affected not only by the difference between the locations of the two populations along the measurement scale but also by difference between the shapes or dispersons of the two populations' distributions (Boneau, 1962). However, the test is typically employed with the desire to conclude only whether there are differences between measurement locations, in which case it must be assumed that the two sampled populations have the same dispersion and shape, a premise that is often ignored, probably in the belief that the test is more robust to unequal dispersion than is the t test. But the Mann-Whitney test is, indeed, adversely affected by sizable differences in the variances or the shapes of the sampled populations, in that the probability of a Type I error is not the specified α (Fligner and Policello, 1981).* As with the two-sample t test, if the two sample sizes are not equal, and if the larger σ^2 is associated with the larger sample, then the probability of a Type I error will be less than α (and the test is called *conservative*); and if the smaller sample came from the population with the larger variance, then this probability will be greater than α (and the test is called *liberal*) (Zimmerman, 1987). The greater the difference between the variances, the greater the departure from α. In situations where the Mann-Whitney test is conservative, it has more power than the t test (Zimmerman, 1987). The power of the Mann-Whitney test may also be decreased, especially in the presence of outliers, to an extent to which the variances differ; but this decrease is far less than it is with t testing (Zimmerman, 1994, 1996, 1998, 2000). But in some cases unequal variances affect the probability of a Type I error using U more severely than if t or t' were employed (Zimmerman, 1998).

The Mann-Whitney test is included in the guidelines (described in Section 8.1d) for when various two-sample statistical procedures are appropriate.

8.12 TESTING FOR DIFFERENCE BETWEEN TWO MEDIANS

The null hypothesis that two samples came from populations having the same median can be tested by the *median test* described by Mood (1950: 394–395). The procedure is to determine the grand median for all the data in both samples and then to tabulate the numbers of data above and below the grand median in a 2×2 contingency table, as shown in Example 8.15. This contingency table can then be analyzed by the chi-square test of Section 23.3b or G test of Section 23.7.

Example 8.15 demonstrates the median test for the data of Example 8.14. In many cases, such as this one, one or more of the data will be equal to the grand median (in this instance a grade of C+) and, therefore, the number of data above

*Fligner and Policello (1981; Hollander and Wolfe, 1999: 135–139) addressed situations where the sampled populations have dissimilar variances (the "Behrens-Fisher problem" discussed in Section 8.1c), in addition to being nonnormal. They presented a modified Mann-Whitney procedure, requiring that the underlying distributions be symmetrical, along with tables of critical values for use with sample sizes ≤ 12 and with a normal approximation good when the n's are much larger than 12.

EXAMPLE 8.15 The Two-Sample Median Test, Using the Data of Example 8.14

H_0: The two samples came from populations with identical medians (i.e., the median performance is the same under the two teaching assistants).

H_A: The medians of the two sampled populations are not equal.

$\alpha = 0.05$

The median of all 25 measurements in Example 8.14 is $X_{(25+1)2} = X_{13}$ = grade of C+. The following 2 × 2 contingency table is then produced:

Number	*Sample 1*	*Sample 2*	*Total*
Above median	6	6	12
Not above median	3	8	11
Total	9	14	23

Analyzing this contingency table (Section 23.3):

$$X_c^2 = \frac{n\left(|f_{11}f_{22} - f_{12}f_{21}| - \frac{n}{2}\right)^2}{(C_1)(C_2)(R_1)(R_2)} \tag{8.59}$$
$$= 0.473.$$

$$X^2_{0.05,1} = 3.841$$

Therefore, do not reject H_0.

$$0.25 < P < 0.50 \quad [P = 0.49]$$

So it is concluded that the two samples did not come from populations with different medians.

and below the median will be less than the number of original data. Some authors and computer programs have preferred to tabulate the row categories as "above median" and "not above median" (that is, "at or below the median") instead of "above median" and "below median." This will retain in the analysis the original number of data, but it does not test the median-comparison hypothesis as well, and it can produce conclusions very different from those resulting from analyzing the same data categorized as "below median" and "not below median" (i.e., "at or above median"). Others have suggested deleting from the analysis any data that are tied at the grand median. This, too, will give results that may be quite different from the other procedures. If there are many data at the grand median, a good option is to place all data in a contingency table with three, instead of two, rows: "above median," "at median," and "below median."

The median test is about 64% as powerful as the two-sample t test when used on data to which the latter is applicable (Mood, 1954), and about 67% as powerful as the Mann-Whitney test of the preceding section.*

If the two sampled populations have equal variances and shapes, then the Mann-Whitney test (Section 8.11) is a test for difference between medians (Fligner and Policello, 1981).

One can also test whether the difference between two population medians is of a specified magnitude. This would be done in a fashion similar to that indicated in Section 8.11f for the Mann-Whitney test. For example, to hypothesize that the median of population 1 is X units greater than the median of population 2, X would be added to each datum in sample 2 (or X would be subtracted from each datum in sample 1) prior to performing the median test.

8.13 TWO-SAMPLE TESTING OF NOMINAL-SCALE DATA

We may compare two samples of nominal data simply by arranging the data in a $2 \times C$ contingency table and proceeding as described in Chapter 23.

8.14 TESTING FOR DIFFERENCE BETWEEN TWO DIVERSITY INDICES

If the Shannon index of diversity, H' (Section 4.7), is obtained for each of two samples, it may be desired to test the null hypothesis that the diversities of the two sampled populations are equal. Hutcheson (1970) proposed a t test for this purpose:

$$t = \frac{H'_1 - H'_2}{s_{H'_1 - H'_2}}, \tag{8.60}$$

where

$$s_{H'_1 - H'_2} = \sqrt{s^2_{H'_1} + s^2_{H'_2}}. \tag{8.61}$$

The variance of each H' may be approximated by

$$s^2_{H'} = \frac{\sum f_i \log^2 f_i - (\sum f_i \log f_i)^2/n}{n^2} \tag{8.62}$$

(Basharin, 1959; Lloyd, Zar, and Karr, 1968),† where s, f_i, and n are as defined in Section 4.7, and $\log^2 f$ signifies $(\log f)^2$. Logarithms to any base may be used for this calculation, but those to base 10 are most commonly employed. The degrees of freedom associated with the preceding t are approximated by

$$\nu = \frac{\left(s^2_{H'_1} + s^2_{H'_2}\right)^2}{\frac{\left(s^2_{H'_1}\right)^2}{n_1} + \frac{\left(s^2_{H'_2}\right)^2}{n_2}} \tag{8.63}$$

(Hutcheson, 1970).

*As the median test refers to a population parameter in hypothesis testing, it is not a nonparametric test; but it is a distribution-free procedure. Although it does not assume a specific underlying distribution (e.g., normal), it does assume that the two populations have the same shape (a characteristic that is addressed by Schlittgen, 1979).

†Bowman et al. (1971) give an approximation [their Equation (11b)] that is more accurate for very small n.

Example 8.16 demonstrates these computations. If one is faced with many calculations of $s^2_{H'_2}$, the tables of $f_i \log^2 f_i$ provided by Lloyd, Zar, and Karr (1968) will be helpful. One-tailed as well as two-tailed hypotheses may be tested by this procedure. Also, the population diversity indices may be hypothesized to differ by some value, μ_0, other than zero, in which case the numerator of t would be $|H'_1 - H'_2| - \mu_0$.

EXAMPLE 8.16 Comparing Two Indices of Diversity

H_0: The diversity of plant food items in the diet of Michigan blue jays is the same as the diversity of plant food items in the diet of Louisiana blue jays.

H_A: The diversity of plant food items in the diet of Michigan blue jays is not the same as in the diet of Louisiana blue jays.

$\alpha = 0.05$

Michigan Blue Jays

Diet item	f_i	$f_i \log f_i$	$f_i \log^2 f_i$
Oak	47	78.5886	131.4078
Corn	35	54.0424	83.4452
Blackberry	7	5.9157	4.9994
Beech	5	3.4949	2.4429
Cherry	3	1.4314	0.6830
Other	2	0.6021	0.1812
$s_1 = 6$	$n_1 = \sum f_i$ = 99	$\sum f_i \log f_i$ = 144.0751	$\sum f_i \log^2 f_i$ = 223.1595

$$H'_1 = \frac{n \log n - \sum f_i \log f_i}{n} = \frac{197.5679 - 144.0751}{99}$$

$$= 0.5403$$

$$s^2_{H'_1} = \frac{\sum f_i \log^2 f_i - (\sum f_i \log f_i)^2 / n}{n^2} = 0.00137602$$

Louisiana Blue Jays

Diet item	f_i	$f_i \log f_i$	$f_i \log^2 f_i$
Oak	48	80.6996	135.6755
Pine	23	31.3197	42.6489
Grape	11	11.4553	11.9294
Corn	13	14.4813	16.1313
Blueberry	8	7.2247	6.5246
Other	2	0.6021	0.1812
$s_2 = 6$	$n_2 = \sum f_i$ = 105	$\sum f_i \log f_i$ = 145.7827	$\sum f_i \log^2 f_i$ = 213.0909

$$H_2' = \frac{n \log n - \sum f_i \log f_i}{n} = \frac{212.2249 - 145.7827}{105} = 0.6328$$

$$s_{H_2'}^2 = \frac{\sum f_i \log^2 f_i - \left(\sum f_i \log f_i\right)^2 / n}{n^2} = 0.00096918$$

$$s_{H_1' - H_2'} = \sqrt{s_{H_1'}^2 + s_{H_2'}^2} = \sqrt{0.00137602 + 0.00096918} = 0.0484$$

$$t = \frac{H_1' - H_2'}{s_{H_1' - H_2'}} = \frac{-0.0925}{0.0484} = -1.911$$

$$\nu = \frac{\left(s_{H_1'}^2 + s_{H_2'}^2\right)^2}{\frac{\left(s_{H_1'}^2\right)^2}{n_1} + \frac{\left(s_{H_2'}^2\right)^2}{n_2}} = \frac{(0.00137602 + 0.00096918)^2}{\frac{(0.00137602)^2}{99} + \frac{(0.00096918)^2}{105}}$$

$$= \frac{0.000005499963}{0.000000028071} = 196$$

$t_{0.05(2),196} = 1.972$

Therefore, do not reject H_0.

$$0.05 < P < 0.10 \quad [P = 0.057]$$

The conclusion is that the diversity of food items is the same in birds from Michigan and Louisiana.

8.15 CODING DATA

As explained in Section 3.5, coding raw data can sometimes simplify computations. Coding will affect the sample statistics of this chapter (i.e., measures of central tendency and of variability, and their confidence limits) as described in Appendix C. The test statistics and hypothesis-test conclusions in Sections 8.1–8.7 will not be altered by coding, except that coding may not be used in performing the Levene test (Section 8.5b). Neither may coding be used in testing for difference between two coefficients of variation (Section 8.8), except that it is permissible if using the F test and coding by addition (or subtraction, but not multiplication or division). There is no effect of coding on the Mann-Whitney test (Section 8.11) or median test (Section 8.12). And, for testing difference between two diversity indices (Section 8.14), coding by multiplication (or division, but not addition or subtraction) may be employed.

Regarding the topics of Chapters 7 and 9, coding affects the sample statistics (and their confidence limits) as indicated in Appendix C. Coding may be employed for any of the hypothesis tests in those chapters, except that only coding by multiplication (or division, but not addition or subtraction) may be used for testing or coefficients of variation (Section 7.14).

EXERCISES

8.1. Using the following data, test the null hypothesis that male and female turtles have the same mean serum cholesterol concentrations.

Serum Cholesterol (mg/100 ml)

Male	*Female*
220.1	223.4
218.6	221.5
229.6	230.2
228.8	224.3
222.0	223.8
224.1	230.8
226.5	

8.2. It is proposed that animals with a northerly distribution have shorter appendages than animals from a southerly distribution. Test an appropriate hypothesis (by computing t), using the following wing-length data for birds (data are in millimeters).

Northern	*Southern*
120	116
113	117
125	121
118	114
116	116
114	118
119	123
	120

8.3. Two populations of animal body weights are randomly sampled, and $\overline{X}_1 = 4.6$ kg, $s_1^2 = 11.02$ kg^2, $n_1 = 18$, $\overline{X}_2 = 6.0$ kg, $s_2^2 = 4.35$ kg^2, and $n_2 = 26$. Test the hypotheses H_0: $\mu_1 \geq \mu_2$ and H_A: $\mu_1 < \mu_2$ using the Behrens-Fisher test.

8.4. If $\overline{X}_1 = 334.6$ g, $\overline{X}_2 = 349.8$ g, $SS_1 = 364.34$ g^2, $SS_2 = 286.78$ g^2, $n_1 = 19$, and $n_2 = 24$, test the hypothesis that the mean weight of population 2 is more than 10 g greater than the mean weight of population 1.

8.5. For the data of Exercise 8.1:

(a) If the null hypothesis is rejected, compute the 95% confidence limits for μ_1, μ_2, and $\mu_1 - \mu_2$. If H_0 is not rejected, compute the 95% confidence limits for the common population mean, μ_p.

(b) Calculate the 95% prediction interval for the difference between the mean of an additional 25 data from the male population and an additional 20 data from the female population.

8.6. A sample is to be taken from each of two populations from which previous samples of size 14 have had $SS_1 = 244.66$ (km/hr)^2 and $SS_2 = 289.18$ (km/hr)^2. What size sample should be taken from each population in order to estimate $\mu_1 - \mu_2$ to within 2.0 km/hr, with 95% confidence?

8.7. Consider the populations described in Exercise 8.6.

(a) How large a sample should we take from each population if we wish to detect a difference between μ_1 and μ_2 of at least 5.0 km/hr, using a 5% significance level and a t test with 90% power?

(b) If we take a sample of 20 from one population and 22 from the other, what is the smallest difference between μ_1 and μ_2 that we have a 90% probability of detecting with a t test using $\alpha = 0.05$?

(c) If $n_1 = n_2 = 50$, and $\alpha = 0.05$, what is the probability of rejecting H_0: $\mu_1 = \mu_2$ when $\mu_1 - \mu_2$ is as small as 2.0 km/hr?

8.8. The experimental data of Exercise 8.1 might have been collected to determine whether serum cholesterol concentrations varied as much in male turtles as in female turtles. With those data, use the variance-ratio test to assess H_0: $\sigma_1^2 = \sigma_2^2$ versus $\sigma_1^2 \neq \sigma_2^2$.

8.9. Let us propose that wings of a particular bird species vary in length more in the northern part of the species' range than in the southern portion. Use the variance ratio test for H_0: $\sigma_1 \leq \sigma_2$ versus H_A: $\sigma_1 > \sigma_2$ with the data of Exercise 8.2.

8.10. A sample of 21 data from one population has a variance of 38.71 g^2, and a sample of 20 data from a second population has a variance of 21.35 g^2.

(a) Calculate the 95% two-tailed confidence interval for the ratio of σ_1^2/σ_2^2.

(b) How large a sample must be taken from each population if we wish to have a 90% chance of rejecting H_0: $\sigma_1^2 \leq \sigma_2^2$ when H_A: $\sigma_1^2 > \sigma_2^2$ is true and we apply the variance-ratio test at the 5% level of significance?

(c) What would be the power of a variance-ratio test of this H_0, with $\alpha = 0.05$, if sample sizes of 20 were used?

8.11. A sample of twenty-nine plant heights of members of a certain species had $\overline{X}_1 = 10.74$ cm and $s^2 = 14.62$ cm^2, and the heights of a sample of twenty-five from a second species had $\overline{X}_2 = 14.32$ cm and $s^2 = 8.45$ cm^2. Test the null hypothesis that the coefficients of variation of the two sampled populations are the same.

8.12. Using the Mann-Whitney test, test the appropriate hypotheses for the data in Exercise 8.1.

8.13. Using the Mann-Whitney procedure, test the appropriate hypotheses for the data in Exercise 8.2.

8.14. The following data are volumes (in cubic microns) of avian erythrocytes taken from normal (diploid) and intersex (triploid) individuals. Test the hypothesis (using the Mann-Whitney test) that the volume of intersex cells is 1.5 times the volume of normal cells.

Normal	*Intersex*
248	380
236	391
269	377
254	392
249	398
251	374
260	
245	
239	
255	

CHAPTER 9

Paired-Sample Hypotheses

The two-sample testing procedures discussed in Chapter 8 apply when the two samples are independent, independence implying that each datum in one sample is in no way associated with any specific datum in the other sample. However, there are instances when each observation in Sample 1 is in some way physically associated with an observation in Sample 2, so that the data may be said to occur in pairs.

For example, we might wish to test the null hypothesis that the left foreleg and left hindleg lengths of deer are equal. We could make these two measurements on a number of deer, but we would have to remember that the variation among the data might be owing to two possible factors. First, the null hypothesis might be false, there being, in fact, a difference between foreleg and hindleg length. Second, deer are of different sizes, and for each deer the hindleg length is correlated with the foreleg length (i.e., a deer with a large front leg is likely to have a large hind leg). Thus, as Example 9.1 shows, the data can be tabulated in pairs, one pair (i.e., one hindleg measurement and one foreleg measurement) per animal.

9.1 TESTING MEAN DIFFERENCE BETWEEN PAIRED SAMPLES

The two-tailed hypotheses implied by Example 9.1 are $H_0: \mu_1 - \mu_2 = 0$ and $H_A: \mu_1 - \mu_2 \neq 0$ (which, as pointed out in Section 8.1, could also be stated $H_0: \mu_1 = \mu_2$ and $H_A: \mu_1 \neq \mu_2$). However, we can define a mean population difference, μ_d, as $\mu_1 - \mu_2$, and write the hypotheses as $H_0: \mu_d = 0$ and $H_A: \mu_d \neq 0$. Although the use of either μ_d or $\mu_1 - \mu_2$ is correct, the former will be used here when it implies the paired-sample situation.

The test statistic for the null hypothesis is

$$t = \frac{\bar{d}}{s_{\bar{d}}}. \tag{9.1}$$

Therefore, we do not use the original measurements for the two samples, but only the difference within each pair of measurements. One deals, then, with a sample of d_j values, whose mean is $\bar{d}$ and whose variance, standard deviation, and standard error are denoted as $s_{\bar{d}}^2$, s_d, and $s_{\bar{d}}$ respectively. Thus, the *paired-sample t test*, as this procedure may be called, is essentially a one-sample t test, analogous to that described

EXAMPLE 9.1 The Two-Tailed Paired-Sample *t* Test

H_0: $\mu_d = 0$
H_A: $\mu_d \neq 0$
$\alpha = 0.05$

Deer (j)	*Hindleg length* (cm) (X_{1j})	*Foreleg length* (cm) (X_{2j})	*Difference* (cm) ($d_j = X_{1j} - X_{2j}$)
1	142	138	4
2	140	136	4
3	144	147	−3
4	144	139	5
5	142	143	−1
6	146	141	5
7	149	143	6
8	150	145	5
9	142	136	6
10	148	146	2

$n = 10$ $\quad \bar{d} = 3.3$ cm

$s_d^2 = 9.3444$ cm^2 $\quad s_{\bar{d}} = 0.97$ cm

$\nu = n - 1 = 9$ $\quad t = \dfrac{\bar{d}}{s_{\bar{d}}} = \dfrac{3.3}{0.97} = 3.402$

$t_{0.05(2),9} = 2.262$ $\quad$ Therefore, reject H_0.

$$0.005 < P(|t| \geq 3.402) < 0.01 \quad [P = 0.008]$$

in Sections 7.1 and 7.2. In the paired-sample t test, n is the number of differences (i.e., the number of pairs of data), and the degrees of freedom are $\nu = n - 1$. Note that the hypotheses used in Example 9.1 are special cases of the general hypotheses H_0: $\mu_d = \mu_0$ and H_A: $\mu_d \neq \mu_0$, where μ_0 is usually, but not always, zero.

For one-tailed hypotheses with paired samples, one can test either H_0: $\mu_d \geq \mu_0$ and H_A: $\mu_d < \mu_0$, or H_0: $\mu_d \leq \mu_0$ and H_A: $\mu_d > \mu_0$, depending on the question to be asked. Example 9.2 presents data from an experiment designed to test whether a new fertilizer results in an increase of more than 250 kg/ha in crop yield over the old fertilizer. For testing this hypothesis, 18 test plots of the crop were set up. It is probably unlikely to find 18 field plots having exactly the same conditions of soil, moisture, wind, and so on, but it should be possible to set up two plots with similar environmental conditions. If so, then the experimenter would be wise to set up nine pairs of plots, applying the new fertilizer randomly to one plot of each pair and the old fertilizer to the other plot of that pair. As Example 9.2 shows, the statistical hypotheses to be tested are H_0: $\mu_d \leq 250$ kg/ha and H_A: $\mu_d > 250$ kg/ha.

Paired-sample t-testing assumes that each datum in one sample is associated with one, *but only one*, datum in the other sample. So, in the last example, each yield using

EXAMPLE 9.2 A One-Tailed Paired-Sample *t* Test

H_0: $\mu_d \leq 250$ kg/ha

H_A: $\mu_d > 250$ kg/ha

$\alpha = 0.05$

	Crop Yield (kg/ha)		
Plot (j)	*With new fertilizer* (X_{1j})	*With old fertilizer* (X_{2j})	d_j
1	2250	1920	330
2	2410	2020	390
3	2260	2060	200
4	2200	1960	240
5	2360	1960	400
6	2320	2140	180
7	2240	1980	260
8	2300	1940	360
9	2090	1790	300

$n = 9$ $\quad \bar{d} = 295.6$ kg/ha

$s_d^2 = 6502.78\ (\text{kg/ha})^2$ $\quad s_{\bar{d}} = 26.9$ kg/ha

$\nu = n - 1 = 8$ $\quad t = \dfrac{\bar{d} - 250}{s_{\bar{d}}} = 1.695$

$t_{0.05(1),8} = 1.860$ $\quad$ Therefore, do not reject H_0.

$0.05 < P < 0.10 \quad [P = 0.064]$

new fertilizer is paired with only one yield using old fertilizer; and it would have been inappropriate to have some tracts of land large enough to collect two or more crop yields using each of the fertilizers.

The paired-sample t test does not have the normality and equality of variances assumptions of the two-sample t test, but it does assume that the differences, d_i, come from a normally distributed population of differences. If a nonnormal distribution of differences is doubted, the nonparametric test of Section 9.5 should be considered.

If there is, in fact, pairwise association of data from the two samples, then analysis by the two-sample t test will often be less powerful than if the paired-sample t test was employed, and the two-sample test will not have a probability of a Type I error equal to the specified significance level, α. It appears that the latter probability will be increasingly less than α for increasingly large correlation between the pairwise data (and, in the less common situation where there is a negative correlation between the data, the probability will be greater than α); and only a small relationship is needed

to make the paired-sample test advantageous (Hines, 1996; Pollak and Cohen, 1981; Zimmerman, 1997). If the data from Example 9.1 were subjected (inappropriately) to the two-sample t test, rather than to the paired-sample t test, a difference would not have been concluded, and a Type II error would have been committed.

9.2 CONFIDENCE LIMITS FOR THE POPULATION MEAN DIFFERENCE

In paired-sample testing we deal with a sample of differences, d_j, so confidence limits for the mean of a population of differences, μ_d, may be determined as in Section 7.3. In the manner of Equation 7.6, the $1 - \alpha$ confidence interval for μ_d is

$$\bar{d} \pm t_{\alpha(2),\nu} s_{\bar{d}}. \tag{9.2}$$

For example, for the data in Example 9.1, we can compute the 95% confidence interval for μ_d to be $3.3\text{ cm} \pm (2.262)(0.97\text{ cm}) = 3.3\text{ cm} \pm 2.2\text{ cm}$; the 95% confidence limits are $L_1 = 1.1$ cm and $L_2 = 5.5$ cm.

