Give Your Students Free Access to the Wiley E-Text for 14 Days and the Choice to Purchase a Format They Prefer

Immediate Access: The free 14-day trial of the Wiley E-Text ensures students have the right materials, right away.

Choice: Students have the option to either rent or purchase the Wiley E-Text, or purchase a print book, all at affordable prices.

A Better User Experience: With the Wiley E-Text capabilities, students can search content, highlight, take and share notes, and study anytime, anywhere, and on the device of their choice.

Help your students access the course materials in the format that best suits their learning style. Send them to **www.wileystudentchoice.com** for a free trial and exceptionally affordable prices. Or contact your Wiley Account Manager to learn more.

www.WileyStudentChoice.com

Cover Design: Wiley
Cover Images: © majcot/Shutterstock;
© Christos Georghiou/Shutterstock

www.wiley.com/go/Daniel/Biostatistics11e

WILEY

This ISBN is for Instructor evaluation only.
For ISBNs for student or
bookstore purchases, go to
www.wileystudentchoice.com

ISBN 978-1-119-49668-7
90000
9 781119 496687

DANIEL • CROSS

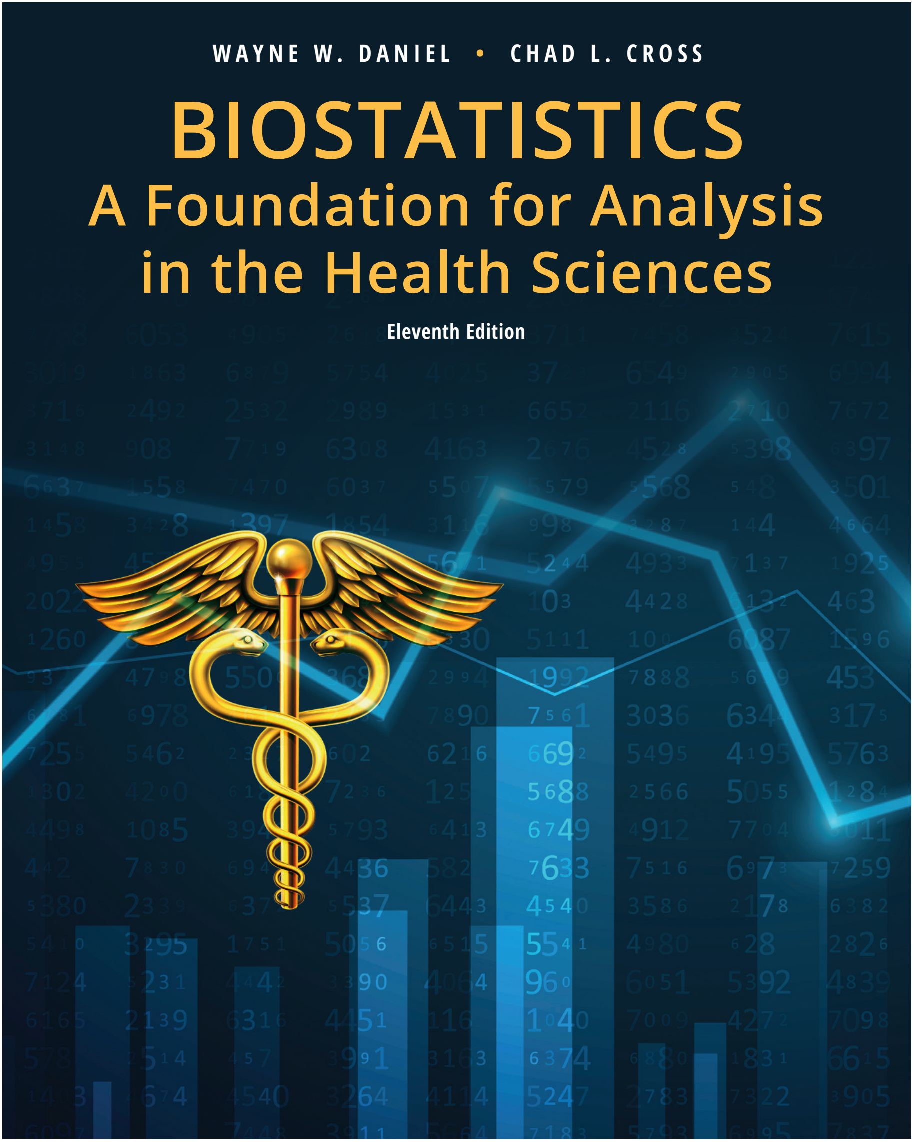
Instructor's Copy

BIOSTATISTICS

Eleventh
Edition

WILEY

Instructor's Copy



WAYNE W. DANIEL • CHAD L. CROSS

BIOSTATISTICS

A Foundation for Analysis in the Health Sciences

Eleventh Edition

Give your students free access to this title for 14 days.

See back for additional information

WILEY

Biostatistics

ELEVENTH
EDITION

A FOUNDATION FOR ANALYSIS
IN THE HEALTH SCIENCES

BIOSTATISTICS

ELEVENTH EDITION

A FOUNDATION FOR ANALYSIS IN THE HEALTH SCIENCES

Wayne W. Daniel, Ph.D.
Professor Emeritus
Georgia State University

Chad L. Cross, Ph.D., PStat®
Biostatistician
Las Vegas, Nevada

WILEY

VP AND DIRECTOR	Laurie Rosatone
EDITORIAL DIRECTOR	Don Fowley
EDITOR	Jennifer Brady
EDITORIAL MANAGER	Judy Howarth
CONTENT MANAGEMENT DIRECTOR	Lisa Wojcik
CONTENT MANAGER	Nichole Urban
SENIOR CONTENT SPECIALIST	Nicole Repasky
PRODUCTION EDITOR	Rajeshkumar Nallusamy
COVER PHOTO CREDIT	© majcot/Shutterstock; Christos Georghiou/Shutterstock

This book was set in 10/12 TimesLTStd by SPi Global and printed and bound by Lightning Source Inc.

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

Evaluation copies are provided to qualified academics and professionals for review purposes only, for use in their courses during the next academic year. These copies are licensed and may not be sold or transferred to a third party. Upon completion of the review period, please return the evaluation copy to Wiley. Return instructions and a free of charge return shipping label are available at: www.wiley.com/go/returnlabel. If you have chosen to adopt this textbook for use in your course, please accept this book as your complimentary desk copy. Outside of the United States, please contact your local sales representative.