Furthermore, we may ask, as in Section 7.5, how large a sample is required to be $1 - \alpha$ confident in estimating μ_d to within $\pm\ d$ (using Equation 7.7).

9.3 POWER, DETECTABLE DIFFERENCE AND SAMPLE SIZE IN PAIRED-SAMPLE TESTING OF MEANS

By considering the paired-sample test to be a one-sample t test for a sample of differences, we may employ the procedures of Section 7.7 to acquire estimates of required sample size (n), minimum detectable difference (δ), and power ($1 - \beta$), using Equations 7.10, 7.11, and 7.12, respectively.

9.4 TESTING FOR DIFFERENCE BETWEEN VARIANCES FROM TWO CORRELATED POPULATIONS

The tests of Section 8.5 address hypotheses comparing σ_1^2 to σ_2^2 when the two samples of data are independent. For example, if we wanted to compare the variance of the lengths of deer forelegs with the variance of deer hindlegs, we could measure a sample of foreleg lengths of several deer and a sample of hindleg lengths from a different group of deer. As these are independent samples, the variance of the foreleg sample could be compared to the variance of the hindleg sample by the procedures of Section 8.5. However, just as the paired-sample comparison of means is more powerful than independent-sample comparison of means when the data are paired (i.e., when there is an association between each member of one sample and a member of the other sample), there is a variance-comparison test more powerful than those of Section 8.5 if the data are paired (as they are in Example 9.1). This test takes into account the amount of association between the members of the pairs of data, as presented by Snedecor and Cochran (1989: 192–193) based upon a procedure of Pitman (1939). We compute:

$$t = \frac{(F - 1)\sqrt{n - 2}}{2\sqrt{F(1 - r^2)}}. \tag{9.3}$$

Here, n is the sample size common to both samples, r is the correlation coefficient described in Section 19.1 (Equation 19.1), and the degrees of freedom associated with t are $\nu = n - 2$. For a two-tailed test (H_0: $\sigma_1^2 = \sigma_2^2$ vs. H_A: $\sigma_1^2 \neq \sigma_2^2$), either $F = s_1^2/s_2^2$ or $F = s_2^2/s_1^2$ may be used, as indicated in Equation 8.29, and H_0 is rejected if $|t| \geq t_{\alpha(2),\nu}$. This is demonstrated in Example 9.3. For the one-tailed hypotheses, H_0: $\sigma_1^2 \leq \sigma_2^2$ versus H_A: $\sigma_1^2 > \sigma_2^2$, use $F = s_1^2/s_2^2$; for H_0: $\sigma_1^2 \geq \sigma_2^2$ versus H_A: $\sigma_1^2 < \sigma_2^2$, use $F = s_2^2/s_1^2$; and a one-tailed test rejects H_0 if $t \geq t_{\alpha(1),\nu}$.

McCulloch (1987) showed that this t test is adversely affected if the two sampled populations do not have a normal distribution, in that the probability of a Type I error can depart greatly from the stated α. He demonstrated a testing procedure that is very little affected by nonnormality and is only slightly less powerful than t when the underlying populations are normal. It utilizes the differences and the sums of the members of the pairs, just as described in the preceding paragraph; but, instead of the parametric correlation coefficient (r) referred to above, it employs the nonparametric correlation coefficient (r_s) and the associated significance testing of Section 19.9. This technique may be used for two-tailed or one-tailed testing.

EXAMPLE 9.3 Testing for Difference Between the Variances of Two Paired Samples

H_0: $\sigma_1^2 = \sigma_2^2$

H_A: $\sigma_1^2 \neq \sigma_2^2$

$\alpha = 0.05$

Using the paired-sample data of Example 9.1:

$n = 10$; $\nu = 8$

$\sum x^2 = 104.10$; $\sum y^2 = 146.40$

$\sum xy = 83.20$

$s_1^2 = 11.57\ \text{cm}^2$; $s_2^2 = 16.27\ \text{cm}^2$

$F = 11.57\ \text{cm}^2/16.27\ \text{cm}^2 = 0.7111$

Using Equation 19.1, $r = 0.6739$.

Using Equation 9.3:

$t = -0.656$ and $t_{0.05(2),8} = 2.306$, so H_0 is not rejected.

$P > 0.50$ [$P = 0.54$]

9.5 PAIRED-SAMPLE TESTING BY RANKS

The *Wilcoxon paired-sample test* (Wilcoxon, 1945; Wilcoxon and Wilcox, 1964: 9) is a nonparametric analogue to the paired-sample t test, just as the Mann-Whitney test is a nonparametric procedure analogous to the two-sample t test. The literature refers to the test by a variety of names, but usually in conjunction with Wilcoxon's name* and some wording such as "paired sample" or "matched pairs," sometimes together with a phrase like "rank sum" or "signed rank."

Whenever the paired-sample t test is applicable, the Wilcoxon paired-sample test is also applicable. Section 7.9 introduced the Wilcoxon procedure as a nonparametric

*Frank Wilcoxon (1892–1965), American (born in Ireland) chemist and statistician, a major developer of statistical methods based on ranks (Bradley and Hollander, 1978).

one-sample test, but it is also very useful for paired-sample testing, just as the one-sample t and the paired-sample t test are basically the same. If the d_j values are from a normal distribution, then the Wilcoxon test has $3/\pi$ (i.e., 95.5%) of the power in detecting differences as the t test has (Conover, 1999: 363; Mood, 1954). But when the d_j's cannot be assumed to be from a normal distribution, the parametric paired-sample t test should be avoided, for with nonnormality, the Wilcoxon paired-sample test will be more powerful, sometimes much more powerful (Blair and Higgins, 1985). However, the Wilcoxon test assumes the population of differences is symmetrical (which the t test also does, for the normal distribution is symmetrical). The sign test of Section 24.6 could also be used for one-sample testing of the d_j's. It has only $2/\pi$ (64%) of the power of the t test, and only 67% of the power of the Wilcoxon test, when the normality assumption of the t test is met (Conover, 1999: 164). But the sign test does not assume symmetry and is therefore preferable to the Wilcoxon test when the differences come from a very asymmetric population.

Example 9.4 demonstrates the use of the Wilcoxon paired-sample test with the ratio-scale data of Example 9.1, and it is best applied to ratio- or interval-scale data. The testing procedure involves the calculation of differences, as does the paired-sample t test. Then one ranks the absolute values of those differences, from low to high, and affixes the sign of each difference to the corresponding rank. As introduced in Section 8.11, the rank assigned to tied observations is the mean of the ranks that would have been assigned to the observations had they not been tied. Differences of zero are ignored in this test.

Then we sum the ranks having a plus sign (calling this sum T_+) and the ranks with a minus sign (labeling this sum T_-). For a two-tailed test (as in Example 9.4), we reject H_0 if either T_+ or T_- is *less than or equal to* the critical value, $T_{\alpha(2),n}$, from Appendix Table B.12. In doing so, n is the number of differences that are not zero.

Having calculated either T_+ or T_-, the other can be determined as

$$T_- = \frac{n(n+1)}{2} - T_+ \tag{9.4}$$

or

$$T_+ = \frac{n(n+1)}{2} - T_- . \tag{9.5}$$

A different value of T_+ (call it T'_+) or T_- (call it T'_-) will be obtained if rank 1 is assigned to the largest, rather than the smallest, d_i (i.e., the absolute values of the d_j's are ranked from high to low). If this is done, the test statistics are obtainable as

$$T_+ = m(n+1) - T'_+ \tag{9.6}$$

and

$$T_- = m(n+1) - T'_- , \tag{9.7}$$

where m is the number of ranks with the sign being considered.

Pratt (1959) recommended maintaining differences of zero until after ranking, and thereafter ignoring the ranks assigned to the zeros. This procedure may yield slightly better results in some circumstances, though worse results in others (Conover, 1973). If used, then the critical values of Rahe (1974) should be consulted or the normal approximation employed (see the following section) instead of using critical values of T from Appendix Table B.12.

If data are paired, the undesirable use of the Mann-Whitney test, instead of the Wilcoxon paired-sample test, may lead to a greater Type II error, with the concomitant inability to detect actual population differences.

EXAMPLE 9.4 The Wilcoxon Paired-Sample Test Applied to the Data of Example 9.1

H_0: Deer hindleg length is the same as foreleg length.

H_A: Deer hindleg length is not the same as foreleg length.

$\alpha = 0.05$

Deer (j)	*Hindleg length* (cm) (X_{1j})	*Foreleg length* (cm) (X_{2j})	*Difference* ($d_j = X_{1j} - X_{2j}$)	*Rank of* $\lvert d_j \rvert$	*Signed rank of* $\lvert d_j \rvert$
1	142	138	4	4.5	4.5
2	140	136	4	4.5	4.5
3	144	147	−3	3	−3
4	144	139	5	7	7
5	142	143	−1	1	−1
6	146	141	5	7	7
7	149	143	6	9.5	9.5
8	150	145	5	7	7
9	142	136	6	9.5	9.5
10	148	146	2	2	2

$n = 10$

$T_+ = 4.5 + 4.5 + 7 + 7 + 9.5 + 7 + 9.5 + 2 = 51$

$T_- = 3 + 1 = 4$

$T_{0.05(2),10} = 8$

Since $T_- < T_{0.05(2),10}$, H_0 is rejected.

$0.01 < P(T_- \text{ or } T_+ \leq 4) < 0.02 \quad [P = 0.014]$

The Wilcoxon paired-sample test has an underlying assumption that the sampled population of d_j's is symmetrical about the median. Another nonparametric test for paired samples is the sign test (described in Section 24.6), which does not have this assumption but is less powerful if the assumption is met.

Section 8.11f discussed the Mann-Whitney test for hypotheses dealing with differences of specified magnitude. The Wilcoxon paired-sample test can be used in a similar fashion. For instance, it can be asked whether the hindlegs in the population sampled in Example 9.4 are 3 cm longer than the lengths of the forelegs. This can be done by applying the Wilcoxon paired-sample test after subtracting 3 cm from each hindleg length in the sample (or adding 3 cm to each foreleg length).

(a) The One-Tailed Wilcoxon Paired-Sample Test. For one-tailed testing we use one-tailed critical values from Appendix Table B.12 and either T_+ or T_- as follows. For the hypotheses

H_0: Measurements in population 1 $\leq$ measurements in population 2
and H_A: Measurements in population 1 $>$ measurements in population 2,

H_0 is rejected if $T_- \leq T_{\alpha(1),n}$. For the opposite hypotheses:

H_0: Measurements in population 1 $\geq$ measurements in population 2
and H_A: Measurements in population 1 $<$ measurements in population 2,

reject H_0 if $T_+ \leq T_{\alpha(1),n}$.

(b) The Normal Approximation to the Wilcoxon Paired-Sample Test. For data consisting of more than 100 pairs* (the limit of Appendix Table B.12), the significance of T (where either T_+ or T_- may be used for T) may be determined by considering that for such large samples the distribution of T is closely approximated by a normal distribution with a mean of

$$\mu_T = \frac{n(n+1)}{4} \tag{9.8}$$

and a standard error of

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}. \tag{9.9}$$

Thus, we can calculate

$$Z = \frac{|T - \mu_T|}{\sigma_T}, \tag{9.10}$$

where for T we may use, with identical results, either T_+ or T_-. Then, for a two-tailed test, Z is compared to the critical value, $Z_{\alpha(2)}$, or, equivalently, $t_{\alpha(2),\infty}$ (which for $\alpha = 0.05$ is 1.9600); if Z is greater than or equal to $Z_{\alpha(2)}$, then H_0 is rejected.

A normal approximation with a correction for continuity employs

$$Z_c = \frac{|T - \mu_T| - 0.5}{\sigma_T}. \tag{9.11}$$

As shown at the end of Appendix Table B.12, the normal approximation is better using Z for $\alpha(2)$ from 0.001 to 0.05 and is better using Z_c for $\alpha(2)$ from 0.10 to 0.50.

If there are tied ranks, then use

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1) - \frac{\sum t}{2}}{24}}, \tag{9.12}$$

where

$$\sum t = \sum (t_i^3 - t_i) \tag{9.13}$$

is the correction for ties introduced in using the normal approximation to the Mann-Whitney test (Equation 8.57), applied here to ties of nonzero differences.

*Fahoome (2002) concluded that the normal approximation also works well for sample sizes smaller than 100. She found that the probability of a Type I error is between 0.045 and 0.055 for two-tailed testing at the 0.05 level of significance with n as small as 10 and is between 0.009 and 0.011 when testing at $\alpha(2) = 0.01$ with n as small as 22. Additional information regarding the accuracy of this approximation is given at the end of Appendix Table B.12.

If we employ the Pratt procedure for handling differences of zero (described above), then the normal approximation is

$$Z = \frac{\left|T - \frac{n(n+1) - m'(m'+1)}{4}\right| - 0.5}{\sqrt{\frac{n(n+1)(2n+1) - m'(m'+1)(2m+1) - \frac{\sum t}{2}}{24}}} \tag{9.14}$$

(Cureton, 1967), where n is the total number of differences (including zero differences), and m' is the number of zero differences; $\sum t$ is as in Equation 9.13, applied to ties other than those of zero differences. We calculate T_+ or T_- by including the zero differences in the ranking and then deleting from considerations both the zero d_j's and the ranks assigned to them. For T in Equation 9.14, either T_+ or T_- may be used. If neither tied ranks nor zero d_j's are present, then Equation 9.14 becomes Equation 9.11.

One-tailed testing may also be performed using the normal approximation (Equation 9.10 or 9.11) or Cureton's procedure (Equation 9.14). The calculated Z is compared to $Z_{\alpha(1)}$ (which is the same as $t_{\alpha(1),\infty}$), and the direction of the arrow in the alternate hypothesis must be examined. If the arrow points to the left ("<"), then H_0 is rejected if $Z \geq Z_{\alpha(1)}$ *and* $T_+ < T_-$; if it points to the right (">"), then reject H_0 if $Z \geq Z_{\alpha(1)}$ *and* $T_+ > T_-$.

Iman (1974a) presents an approximation based on Student's t:

$$t = \frac{T - \mu_T}{\sqrt{\frac{n^2(n+1)(2n+1)}{2(n-1)} - \frac{(T-\mu_T)^2}{n-1}}}, \tag{9.15}$$

with $n - 1$ degrees of freedom. As shown at the end of Appendix Table B.12, this performs slightly better than the normal approximation (Equation 9.10). The test with a correction for continuity is performed by subtracting 0.5 from $|T - \mu_T|$ in both the numerator and denominator of Equation 9.15. This improves the test for $\alpha(2)$ from 0.001 to 0.10, but the uncorrected t is better for $\alpha(2)$ from 0.20 to 0.50. One-tailed t-testing is effected in a fashion similar to that described for Z in the preceding paragraph.*

Fellingham and Stoker (1964) discuss a more accurate approximation, but it requires more computation, and for sample sizes beyond those in Table B.12 the increased accuracy is of no great consequence.

(c) The Wilcoxon Paired-Sample Test for Ordinal Data. The Wilcoxon test nonparametrically examines differences between paired samples when the samples consist of interval-scale or ratio-scale data (such as in Example 9.4), which is legitimate because the paired differences can be meaningfully ordered. However, it may not work well with samples comprising ordinal-scale data because the differences between ordinal scores may not have a meaningful ordinal relationship to each other. For example, each of several frogs could have the intensity of its green skin color recorded on a scale of 1 (very pale green) to 10 (very deep green). Those data would represent an ordinal scale of measurement because a score of 10 indicates a more intense green than a score of 9, a 9 represents an intensity greater than an 8, and so on. Then the skin-color

*When Appendix Table B.12 cannot be used, a slightly improved approximation is effected by comparing the mean of t and Z to the mean of the critical values of t and Z (Iman, 1974a).

intensity could be recorded for these frogs after they were administered hormones for a period of time, and those data would also be ordinal. However, the *differences* between skin-color intensities before and after the hormonal treatment would not necessarily be ordinal data because, for example, the difference between a score of 5 and a score of 2 (a difference of 3) cannot be said to represent a difference in skin color that is greater than the difference between a score of 10 and a score of 8 (a difference of 2).

To deal with such a situation, Kornbrot (1990) presented a modification of the Wilcoxon paired-sample test (which she called the "rank difference test"), along with tables to determine statistical significance of its results.

(d) Wilcoxon Paired-Sample Test Hypotheses about a Specified Difference Other Than Zero. As indicated in Section 9.1, the paired-sample t test can be used for hypotheses proposing that the mean difference is something other than zero. Similarly, the Wilcoxon paired-sample test can examine whether paired differences are centered around a quantity other than zero. Thus, for data such as in Example 9.2, the nonparametric hypotheses could be stated as H_0: crop yield does not increase more than 250 kg/ha with the new fertilizer, versus H_A: crop yield increases more than 250 kg/ha with the new fertilizer. In that case, each datum for the old-fertilizer treatment would be increased by 250 kg/ha (resulting in nine data of 1920, 2020, 2060 kg/ha, etc.) to be paired with the nine new-fertilizer data of 2250, 2410, 2260 kg/ha, and so on. Then, the Wilcoxon paired-sample test would be performed on those nine pairs of data.

With ratio- or interval-scale data, it is also possible to propose hypotheses considering a multiplication, rather than an addition, constant. This concept is introduced at the end of Section 8.11f.

9.6 CONFIDENCE LIMITS FOR THE POPULATION MEDIAN DIFFERENCE

In Section 9.2, confidence limits were obtained for the mean of a population of differences. Given a population of differences, one can also determine confidence limits for the population median. This is done exactly as indicated in Section 7.10; simply consider the observed differences between members of pairs (d_j) as a sample from a population of such differences.

EXERCISES

9.1. Concentrations of nitrogen oxides and of hydrocarbons (recorded in $\mu g/m^3$) were determined in a certain urban area.

(a) Test the hypothesis that both classes of air pollutants were present in the same concentration.

Day	*Nitrogen oxides*	*Hydrocarbons*
1	104	108
2	116	118
3	84	89
4	77	71
5	61	66
6	84	83
7	81	88
8	72	76
9	61	68
10	97	96
11	84	81

(b) Calculate the 95% confidence interval for μ_d.

9.2. Using the data of Exercise 9.1, test the appropriate hypotheses with Wilcoxon's paired-sample test.

9.3. Using the data of Exercise 9.1, test for equality of the variances of the two kinds of air pollutants.

CHAPTER 10

Multisample Hypotheses and the Analysis of Variance

When measurements of a variable are obtained for each of two independently collected samples, hypotheses such as those described in Chapter 8 are appropriate. However, biologists often obtain data in the form of three or more samples, which are from three or more populations, a situation calling for multisample analyses, as introduced in this chapter.

It is tempting to some to test multisample hypotheses by applying two-sample tests to all possible pairs of samples. In this manner, for example, one might proceed to test the null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3$ by testing each of the following hypotheses by the two-sample t test: H_0: $\mu_1 = \mu_2$, H_0: $\mu_1 = \mu_3$, H_0: $\mu_2 = \mu_3$. But such a procedure, employing a series of two-sample tests to address a multisample hypothesis, is invalid.

The calculated test statistic, t, and the critical values we find in the t table are designed to test whether the two sample statistics, $\overline{X}_1$ and $\overline{X}_2$, are likely to have come from the same population (or from two populations with identical means). In properly employing the two-sample test, we could randomly draw two sample means from the same population and wrongly conclude that they are estimates of two different populations' means; but we know that the probability of this error (the Type I error) will be no greater than α. However, consider that three random samples were taken from a single population. In performing the three possible two-sample t tests indicated above, with $\alpha = 0.05$, the probability of wrongly concluding that two of the means estimate different parameters is 14%, considerably greater than α. Similarly, if α is set at 5% and four means are tested, two at a time, by the two-sample t test, there are six pairwise H_0's to be tested in this fashion, and there is a 26% chance of wrongly concluding a difference between one or more of the means. Why is this?

For each two-sample t test performed at the 5% level of significance, there is a 95% probability that we shall correctly conclude not to reject H_0 when the two population means are equal. For the set of three hypotheses, the probability of *correctly* declining to reject all of them is only $0.95^3 = 0.86$. This means that the probability of *incorrectly* rejecting at least one of the H_0's is $1 - (1 - \alpha)^C = 1 - (0.95)^3 = 0.14$, where C

is the number of possible different pairwise combinations of k samples (see footnote to Table 10.1). As the number of means increases, it becomes almost certain that performing all possible two-sample t tests will conclude that some of the sample means estimate different values of μ, even if all of the samples came from the same population or from populations with identical means. Table 10.1 shows the probability of committing a Type I error if multiple t tests are employed to assess differences among more than two means. If, for example, there are 10 sample means, then $k = 10$, $C = 45$, and $1 - 0.95^{45} = 1 - 0.10 = 0.90$ is the probability of at least one Type I error when testing at the 0.05 level of significance. Two-sample tests, it must be emphasized, should *not* be applied to multisample hypotheses. The appropriate procedures are introduced in the following sections.

TABLE 10.1: Probability of Committing at Least One Type I Error by Using Two-Sample t Tests for All C Pairwise Comparisons of k Means*

		Level of Significance, α, Used in the t Tests				
k	C	0.10	0.05	0.01	0.005	0.001
2	1	0.10	0.05	0.01	0.005	0.001
3	3	0.27	0.14	0.03	0.015	0.003
4	6	0.47	0.26	0.06	0.030	0.006
5	10	0.65	0.40	0.10	0.049	0.010
6	15	0.79	0.54	0.14	0.072	0.015
10	45	0.99	0.90	0.36	0.202	0.044
	∞	1.00	1.00	1.00	1.000	1.000

*There are $C = k(k - 1)/2$ pairwise comparisons of k means. This is the number of combinations of k items taken two at a time; see Equation 5.10.

10.1 SINGLE-FACTOR ANALYSIS OF VARIANCE

To test the null hypothesis H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$, where k is the number of experimental groups, or samples, we need to become familiar with the topic of *analysis of variance*, often abbreviated ANOVA (or less commonly, ANOV or AOV). Analysis of variance is a large area of statistical methods, owing its name and much of its early development to R. A. Fisher;* in fact, the F statistic was named in his honor by G. W. Snedecor† (1934: 15). There are many ramifications of analysis of variance considerations, the most common of which will be discussed in this and subsequent chapters. More complex applications and greater theoretical coverage are found in the many books devoted specifically to analysis of variance and experimental design. At this point, it may appear strange that a procedure used for testing the equality of *means* should be named analysis of *variance*, but the reason for this terminology soon will become apparent.

*Sir Ronald Aylmer Fisher (1890–1962), British statistician and geneticist, who introduced the name and basic concept of the technique in 1918 (David, 1995; Street, 1990) and stressed the importance of randomness as discussed in this section. When he introduced analysis of variance, he did so by way of intraclass correlation (Box, 1978: 101), to which it is related (Section 19.12).

†George W. Snedecor (1881–1974), American statistician.

Let us assume that we wish to test whether four different feeds result in different body weights in pigs. Since we are to test for the effect of only one *factor* (feed type) on the variable in question (body weight), the appropriate analysis is termed a single-factor (or "single-criterion" or "single-classification" or "one-way") analysis of variance.* Furthermore, each type of feed is said to be a *level* of the factor. The design of this experiment should have each experimental animal being assigned at random to receive one of the four feeds, with approximately equal numbers of pigs receiving each feed.

As with other statistical testing, it is of fundamental importance that each sample is composed of a random set of data from a population of interest. In Example 10.1, each of four populations consists of body weights of pigs on one of four experimental diets, and 19 pigs were assigned, at random, to the four diets. In other instances, researchers do not actually perform an experiment but, instead, collect data from populations defined other than by the investigator. For example, the interest might be in comparing body weights of four strains of pigs. If that were the case, the strain to which each animal belonged would not have been under the control of the researcher. Instead, he or she would measure a sample of weights for each strain, and the important consideration would be having each of the four samples consist of data assumed to have come at random from one of the four populations of data being studied.

EXAMPLE 10.1 A Single-Factor Analysis of Variance (Model I)

Nineteen pigs are assigned at random among four experimental groups. Each group is fed a different diet. The data are pig body weights, in kilograms, after being raised on these diets. We wish to ask whether pig weights are the same for all four diets.

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$.

H_A: The mean weights of pigs on the four diets are not all equal.

$\alpha = 0.05$

	Feed 1	*Feed 2*	*Feed 3*	*Feed 4*
	60.8	68.7	69.6	61.9
	67.0	67.7	77.1	64.2
	65.0	75.0	75.2	63.1
	68.6	73.3	71.5	66.7
	61.7	71.8		60.3
i	1	2	3	4
n_i	5	5	4	5
$\sum_{j=1}^{n_i} X_{ij}$	323.1	356.5	293.4	316.2
$\overline{X_i}$	64.62	71.30	73.35	63.24

Because the pigs are assigned to the feed groups at random (as with the aid of a random-number table, such as Appendix Table B.41, described in Section 2.3),

*Some authors would here refer to the feed as the "independent variable" and to the weight as the "dependent variable."

the single factor ANOVA is said to represent a *completely randomized experimental design*, or "completely randomized design" (sometimes abbreviated "CRD"). In general, statistical comparison of groups of data works best if each group has the same number of data (a situation referred to as being a *balanced*, or *orthogonal, experimental design*), and the power of the test is heightened by having sample sizes as nearly equal as possible. The present hypothetical data might represent a situation where there were, in fact, five experimental animals in each of four groups, but the body weight of one of the animals (in group 3) was not used in the analysis for some appropriate reason. (Perhaps the animal died, or perhaps it became ill or was discovered to be pregnant, thus introducing a factor other than feed into the experiment.) The performance of the test is also enhanced by having all pigs as similar as possible in all respects except for the experimental factor, diet (i.e., the animals should be of the same breed, sex, and age, should be kept at the same temperature, etc.).

Example 10.1 shows the weights of 19 pigs subjected to this feed experiment, and the null hypothesis to be tested would be H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$. Each datum in the experiment may be uniquely represented by the double subscript notation, where X_{ij} denotes datum j in experimental group i. For example, X_{23} denotes the third pig weight in feed group 2, that is, $X_{23} = 74.0$ kg. Similarly, $X_{34} = 96.5$ kg, $X_{41} = 87.9$ kg, and so on. We shall let the mean of group i be denoted by $\overline{X}_i$, and the grand mean of all observations will be designated by $\overline{X}$. Furthermore, n_i will represent the size of sample i, and $N = \sum_{i=1}^{k} n_i$ will be the total number of data in the experiment. The alternate hypothesis for this experiment is H_A: The mean weight of pigs is not the same on these four diets. Note that H_A is *not* $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$ nor $\mu_1 \neq \mu_2 = \mu_3 = \mu_4$ nor any other specification of which means are different from which; we can only say that, if H_0 is rejected, then there is at least one difference among the four means.*

The four groups in this example (namely, types of feed) represent a nominal-scale variable (see Section 1.1d) in that the groups could be arranged in any sequence. However, in some situations the groups represent a measurement made on a ratio or interval scale (Sections 1.1a and 1.1b). For example, the animal weights of Example 10.1 might have been measured at each of four environmental temperatures or, instead of the groups being different types of feed, the groups might have been different daily amounts of the same kind of feed. In other situations the groups might be expressed on an ordinal scale (Section 1.1c); for example, body weights could be measured at four environmental temperatures defined as cold, medium, warm, and hot, or as quantities of feed defined as very low, low, medium, and high. The analysis of variance of this section is appropriate when the groups are defined on a nominal or ordinal scale. If they represent a ratio or interval scale, the regression procedures of Section 17.7 may be more appropriate, but the latter methods require more levels of the factor than are generally present in an analysis-of-variance experimental design.

(a) Sources of Variation. The statistical technique widely known as *analysis of variance* (ANOVA) examines the several sources of variation among all of the data

*There may be situations where the desired hypothesis is not whether k means are equal to each other, but whether they are all equal to some particular value. Mee, Shah, and Lefante (1987) proposed a procedure for testing H_0: $\mu_1 = \mu_2 = \cdots = \mu_k = \mu_0$, where μ_0 is the specified mean to which all of the other means are to be compared. The alternate hypothesis would be that at least one of the means is different from μ_0 (i.e., H_A: $\mu_i \neq \mu_0$ for at least one i).

in an experiment, by determining a *sum of squares* (meaning "sum of squares of deviations from the mean," a concept introduced in Section 4.4) for each source. Those sums of squares are as shown below.

In an experimental design with k groups, there are n_i data in group i; that is, n_1 designates the number of data in group 1, n_2 the number in group 2, and so on. The total number of data in all k groups will be called N; that is,

$$N = \sum_{i=1}^{k} n_i, \tag{10.1}$$

which in Example 10.1 is $N = 5 + 5 + 4 + 5 = 19$. The sum of squares for all N data is

$$\text{total SS} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X})^2, \tag{10.2}$$

where X_{ij} is datum j in group i and $\overline{X}$ is the mean of all N data:

$$\overline{X} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}}{N}. \tag{10.2a}$$

This is the same as considering the N data in all the groups to compose a single group for which the sum of squares is as shown in Equation 4.12. For the data in Example 10.1, these calculations are demonstrated in Example 10.1a.

EXAMPLE 10.1a Sums of Squares and Degrees of Freedom for the Data of Example 10.1.

	Feed 1	*Feed 2*	*Feed 3*	*Feed 4*
$\sum_{j=1}^{n_i} X_{ij}$	323.1	356.5	293.4	316.2
$\overline{X}_i$	64.62	71.30	73.35	63.24

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij} = 60.8 + 67.0 + 65.0 + \cdots + 63.1 + 66.7 + 60.3 = 1289.2$$

$$\overline{X} = \frac{1289.2}{19} = 67.8526$$

$$\begin{aligned}\text{Total SS} &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X})^2 \\ &= (60.8 - 67.8526)^2 + (67.0 - 67.8526)^2 \\ &\quad + \cdots + (66.7 - 67.8526)^2 + (60.3 - 6.8526)^2 \\ &= 49.7372 + 0.7269 + \cdots + 1.3285 + 57.0418 = 479.6874.\end{aligned}$$

$$\text{total DF} = N - 1 = 19 - 1 = 18$$

$$\text{groups SS} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$$

$$= 5(64.62 - 67.8526)^2 + 5(71.30 - 67.8526)^2$$
$$+ 4(73.35 - 67.8526)^2 + 5(63.24 - 67.8526)^2$$
$$= 52.2485 + 59.4228 + 120.8856 + 106.3804 = 338.9372$$

$$\text{groups DF} = k - 1$$

$$\text{within-groups (error) SS} = \sum_{i=1}^{k} \left[\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 \right]$$
$$= (60.8 - 64.62)^2 + (67.0 - 64.62)^2$$
$$+ \cdots + (66.7 - 63.24)^2 + (60.3 - 63.24)^2$$
$$= 14.5924 + 5.6644 + \cdots + 11.9716 + 8.6436$$
$$= 140.7500$$

or, alternatively,

$$\text{within-groups (error) SS} = \text{Total SS} - \text{Groups SS}$$
$$= 479.6874 - 338.9373 = 140.7501.$$

$$\text{within-groups (error) DF} = \sum_{i=1}^{k} (n_i - 1)$$
$$= (5 - 1) + (5 - 1) + (4 - 1) + (5 - 1) = 15$$

or within-groups (error) DF $= N - k = 19 - 4 = 15$

or within-groups (error) DF = Total DF − Groups DF $= 18 - 3 = 15$.

Note: The quantities involved in the sum-of-squares calculations are carried to several decimal places (as computers typically do) to avoid rounding errors. All of these sums of squares (and the subsequent mean squares) have $(\text{kg})^2$ as units. However, for typographic convenience and ease in reading, the units for ANOVA computations are ordinarily not printed.

The degrees of freedom associated with the total sum of squares are

$$\text{total DF} = N - 1, \tag{10.3}$$

which for the data in Example 10.1 are $19 - 1 = 18$.

A portion of this total amount of variability of the N data is attributable to differences among the means of the k groups; this is referred to as the *among-groups sum of squares* or, simply, as the *groups sum of squares*:

$$\text{groups SS} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2, \tag{10.4}$$

where $\overline{X}_i$ is the mean of the n_i data in sample i and $\overline{X}$ is the mean of all N data. Example 10.1a shows this computation for the data in Example 10.1. Associated with this sum of squares are these degrees of freedom:

$$\text{groups DF} = k - 1, \tag{10.5}$$

which for Example 10.1 are $4 - 1 = 3$.

Furthermore, the portion of the total sum of squares that is not explainable by differences *among* the group means is the variability *within* the groups:

$$\text{within-groups SS} = \sum_{i=1}^{k} \left[\sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)^2 \right] \tag{10.6}$$

and is commonly called the *error sum of squares*. Within the brackets Equation 4.12 is applied to the data in each one of the k groups, and the within-groups sum of squares is the sum of all k of these applications of Equation 4.12. For the data of Example 10.1, this is shown in Example 10.1a.

The degrees of freedom associated with the within-groups sum of squares are

$$\text{within-groups DF} = \sum_{i=1}^{k} (n_i - 1) = N - k, \tag{10.7}$$

also called the *error DF*, which for Example 10.1 are $4 + 4 + 3 + 4 = 15$ or, equivalently, $19 - 4 = 15$.

The within-groups SS and DF may also be obtained by realizing that they represent the difference between the total variability among the data and the variability among groups:

$$\text{within-groups SS} = \text{total SS} - \text{groups SS} \tag{10.8}$$

and

$$\text{within-groups DF} = \text{total DF} - \text{groups DF}. \tag{10.8a}$$

In summary, each deviation of an observed datum from the grand mean of all data is attributable to a deviation of that datum from its group mean plus the deviation of that group mean from the grand mean; that is,

$$(X_{ij} - \overline{X}) = (X_{ij} - \overline{X}_i) + (\overline{X}_i - \overline{X}). \tag{10.9}$$

Furthermore, sums of squares and degrees of freedom are additive, so

$$\text{total SS} = \text{groups SS} + \text{error SS} \tag{10.10}$$

and

$$\text{total DF} = \text{groups DF} + \text{error DF}. \tag{10.11}$$

(b) "Machine Formulas." The total sum of squares (Equation 10.2) may be calculated readily by a "machine formula" analogous to Equation 4.16:

$$\text{total SS} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}^2 - C, \tag{10.12}$$

where*

$$C = \frac{\left(\sum\sum X_{ij}\right)^2}{N}. \tag{10.13}$$

Example 10.1b demonstrates these calculations.

EXAMPLE 10.1b Sums of Squares and Degrees of Freedom for the Data of Example 10.1, Using Machine Formulas

	Feed 1	*Feed 2*	*Feed 3*	*Feed 4*	
i	1	2	3	4	
n_i	5	5	4	5	
$\sum_{j=1}^{n_i} X_{ij}$	323.1	356.5	293.4	316.2	
$\frac{\left(\sum_{j=1}^{n_i} X_{ij}\right)^2}{n_i}$	20878.7220	25418.4500	21520.8900	19996.4480	$\sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n_i} X_{ij}\right)}{n_i} = 87814.5500$

$\sum_i \sum_j X_{ij} = 1289.2$ total DF $= N - 1 = 19 - 1 = 18$

$\sum_i \sum_j X_{ij}^2 = 87955.30$ groups DF $= k - 1 = 4 - 1 = 3$

error DF $= N - k = 19 - 4 = 15$

$$C = \frac{\left(\sum_i \sum_j X_{ij}\right)^2}{N} = \frac{(1289.2)^2}{19} = 87475.6126$$

$$\text{total SS} = \sum_i \sum_j X_{ij}^2 - C = 87955.3000 - 87475.6126 = 479.6874$$

$$\text{groups SS} = \sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n_i}\right)^2}{n_i} - C = 87814.5500 - 87475.6126 = 338.9374$$

$$\text{error SS} = \text{total SS} - \text{groups SS} = 479.68747 - 338.9374 = 140.7500$$

A machine formula for the groups sum of squares (Equation 10.4) is

$$\text{groups SS} = \sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n_i} X_{ij}\right)^2}{n_i} - C, \tag{10.14}$$

where $\sum_{j=1}^{n_i} X_{ij}$ is the sum of the n_i data from group i.

*The term "machine formula" derives from the formula's utility when using calculating machines. The quantity C is often referred to as a "correction term"—an unfortunate expression, for it implies that some miscalculation needs to be rectified.

The error SS may be calculated as

$$\text{error SS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n_i} X_{ij}\right)^2}{n_i}, \tag{10.15}$$

which is the machine formula for Equation 10.6.

As shown in Example 10.1b, machine formulas such as these can be very convenient when using simple calculators; but they are of less importance if the statistical computations are performed by computer. As demonstrated in Examples 10.1a and 10.1b, the sums of squares are the same using the two computational formulas.

(c) Testing the Null Hypothesis. Dividing the groups SS or the error SS by the respective degrees of freedom results in a variance, referred to in ANOVA terminology as a *mean square* (abbreviated MS and short for *mean squared deviation from the mean*). Thus,

$$\text{groups MS} = \frac{\text{groups SS}}{\text{groups DF}} \tag{10.16}$$

and

$$\text{error MS} = \frac{\text{error SS}}{\text{error DF}}, \tag{10.17}$$

and the latter quantity, which may also be called the *within-groups mean square*, is occasionally abbreviated as MSE (for "mean square error"). As will be seen below, testing the null hypothesis of equality among population means involves the examination of the groups mean square and the error mean square. Because a mean square is a kind of variance, this procedure is named *analysis of variance*. A total mean square could also be calculated, as (total SS)/(total DF), but it is not used in the ANOVA.

Statistical theory informs us that if the null hypothesis is a true statement about the populations, then the groups MS and the error MS will each be an estimate of σ^2, the variance common to all k populations. But if the k population means are not equal, then the groups MS in the population will be greater than the population's error MS.* Therefore, the test for the equality of means is a one-tailed variance ratio test (introduced in Section 8.5), where the groups MS is always placed in the numerator so as to ask whether it is significantly larger than the error MS:†

$$F = \frac{\text{groups MS}}{\text{error MS}}. \tag{10.18}$$

*Two decades before R. A. Fisher developed analysis of variance techniques, the Danish applied mathematician, Thorvald Nicolai Thiele (1838–1910) presented the concept of comparing the variance among groups to the variance within groups (Thiele, 1897: 41–44). Stigler (1986: 244) reported that an 1860 book by Gustav Theodor Fechner included the most extensive discussion of the concepts of experimental design prior to R. A. Fisher.

†An equivalent computation of F is

$$F = \left(\frac{\text{error DF}}{\text{groups DF}}\right)\left(\frac{\text{(groups SS)/(total SS)}}{1 - \text{(groups SS)/(total SS)}}\right), \tag{10.18a}$$

and Levin, Serlin, and Webne-Behrman (1989) show how ANOVA can be performed by considering the correlation (the topic of Section 19.1) between observations and their group means.

This quantity expresses how the variability of data among groups compares to the variability of data within groups.

The critical value for this test is $F_{\alpha(1),(k-1),(N-k)}$, which is the value of F at the one-tailed significance level α and with numerator degrees of freedom (ν_1 = groups DF) of $k - 1$ and denominator degrees of freedom (ν_2 = error DF) of $N - k$. If the calculated F is as large as, or larger than, the critical value (see Appendix Table B.4), then we reject H_0; and a rejection indicates that the probability is $\leq \alpha$ that the observed data came from populations described by H_0. But remember that all we conclude in such a case is that all the k population means are not equal. To conclude between which means the equalities or inequalities lie, we must turn to the procedures of Chapter 11.

Example 10.1c shows the conclusion of the analysis of variance performed on the data and hypotheses of Example 10.1. Table 10.2 summarizes the single-factor ANOVA calculations.*

EXAMPLE 10.1c The Conclusion of the ANOVA of Example 10.1, Using the Results of Either Example 10.1a or 10.1b

Summary of the Analysis of Variance

Source of variation	*SS*	*DF*	*MS*
Total	479.6874	18	
Groups	338.9374	3	112.9791
Error	140.7500	15	9.3833

$$F = \frac{\text{groups MS}}{\text{error MS}} = \frac{112.9791}{9.3833} = 12.04$$

$F_{0.05(1),3,15} = 3.29$, so reject H_0.

$P < 0.0005$ [$P = 0.00029$]

(d) The Case where $k = 2$. If $k = 2$, then H_0: $\mu_1 = \mu_2$, and either the two-sample t test (Section 8.1) or the single-factor ANOVA may be applied; the conclusions obtained from these two procedures will be identical. The error MS will, in fact, be identical to the pooled variance, s_p^2, in the t test; the groups DF will be $k - 1 = 1$; the F value determined by the analysis of variance will be the square of the t value from the t test; and $F_{\alpha(1),1,(N-2)} = (t_{\alpha(2),(N-2)})^2$. If a one-tailed test between means is required, or if the hypothesis H_0: $\mu_1 - \mu_2 = \mu_0$ is desired for a μ_0 not equal to zero, then the t test is applicable, whereas the ANOVA is not.

*Occasionally the following quantity (or its square root) is called the *correlation ratio*:

$$\eta^2 = \frac{\text{groups SS}}{\text{total SS}}. \tag{10.18b}$$

This is also called *eta squared*, for it is represented using the lowercase Greek letter eta. It is always between 0 and 1, it has no units of measurement, and it expresses the proportion of the total variability of X that is accounted for by the effect of differences among the groups (and is, therefore, reminiscent of the coefficient of determination introduced in Section 17.3a). For Example 10.1, $\eta^2 = 338.9374/479.6874 = 0.71$, or 71%.

TABLE 10.2: Summary of the Calculations for a Single-Factor Analysis of Variance

Source of variation	Sum of squares (SS)	Degrees of freedom (DF)	Mean square (MS)
Total $[X_{ij} - \overline{X}]$	Equation 10.2 or 10.12	$N - 1$	
Groups (i.e., among groups) $[\overline{X}_i - \overline{X}]$	Equation 10.4 or 10.14	$k - 1$	$\frac{\text{groups SS}}{\text{groups DF}}$
Error (i.e., within groups) $[X_{ij} - \overline{X}_i]$	Equation 10.6 or 10.8	$N - k$ or Equation 10.8a	$\frac{\text{error SS}}{\text{error DF}}$

Note: For each source of variation, the bracketed quantity indicates the variation being assessed: k is the number of groups; X_{ij} is datum j in group i; $\overline{X}_i$ is the mean of the data in group i; $\overline{X}$ is the mean of all N data.

(e) ANOVA Using Means and Variances. The above discussion assumes that all the data from the experiment to be analyzed are in hand. It may occur, however, that all we have for each of the k groups is the mean and some measure of variability based on the variances of each group. That is, we may have $\overline{X}_i$ and either SS_i, s_i^2, s_i, or $s_{\overline{X}_i}$ for each group, rather than all the individual values of X_{ij}. For example, we might encounter presentations such as Tables 7.1, 7.2, or 7.3a. If the sample sizes, n_i, are also known, then the single-factor analysis of variance may still be performed, in the following manner.

First, determine the sum of squares or sample variance for each group: recall that

$$SS_i = (n_i - 1)s_i^2 \text{ and } s_i^2 = (s_i)^2 = n_i(s_{\overline{X}_i})^2. \tag{10.19}$$

Then calculate

$$\text{error SS} = \sum_{i=1}^{k} SS_i = \sum_{i=1}^{k} (n_i - 1)s_i^2 \tag{10.20}$$

and

$$\text{groups SS} = \sum_{i=1}^{k} n_i \overline{X}_i^2 - \frac{\left(\sum_{i=1}^{k} n_i \overline{X}_i\right)^2}{\sum_{i=1}^{k} n_i}. \tag{10.21}$$

Knowing the groups SS and error SS, the ANOVA can proceed in the usual fashion.

(f) Fixed-Effects amd Random-Effects ANOVA. In Example 10.1, the biologist designing the experiment was interested in whether all of these particular four feeds have the same effect on pig weight. That is, these four feeds were not randomly selected from a feed catalog but were specifically chosen. When the levels of a factor are specifically chosen one is said to have designed a *fixed-effects model*, or a *Model I*, ANOVA. In such a case, the null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$ is appropriate.

However, there are instances where the levels of a factor to be tested are indeed chosen at random. For example, we might have been interested in the effect of geographic location of the pigs, rather than the effect of their feed. It is possible that our concern might be with certain specific locations, in which case we would

be employing a fixed-effects mode ANOVA. But we might, instead, be interested in testing the statement that in general there is a difference in pig weights in animals from different locations. That is, instead of being concerned with only the particular locations used in the study, the intent might be to generalize, considering the locations in our study to be a random sample from all possible locations. In this *random-effects model*, or *Model II*, ANOVA,* all the calculations are identical to those for the fixed-effects model, but the null hypothesis is better stated as H_0: there is no difference in pig weight among geographic locations (or H_0: there is no variability in weights among locations). Examination of Equation 10.18 shows that what the analysis asks is whether the variability among locations is greater than the variability within locations. Example 10.2 demonstrates the ANOVA for a random-effects model. The relevant sums of squares could be computed as in Section 10.1a or 10.1b; the machine formulas of Section 10.1b are used in this example. Most biologists will encounter Model I analyses more commonly than Model II situations. When dealing with more than one experimental factor (as in Chapters 12 and 14), the distinction between the two models becomes essential, as it will determine the calculation of F.

(g) Violation of Underlying Assumptions. Recall from Section 8.1b that to test H_0: $\mu_1 = \mu_2$ by the two-sample t test, we assume that $\sigma_1^2 = \sigma_2^2$ and that each of the two samples came at random from a normal population. Similarly, in order to apply the analysis of variance to $\mu_1 = \mu_2 = \cdots = \mu_k$, we assume that $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$ and that each of the k samples came at random from a normal population.

However, these conditions are never *exactly* met, so the question becomes how serious the consequences are when there are departures from these underlying assumptions. Fortunately, under many circumstances the analysis of variance is a robust test, meaning that its Type I and Type II error probabilities are not always seriously altered by violation of the test's assumptions. Reports over several decades of research have not agreed on every aspect of this issue, but the following general statements can be made about fixed-effects (i.e., Model I) ANOVA:

As with the two-sample t test (Section 8.1b), the adverse effect of nonnormality is greater with greater departures from normality, but the effect is relatively small if samples sizes are equal, or if the n_i's are unequal but large (with the test less affected by nonnormality as the n_i's increase), or if the variances are equal; and asymmetric distributions have a greater adverse effect than do symmetric distributions (Box and Anderson, 1955; Büning, 1997; Donaldson, 1968; Glass, Peckham, and Sanders, 1972; Harwell et al., 1992; Lix, Keselman, and Keselman, 1996; Srivastava, 1959; Tiku, 1971).

If the variances of the k populations are not equal, the analysis of variance is generally liberal for equal sample sizes, and the extent to which the test is liberal (i.e., the probability of a Type I error exceeds α) increases with greater variance heterogeneity (Büning, 1997; Clinch and Keselman, 1982; Rogan and Keselman, 1977) and decreases with increased sample size (Rogan and Keselman, 1977). Myers and Well (2003: 221) report that this inflation of P(Type I error) is usually less than 0.02 at the 0.05 significance level and less than 0.005 when using $\alpha = 0.01$, when n is at least 5 and the largest variance is no more than four times the smallest variance.

If the group variances are not equal and the n_i's also are unequal, then there can be very serious effects on P(Type I error). The effect will be greater for greater variance heterogeneity (Box, 1954), and if the larger variances are associated with the larger sample sizes (what we shall call a "direct" relationship), the test is conservative

*Also referred to as a components of variance model. The terms *components of variance*, *fixed effects*, *random effects*, *Class I*, and *Class II* for analysis of variance were introduced by Eisenhart (1947).

EXAMPLE 10.2 A Single-Factor Analysis of Variance for a Random-Effects Model (i.e., Model II) Experimental Design

A laboratory employs a technique for determining the phosphorus content of hay. The question arises: "Do phosphorus determinations differ among the technicians performing the analysis?" To answer this question, each of four randomly selected technicians was given five samples from the same batch of hay. The results of the 20 phosphorus determinations (in mg phosphorus/g of hay) are shown.

H_0: Determinations of phosphorus content do not differ among technicians.

H_A: Determinations of phosphorus content do differ among technicians.

$\alpha = 0.05$

	Technician			
	1	2	3	4
	34	37	34	36
	36	36	37	34
	34	35	35	37
	35	37	37	34
	34	37	36	35
Group sums:	173	182	179	176

$$\sum_i \sum_j X_{ij} = 710$$

$$\sum_i \sum_j X_{ij}^2 = 25234$$

$N = 20$

$$C = \frac{(710)^2}{20} = 25205.00$$

total SS $= 25234 - 25205.00 = 29.00$

$$\text{groups (i.e., technicians) SS} = \frac{(173)^2}{5} + \frac{a(182)^2}{5} + \frac{(179)^2}{5} + \frac{(176)^2}{5} - 25205.00$$

$$= 25214.00 - 25205.00 = 9.00$$

error SS $= 29.00 - 9.00 = 20.00$

Source of variation	SS	DF	MS
Total	29.00	19	
Groups (technicians)	9.00	3	3.00
Error	20.00	16	1.25

$$F = \frac{3.00}{1.25} = 2.40$$

$F_{0.05(1),3,16} = 3.24$

Do not reject H_0.

$0.10 < P < 0.25 \quad [P = 0.11]$

(i.e., the probability of a Type I error is less than α), while larger variances affiliated with smaller samples (what we'll call an "inverse" relationship) cause the test to be liberal—that is, P(Type I error) $> \alpha$ (Brown and Forsythe, 1974a; Büning, 1997; Clinch and Keselman, 1982; Donaldson, 1968; Glass, Peckham, and Sanders, 1972; Harwell et al., 1992; Kohr and Games, 1974; Maxwell and Delaney, 2004: 131; Stonehouse and Forrester, 1998; Tomarkin and Serlin, 1986). For example, if testing at $\alpha = 0.05$ and the largest n is twice the size of the smallest, the probability of a Type I error can be as small as 0.006 for a direct relationship between variances and sample sizes and as large as 0.17 for an inverse relationship; if the ratio between the largest and smallest variances is 5, P(Type I error) can be as small as 0.00001 or as large as 0.38, depending upon whether the relationship is direct or inverse, respectively (Scheffé, 1959: 340). This huge distortion of P(Type I error) may be reason to avoid employing the analysis of variance when there is an inverse relationship (if the researcher is primarily concerned about avoiding a Type I error) or when there is a direct relationship (if the principal concern is to evade a Type II error). The adverse effect of heterogeneous variances appears to increase as k increases (Tomarkin and Serlin, 1986).

Recall (Section 6.3b) that a decrease in the probability of the Type I error (α) is associated with an increase in the Type II error (β), and an increase in β means a decrease in the power of the test ($1 - \beta$). Therefore, for situations described above as conservative [i.e., P(Type I error) $< \alpha$], there will generally be less power than if the population variances were all equal; and when the test is liberal [i.e., P(Type I error) $> \alpha$], there will generally be more power than if the variances were equal.

If the sample sizes are all equal, nonnormality generally affects the power of the analysis of variance to only a small extent (Clinch and Keselman, 1982; Glass, Peckham, and Sanders, 1972; Harwell et al., 1992; Tan, 1982), and the effect decreases with increased n (Donaldson, 1968). However, extreme skewness or kurtosis can severely alter (and reduce) the power (Games and Lucas, 1966), and nonnormal kurtosis generally has a more adverse effect than skewness (Sahai and Ageel, 2000: 85). With small samples, for example, very pronounced platykurtosis in the sampled populations will decrease the test's power, and strong leptokurtosis will increase it (Glass, Peckham, and Sanders, 1972). When sample sizes are not equal, the power is much reduced, especially when the large samples have small means (Bochnke, 1984).

The robustness of random-effects (i.e., Model II) analysis of variance (Section 10.1f) has not been studied as much as that of the fixed-effects (Model I) ANOVA. However, the test appears to be robust to departures of normality within the k populations, though not as robust as Model I ANOVA (Sahai, 2000: 86), provided the k groups (levels) of data can be considered to have been selected at random from all possible groups and that the effect of each group on the variable can be considered to be from a normally distributed set of group effects. When the procedure is nonrobust, power appears to be affected more than the probability of a Type I error; and the lack of robustness is not very different if sample sizes are equal or unequal (Tan, 1982; Tan and Wong, 1980). The Model II analysis of variance (as is the case with the Model I ANOVA) also assumes that the k sampled populations have equal variances.

(h) Testing of Multiple Means when Variances Are Unequal. Although testing hypotheses about means via analysis of variance is tolerant to small departures from the assumption of variance homogeneity when the sample sizes are equal, it can yield

very misleading results in the presence of more serious heterogeneity of variances and/or unequal sample sizes. Having unequal variances represents a multisample Behrens-Fisher problem (i.e., an extension of the two-sample Behrens-Fisher situation discussed in Section 8.1c).

Several approaches to this analysis have been proposed (e.g., see Keselman et al., 2000; Lix, Keselman, and Keselman, 1996). A very good one is that described by Welch (1951), which employs

$$F' = \frac{\sum_{i=1}^{k} c_i(\overline{X}_i - \overline{X}_w)^2}{(k-1)\left[1 + \frac{2A(k-2)}{k^2-1}\right]}, \tag{10.22}$$

where

$$c_i = \frac{n_i}{s_i^2} \tag{10.23}$$

$$C = \sum_{i=1}^{k} c_i \tag{10.24}$$

$$\overline{X}_w = \frac{\sum_{i=1}^{k} c_i \overline{X}_i}{C} \tag{10.25}$$

$$A = \sum_{i=1}^{k} \frac{(1 - c_i/C)^2}{\nu_i}, \text{ where } \nu_i = n_i - 1 \tag{10.26}$$

and F' is associated with degrees of freedom of $\nu_1 = k - 1$ and

$$\nu_2 = \frac{k^2 - 1}{3A}, \tag{10.27}$$

which should be rounded to the next lower integer when using Appendix Table B.4. This procedure is demonstrated in Example 10.3.

A modified ANOVA advanced by Brown and Forsythe (1974a, b) also works well:

$$F'' = \frac{\text{groups SS}}{B}, \tag{10.28}$$

EXAMPLE 10.3 Welch's Test for an Analysis-of-Variance Experimental Design with Dissimilar Group Variances

The potassium content (mg of potassium per 100 mg of plant tissue) was measured in five seedlings of each of three varieties of wheat.

H_0: $\mu_1 = \mu_2 = \mu_3$.

H_A: The mean potassium content is not the same for seedlings of all three wheat varieties.

$\alpha = 0.05$

	Variety G	*Variety A*	*Variety L*	
	27.9	24.2	29.1	
	27.0	24.7	27.7	
	26.0	25.6	29.9	
	26.5	26.0	30.7	
	27.0	27.4	28.8	
	27.5	26.1	31.1	
i	1	2	3	
n_i	6	6	6	
ν_i	5	5	5	
$\overline{X}_i$	26.98	25.67	29.55	
s_i^2	0.4617	1.2787	1.6070	
$c_i = n_i/s_i^2$	12.9955	4.6923	3.7337	$C = \sum_i c_i = 21.4215$
$c_i\overline{X}_i$	350.6186	120.4513	110.3308	$\sum_i c_i\overline{X}_i = 581.4007$
$\frac{\left(1 - \frac{c_i}{C}\right)^2}{\nu_i}$	0.0309	0.1220	0.1364	$A = \sum_i \frac{\left(1 - \frac{c_i}{C}\right)^2}{\nu_i} = 0.2893$

$$\overline{X}_w = \frac{\sum_i c_i\overline{X}_i}{C} = \frac{581.4007}{21.4215} = 27.14$$

$$F' = \frac{\sum c_i(\overline{X}_i - \overline{X}_w)^2}{(k-1)\left[1 + \frac{2A(k-2)}{k^2-1}\right]}$$

$$= \frac{12.9955(26.98 - 27.14)^2 + 4.6923(25.67 - 27.14)^2 + 3.7337(29.55 - 27.14)}{(3-1)\left[1 + \frac{2(0.2893)(3-2)}{3^2-1}\right]}$$

$$= \frac{0.3327 + 10.1396 + 21.6857}{2(0.9268)} = \frac{32.4144}{1.8536} = 17.5$$

For critical value of F:

$$\nu_1 = k - 1 = 3 - 1 = 2$$

$$\nu_2 = \frac{k^2 - 1}{3A} = \frac{3^2 - 1}{3(0.2893)} = \frac{8}{0.8679} = 9.22$$

By harmonic interpolation in Appendix Table B.4 or by computer program:

$F_{0.05(1),2,9.22} = 4.22$. So, reject H_0.

$$0.0005 < P < 0.001 \quad [P = 0.0073]$$

where

$$b_i = \left(1 - \frac{n_i}{N}\right)s_i^2, \tag{10.28a}$$

and

$$B = \sum_{i=1}^{k} b_i. \tag{10.29}$$

F'' has degrees of freedom of $\nu_1 = k - 1$ and

$$\nu_2 = \frac{B^2}{\sum_{i=1}^{k} \frac{b_i^2}{\nu_i}}. \tag{10.30}$$

If $k = 2$, both the F' and F'' procedures are equivalent to the t' test. The Welch method (F') has been shown (by Brown and Forsythe, 1974a; Büning, 1997; Dijkstra and Werter, 1981; Harwell et al., 1992; Kohr and Games, 1974; Levy, 1978a; Lix, Keselman, and Keselman, 1996) to generally perform better than F or F'' when population variances are unequal, especially when n_i's are equal. However, the Welch test is liberal if the data come from highly skewed distributions (Clinch and Keselman, 1992; Lix, Keselman, and Keselman, 1996).

Browne and Forsythe (1974a) reported that when variances are equal, the power of F is a little greater than the power of F', and that of F' is a little less than that of F''. But if variances are not equal, F' has greater power than F'' in cases where extremely low and high means are associated with low variances, and the power of F'' is greater than that of F' when extreme means are associated with large variances. Also, in general, F' and F'' are good if all $n_i \geq 10$ and F' is reasonably good if all $n_i \geq 5$.

(i) Which Multisample Test to Use. As with all research reports, the reader should be informed of explicitly what procedures were used for any statistical analysis. And, when results involve the examination of means, that reporting should include the size (n), mean ($\overline{X}$), and variability (e.g., standard deviation or standard error) of each sample. If the samples came from populations having close to normal distributions, then presentation of each sample's confidence limits (Section 10.2) might also be included. Additional interpretation of the results could include displaying the means and measures of variability via tables or graphs such as those described in Section 7.4.

Although it not possible to generalize to all possible situations that might be encountered, the major approaches to comparing the means of k samples, where k is more than two, are as follows:

- If the k sampled populations are normally distributed and have identical variances (or if they are only slightly to moderately nonnormal and have similar variances): The analysis of variance, using F, is appropriate and preferable to test for difference among the means. (However, samples nearly always come from distributions that are not *exactly* normal with *exactly* the same variances, so conclusions to reject or not reject a null hypothesis should not be considered definitive when the probability associated with F is very near the α specified for the hypothesis test; in such a situation the statistical conclusion should be expressed with some caution and, if feasible, the experiment should be repeated (perhaps with more data).

- If the k sampled populations are distributed normally (or are only slightly to moderately nonnormal), but they have very dissimilar variances: The Behrens-Fisher testing of Section 10.1h is appropriate and preferable to compare the k means. If extremely high and low means are associated with small variances, F' is preferable; but if extreme means are associated with large variances, then F'' works better.
- If the k sampled populations are very different from normally distributed, but they have similar distributions and variances: The Kruskal-Wallis test of Section 10.4 is appropriate and preferable.
- If the k sampled populations have distributions greatly different from normal and do not have similar distributions and variances: (1) Consider the procedures of Chapter 13 for data that do not exhibit normality and variance equality but that can be transformed into data that are normal and homogeneous of variance; or (2) report the mean and variability for each of the k samples, perhaps also presenting them in tables and/or graphs (as in Section 7.4), but do not perform hypothesis testing.

(j) Outliers. A small number of data that are much more extreme than the rest of the measurements are called *outliers* (introduced in Section 2.5), and they may cause a sample to depart seriously from the assumptions of normality and variance equality. If, in the experiment of Example 10.1, a pig weight of 652 kg, or 7.12 kg, or 149 kg was reported, the researcher would likely suspect an error. Perhaps the first two of these measurements were the result of the careless reporting of weights of 65.2 kg and 71.2 kg, respectively; and perhaps the third was a weight measured in pounds and incorrectly reported as kilograms. If there is a convincingly explained error such as this, then an offending datum might be readily corrected. Or, if it is believed that a greatly disparate datum is the result of erroneous data collection (e.g., an errant technician, a contaminated reagent, or an instrumentation malfunction), then it might be discarded or replaced. In other cases outliers might be valid data, and their presence may indicate that one should not employ statistical analyses that require population normality and variance equality. There are statistical methods that are sometimes used to detect outliers, some of which are discussed by Barnett and Lewis (1994: Chapter 6), Snedecor and Cochran (1989: 280–281), and Thode (2002: Chapter 6).

True outliers typically will have little or no influence on analyses employing nonparametric two-sample tests (Sections 8.11 and 8.12) or multisample tests (Section 10.4).

10.2 CONFIDENCE LIMITS FOR POPULATION MEANS

When $k > 2$, confidence limits for each of the k population means may be computed in a fashion analogous to that for the case where $k = 2$ (Section 8.2, Equation 8.13), under the same assumptions of normality and homogeneity of variances applicable to the ANOVA. The $1 - \alpha$ confidence interval for μ_i is

$$\overline{X}_i \pm t_{\alpha(2),\nu}\sqrt{\frac{s^2}{n_i}}, \tag{10.31}$$

where s^2 is the error mean square and ν is the error degrees of freedom from the analysis of variance. For example, let us consider the 95% confidence interval for μ_4

in Example 10.1. Here, $\overline{X}_4 = 63.24$ kg, $s^2 = 9.383\ \text{kg}^2$, $n_4 = 5$, and $t_{0.05(2),15} = 2.131$. Therefore, the lower 95% confidence limit, L_1, is $63.24\ \text{kg} - 2.131\sqrt{9.383\ \text{kg}^2/5} = 63.24\ \text{kg} - 3.999\ \text{kg} = 59.24$ kg, and L_2 is $63.24\ \text{kg} + 2.131\sqrt{9.383\ \text{kg}^2/5} = 63.24\ \text{kg} + 3.999\ \text{kg} = 67.24$ kg.

Computing a confidence interval for μ_i would only be warranted if that population mean was concluded to be different from each other population mean. And calculation of a confidence interval for each of the k μ's may be performed only if it is concluded that $\mu_1 \neq \mu_2 \neq \cdots \neq \mu_k$. However, the analysis of variance does not enable conclusions as to which population means are different from which. Therefore, we must first perform multiple comparison testing (Chapter 11), after which confidence intervals may be determined for each different population mean. Confidence intervals for differences between means may be calculated as shown in Section 11.2.

10.3 SAMPLE SIZE, DETECTABLE DIFFERENCE, AND POWER IN ANALYSIS OF VARIANCE

In Section 8.3, dealing with the difference between two means, we saw how to estimate the sample size required to predict a population difference with a specified level of confidence. When dealing with more than two means, we may also wish to determine the sample size necessary to estimate difference between any two population means, and the appropriate procedure will be found in Section 11.2.

In Section 8.4, methods were presented for estimating the power of the two-sample t test, the minimum sample size required for such a test, and the minimum difference between population means that is detectable by such a test. There are also procedures for analysis-of-variance situations, namely for dealing with more than two means. (The following discussion begins with consideration of Model I—fixed-effects model—analyses of variance.)

If H_0 is true for an analysis of variance, then the variance ratio of Equation 10.18 follows the F distribution, this distribution being characterized by the numerator and denominator degrees of freedom (ν_1 and ν_2, respectively). If, however, H_0 is false, then the ratio of Groups MS to error MS follows instead what is known as the *noncentral F distribution*, which is defined by ν_1, ν_2, and a third quantity known as the *noncentrality parameter*. As power refers to probabilities of detecting a false null hypothesis, statistical discussions of the power of ANOVA testing depend upon the noncentral F distribution.

A number of authors have described procedures for estimating the power of an ANOVA, or the required sample size, or the detectable difference among means (e.g., Bausell and Li, 2002; Cohen, 1988: Ch. 8; Tiku, 1967, 1972), but the charts prepared by Pearson and Hartley (1951) provide one of the best of the methods and will be described below.

(a) Power of the Test. Prior to performing an experiment and collecting data from it, it is appropriate and desirable to estimate the power of the proposed test. (Indeed, it is possible that on doing so one would conclude that the power likely will be so low that the experiment needs to be run with many more data or with fewer groups or, perhaps, not run at all.)

Let us specify that an ANOVA involving k groups will be performed at the α significance level, with n data (i.e., replications) per group. We can then estimate the power of the test if we have an estimate of σ^2, the variability within the k populations (e.g., this estimate typically is s^2 from similar experiments, where s^2 is the error MS), and an estimate of the variability among the populations. From this information we

may calculate a quantity called ϕ (lowercase Greek phi), which is related to the noncentrality parameter.

The variability among populations might be expressed in terms of deviations of the k population means, μ_i, from the overall mean of all populations, μ, in which case

$$\phi = \sqrt{\frac{n\sum_{i=1}^{k}(\mu_i - \mu)^2}{k\sigma^2}} \tag{10.32}$$

(e.g., Guenther, 1964: 47; Kirk, 1995: 182). The grand population mean is

$$\mu = \frac{\sum_{i=1}^{k}\mu_i}{k} \tag{10.33}$$

if all the samples are the same size. In practice, we employ the best available estimates of these population means.

Once ϕ has been obtained, we consult Appendix Figure B.1. This figure consists of several pages, each with a different ν_1 (i.e., groups DF) indicated at the upper left of the graph. Values of ϕ are indicated on the lower axis of the graph for both $\alpha = 0.01$ and $\alpha = 0.05$. Each of the curves on a graph is for a different ν_2 (i.e., error DF), for $\alpha = 0.01$ or 0.05, identified on the top margin of a graph. After turning to the graph for the ν_1 at hand, one locates the point at which the calculated ϕ intersects the curve for the given ν_2 and reads horizontally to either the right or left axis to determine the power of the test. This procedure is demonstrated in Example 10.4.

EXAMPLE 10.4 Estimating the Power of an Analysis of Variance When Variability among Population Means Is Specified

A proposed analysis of variance of plant root elongations is to comprise ten roots at each of four chemical treatments. From previous experiments, we estimate σ^2 to be 7.5888 mm^2 and estimate that two of the population means are 8.0 mm, one is 9.0 mm, and one is 12.0 mm. What will be the power of the ANOVA if we test at the 0.05 level of significance?

$$k = 4$$
$$n = 10$$
$$\nu_1 = k - 1 = 3$$
$$\nu_2 = k(n - 1) = 4(9) = 36$$
$$\mu = \frac{8.0 + 8.0 + 9.0 + 12.0}{4} = 9.25$$

$$\phi = \sqrt{\frac{n\sum(\mu_i - \mu)^2}{ks^2}}$$
$$= \sqrt{\frac{10[(8.0 - 9.25)^2 + (8.0 - 9.25)^2 + (9.0 - 9.25)^2 + (12.0 - 9.25)^2]}{4(7.5888)}}$$

$$= \sqrt{\frac{10(10.75)}{4(7.5888)}}$$

$$= \sqrt{3.5414}$$

$$= 1.88$$

In Appendix Figure B.1c, we enter the graph for $\nu_1 = 3$ with $\phi = 1.88, \alpha = 0.05$, and $\nu_2 = 36$ and read a power of about 0.88. Thus, there will be a 12% chance of committing a Type II error in the proposed analysis.

An alternative, and common, way to estimate power is to specify the smallest difference we wish to detect between the two most different population means. Calling this minimum detectable difference δ, we compute

$$\phi = \sqrt{\frac{n\delta^2}{2ks^2}} \tag{10.34}$$

and proceed to consult Appendix Figure B.1 as above, and as demonstrated in Example 10.5. This procedure leads us to the statement that the power will be at least that determined from Appendix Figure B.1 (and, indeed, it typically is greater).

EXAMPLE 10.5 Estimating the Power of an Analysis of Variance When Minimum Detectable Difference Is Specified

For the ANOVA proposed in Example 10.3, we do not estimate the population means, but rather specify that, using ten data per sample, we wish to detect a difference between population means of at least 4.0 mm.

$k = 4$
$\nu_1 = 3$
$n = 10$
$\nu_2 = 36$
$\delta = 4.0$ mm
$s^2 = 7.5888$ mm^2

$$\phi = \sqrt{\frac{n\delta^2}{2ks^2}}$$

$$= \sqrt{\frac{10(4.0)^2}{2(4)(7.5888)}}$$

$$= \sqrt{2.6355}$$

$$= 1.62$$

In Appendix Figure B.1, we enter the graph for $\nu_1 = 3$ with $\phi = 1.62$, $\alpha = 0.05$, and $\nu_2 = 36$ and read a power of about 0.72. That is, there will be a 28% chance of committing a Type II error in the proposed analysis.

It can be seen in Appendix Figure B.1 that power increases rapidly as ϕ increases, and Equations 10.32 and 10.34 show that the power is affected in the following ways:

- Power is greater for greater differences among group means (as expressed by $\Sigma(\mu_i - \mu)^2$ or by the minimum detectable difference, δ).
- Power is greater for larger sample sizes, n_i (and it is greater when the sample sizes are equal).

- Power is greater for fewer groups, k.
- Power is greater for smaller within-group variability, σ^2 (as estimated by s^2, which is the error mean square).
- Power is greater for larger significance levels, α.

These relationships are further demonstrated in Table 10.3a (which shows that for a given total number of data, N, power increases with increased δ and decreases with increased k) and Table 10.3b (in which, for a given sample size, n_i, power is greater for larger δ's and is less for larger k's).

The desirable power in performing a hypothesis test is arbitrary, just as the significance level (α) is arbitrary. A goal of power between 0.75 and 0.90 is often used, with power of 0.80 being common.

TABLE 10.3a: Estimated Power of Analysis of Variance Comparison of Means, with k Samples, with Each Sample of Size $n_i = 20$, with $N = \sum_{i=1}^{k} n_i$ Total Data, and with a Pooled Variance (s^2) of 2.00, for Several Different Minimum Detectable Differences (δ)

	k:	2	3	4	5	6
δ	N:	40	60	80	100	120
1.0		0.59	0.48	0.42	0.38	0.35
1.2		0.74	0.64	0.58	0.53	0.49
1.4		0.86	0.78	0.73	0.68	0.64
1.6		0.94	0.89	0.84	0.81	0.78
1.8		0.97	0.95	0.92	0.90	0.88
2.0		0.99	0.98	0.97	0.95	0.94
2.2		>0.99	0.99	0.99	0.98	0.98
2.4		>0.99	>0.99	>0.99	0.99	0.99

The values of power were obtained from UNISTAT (2003: 473–474).

TABLE 10.3b: Estimated Power of Analysis of Variance Comparison of Means, with k Samples, with the k Sample Sizes (n_i) Totaling $N = \sum_{i=1}^{k} n_i = 60$ Data, and with a Pooled Variance (s^2) of 2.00, for Several Different Minimum Detectable Differences (δ)

	k:	2	3	4	5	6
δ	n_i:	30	20	15	12	10
1.0		0.77	0.48	0.32	0.23	0.17
1.2		0.90	0.64	0.44	0.32	0.24
1.4		0.96	0.78	0.58	0.43	0.32
1.6		0.99	0.89	0.71	0.54	0.42
1.8		>0.99	0.95	0.82	0.66	0.52
2.0		>0.99	0.98	0.90	0.76	0.63
2.2		>0.99	0.99	0.95	0.85	0.72
2.4		>0.99	>0.99	0.98	0.91	0.81

The values of power were obtained from UNISTAT (2003: 473–474).

Estimating the power of a proposed ANOVA may effect considerable savings in time, effort, and expense. For example, such an estimation might conclude that the power is so very low that the experiment, as planned, ought not to be performed. The proposed experimental design might be revised, perhaps by increasing n, or decreasing k, so as to render the results more likely to be conclusive. One may also strive to increase power by decreasing s^2, which may be possible by using experimental subjects that are more homogeneous. For instance, if the 19 pigs in Example 10.1 were not all of the same age and breed and not all maintained at the same temperature, there might well be more weight variability within the four dietary groups than if all 19 were the same in all respects except diet.

As noted for one-sample (Section 7.7) and two-sample (Section 8.4) testing, calculations of power (and of minimum required sample size and minimum detectable difference) and estimates apply to future samples, not to the samples already subjected to the ANOVA. There are both theoretical and practical reasons for this (Hoenig and Heisey, 2001).

(b) Sample Size Required. Prior to performing an analysis of variance, we might ask how many data need to be obtained in order to achieve a desired power. We can specify the power with which we wish to detect a particular difference (say, a difference of biological significance) among the population means and then ask how large the sample from each population must be. This is done, with Equation 10.34, by iteration (i.e., by making an initial guess and repeatedly refining that estimate), as shown in Example 10.6.

How well Equation 10.34 performs depends upon how good an estimate s^2 is of the population variance common to all groups. As the excellence of s^2 as an estimate improves with increased sample size, one should strive to calculate this statistic from a sample with a size that is not a very small fraction of the n estimated from Equation 10.34.

EXAMPLE 10.6 Estimation of Required Sample Size for a One-Way Analysis of Variance

Let us propose an experiment such as that described in Example 10.1. How many replicate data should be collected in each of the four samples so as to have an 80% probability of detecting a difference between population means as small as 3.5 kg, testing at the 0.05 level of significance?

In this situation, $k = 4$, $\nu_1 = k - 1 = 3$, $\delta = 3.5$ kg, and we shall assume (from the previous experiment in Example 10.1) that $s^2 = 9.383$ kg^2 is a good estimate of σ^2.

We could begin by guessing that $n = 15$ is required. Then, $\nu_2 = 4(15 - 1) = 56$, and by Equation 10.34,

$$\phi = \sqrt{\frac{n\delta^2}{2ks^2}} = \sqrt{\frac{15(3.5)^2}{2(4)(9.383)}} = 1.56.$$

Consulting Appendix Figure B.1, the power for the above ν_1, ν_2, α, and ϕ is approximately 0.73. This is a lower power than we desire, so we guess again with a larger n, say $n = 20$:

$$\phi = \sqrt{\frac{20(3.5)^2}{2(4)(9.383)}} = 1.81.$$

Appendix Figure B1 indicates that this ϕ, for $\nu_2 = 4(20 - 1) = 76$, is associated with a power of about 0.84. This power is somewhat higher than we specified, so we could recalculate power using $n = 18$:

$$\phi = \sqrt{\frac{18(3.5)^2}{2(4)(9.383)}} = 1.71$$

and, for $\nu_2 = 4(18 - 1) = 68$, Appendix Figure B.1 indicates a power slightly above 0.80.

Thus, we have estimated that using sample sizes of at least 18 will result in an ANOVA of about 80% for the described experiment. (It will be seen that the use of Appendix Figure B.1 allows only approximate determinations of power; therefore, we may feel more comfortable in specifying that n should be at least 19 for each of the four samples.)

(c) Minimum Detectable Difference. If we specify the significance level and sample size for an ANOVA and the power that we desire the test to have, and if we have an estimate of σ^2, then we can ask what the smallest detectable difference between population means will be. This is sometimes called the "effect size." By entering on Appendix Figure B.1 the specified α, ν_1, and power, we can read a value of ϕ on the bottom axis. Then, by rearrangement of Equation 10.34, the minimum detectable difference is

$$\delta = \sqrt{\frac{2ks^2\phi^2}{n}}. \tag{10.35}$$

Example 10.7 demonstrates this estimation procedure.

EXAMPLE 10.7 Estimation of Minimum Detectable Difference in a One-Way Analysis of Variance

In an experiment similar to that in Example 10.1, assuming that $s^2 = 9.3833\ (\text{kg})^2$ is a good estimate of σ^2, how small a difference between μ's can we have 90% confidence of detecting if $n = 10$ and $\alpha = 0.05$ are used?

As $k = 4$ and $n = 10, \nu_2 = 4(10 - 1) = 36$. For $\nu_1 = 3, \nu_2 = 36, 1 - \beta = 0.90$, and $\alpha = 0.05$, Appendix Figure B.1c gives a ϕ of about 2.0, from which we compute an estimate of

$$\delta = \sqrt{\frac{2ks^2\phi^2}{n}} = \sqrt{\frac{2(4)(9.3833)(2.0)^2}{10}} = 5.5\ \text{kg}.$$

(d) Maximum Number of Groups Testable. For a given α, n, δ, and σ^2, power will decrease as k increases. It may occur that the total number of observations, N, will be limited, and for given ANOVA specifications the number of experimental groups, k, may have to be limited. As Example 10.8 illustrates, the maximum k can be determined by trial-and-error estimation of power, using Equation 10.34.

EXAMPLE 10.8 Determination of Maximum Number of Groups to be Used in a One-Way Analysis of Variance

Consider an experiment such as that in Example 10.1. Perhaps we have six feeds that might be tested, but we have only space and equipment to examine a total of 50 pigs. Let us specify that we wish to test, with $\alpha = 0.05$ and $\beta \leq 0.20$ (i.e., power of at least 80%), and to detect a difference as small as 4.5 kg between population means.

If $k = 6$ were used, then $n = 50/6 = 8.3$ (call it 8), $\nu_1 = 5, \nu_2 = 6(8 - 1) = 42$, and (by Equation 10.34)

$$\phi = \sqrt{\frac{(8)(4.5)^2}{2(6)(9.3833)}} = 1.20,$$

for which Appendix Figure B.1e indicates a power of about 0.55.

If $k = 5$ were used, $n = 50/5 = 10, \nu_1 = 4, \nu_2 = 5(10 - 1) = 45$, and

$$\phi = \sqrt{\frac{(10)(4.5)^2}{2(5)(9.3833)}} = 1.47,$$

for which Appendix Figure B.1d indicates a power of about 0.70.

If $k = 4$ were used, $n = 50/4 = 12.5$ (call it 12), $\nu_1 = 3, \nu_2 = 4(12 - 1) = 44$, and

$$\phi = \sqrt{\frac{(12)(4.5)^2}{2(4)(9.3833)}} = 1.80,$$

for which Appendix Figure B.1c indicates a power of about 0.84.

Therefore, we conclude that no more than four of the feeds should be tested in an analysis of variance if we are limited to a total of 50 experimental pigs.

(e) Random-Effects Analysis of Variance. If the analysis of variance is a random-effects model (described in Section 10.1f), the power, $1 - \beta$, may be determined from

$$F_{(1-\beta),\nu_1,\nu_2} = \frac{\nu_2 s^2 F_{\alpha(1),\nu_1,\nu_2}}{(\nu_2 - 2)(\text{groups MS})} \tag{10.36}$$

(after Scheffé, 1959: 227; Winer, Brown, and Michels, 1979: 246). This is shown in Example 10.9. As with the fixed-effects ANOVA, power is greater with larger n, larger differences among groups, larger α, and smaller s^2.

EXAMPLE 10.9 Estimating the Power of the Random-Effects Analysis of Variance of Example 10.2

$$\text{Groups MS} = 3.00; s^2 = 1.25; \nu_1 = 3, \nu_2 = 16$$

$$F_{\alpha(1),\nu_1,\nu_2} = F_{0.05(1),3,16} = 3.24$$

$$F_{(1-\beta),\nu_1,\nu_2} = \frac{\nu_2 s^2 F_{\alpha(1),\nu_1,\nu_2}}{(\nu_2 - 2)(\text{groups MS})}$$

$$= \frac{(16)(1.25)(3.24)}{(14)(3.00)} = 1.54$$

By consulting Appendix Table B.4, it is seen that an F of 1.54, with degrees of freedom of 3 and 16, is associated with a one-tailed probability between 0.10 and 0.25. (The exact probability is 0.24.) This probability is the power.

To determine required sample size in a random-effects analysis, one can specify values of α, groups MS, s^2, and k. Then, $\nu_1 = k - 1$ and $\nu_2 = k(n - 1)$; and, by iterative trial and error, one can apply Equation 10.36 until the desired power (namely, $1 - \beta$) is obtained.

10.4 NONPARAMETRIC ANALYSIS OF VARIANCE

If a set of data is collected according to a completely randomized design where $k > 2$, it is possible to test nonparametrically for difference among groups. This may be done by the *Kruskal-Wallis test** (Kruskal and Wallis, 1952), often called an "analysis of variance by ranks."† This test may be used in any situation where the parametric single-factor ANOVA (using F) of Section 10.1 is applicable, and it will be $3/\pi$ (i.e., 95.5%) as powerful as the latter; and in other situations its power, relative to F, is never less than 86.4% (Andrews, 1954; Conover 1999: 297). It may also be employed in instances where the latter is not applicable, in which case it may in fact be the more powerful test. The nonparametric analysis is especially desirable when the k samples do not come from normal populations (Keselman, Rogan, and Feir-Walsh, 1977; Krutchkoff, 1998). It also performs acceptably if the populations have no more than slightly different dispersions and shapes; but if the k variances are not the same, then (as with the Mann-Whitney test) the probability of a Type I error departs from the specified α in accordance with the magnitude of those differences (Zimmerman, 2000).‡

As with the parametric analysis of variance (Section 10.1), the Kruskal-Wallis test tends to be more powerful with larger sample sizes, and the power is less when the n_i's are not equal, especially if the large means are associated with the small n_i's (Boehnke, 1984); and it tends to be conservative if the groups with large n_i's have high within-groups variability and liberal if the large samples have low variability (Keselman, Rogan, and Feir-Walsh, 1997). Bochnke (1984) advises against using the Kruskal-Wallis test unless $N > 20$.

If $k = 2$, then the Kruskal-Wallis test is equivalent to the Mann-Whitney test of Section 8.11. Like the Mann-Whitney test, the Kruskal-Wallis procedure does

*William Henry Kruskal (b. 1919), American statistician, and Wilson Allen Wallis (b. 1912), American statistician and econometrician.

†As will be seen, this procedure does not involve variances, but the term *nonparametric analysis of variance* is commonly applied to it in recognition that the test is a nonparametric analog to the parametric ANOVA.

‡Modifications of the Kruskal-Wallis test have been proposed for nonparametric situations where the k variances are not equal (the "Behrens-Fisher problem" addressed parametrically in Section 10.1h) but the k populations are symmetrical (Rust and Fligner, 1984; Conover 1999: 223–224).

not test whether means (or medians or other parameters) may be concluded to be different from each other, but instead addresses the more general question of whether the sampled populations have different distrubutions. However, if the shapes of the distributions are very similar, then the test does become a test for central tendency (and is a test for means if the distributions are symmetric).

The Type I error rate with heterogeneous variances is affected less with the Kruskal-Wallis test than with the parametric analysis of variance if the groups with large variances have small sample sizes (Keselman, Rogan, and Feir-Walsh, 1977; Tomarkin and Serlin, 1986).

Example 10.10 demonstrates the Kruskal-Wallis test procedure. As in other nonparametric tests, we do not use population parameters in statements of hypotheses, and neither parameters nor sample statistics are used in the test calculations. The Kruskal-Wallis test statistic, H, is calculated as

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1), \tag{10.37}$$

where n_i is the number of observations in group i, $N = \sum_{i=1}^{k} n_i$ (the total number of observations in all k groups), and R_i is the sum of the ranks of the n_i observations in group i.* The procedure for ranking data is as presented in Section 8.11 for the Mann-Whitney test. A good check (but not a guarantee) of whether ranks have been assigned correctly is to see whether the sum of all the ranks equals $N(N+1)/2$.

Critical values of H for small sample sizes where $k \leq 5$ are given in Appendix Table B.13. For larger samples and/or for $k > 5$, H may be considered to be approximated by χ^2 with $k - 1$ degrees of freedom. Chi-square, χ^2, is a statistical distribution that is shown in Appendix Table B.1, where probabilities are indicated as column headings and degrees of freedom (ν) designate the rows.

If there are tied ranks, as in Example 10.11, H is a little lower than it should be, and a correction factor may be computed as

$$C = 1 - \frac{\sum t}{N^3 - N}, \tag{10.40}$$

and the corrected value of H is

$$H_c = \frac{H}{C}. \tag{10.41}$$

*Interestingly, H (or H_c of Equation 10.41) could also be computed as

$$H = \frac{\text{groups SS}}{\text{total MS}}, \tag{10.38}$$

applying the procedures of Section 10.1 to the ranks of the data in order to obtain the Groups SS and Total MS. And, because the Total MS is the variance of all N ranks, if there are no ties the Total MS is the variance of the integers from 1 to N, which is

$$\frac{N(N+1)(2N+1)/6 - N^2(N+1)^2/4N}{N-1} \tag{10.38a}$$

The following alternate formula (Pearson and Hartley, 1976: 49) shows that H is expressing the differences among the κ groups' mean ranks ($\bar{R}_i = R_i/n_i$) and the mean of all N ranks, which is $\bar{R} = N(N-1)/2$:

$$H = \frac{12 \sum_{i=1}^{\kappa} n_i(\bar{R}_i - \bar{R})^2}{N(N-1)}. \tag{10.39}$$

EXAMPLE 10.10 The Kruskal-Wallis Single-Factor Analysis of Variance by Ranks

An entomologist is studying the vertical distribution of a fly species in a deciduous forest and obtains five collections of the flies from each of three different vegetation layers: herb, shrub, and tree.

H_0: The abundance of the flies is the same in all three vegetation layers.
H_A: The abundance of the flies is not the same in all three vegetation layers.

$\alpha = 0.05$
The data are as follows (with ranks of the data in parentheses):*

Numbers of Flies/m^3 of Foliage		
Herbs	*Shrubs*	*Trees*
14.0 (15)	8.4 (11)	6.9 (8)
12.1 (14)	5.1 (2)	7.3 (9)
9.6 (12)	5.5 (4)	5.8 (5)
8.2 (10)	6.6 (7)	4.1 (1)
10.2 (13)	6.3 (6)	5.4 (3)
$n_1 = 5$	$n_2 = 5$	$n_3 = 5$
$R_1 = 64$	$R_2 = 30$	$R_3 = 26$

$$N = 5 + 5 + 5 = 15$$

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_1} - 3(N+1)$$

$$= \frac{12}{15(16)} \left[\frac{64^2}{5} + \frac{30^2}{5} + \frac{26^2}{5} \right] - 3(16)$$

$$= \frac{12}{240}[1134.400] - 48$$

$$= 56.720 - 48$$

$$= 8.720$$

$H_{0.05,5,5,5} = 5.780$
Reject H_0.

$$0.005 < P < 0.01$$

*To check whether ranks were assigned correctly, the sum of the ranks (or sum of the rank sums: $64 + 30 + 26 = 120$) is compared to $N(N+1)/2 = 15(16)/2 = 120$. This check will not guarantee that the ranks were assigned properly, but it will often catch errors of doing so.

EXAMPLE 10.11 The Kruskal-Wallis Test with Tied Ranks

A limnologist obtained eight containers of water from each of four ponds. The *p*H of each water sample was measured. The data are arranged in ascending order within each pond. (One of the containers from pond 3 was lost, so $n_3 = 7$, instead of 8; but the test procedure does not require equal numbers of data in each group.) The rank of each datum is shown parenthetically.

H_0: *p*H is the same in all four ponds.
H_A: *p*H is not the same in all four ponds.

$\alpha = 0.05$

Pond 1	*Pond 2*	*Pond 3*	*Pond 4*
7.68 (1)	7.71 (6*)	7.74 (13.5*)	7.71 (6*)
7.69 (2)	7.73 (10*)	7.75 (16)	7.71 (6*)
7.70 (3.5*)	7.74 (13.5*)	7.77 (18)	7.74 (13.5*)
7.70 (3.5*)	7.74 (13.5*)	7.78 (20*)	7.79 (22)
7.72 (8)	7.78 (20*)	7.80 (23.5*)	7.81 (26*)
7.73 (10*)	7.78 (20*)	7.81 (26*)	7.85 (29)
7.73 (10*)	7.80 (23.5*)	7.84 (28)	7.87 (30)
7.76 (17)	7.81 (26*)		7.91 (31)
*Tied ranks.			
$n_1 = 8$	$n_2 = 8$	$n_3 = 7$	$n_4 = 8$
$R_1 = 55$	$R_2 = 132.5$	$R_3 = 145$	$R_4 = 163.5$

$$N = 8 + 8 + 7 + 8 = 31$$

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

$$= \frac{12}{31(32)} \left[\frac{55^2}{8} + \frac{132.5^2}{8} + \frac{145^2}{7} + \frac{163.5^2}{8} \right] - 3(32)$$

$$= 11.876$$

Number of groups of tied ranks = $m = 7$.

$$\Sigma t = \Sigma(t_i^3 - t_i)$$

$$= (2^3 - 2) + (3^3 - 3) + (3^3 - 3) + (4^3 - 4) + (3^3 - 3) + (2^3 - 2) + (3^3 - 3)$$

$$= 168$$

$$C = 1 - \frac{\Sigma t}{N^3 - N} = 1 - \frac{168}{31^3 - 31} = 1 - \frac{168}{29760} = 0.9944$$

$$H_c = \frac{H}{C} = \frac{11.876}{0.9944} = 11.943$$

$$\nu = k - 1 = 3$$

$\chi^2_{0.05,3} = 7.815$

Reject H_0.

$$0.005 < P < 0.01 \quad [P = 0.0076]$$

or, by Equation 10.43,

$$F = \frac{(N - k)H_c}{(k - 1)(N - 1 - H_c)} = \frac{(31 - 4)(11.943)}{(4 - 1)(31 - 1 - 11.943)} = 5.95$$

$F_{0.05(1),3,26} = 2.98$

Reject H_0.

$$0.0025 < P < 0.005 \quad [P = 0.0031]$$

Here,

$$\sum t = \sum_{i=1}^{m} (t_i^3 - t_i), \tag{10.42}$$

where t_i is the number of ties in the ith group of ties, and m is the number of groups of tied ranks. H_c will differ little from H when the t_i's are very small compared to N.

Kruskal and Wallis (1952) give two approximations that are better than chi-square when the n_i's are small or when significance levels less than 1% are desired; but they are relatively complicated to use. The chi-square approximation is slightly conservative for $\alpha = 0.05$ or 0.10 (i.e., the true Type I probability is a little less than α) and more conservative for $\alpha = 0.01$ (Gabriel and Lachenbruch, 1969): it performs better with larger n_i's. Fahoome (2002) found the probability of a Type I error to be between 0.045 and 0.055 when employing this approximation at the 0.05 significance level if each sample size is at least 11, and between 0.009 and 0.011 when testing at $\alpha = 0.01$ when each $n_i \geq 22$.

Because the χ^2 approximation tends to be conservative, other approximations have been proposed that are better in having Type I error probabilities closer to α. A good alternative is to calculate

$$F = \frac{(N - k)H}{(k - 1)(N - 1 - H)}, \tag{10.43}$$

which is also the test statistic that would be obtained by applying the parametric ANOVA of Section 10.1 to the ranks of the data (Iman, Quade and Alexander, 1975). For the Kruskal-Wallis test, this F gives very good results, being only slightly liberal (with the probability of a Type I error only a little larger than the specified α), and the preferred critical values are F for the given α and degrees of freedom of $\nu_1 = k - 1$ and $\nu_2 = N - k - 1$.* This is demonstrated at the end of Example 10.11.

*A slightly better approximation in some, but not all, cases is to compare

$$\frac{H}{2}\left[1 + \frac{N - k}{N - 1 - H}\right] \text{ to } \frac{(k - 1)F_{\alpha(1),k-1,N-k} + \chi^2_{\alpha,k-1}}{2}. \tag{10.43a}$$

5 TESTING FOR DIFFERENCE AMONG SEVERAL MEDIANS

Section 8.12 presented the *median test* for the two-sample case. This procedure may be expanded to multisample considerations (Mood, 1950: 398–399). The method requires the determination of the grand median of all observations in all k samples considered together. The numbers of data in each sample that are above and below this median are tabulated, and the significance of the resultant $2 \times k$ contingency table is then analyzed, generally by chi-square (Section 23.1), alternatively by the G test (Section 23.7). For example, if there were four populations being compared, the statistical hypotheses would be H_0: all four populations have the same median, and H_A: all four populations do not have the same median. The median test would be the testing of the following contingency table:

	Sample 1	*Sample 2*	*Sample 3*	*Sample 4*	*Total*
Above median	f_{11}	f_{12}	f_{13}	f_{14}	R_1
Below median	f_{21}	f_{22}	f_{23}	f_{24}	R_2
Total	C_1	C_2	C_3	C_4	n

This multisample median test is demonstrated in Example 10.12. Section 8.12 discusses situations where one or more data in the sample are equal to the grand median. Recommended sample sizes are those described is Section 23.4. If H_0 is rejected, than the method of Section 11.7 can be used to attempt to conclude which population medians are different from which.

EXAMPLE 10.12 The Multisample Median Test

H_0: Median elm tree height is the same on all four sides of a building.

H_A: Median elm tree height is not the same on all four sides of a building.

A total of 48 seedlings of the same size were planted at the same time, 12 on each of a building's four sides. The heights, after several years of growth, were as follows:

North	*East*	*South*	*West*
7.1 m	6.9 m	7.8 m	6.4 m
7.2	7.0	7.9	6.6
7.4	7.1	8.1	6.7
7.6	7.2	8.3	7.1
7.6	7.3	8.3	7.6
7.7	7.3	8.4	7.8
7.7	7.4	8.4	8.2
7.9	7.6	8.4	8.4
8.1	7.8	8.6	8.6
8.4	8.1	8.9	8.7
8.5	8.3	9.2	8.8
8.8	8.5	9.4	8.9
medians: 7.7 m	7.35 m	8.4 m	8.0 m

grand median = 7.9 m

The 2×4 contingency table is as follows, with expected frequencies (see Section 23.1) in parentheses:

	North	*East*	*South*	*West*	
Above median	4 (5.5000)	3 (6.0000)	10 (5.5000)	6 (6.0000)	23
Below median	7 (5.5000)	9 (6.0000)	1 (5.5000)	6 (6.0000)	23
Total	11	12	11	12	46

$\chi^2 = 11.182$

$\chi^2_{0.05,3} = 7.815$

Reject H_0.

$0.0005 < P < 0.001$ [$P = 0.00083$]

If the k samples came from populations having the same variance and shape, then the Kruskal-Wallis test may be used as a test for difference among the k population medians.

10.6 HOMOGENEITY OF VARIANCES

Section 8.5 discussed testing the null hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ against the alternate, H_A: $\sigma_1^2 \neq \sigma_2^2$. This pair of two-sample hypotheses can be extended to more than two samples (i.e., $k > 2$) to ask whether all k sample variances estimate the same population variance. The null and alternate hypotheses would then be H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$ and H_0: the k population variances are not all the same. The equality of variances is called homogeneity of variances, or *homoscedasticity*; variance heterogeneity is called *heteroscedasticity*.*

(a) Bartlett's Test. A commonly encountered method employed to test for homogeneity of variances is *Bartlett's test*† (Bartlett, 1937a, 1937b; based on a principle of Neyman and Pearson, 1931). In this procedure, the test statistic is

$$B = (\ln s_p^2)\left(\sum_{i=1}^{k} \nu_i\right) - \sum_{i=1}^{k} \nu_i \ln s_i^2, \tag{10.44}$$

where $\nu_i = n_i - 1$ and n_i is the size of sample i. The pooled variance, s_p^2, is calculated as before as $\sum_{i=1}^{k} SS_i / \sum_{i=1}^{k} \nu_i$. Many researchers prefer to operate with common logarithms (base 10), rather than with natural logarithms (base e);‡ so Equation 10.44 may be written as

$$B = 2.30259[(\log s_p^2)\left(\sum_{i=1}^{k} \nu_i\right) - \sum_{i=1}^{k} \nu_i \log s_i^2]. \tag{10.45}$$

The distribution of B is approximated by the chi-square distribution,§ with $k - 1$ degrees of freedom (Appendix Table B.1), but a more accurate chi-square

*The two terms were introduced by K. Pearson in 1905 (Walker, 1929: 181); since then they have occasionally been spelled *homoskedasticity* and *heteroskedasticity*, respectively.

†Maurice Stevenson Bartlett (1910–2002), English statistician.

‡See footnote in Section 8.7.

§A summary of approximations is given by Nagasenker (1984).

approximation is obtained by computing a correction factor,

$$C = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{\nu_i} - \frac{1}{\sum_{i=1}^{k}\nu_i}\right), \tag{10.46}$$

with the corrected test statistic being

$$B_c = \frac{B}{C}. \tag{10.47}$$

Example 10.13 demonstrates these calculations. The null hypothesis for testing the homogeneity of the variances of four populations may be written symbolically as H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$, or, in words, as "the four population variances are homogeneous (i.e., are equal)." The alternate hypothesis can be stated as "The four population variances are not homogeneous (i.e., they are not all equal)," or "There is difference (or heterogeneity) among the four population variances." If H_0 is rejected, the further testing of Section 11.8 will allow us to ask which population variances are different from which.

Bartlett's test is powerful if the sampled populations are normal, but it is very badly affected by nonnormal populations (Box, 1953; Box and Anderson, 1955; Gartside, 1972). If the population distribution is platykurtic, the true α is less than the stated α (i.e., the test is conservative and the probability of a Type II error is increased); if it is leptokurtic, the true α is greater than the stated α (i.e., the probability of a Type I error is increased).

When $k = 2$ and $n_1 = n_2$, Bartlett's test is equivalent to the variance-ratio test of Section 8.5a. However, with two samples of unequal size, the two procedures may yield different results; one will be more powerful in some cases, and the other more powerful in others (Maurais and Ouimet, 1986).

(b) Other Multisample Tests for Variances. Section 8.5b noted that there are other tests for heterogeneity (Levene's test and others) but that all are undesirable in many situations. The Bartlett test remains commendable when the sampled populations are normal, and no procedure is especially good when they are not.

Because of the poor performance of tests for variance homogeneity and the robustness of analysis of variance for multisample testing among means (Section 10.1), it is not recommended that the former be performed as tests of the underlying assumptions of the latter.

HOMOGENEITY OF COEFFICIENTS OF VARIATION

The two-sample procedure of Section 8.8 has been extended by Feltz and Miller (1996) for hypotheses where $k \geq 3$ and each coefficient of variation (V_i) is positive:

$$\chi^2 = \frac{\sum_{i=1}^{k}\nu_i V_i^2 - \dfrac{\left(\sum_{i=1}^{k}\nu_i V_i\right)^2}{\sum_{i=1}^{k}\nu_i}}{V_p^2(0.5 + V_p^2)}, \tag{10.48}$$

EXAMPLE 10.13 Bartlett's Test for Homogeneity of Variances

Nineteen pigs were divided into four groups, and each group was raised on a different food. The data, which are those of Example 10.1, are weights, in kilograms, and we wish to test whether the variance of weights is the same for pigs fed on all four feeds.

H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$

H_A: The four population variances are not all equal (i.e., are heterogeneous).

$\alpha = 0.05$

	Feed 1	*Feed 2*	*Feed 3*	*Feed 4*	
	60.8	68.7	69.6	61.9	
	67.0	67.7	77.1	64.2	
	65.0	75.0	75.2	63.1	
	68.6	73.3	71.5	66.7	
	61.7	71.8		60.3	
i	1	2	3	4	
n_i	5	5	4	5	
ν_i	4	4	3	4	$\sum_{i=1}^{k} \nu_i = 15$
SS_i	44.768	37.660	34.970	23.352	$\sum_{i=1}^{k} SS_i = 140.750$
s_i^2	11.192	9.415	11.657	5.838	
$\log s_i^2$	1.0489	0.9738	1.0666	0.7663	
$\nu_i \log s_i^2$	4.1956	3.8952	3.1998	3.0652	$\sum_{i=1}^{k} \nu_i \log s_i^2 = 14.3558$
$1/\nu_i$	0.250	0.250	0.333	0.250	$\sum_{i=1}^{k} 1/\nu_i = 1.083$

$$s_p^2 = \frac{\sum SS_i}{\sum \nu_i} = \frac{140.750}{15} = 9.3833$$

$$\log s_p^2 = 0.9724$$

$$B = 2.30259\left[(\log s_p^2)(\sum \nu_i) - \sum \nu_i \log s_i^2)\right]$$

$$= 2.30259[(0.9724)(15) - 14.3558]$$

$$= 2.30259(0.2302)$$

$$= 0.530$$

$$C = 1 + \frac{1}{3(k-1)} \times \left(\sum \frac{1}{\nu_i} - \frac{1}{\sum \nu_i}\right)$$

$$= 1 + \frac{1}{3(3)}\left(1.083 - \frac{1}{15}\right)$$

$$= 1.113$$

$$B_c = \frac{B}{C} = \frac{0.530}{1.113} = 0.476$$

$$\chi^2_{0.05,3} = 7.815$$

Do not reject H_0.

$0.90 < P < 0.95$ $[P = 0.92]$

EXERCISES

10.1. The following data are weights of food (in kilograms) consumed per day by adult deer collected at different times of the year. Test the null hypothesis that food consumption is the same for all the months tested.

Feb.	*May*	*Aug.*	*Nov.*
4.7	4.6	4.8	4.9
4.9	4.4	4.7	5.2
5.0	4.3	4.6	5.4
4.8	4.4	4.4	5.1
4.7	4.1	4.7	5.6
	4.2	4.8	

10.2. An experiment is to have its results examined by analysis of variance. The variable is temperature (in degrees Celsius), with 12 measurements to be taken in each of five experimental groups. From previous experiments, we estimate the within-groups variability, σ^2, to be $1.54(^\circ C)^2$. If the 5% level of significance is employed, what is the probability of the ANOVA detecting a difference as small as 2.0°C between population means?

10.3. For the experiment of Exercise 10.2, how many replicates are needed in each of the five groups to detect a difference as small as 2.0°C between population means, with 95% power?

10.4. For the experiment of Exercise 10.2, what is the smallest difference between population means that we are 95% likely to detect with an ANOVA using 10 replicates per group?

10.5. Using the Kruskal-Wallis test, test nonparametrically the appropriate hypotheses for the data of Exercise 10.1.

10.6. Three different methods were used to determine the dissolved-oxygen content of lake water. Each of the three methods was applied to a sample of water six times, with the following results. Test the null hypothesis that the three methods yield equally variable results ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2$).

Method 1 (mg/kg)	*Method 2* (mg/kg)	*Method 3* (mg/kg)
10.96	10.88	10.73
10.77	10.75	10.79
10.90	10.80	10.78
10.69	10.81	10.82
10.87	10.70	10.88
10.60	10.82	10.81

10.7. The following statistics were obtained from measurements of the circumferences of trees of four species. Test whether the coefficients of variation of circumferences are the same among the four species.

	Species B	*Species A*	*Species Q*	*Species H*
n:	40	54	58	32
$\overline{X}$ (m):	2.126	1.748	1.350	1.392
s^2 (m^2):	0.488219	0.279173	0.142456	0.203208

where the common coefficient of variation is

$$V_p = \frac{\sum_{i=1}^{k} \nu_i V_i}{\sum_{i=1}^{k} \nu_i}. \tag{10.49}$$

This test statistic approximates the chi-square distribution with $k - 1$ degrees of freedom (Appendix Table B.1) and its computation is shown in Example 10.14. When $k = 2$, the test yields results identical to the two-sample test using Equation 8.42 (and $\chi^2 = Z^2$). As with other tests, the power is greater with larger sample size; for a given sample size, the power is greater for smaller coefficients of variation and for greater differences among coefficients of variation. If the null hypothesis of equal population coefficients of variation is not rejected, then V_p is the best estimate of the coefficient of variation common to all k populations.

EXAMPLE 10.14 Testing for Homogeneity of Coefficients of Variation

For the data of Example 10.1:

H_0: The coefficients of the four sampled populations are the same; i.e., $\sigma_1^2/\mu_1 = \sigma_2^2/\mu_2 = \sigma_3^2/\mu_3 = \sigma_4^2/\mu_4$.

H_A: The coefficients of variation of the four populations are not all the same.

	Feed 1	*Feed 2*	*Feed 3*	*Feed 4*
n_i	5	5	4	5
ν_i	4	4	3	4
$\overline{X}_i$ (kg)	64.62	68.30	73.35	66.64
s_i^2 (kg^2)	11.192	16.665	11.657	9.248
s_i (kg)	3.35	4.08	3.41	3.04
V_i	0.0518	0.0597	0.0465	0.0456

$$\sum_{i=1}^{n} \nu_i = 4 + 4 + 3 + 4 = 15$$

$$\sum_{i=1}^{n} \nu_i V_i = (4)(0.0518) + (4)(0.0597) + (3)(0.0465) + (4)(0.0456) = 0.7679$$

$$V_p = \frac{\sum \nu_i V_i}{\sum \nu_i} = \frac{0.7679}{15} = 0.05119 \qquad V_p^2 = (0.7679)^2 = 0.002620$$

$$\sum_{i=1}^{n} \nu V_i^2 = (4)(0.0518)^2 + (4)(0.0597)^2 + (3)(0.0465)^2 + (4)(0.0456)^2$$
$$= 0.03979$$

$$\chi^2 = \frac{\sum \nu_i V_i^2 - \frac{(\sum \nu_i V_i)^2}{\sum \nu_i}}{V_p^2(0.5 + V_p^2)} = \frac{0.03979 - \frac{(0.7679)^2}{15}}{0.002620(0.5 + 0.002620)} = \frac{0.0004786}{0.001317} = 0.363$$

For chi-square: $\nu = 4 - 1 = 3$; $\chi^2_{0.05,3} = 7.815$. Do not reject H_0.

$$0.90 < P < 0.95 \quad [P = 0.948]$$

Miller and Feltz (1997) reported that this test works best if each sample size (n_i is at least 10 and each coefficient of variation (V_i) is no greater than 0.33; and they describe how the power of the test (and, from such a calculation, the minimum detectable difference and the required sample size) may be estimated.

10.8 CODING DATA

In the parametric ANOVA, coding the data by addition or subtraction of a constant causes no change in any of the sums of squares or mean squares (recall Section 4.8) so the resultant F and the ensuing conclusions are not affected at all. If the coding is performed by multiplying or dividing all the data by a constant, the sums of squares and the mean squares in the ANOVA each will be altered by an amount equal to the square of that constant, but the F value and the associated conclusions will remain unchanged.

A test utilizing ranks (such as the Kruskal-Wallis procedure) will not be affected at all by coding of the raw data. Thus, the coding of data for analysis of variance either parametric or nonparametric, may be employed with impunity, and coding frequently renders data easier to manipulate. Neither will coding of data alter the conclusions from the hypothesis tests in Chapter 11 (multiple comparisons) or Chapters 12, 14, 15, or 16 (further analysis-of-variance procedures). Bartlett's test is also unaffected by coding. The testing of coefficients of variation is unaffected by coding by multiplication or division, but coding by addition or subtraction may not be used. The effect of coding is indicated in Appendix C for many statistics.

10.9 MULTISAMPLE TESTING FOR NOMINAL-SCALE DATA

A $2 \times c$ contingency table may be analyzed to compare frequency distributions of nominal data for two samples. In a like fashion, an $r \times c$ contingency table may be set up to compare frequency distributions of nominal-scale data from r samples. Contingency table procedures are discussed in Chapter 23.

Other procedures have been proposed for multisample analysis of nominal-scale data (e.g., Light and Margolin, 1971; Windsor, 1948).

CHAPTER 11

Multiple Comparisons

The Model I single-factor analysis of variance (ANOVA) of Chapter 10 tests the null hypothesis H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$. However, the rejection of H_0 does not imply that all k population means are different from one another, and we don't know how many differences there are or where differences lie among the k means. For example, if $k = 3$ and H_0: $\mu_1 = \mu_2 = \mu_3$ is rejected, we are not able to conclude whether there is evidence of $\mu_1 \neq \mu_2 = \mu_3$ or of $\mu_1 = \mu_2 \neq \mu_3$ or of $\mu_1 \neq \mu_2 \neq \mu_3$.

The introduction to Chapter 10 explained that it is invalid to employ multiple two-sample t tests to examine the difference among more than two means, for to do so would increase the probability of a Type I error (as shown in Table 10.1). This chapter presents statistical procedures that may be used to compare k means with each other; they are called *multiple-comparison procedures** (MCPs). Except for the procedure known as the least significance difference test, all of the tests referred to in this chapter may be performed even without a preliminary analysis of variance. Indeed, power may be lost if a multiple-comparison test is performed only if the ANOVA concludes a significant difference among means (Hsu, 1996: 177–178; Myers and Well, 2003: 261). And all except the Scheffé test of Section 11.4 are for a set of comparisons to be specified before the collection of data.

The most common principle for multiple-comparison testing is that the significance level, α, is the probability of committing at least one Type I error when making all of the intended comparisons for a set of data. These are said to be a *family* of comparisons, and this error is referred to as *familywise error* (FWE) or, sometimes, *experimentwise error*. Much less common are tests designed to express *comparisonwise error*, the probability of a Type I error in a single comparison.

A great deal has been written about numerous multiple-comparison tests with various objectives, and the output of many statistical computer packages enhances misuse of them (Hsu, 1996: xi). Although there is not unanimity regarding what the "best" procedure is for a given situation, this chapter will present some frequently encountered highly regarded tests for a variety of purposes.

If the desire is to test for differences between members of all possible pairs of means, then the procedures of Section 11.1 would be appropriate, using Section 11.1a

*The term *multiple comparisons* was introduced by D. E. Duncan in 1951 (David, 1995).

if sample sizes are unequal and Section 11.1b if variances are not the same. If the data are to be analyzed to compare the mean of one group (typically called the control) to each of the other group means, then Section 11.3 would be applicable. And if the researcher wishes to examine sample means after the data are collected and compare specific means, or groups of means, of interest, then the testing in Section 11.4 is called for.

Just as with the parametric analysis of variance, the testing procedures of Sections 11.1–11.4 are premised upon there being a normal distribution of the population from which each of the k samples came; but, like the ANOVA, these tests are somewhat robust to deviations from that assumption. However, if it is suspected that the underlying distributions are far from normal, then the analyses of Section 11.5, or data transformations (Chapter 13), should be considered. Multiple-comparison tests are adversely affected by heterogeneous variances among the sampled populations, in the same manner as in ANOVA (Section 10.1g) (Keselman and Toothaker, 1974; Petrinovich and Hardyck, 1969), though to a greater extent (Tukey, 1993).

In multiple-comparison testing—except when comparing means to a control—equal sample sizes are desirable for maximum power and robustness, but the procedures presented can accommodate unequal n's. Petrinovich and Hardyck (1969) caution that the power of the tests is low when sample sizes are less than 10.

This chapter discusses multiple comparisons for the single-factor ANOVA experimental design (Chapter 10).* Applications for other situations are found in Section 12.5 (for the two-factor ANOVA design), 12.7b (for the nonparametric randomized-block ANOVA design), 12.9 (for dichotomous data in randomized blocks), 14.6 (for the multiway ANOVA design), 18.6 and 18.7 (for regression), and 19.8 (for correlation).

11.1 TESTING ALL PAIRS OF MEANS

There are $k(k-1)/2$ different ways to obtain pairs of means from a total of k means.† For example, if $k = 3$, the $k(k-1)/2 = 3(2)/2 = 3$ pairs are μ_1 and μ_2, μ_1 and μ_3, and μ_2 and μ_3; and for $k = 4$, the $k(k-1)/2 = 4(3)/2 = 6$ pairs are μ_1 and μ_2, μ_1 and μ_3, μ_1 and μ_4, μ_2 and μ_3, μ_2 and μ_4, and μ_3 and μ_4. So each of $k(k-1)/2$ null hypotheses may be tested, referring to them as H_0: $\mu_B = \mu_A$, where the subscripts A and B represent each pair of subscripts; each corresponding alternate hypothesis is H_0: $\mu_B \neq \mu_A$.

An excellent way to address these hypotheses is with the *Tukey test* (Tukey, 1953), also known as the honestly significant difference test (HSD test) or wholly significant difference test (WSD test). Example 11.1 demonstrates the Tukey test, utilizing an ANOVA experimental design similar to that in Example 10.1, except that all groups have equal numbers of data (i.e., all of the n_i's are equal). The first step in examining these multiple-comparison hypotheses is to arrange and number all five sample means in order of increasing magnitude. Then pairwise differences between the means, $\overline{X}_A - \overline{X}_B$, are tabulated. Just as a difference between means, divided by

*For nonparametric testing, Conover and Iman (1981) recommend applying methods as those in Sections 11.1–11.4 on the ranks of the data. However Hsu (1996: 177); Sawilowsky, Blair, and Higgins (1999); and Toothaker (1991: 109) caution against doing so.

†The number of combinations of k groups taken 2 at a time is (by Equation 5.10):

$$ {}_kC_2 = \frac{k!}{2!(k-2)!} = \frac{k(k-1)(k-2)!}{2!(k-2)!} = \frac{k(k-1)}{2}. \qquad (11.1) $$

EXAMPLE 11.1 Tukey Multiple Comparison Test with Equal Sample Sizes.

The data are strontium concentrations (mg/ml) in five different bodies of water. First an analysis of variance is performed.

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$.

H_A: Mean strontium concentrations are not the same in all five bodies of water.

$\alpha = 0.05$

Grayson's Pond	*Beaver Lake*	*Angler's Cove*	*Appletree Lake*	*Rock River*
28.2	39.6	46.3	41.0	56.3
33.2	40.8	42.1	44.1	54.1
36.4	37.9	43.5	46.4	59.4
34.6	37.1	48.8	40.2	62.7
29.1	43.6	43.7	38.6	60.0
31.0	42.4	40.1	36.3	57.3
$\overline{X}_1 = 32.1$ mg/ml	$\overline{X}_2 = 40.2$ mg/ml	$\overline{X}_3 = 44.1$ mg/ml	$\overline{X}_4 = 41.1$ mg/ml	$\overline{X}_5 = 58.3$ mg/ml
$n_1 = 6$	$n_2 = 6$	$n_3 = 6$	$n_4 = 6$	$n_5 = 6$

Source of variation	SS	DF	MS
Total	2437.5720	29	
Groups	2193.4420	4	548.3605
Error	244.1300	25	9.7652

$k = 5, n = 6$

Samples number (i) of ranked means:	1	2	4	3	5
Ranked sample mean ($\overline{X}_i$):	32.1	40.2	41.1	44.1	58.3

To test each H_0: $\mu_B = \mu_A$,

$$SE = \sqrt{\frac{9.7652}{6}} = \sqrt{1.6275} = 1.28.$$

As $q_{0.05,25,k}$ does not appear in Appendix Table B.5, the critical value with the next lower DF is used: $q_{0.05,24,5} = 4.166$.

Comparison B vs. A	*Difference* $(\overline{X}_B - \overline{X}_A)$	SE	q	*Conclusion*
5 vs. 1	58.3 − 32.1 = 26.2	1.28	20.47	Reject H_0: $\mu_5 = \mu_1$
5 vs. 2	58.3 − 40.2 = 18.1	1.28	14.14	Reject H_0: $\mu_5 = \mu_2$
5 vs. 4	58.3 − 41.1 = 17.2	1.28	13.44	Reject H_0: $\mu_5 = \mu_4$
5 vs. 3	58.3 − 44.1 = 14.2	1.28	11.09	Reject H_0: $\mu_5 = \mu_3$
3 vs. 1	44.1 − 32.1 = 12.0	1.28	9.38	Reject H_0: $\mu_3 = \mu_1$
3 vs. 2	44.1 − 40.2 = 3.9	1.28	3.05	Do not reject H_0: $\mu_3 = \mu_2$
3 vs. 4	Do not test			
4 vs. 1	44.1 − 32.1 = 9.0	1.28	7.03	Reject H_0: $\mu_4 = \mu_1$
4 vs. 2	Do not test			
2 vs. 1	40.2 − 32.1 = 8.1	1.28	6.33	Reject H_0: $\mu_2 = \mu_1$

Thus, we conclude that μ_1 is different from the other means, that μ_5 is different from the other means, and that μ_2, μ_4, and μ_3 are indistinguishable from each other: $\mu_1 \neq \mu_2 = \mu_4 = \mu_3 \neq \mu_5$.

the appropriate standard error, yields a t value (Section 8.1), the Tukey test statistic, q, is calculated by dividing a difference between two means by

$$\text{SE} = \sqrt{\frac{s^2}{n}}, \tag{11.2}$$

where n is the number of data in each of groups B and A, and s^2 is the error mean square by ANOVA computation (Equation 10.14). Thus

$$q = \frac{\overline{X}_B - \overline{X}_A}{\text{SE}}, \tag{11.3}$$

which is known as the *studentized range** (and is sometimes designated as T). The null hypothesis H_0: $\overline{X}_B = \overline{X}_A$ is rejected if q is equal to or greater than the critical value, $q_{\alpha,\nu,k}$, from Appendix Table B.5, where ν is the error degrees of freedom (via Equation 10.15, which is $N - k$).

The significance level, α, is the probability of committing at least one Type I error (i.e., the probability of incorrectly rejecting at least one H_0) during the course of comparing all pairs of means. And the Tukey test has good power and maintains the probability of the familywise Type I error at or below the stated α.

The conclusions reached by this multiple-comparison testing may depend upon the order in which the pairs of means are compared. The proper procedure is to compare first the largest mean against the smallest, then the largest against the next smallest, and so on, until the largest has been compared with the second largest. Then one compares the second largest with the smallest, the second largest with the next smallest, and so on. Another important procedural rule is that if no significant difference is found

*E. S. Pearson and H. O. Hartley first used this term in 1953 (David, 1995).

between two means, then it is concluded that no significant difference exists between any means enclosed by those two, and no differences between enclosed means are tested for. Thus, in Example 11.1, because we conclude no difference between population means 3 and 2, no testing is performed to judge the difference between means 3 and 4, or between means 4 and 2. The conclusions in Example 11.1 are that Sample 1 came from a population having a mean different from that of any of the other four sampled populations; likewise, it is concluded that the population mean from which Sample 5 came is different from any of the other population means, and that samples 2, 4, and 3 came from populations having the same means. Therefore, the overall conclusion is that $\mu_1 \neq \mu_2 = \mu_4 = \mu_3 \neq \mu_5$. As a visual aid in Example 11.1, each time a null hypothesis was not rejected, a line was drawn beneath means to connect the two means tested and to encompass any means between them.

The null hypothesis H_0: $\mu_B = \mu_A$ may also be written as $\mu_B - \mu_A = 0$. The hypothesis $\mu_B - \mu_A = \mu_0$, where $\mu_0 \neq 0$, may also be tested; this is done by replacing $\overline{X}_B - \overline{X}_A$ with $|\overline{X}_B - \overline{X}_A| - \mu_0$ in the numerator of Equation 11.3.

Occasionally, a multiple-comparison test, especially if $n_B \neq n_A$, will yield ambiguous results in the form of conclusions of overlapping spans of nonsignificance. For example, one might arrive at the following:

$$\underline{\overline{X}_1 \quad \overline{X}_2 \quad \overline{X}_3 \quad \overline{X}_4}$$

for an experimental design consisting of four groups of data. Here the four samples seem to have come from populations among which there were two different population means: Samples 1 and 2 appear to have been taken from one population, and Samples 2, 3, and 4 from a different population. But this is clearly impossible, for Sample 2 has been concluded to have come from both populations. Because the statistical testing was not able to conclude decisively from which population Sample 2 came, at least one Type II error has been committed. Therefore, it can be stated that $\mu_1 \neq \mu_3 \neq \mu_4$, but it cannot be concluded from which of the two populations Sample 2 came (or if it came from a third population). Repeating the data collection and analysis with a larger number of data might yield more conclusive results.

(a) Multiple Comparisons with Unequal Sample Sizes. If the sizes of the k samples are not equal, the Tukey-Kramer procedure (Kramer, 1956; supported by Dunnett, 1980a; Stoline, 1981; Jaccard, Becker, and Wood, 1984)* is desirable to maintain the probability of a Type I error near α and to operate with good power. For each comparison involving unequal n's, the standard error for use in Equation 11.3 is calculated as

$$\text{SE} = \sqrt{\frac{s^2}{2}\left(\frac{1}{n_B} + \frac{1}{n_A}\right)}, \tag{11.4}$$

which is inserting the harmonic mean of n_B and n_A (Section 3.4b) in place of n in Equation 11.2;† and Equation 11.4 is equivalent to 11.2 when $n_B = n_A$. This test is shown in Example 11.2, using the data of Example 10.1.

*This procedure has been shown to be excellent (e.g., Dunnett, 1980a; Hayter, 1984; Keselman, Murray, and Rogan, 1976; Smith, 1971; Somerville, 1993; Stoline, 1981), with the probability of a familywise Type I error no greater than the stated α.

†Some researchers have replaced n in Equation 11.2 with the harmonic mean of all k samples or with the median or arithmetic mean of the pair of means examined. Dunnett (1980a); Keselman, Murray, and Rogan (1976); Keselman and Rogan (1977); and Smith (1971) concluded the Kramer approach to be superior to those methods, and it is analogous to Equation 8.7a, which is used for two-sample t testing.

EXAMPLE 11.2 The Tukey-Kramer Test with Unequal Sample Sizes

The data (in kg) are those from Equation 10.1.

$k = 4$

s^2 = Error MS = 9.383

Error DF = 15

$q_{0.05,15,4} = 4.076$

Sample number (i) of ranked means:	4	1	2	3
Ranked sample mean ($\overline{X}_i$):	63.24	64.62	71.30	73.35
Sample sizes (n_i):	4	5	5	5

If $n_B = n_A$ (call it n), then $\text{SE} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{9.383}{5}} = \sqrt{2.111} = 1.453.$

If $n_B \neq n_A$, then $\text{SE} = \sqrt{\frac{s^2}{2}\left(\frac{1}{n_B} + \frac{1}{n_A}\right)} = \sqrt{\frac{0.383}{2}\left(\frac{1}{5} + \frac{1}{4}\right)}$

$= \sqrt{1.877} = 1.370.$

Comparison B vs. A	*Difference* $(\overline{X}_B - \overline{X}_A)$	SE	q	*Conclusion*
3 vs. 4	73.35 − 63.24 = 10.11	1.453	6.958	Reject H_0: $\mu_3 = \mu_4$
3 vs. 1	73.35 − 64.62 = 8.73	1.370	6.371	Reject H_0: $\mu_3 = \mu_1$
3 vs. 2	73.35 − 71.30 = 2.05	1.370	1.496	Do not reject H_0: $\mu_3 = \mu_2$
2 vs. 4	71.30 − 63.24 = 8.06	1.453	5.547	Reject H_0: $\mu_2 = \mu_4$
2 vs. 1	71.30 − 64.62 = 6.68	1.370	4.876	Reject H_0: $\mu_2 = \mu_1$
1 vs. 4	64.62 − 63.24 = 1.38	1.453	0.950	Do not reject H_0: $\mu_1 = \mu_4$

Thus, we conclude that μ_4 and μ_1 are indistinguishable, that μ_2 and μ_3 are indistinguishable, and that μ_4 and μ_1 are different from μ_2 and μ_3: $\mu_4 = \mu_1 \neq \mu_2 = \mu_3$.

(b) Multiple Comparisons with Unequal Variances. Although the Tukey test can withstand some deviation from normality (e.g., Jaccard, Becker, and Wood, 1984), it is less resistant to heterogeneous variances, especially if the sample sizes are not equal. The test is conservative if small n's are associated with small variances and undesirably liberal if small samples come from populations with large variances, and in the presence of both nonnormality and heteroscedasticity the test is very liberal. Many investigations* have determined that the Tukey-Kramer test is also adversely affected by heterogeneous variances.

*These include those of Dunnett (1980b, 1982); Games and Howell (1976); Jaccard, Becker, and Wood (1984); Keselman, Games, and Rogan (1979); Keselman and Rogan (1978); Keselman and Toothaker (1974); Keselman, Toothaker, and Shooter (1975); Ramseyer and Tcheng (1973); Jenkdon and Tamhane (1979).

As a solution to this problem, Games and Howell (1976) proposed the use of the Welch approximation (Section 8.1b) to modify Equation 11.4 to be appropriate when the k population variances are not assumed to be the same or similar:

$$\text{SE} = \sqrt{\frac{1}{2}\left(\frac{s_B^2}{n_B} + \frac{s_A^2}{n_A}\right)}. \tag{11.5}$$

and q will be associated with the degrees of freedom of Equation 8.12; but each sample size should be at least 6. This test maintains the probability of a familywise Type I error around α (though it is sometimes slightly liberal) and it has good power (Games, Keselman, and Rogan, 1981; Keselman, Games, and Rogan, 1979; Keselman and Rogan, 1978; Tamhane, 1979). If the population variances are the same, then the Tukey or Tukey-Kramer test is preferable (Kirk, 1995: 147–148). If there is doubt about whether there is substantial heteroscedacity, it is safer to use the Games and Howell procedure, for if the underlying populations do not have similar variances, that test will be far superior to the Tukey-Kramer test; and if the population variances are similar, the former will have only a little less power than the latter (Levy, 1978c).

(c) Other Multiple-Comparison Methods. Methods other than the Tukey and Tukey-Kramer tests have been employed by statisticians to examine pairwise differences for more than two means. The *Newman-Keuls test* (Newman, 1939; Keuls, 1952), also referred to as the *Student-Newman-Keuls test*, is employed as is the Tukey test, except that the critical values from Appendix Table B.5 are those for $q_{\alpha,\nu,p}$ instead of $q_{\alpha,\nu,k}$, where p is the range of means for a given H_0. So, in Example 11.2, comparing means 3 and 4 would use $p = 4$, comparing means 3 and 1 would call for $p = 3$, and so on (with p ranging from 2 to k). This type of multiple-comparison test is called a *multiple-range test*. There is considerable opinion against using this procedure (e.g., by Einot and Gabriel, 1975; Ramsey, 1978) because it may falsely declare differences with a probability undesirably greater than α.

The *Duncan test* (Duncan, 1955) is also known as the *Duncan new multiple range test* because it succeeds an earlier procedure (Duncan, 1951). It has a different theoretical basis, one that is not as widely accepted as that of Tukey's test, and it has been declared (e.g., by Carmer and Swanson, 1973; Day and Quinn, 1899) to perform poorly. This procedure is executed as is the Student-Newman-Keuls test, except that different critical-value tables are required.

Among other tests, there is also a procedure called the *least significant difference test* (LSD), and there are other tests, such as with *Dunn* or *Bonferroni* in their names (e.g., Howell, 2007: 356–363). The name *wholly significant difference test* (WSD test) is sometimes applied to the Tukey test (Section 11.1) and sometimes as a compromise between the Tukey and Student-Newman-Keuls procedures by employing a critical value midway between $q_{\alpha,\nu,k}$ and $q_{\alpha,\nu,p}$. The Tukey test is preferred here because of its simplicity and generally good performance with regard to Type I and Type II errors.

11.2 CONFIDENCE INTERVALS FOR MULTIPLE COMPARISONS

Expressing a $1 - \alpha$ confidence interval using a sample mean denotes that there is a probability of $1 - \alpha$ that the interval encloses its respective population mean. Once multiple-comparison testing has concluded which of three or more sample

means are significantly different, confidence intervals may be calculated for each different population mean. If one sample mean ($\overline{X}_i$) is concluded to be significantly different from all others, then Equation 10.31 (introduced in Section 10.2) is used:

$$\overline{X}_i \pm t_{\alpha(2),\nu}\sqrt{\frac{s^2}{n_i}}. \tag{10.31}$$

In these calculations, s^2 (essentially a pooled variance) is the same as the error mean square would be for an analysis of variance for these groups of data. If two or more sample means are not concluded to be significantly different, then a pooled mean of those samples is the best estimate of the mean of the population from which those samples came:

$$\overline{X}_p = \frac{\sum n_i \overline{X}_i}{\sum n_i}, \tag{11.6}$$

where the summation is over all samples concluded to have come from the same population. Then the confidence interval is

$$\overline{X}_p \pm t_{\alpha(2),\nu}\sqrt{\frac{s^2}{\sum n_i}}, \tag{11.6a}$$

again summing over all samples whose means are concluded to be indistinguishable. This is analogous to the two-sample situation handled by Equation 8.16, and it is demonstrated in Example 11.3.

If a pair of population means, μ_B and μ_A, are concluded to be different, the $1 - \alpha$ confidence interval for the difference $(\mu_B - \mu_A)$ may be computed as

$$(\overline{X}_B - \overline{X}_A) \pm (q_{\alpha,\nu,k})(\text{SE}). \tag{11.7}$$

Here, as in Section 11.1, ν is the error degrees of freedom appropriate to an ANOVA, k is the total number of means, and SE is obtained from either Equation 11.2 or Equation 11.4, depending upon whether n_B and n_A are equal, or Equation 11.5 if the underlying population variances are not assumed to be equal. This calculation is demonstrated in Example 11.3 for the data in Example 11.1.

(a) Sample Size and Estimation of the Difference between Two Population Means. Section 8.3 showed how to estimate the sample size required to obtain a confidence interval of specified width for a difference between the two population means associated with the two-sample t test. In a multisample situation, a similar procedure may be used with the difference between population means, employing q instead of the t statistic. As in Section 8.3, iteration is necessary, whereby n is determined such that

$$n = \frac{s^2(q_{\alpha,\nu,k})^2}{d^2}. \tag{11.8}$$

Here, d is the half-width of the $1 - \alpha$ confidence interval, s^2 is the estimate of error variance, and k is the total number of means; ν is the error degrees of freedom with the estimated n, namely $\nu = k(n - 1)$.

EXAMPLE 11.3 Confidence Intervals (CI) for the Population Means from Example 11.1

It was concluded in Example 11.1 that $\mu_1 \neq \mu_2 = \mu_4 = \mu_3 \neq \mu_5$. Therefore, we may calculate confidence intervals for μ_1 for $\mu_{2,4,3}$ and for μ_5 (where $\mu_{2,4,3}$ indicates the mean of the common population from which Samples 2, 4, and 3 came).

Using Equation 10.31:

$$95\%\ \text{CI for } \mu_1 = \overline{X}_1 \pm t_{0.05(2),25}\sqrt{\frac{s^2}{n_1}} = 32.1 \pm (2.060)\sqrt{\frac{9.7652}{6}} = 32.1 \text{ mg/ml} \pm 2.6 \text{ mg/ml}.$$

Again using Equation 10.31:

$$95\%\ \text{CI for } \mu_5 = \overline{X}_5 \pm t_{0.05(2),25}\sqrt{\frac{s^2}{n_5}} = 58.3 \text{ mg/ml} \pm 2.6 \text{ mg/ml}.$$

Using Equation 11.6:

$$\overline{X}_p = \overline{X}_{2,4,3} = \frac{n_2\overline{X}_2 + n_4\overline{X}_4 + n_3\overline{X}_3}{n_2 + n_4 + n_3} = \frac{(6)(40.2) + (6)(41.1) + (6)(44.1)}{6 + 6 + 6} = 41.8 \text{ mg/ml}.$$

Using Equation 11.6a:

$$95\%\ \text{CI for } \mu_{2,4,3} = \overline{X}_{2,4,3} \pm t_{0.05(2),25}\sqrt{\frac{s^2}{6 + 6 + 6}} = 41.8 \text{ mg/ml} \pm 1.5 \text{ mg/ml}.$$

Using Equation 11.7:

$$95\%\ \text{CI for } \mu_5 - \mu_{2,4,3} = \overline{X}_5 - \overline{X}_{2,4,3} \pm q_{0.05,25,5}\sqrt{\frac{s^2}{2}\left(\frac{1}{n_5} + \frac{1}{n_2 + n_4 + n_3}\right)}$$
$$= 58.3 - 41.8 \pm (4.166)(1.04)$$
$$= 16.5 \text{ mg/ml} \pm 4.3 \text{ mg/ml}.$$

Using Equation 11.7:

$$95\%\ \text{CI for } \mu_{2,4,3} - \mu_1 = \overline{X}_{2,4,3} - \overline{X}_1 \pm q_{0.05,25,5}\sqrt{\frac{s^2}{2}\left(\frac{1}{n_2 + n_4 + n_3} + \frac{1}{n_1}\right)}$$
$$= 41.8 - 32.1 \pm (4.166)(1.04)$$
$$= 9.7 \text{ mg/ml} \pm 4.3 \text{ mg/ml}.$$

11.3 TESTING A CONTROL MEAN AGAINST EACH OTHER MEAN

Sometimes means are obtained from k groups with the a priori objective of concluding whether the mean of one group, commonly designated as a control, differs significantly from each of the means of the other $k - 1$ groups. Dunnett (1955) provided

an excellent procedure for such testing. Thus, whereas the data described in Section 11.1 were collected with the intent of comparing each sample mean with each other sample mean, the Dunnett test is for multisample data where the objective of the analysis was stated as comparing the control group's mean to the mean of each other group. Tukey's test could be used for this purpose, but it would be less powerful (Myers and Well, 2003: 255). If $k = 2$, Dunnett's test is equivalent to the two-sample t test (Section 8.1).

As in the previous section, s^2 denotes the error mean square, which is an estimate of the common population variance underlying each of the k samples. The Dunnett's test statistic (analogous to that of Equation 11.3) is

$$q' = \frac{\overline{X}_{\text{control}} - \overline{X}_A}{\text{SE}}, \tag{11.9}$$

where the standard error, when the sample sizes are equal, is

$$\text{SE} = \sqrt{\frac{2s^2}{n}}, \tag{11.10}$$

and when the sample sizes are not equal, it is

$$\text{SE} = \sqrt{s^2\left(\frac{1}{n_A} + \frac{1}{n_{\text{control}}}\right)}, \tag{11.11}$$

and when the variances are not equal:

$$\text{SE} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_{\text{control}}^2}{n_{\text{control}}}}. \tag{11.11a}$$

For a two-tailed test, critical values, $q'_{\alpha(2),\nu,k}$, are given in Appendix Table B.7. If $|q'| \geq q'_{\alpha(2),\nu,k}$, then H_0: $\mu_{\text{control}} = \mu_A$ is rejected. Critical values for a one-sample test, $q'_{\alpha(1),\nu,k}$, are given in Appendix Table B.6. In a one-tailed test, H_0: $\mu_{\text{control}} \leq \mu_A$ is rejected if $q' \geq q'_{\alpha(1),\nu,k}$; and H_0: $\mu_{\text{control}} \geq \mu_A$ is rejected if $|q'| \geq q'_{\alpha(1),\nu,k}$ and $\overline{X}_{\text{control}} < \mu_A$ (i.e., if $q \leq -q'_{\alpha(1),\nu,k}$). This is demonstrated in Example 11.4. These critical values ensure that the familywise Type I error = α.

The null hypothesis H_0: $\mu_{\text{control}} = \mu_A$ is a special case of H_0: $\mu_{\text{control}} - 0 = \mu_0$ where $\mu_0 = 0$. However, other values of μ_0 may be placed in the hypothesis, and Dunnett's test would proceed by placing $|\overline{X}_{\text{control}} - \overline{X}_A - \mu_0|$ in the numerator of the q' calculation. In an analogous manner, H_0: $\mu_{\text{control}} - \mu_0 \leq \mu$ (or H_0: $\mu_{\text{control}} - \mu_0 \geq \mu$) may be tested.

When comparison of group means to a control mean is the researcher's stated desire, the sample from the group designated as the control ought to contain more observations than the samples representing the other groups. Dunnett (1955) showed that the optimal size of the control sample typically should be a little less than $\sqrt{k-1}$ times the size of each other sample.

(a) Sample Size and Estimation of the Difference between One Population Mean and the Mean of a Control Population. This situation is similar to that discussed in Section 11.2a, but it pertains specifically to one of the k means being designated as

EXAMPLE 11.4 Dunnett's Test for Comparing the Mean of a Control Group to the Mean of Each Other Group

The yield (in metric tons per hectare) of each of several plots (24 plots, as explained below) of potatoes has been determined after a season's application of a standard fertilizer. Likewise, the potato yields from several plots (14 of them) were determined for each of four new fertilizers. A manufacturer wishes to promote at least one of these four fertilizers by claiming a resultant increase in crop yield. A total of 80 plots is available for use in this experiment.

Optimum allocation of plots among the five fertilizer groups will be such that the control group (let us say that it is group 2) has a little less than $\sqrt{k-1} = \sqrt{4} = 2$ times as many data as each of the other groups. Therefore, it was decided to use $n_2 = 24$ and $n_1 = n_3 = n_4 = m_5 = 14$, for a total N of 80.

Using analysis-of-variance calculations, the error MS (s^2) was found to be 10.42 (metric tons/ha)^2 and the error DF = 75.

$$\text{SE} = \sqrt{10.42\left(\frac{1}{14} + \frac{1}{24}\right)} = 1.1 \text{ metric tons/acre}$$

Group number (i) of ranked means:	1	2	3	4	5
Ranked group mean ($\overline{X}_i$):	17.3	21.7	22.1	23.6	27.8

As the control group (i.e., the group with the standard fertilizer) is group 2, each H_0: $\mu_2 \geq \mu_A$ will be tested against H_A: $\mu_2 < \mu_A$. And for each hypothesis test, $q'_{\alpha,\nu,k} = q'_{0.05(1),75,5}$.

Comparison B vs. A	*Difference* $(\overline{X}_2 - \overline{X}_A)$	SE	$\lvert q' \rvert$	*Conclusion*
2 vs. 1	$21.7 - 17.3 = 4.4$			Because $\overline{X}_2 > \overline{X}_1$, do not reject H_0: $\mu_2 \geq \mu_1$
2 vs. 5	$21.7 - 27.8 = -6.1$	1.1	5.55	Reject H_0: $\mu_2 \geq \mu_5$
2 vs. 4	$21.7 - 23.6 = -1.9$	1.1	1.73	Reject H_0: $\mu_2 \geq \mu_4$
2 vs. 3	Do not test			

We conclude that only fertilizer 5 produces a yield greater than the yield from the control fertilizer (fertilizer 2).

from a control group. The procedure uses this modification of Equation 11.8:

$$n = \frac{2s^2(q'_{\alpha,\nu,k})^2}{d^2}. \tag{11.12}$$

(b) Confidence Intervals for Differences between Control and Other Group Means. Using Dunnett's q' statistic and the SE of Equation 11.10, 11.11, or 11.11a, two-tailed confidence limits can be calculated for the difference between the control mean and each of the other group means:

$$1 - \alpha \text{ CI for } \mu_{\text{control}} - \mu_A = (\overline{X}_{\text{control}} - \overline{X}_A) \pm (q'_{\alpha(2),\nu,k})(\text{SE}). \tag{11.13}$$

One-tailed confidence limits are also possible. The $1 - \alpha$ confidence can be expressed that a difference, $\mu_{\text{control}} - \mu_A$, is not less than (i.e., is at least as large as)

$$(\overline{X}_{\text{control}} - \overline{X}_A) - (q'_{\alpha(1),\nu,k})(\text{SE}), \tag{11.14}$$

or it might be desired to state that the difference is no greater than

$$(\overline{X}_{\text{control}} - \overline{X}_A) + (q'_{\alpha(1),\nu,k})(\text{SE}). \tag{11.15}$$

MULTIPLE CONTRASTS

Inspecting the sample means after performing an analysis of variance can lead to a desire to compare combinations of samples to each other, by what are called *multiple contrasts*. The method of Scheffé* (1953; 1959: Sections 3.4, 3.5) is an excellent way to do this while ensuring a familywise Type I error rate no greater than α.

The data in Example 11.1 resulted in ANOVA rejection of the null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$; and, upon examining the five sample means, perhaps by arranging them in order of magnitude ($\overline{X}_1 < \overline{X}_2 < \overline{X}_4 < \overline{X}_3 < \overline{X}_5$), the researcher might then want to compare the mean strontium concentration in the river (group 5) with that of the bodies of water represented by groups 2, 4, and 3. The relevant null hypothesis would be H_0: $(\mu_2 + \mu_4 + \mu_3)/3 = \mu_5$, which can also be expressed as H_0: $\mu_2/3 + \mu_4/3 + \mu_3/3 - \mu_5 = 0$. The Scheffé test considers that each of the four μ's under consideration is associated with a coefficient, c_i: $c_2 = \frac{1}{3}$, $c_4 = \frac{1}{3}$, $c_3 = \frac{1}{3}$, and $c_5 = -1$ (and the sum of these coefficients is always zero). The test statistic, S, is calculated as

$$S = \frac{\left|\sum c_i \overline{X}_i\right|}{\text{SE}}, \tag{11.16}$$

where

$$\text{SE} = \sqrt{s^2\left(\sum \frac{c_i^2}{n_i}\right)}, \tag{11.17}$$

and the critical value of the test is

$$S_\alpha = \sqrt{(k-1)F_{\alpha(1),k-1,N-k}}\ . \tag{11.18}$$

Also, with these five groups of data, there might be an interest in testing H_0: $\mu_1 - (\mu_2 + \mu_4 + \mu_3)/3 = 0$ or H_0: $(\mu_1 + \mu_5)/2 - (\mu_2 + \mu_4 + \mu_3)/3 = 0$ or H_0: $(\mu_1 + \mu_4)/2 - (\mu_2 + \mu_3)/2 = 0$, or other contrasts. Any number of such hypotheses may be tested, and the familywise Type I error rate is the probability of falsely rejecting at least one of the possible hypotheses. A significant F from an ANOVA of the k groups indicates that there is at least one significant contrast among the groups, although the contrasts that are chosen may not include one to be rejected by the Scheffé test. And if F is not significant, testing multiple contrasts need not be done, for the probability of a Type I error in that testing is not necessarily at α (Hays, 1994: 458). The testing of several of these hypotheses is shown in Example 11.5. In employing the Scheffé test, the decision of which means to compare with which others occurs after inspecting the data, so this is referred to as an a posteriori, or post hoc test.

*Henry Scheffé (1907–1977), American statistician.

EXAMPLE 11.5 Scheffé's Test for Multiple Contrasts, Using the Data of Example 11.1

For $\alpha = 0.05$, the critical value, S_α, for each contrast is (via Equation 11.18) $\sqrt{(k-1)F_{0.05(1),k-1,N-k}}$

$$= \sqrt{(5-1)F_{0.05(1),4,25}}$$
$$= \sqrt{4(2.76)}$$
$$= 3.32.$$

Example 11.1 showed $s^2 = 9.7652$ and $n = 6$.

Contrast	SE	S	*Conclusion*
$\frac{\bar{X}_2 + \bar{X}_3 + \bar{X}_4}{3} - \bar{X}_5$ $= 41.8 - 58.3$ $= -16.5$	$\sqrt{9.7652\left[\frac{\left(\frac{1}{3}\right)^2}{6} + \frac{\left(\frac{1}{3}\right)^2}{6} + \frac{\left(\frac{1}{3}\right)^2}{6} + \frac{(1)^2}{6}\right]} = 1.47$	11.22	Reject H_0: $\frac{\mu_2 + \mu_3 + \mu_4}{3} - \mu_5 = 0$
$\bar{X}_1 - \frac{\bar{X}_2 + \bar{X}_3 + \bar{X}_4}{3}$ $= 32.1 - 41.8$ $= -9.7$	$\sqrt{9.7652\left[\frac{(1)^2}{6} + \frac{\left(\frac{1}{3}\right)^2}{6} + \frac{\left(\frac{1}{3}\right)^2}{6} + \frac{\left(\frac{1}{3}\right)^2}{6}\right]} = 1.47$	6.60	Reject H_0: $\mu_1 - \frac{\mu_2 + \mu_3 + \mu_4}{3} = 0$
$\frac{\bar{X}_1 + \bar{X}_5}{2} - \frac{\bar{X}_2 + \bar{X}_3 + \bar{X}_4}{3}$ $= 45.2 - 41.8$ $= 3.4$	$\sqrt{9.7652\left[\frac{\left(\frac{1}{2}\right)^2}{6} + \frac{\left(\frac{1}{2}\right)^2}{6} + \frac{\left(\frac{1}{3}\right)^2}{6} + \frac{\left(\frac{1}{3}\right)^2}{6} + \frac{\left(\frac{1}{3}\right)^2}{6}\right]} = 1.16$	2.93	Accept H_0: $\frac{\mu_1 + \mu_5}{2} - \frac{\mu_2 + \mu_3 + \mu_4}{3} = 0$
$\frac{\bar{X}_1 + \bar{X}_4}{2} - \frac{\bar{X}_2 + \bar{X}_3}{2}$ $= 36.6 - 42.15$ $= -5.55$	$\sqrt{9.7652\left[\frac{\left(\frac{1}{2}\right)^2}{6} + \frac{\left(\frac{1}{2}\right)^2}{6} + \frac{\left(\frac{1}{2}\right)^2}{6} + \frac{\left(\frac{1}{2}\right)^2}{6}\right]} = 1.28$	4.34	Reject H_0: $\frac{\mu_1 + \mu_4}{2} - \frac{\mu_2 + \mu_3}{2} = 0$

The Scheffé test may also be used to compare one mean with one other. It is then testing the same hypotheses as is the Tukey test. It is less sensitive than the Tukey test to nonnormality and heterogeneity of variances (Hays, 1994: 458, 458; Sahai and Ageel, 2000: 77); but is less powerful and it is recommended that it not be used for pairwise comparisons (e.g., Carmer and Swanson, 1973; Kirk, 1995: 154; Toothaker, 1991: 51, 77, 89–90). Shaffer (1977) described a procedure, more powerful than Scheffé's, specifically for comparing a combination of groups to a group specified as a control.

(a) Multiple Contrasts with Unequal Variances. The Scheffé test is suitable when the samples in the contrast each came from populations with the same variance.

When the population variances differ but the sample sizes are equal, the probability of a Type I error can be different from 0.05 when α is set at 0.05. If the variances *and* the sample sizes are unequal, then (as with the ANOVA, Section 10.1g) the test will be very conservative if the large variances are associated with large sample sizes and very liberal if the small samples come from the populations with the large variances (Keselman and Toothaker, 1974). If the σ_i^2's cannot be assumed to be the same or similar, the procedure of Brown and Forsythe (1974b) may be employed. This is done in a fashion analogous to the two-sample Welch modification of the t test (Section 8.1c), using

$$t' = \frac{\left|\sum c_i \overline{X}_i\right|}{\sqrt{\sum \frac{c_i^2 s_i^2}{n_i}}} \tag{11.19}$$

with degrees of freedom of

$$\nu' = \frac{\left(\sum \frac{c_i^2 s_i^2}{n_i}\right)^2}{\sum \frac{(c_i^2 s_i^2)^2}{n_i^2 (n_i - 1)}}. \tag{11.20}$$

(b) Confidence Intervals for Contrasts. The Scheffé procedure enables the establishment of $1 - \alpha$ confidence limits for a contrast:

$$\sum c_i \overline{X}_i \pm S_\alpha \text{SE} \tag{11.21}$$

(with SE from Equation 11.17). Shaffer's (1977) method produces confidence intervals for a different kind of contrast, that of a group of means with the mean of a control group.

Example 11.6 demonstrates the determination of confidence intervals for two of the statistically significant contrasts of Example 11.5.

11.5 NONPARAMETRIC MULTIPLE COMPARISONS

In the multisample situation where the nonparametric Kruskal-Wallis test (Section 10.4) is appropriate, the researcher usually will desire to conclude which of the samples are significantly different from which others, and the experiment will be run with that goal. This may be done in a fashion paralleling the Tukey test of Section 11.1, by using rank sums instead of means, as demonstrated in Example 11.7. The rank sums, determined as in the Kruskal-Wallis test, are arranged in increasing order of magnitude. Pairwise differences between rank sums are then tabulated, starting with the difference between the largest and smallest rank sums, and proceeding in the same sequence as described in Section 11.1. The standard error is calculated as

$$\text{SE} = \sqrt{\frac{n(nk)(nk + 1)}{12}} \tag{11.22}$$

(Nemenyi, 1963; Wilcoxon and Wilcox, 1964: 10),* and the Studentized range (Appendix Table B.5 to be used is $q_{\alpha,\infty,k}$.)

EXAMPLE 11.6 Confidence Intervals for Multiple Contrasts

The critical value, S_α, for each confidence interval is that of Equation 11.8: $\sqrt{(k-1)F_{\alpha(1),k-1,N-k}}$, and for $\alpha = 0.05$, $S_\alpha = 3.32$ and $s^2 = 9.7652$ as in Example 11.5.

(a) A confidence interval for $\frac{\mu_2 + \mu_3 + \mu_4}{3} - \mu_5$ would employ SE = 1.47 from Example 11.5, and the 95% confidence interval is

$$\left(\frac{\overline{X}_2 + \overline{X}_3 + \overline{X}_4}{3} - \overline{X}_5\right) \pm S_\alpha \text{SE} = -16.5 \pm (3.32)(1.47)$$

$$= -16.5 \text{ mg/ml} \pm 4.9 \text{ mg/ml}$$

$$L_1 = -21.4 \text{ mg/ml}$$

$$L_2 = -11.6 \text{ mg/ml}.$$

(b) A confidence interval for $\mu_1 - \frac{\mu_2 + \mu_3 + \mu_4}{3}$ would employ SE = 1.47 from Example 11.5, and the 95% confidence interval is

$$\left(\overline{X}_1 - \frac{\overline{X}_2 + \overline{X}_3 + \overline{X}_4}{3}\right) \pm S_\alpha \text{SE} = -9.7 \pm (3.32)(1.47)$$

$$= -9.7 \text{ mg/ml} \pm 4.9 \text{ mg/ml}$$

$$L_1 = -14.6 \text{ mg/ml}$$

$$L_2 = -4.8 \text{ mg/ml}.$$

(a) Nonparametric Multiple Comparisons with Unequal Sample Sizes. Multiple-comparison testing such as in Example 11.7 requires that there be equal numbers of data in each of the k groups. If such is not the case, then we may use the procedure of Section 11.7, but a more powerful test is that proposed by Dunn (1964), using a standard error of

$$\text{SE} = \sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} \tag{11.24}$$

for a test statistic we shall call

$$Q = \frac{\overline{R}_B - \overline{R}_A}{\text{SE}}, \tag{11.25}$$

*Some authors (e.g., Miller 1981: 166) perform this test in an equivalent fashion by considering the difference between mean ranks ($\overline{R}_A$ and $\overline{R}_B$) rather than rank sums (R_A and R_B), in which case the appropriate standard error would be

$$\text{SE} = \sqrt{\frac{k(nk+1)}{12}}. \tag{11.23}$$

EXAMPLE 11.7 Nonparametric Tukey-Type Multiple Comparisons, Using the Nemenyi Test

The data are those from Example 10.10.

$$SE = \sqrt{\frac{n(nk)(nk + 1)}{12}} = \sqrt{\frac{5(15)(16)}{12}} = \sqrt{100} = 10.00$$

Sample number (i) of ranked rank sums:	3	2	1
Rank sum (R_i):	26	30	64

Comparison (B vs. A)	*Difference* ($R_B - R_A$)	SE	q	$q_{0.05,\infty,3}$	*Conclusion*
1 vs. 3	64 − 26 = 38	10.00	3.80	3.314	Reject H_0: Fly abundance is the same at vegetation heights 3 and 1.
1 vs. 2	64 − 30 = 34	10.00	3.40	3.314	Reject H_0: Fly abundance is the same at vegetation heights 2 and 1.
2 vs. 3	30 − 26 = 4	10.00	0.40	3.314	Do not reject H_0: Fly abundance is the same at vegetation heights 3 and 2.

Overall conclusion: Fly abundance is the same at vegetation heights 3 and 2 but is different at height 1.

where $\overline{R}$ indicates a mean rank (i.e., $\overline{R}_A = R_A/n_A$ and $\overline{R}_B = R_B/n_B$). Critical valucs for this test, $Q_{\alpha,k}$, are given in Appendix Table B.15. Applying this procedure to the situation of Example 11.7 yields the same conclusions, but this will not always be the case as this is only an approximate method and conclusions based upon a test statistic very near the critical value should be expressed with reservation. It is advisable to conduct studies that have equal sample sizes so Equation 11.22 or 11.23 may be employed.

If tied ranks are present, then the following is an improvement over Equation 11.24 (Dunn, 1964):

$$SE = \sqrt{\left(\frac{N(N + 1)}{12} - \frac{\sum t}{12(N - 1)}\right)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}. \qquad (11.26)$$

In the latter equation, $\sum t$ is used in the Kruskal-Wallis test when ties are present and is defined in Equation 10.42. The testing procedure is demonstrated in Example 11.8; note that it is the mean ranks ($\overline{R}_i$), rather than the ranks sums (R_i), that are arranged in order of magnitude.

A procedure developed independently by Steel (1960, 1961b) and Dwass (1960) is somewhat more advantageous than the tests of Nemenyi and Dunn (Critchlow and Fligner, 1991; Miller, 1981: 168–169), but it is less convenient to use and it tends to be very conservative and less powerful (Gabriel and Lachenbruch, 1969). And

EXAMPLE 11.8 Nonparametric Multiple Comparisons with Unequal Sample Sizes

The data are those from Example 10.11, where the Kruskal-Wallis test rejected the null hypothesis That water *p*H was the same in all four ponds examined

$\sum t = 168$, as in Example 10.11.

For $n_A = 8$ and $n_B = 8$,

$$\text{SE} = \sqrt{\left(\frac{N(N+1)}{12} - \frac{\sum t}{12(N-1)}\right)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

$$= \sqrt{\left(\frac{31(32)}{12} - \frac{168}{12(30)}\right)\left(\frac{1}{8} + \frac{1}{8}\right)}$$

$$= \sqrt{20.5500} = 4.53$$

For $n_A = 7$ and $n_B = 8$,

$$\text{SE} = \sqrt{\left(\frac{31(32)}{12} - \frac{168}{12(30)}\right)\left(\frac{1}{7} + \frac{1}{8}\right)} = \sqrt{22.0179} = 4.69.$$

Sample number (i) of ranked means:	1	2	4	3
Rank sum (R_i):	63.24	64.62	71.30	73.35
Sample size (n_i):	8	5	8	7
Mean rank ($\overline{R}_i$)	6.88	16.56	20.44	20.71

To test at the 0.05 significance level, the critical value is $Q_{0.05,4} = 2.639$.

Comparison B vs. A	*Difference* $(\overline{R}_B - \overline{R}_A)$	SE	Q	*Conclusion*
3 vs. 1	20.71 − 6.88 = 13.831	4.69	2.95	Reject H_0: Water *p*H is the same in ponds 3 and 1.
3 vs. 2	20.71 − 16.56 = 4.15	4.69	0.88	Do not reject H_0: Water *p*H is the same in ponds 3 and 2.
3 vs. 4	Do not test			
4 vs. 1	20.44 − 6.88 = 13.56	4.53	2.99	Reject H_0: Water *p*H is the same in ponds 4 and 1.
4 vs. 2	Do not test			
2 vs. 1	16.56 − 6.88 = 9.68	4.53	2.14	Do not reject H_0: Water *p*H is the same in ponds 2 and 1.

Overall conclusion: Water *p*H is the same in ponds 4 and 3 but is different in pond 1, and the relationship of pond 2 to the others is unclear.

this test can lose control of Type I error if the data come from skewed populations (Toothaker, 1991: 108).

(b) Nonparametric Comparisons of a Control to Other Groups. Subsequent to a Kruskal-Wallis test in which H_0 is rejected, a nonparametric analysis may be performed to seek either one-tailed or two-tailed significant differences between one group (designated as the "control") and each of the other groups of data. This is done in a manner paralleling that of the procedure of Section 11.4, but using group rank sums instead of group means. The standard error to be calculated is

$$\text{SE} = \sqrt{\frac{n(nk)(nk + 1)}{6}} \tag{11.27}$$

(Wilcoxon and Wilcox, 1964: 11), and one uses as critical values either $q'_{\alpha(1),\infty,k}$ or $q'_{\alpha(2),\infty,k}$ (from Appendix Table B.6 or Appendix Table B.7, respectively) for one-tailed or two-tailed hypotheses, respectively.*

The preceding nonparametric test requires equal sample sizes. If the n's are not all equal, then the procedure suggested by Dunn (1964) may be employed. By this method, group B is considered to be the control and uses Equation 11.27, where the appropriate standard error is that of Equation 11.26 or 11.28, depending on whether there are ties or no ties, respectively. We shall refer to critical values for this test, which may be two tailed or one tailed, as $Q'_{\alpha,k}$; and they are given in Appendix Table B.16. The test presented by Steel (1959) has drawbacks compared to the procedures above (Miller, 1981: 133).

11.6 NONPARAMETRIC MULTIPLE CONTRASTS

Multiple contrasts, introduced in Section 11.4, can be tested nonparametrically using the Kruskal-Wallis H statistic instead of the F statistic. As an analog of Equation 11.16, we compute

$$S = \frac{\left|\sum c_i \overline{R}_i\right|}{\text{SE}}, \tag{11.29}$$

where c_i is as in Section 11.4, and

$$\text{SE} = \sqrt{\left(\frac{N(N + 1)}{12}\right)\left(\sum \frac{c_i^2}{n_i}\right)}, \tag{11.30}$$

unless there are tied ranks, in which cases we use

$$\text{SE} = \sqrt{\left(\frac{N(N + 1)}{12} - \frac{\sum t}{12(N - 1)}\right)\left(\sum \frac{c_i^2}{n_i}\right)}, \tag{11.31}$$

*If mean ranks, instead of rank sums, are used, then

$$\text{SE} = \sqrt{\frac{k(nk + 1)}{6}}. \tag{11.28}$$

where $\sum t$ is as in Equation 10.42. The critical value for these multiple contrasts is $\sqrt{H_{\alpha,n_1,n_2,\ldots}}$, using Appendix Table B.13 to obtain the critical value of H. If the needed critical value of H is not on that table, then $\chi^2_{\alpha,(k-1)}$ may be used.

11.7 MULTIPLE COMPARISONS AMONG MEDIANS

If the null hypothesis is rejected in a multisample median test (Section 10.5), then it is usually desirable to ascertain among which groups significant differences exist. A Tukey-type multiple comparison test has been provided by Levy (1979), using

$$q = \frac{f_{1B} - f_{1A}}{\text{SE}}. \tag{11.32}$$

As shown in Example 11.19, we employ the values of f_{1j} for each group, where f_{1j} is the number of data in group j that are greater than the grand median. (The values of f_{1j} are the observed frequencies in the first row in the contingency table used in the multisample median test of Section 10.5.) The values of f_{1j} are ranked, and pairwise differences among the ranks are examined as in other Tukey-type tests. The appropriate standard error, when N (the total number of data in all groups) is an even number, is

$$\text{SE} = \sqrt{\frac{n(N+1)}{4N}}, \tag{11.33}$$

and, when N is an odd number, the standard error is

$$\text{SE} = \sqrt{\frac{nN}{4(N-1)}}. \tag{11.34}$$

The critical values to be used are $q_{\alpha,\infty,k}$. This multiple-comparison test appears to possess low statistical power. If the sample sizes are slightly unequal, as in Example 11.9, the test can be used by employing the harmonic mean (see Section 3.4b) of the sample sizes,

$$n = \frac{k}{\sum_{j=1}^{k} \frac{1}{n_j}}, \tag{11.35}$$

for an approximate result.

11.8 MULTIPLE COMPARISONS AMONG VARIANCES

If the null hypothesis that k population variances are all equal (see Section 10.6) is rejected, then we may wish to determine which of the variances differ from which others. Levy (1975a, 1975c) suggests multiple-comparison procedures for this purpose based on a logarithmic transformation of sample variances.

A test analogous to the Tukey test of Section 11.1 is performed by calculating

$$q = \frac{\ln s_B^2 - \ln s_A^2}{\text{SE}}. \tag{11.36}$$

EXAMPLE 11.9 Tukey-Type Multiple Comparison for Differences among Medians, Using the Data of Example 10.12

Sample number (j) of samples ranked by f_{1j}:	2	1	4	3
Ranked f_{1j}:	2	3	6	9
Sample size (n_j):	12	11	12	11

$k = 4$

$N = 11 + 12 + 11 + 12 = 46$

$n = 12$

By Equation 11.35,

$$n = \frac{4}{\frac{1}{12} + \frac{1}{11} + \frac{1}{12} + \frac{1}{11}} = 11.48$$

By Equation 11.36,

$$\text{SE} = \sqrt{\frac{(11.48)(46)}{4(46 - 1)}} = 1.713.$$

H_0: Median of population B = Median of population A.

H_A: Median of population $B \neq$ Median of population A.

Comparison	$f_{1B} - f_{1A}$	SE	q	$q_{0.05,4,\infty}$	*Conclusion*
3 vs. 2	$9 - 2 = 7$	1.713	4.086	3.633	Reject H_0.
3 vs. 1	$9 - 3 = 6$	1.713	3.503	3.633	Do not reject H_0.
3 vs. 4	Do not test				
4 vs. 2	$6 - 2 = 4$	1.713	2.335	3.633	Do not reject H_0.
4 vs. 1	Do not test				
1 vs. 2	Do not test				

Overall conclusion: The medians of populations 3 and 2 (i.e., south and east—see Example 10.12) are not the same; but the test lacks the power to allow clear conclusions about the medians of populations 4 and 1.

where

$$\text{SE} = \sqrt{\frac{2}{\nu}}, \tag{11.37}$$

if both samples being compared are of equal size. If $\nu_A \neq \nu_B$, we can employ

$$\text{SE} = \sqrt{\frac{1}{\nu_B} + \frac{1}{\nu_A}}. \tag{11.38}$$

EXAMPLE 11.10 Tukey-Type Multiple Comparison Test for Differences among Four Variances (i.e., $k = 4$)

i	s_i^2	n_i	ν_i	$\ln s_i^2$
1	2.74 g^2	50	49	1.0080
2	2.83 g^2	48	47	1.0403
3	2.20 g^2	50	49	0.7885
4	6.42 g^2	50	49	1.8594

Sample ranked by variances (i):	3	1	2	4
Logorithm of ranked sample variance ($\ln s_i^2$):	0.7885	1.0080	1.0403	1.8594
Sample degrees of freedom (ν_i):	49	49	47	49

Comparison (B vs. A)	*Difference* ($\ln s_B^2 - \ln s_A^2$)	SE	q	$q_{0.05,\infty,4}$	*Conclusions*
4 vs. 3	1.8594 − 0.7885 = 1.0709	0.202*	5.301	3.633	Reject H_0: $\sigma_4^2 = \sigma_3^2$
4 vs. 1	1.8594 − 1.0080 = 0.8514	0.202	4.215	3.633	Reject H_0: $\sigma_4^2 = \sigma_1^2$
4 vs. 2	1.8594 − 1.0403 = 0.8191	0.204†	4.015	3.633	Reject H_0: $\sigma_4^2 = \sigma_2^2$
2 vs. 3	1.0403 − 0.7885 = 0.2518	0.204	1.234	3.633	Do not reject H_0: $\sigma_2^2 = \sigma_3^2$
2 vs. 1	Do not test				
1 vs. 3	Do not test				

*As $\nu_4 = \nu_3$: $\text{SE} = \sqrt{\frac{2}{\nu}} = \sqrt{\frac{2}{49}} = 0.202.$

†As $\nu_4 \neq \nu_2$: $\text{SE} = \sqrt{\frac{1}{\nu_4} + \frac{1}{\nu_2}} = \sqrt{\frac{1}{49} + \frac{1}{47}} = 0.204.$

Overall conclusion: $\sigma_3^2 = \sigma_1^2 = \sigma_2^2 \neq \sigma_4^2$.

Just as in Sections 11.1 and 11.2, the subscripts A and B refer to the pair of groups being compared; and the sequence of pairwise comparisons must follow that given in those sections. This is demonstrated in Example 11.10.* The critical value for this test is $q_{\alpha,\infty,k}$ (from Appendix Table B.5).

A Newman-Keuls-type test can also be performed using the logarithmic transformation. For this test, we calculate q using Equation 11.36; but the critical value,

*Recall (as in Section 10.6) that "ln" refers to natural logarithms (i.e., logarithms using base e). If one prefers using common logarithms ("log"; logarithms in base 10), then

$$q = \frac{2.30259(\log s_B^2 - \log s_A^2)}{\text{SE}}. \tag{11.39}$$

4 are the same as the means of groups 2 and 3.

(b) Test the hypothesis that the means of groups 2 and 4 are the same as the mean of group 3.

11.6. The following ranks result in a significant Kruskal-Wallis test. Employ nonparametric multiple-range testing to conclude between which of the three groups population differences exist.

Group 1	*Group 2*	*Group 3*
8	10	14
4	6	13
3	9	7
5	11	12
1	2	15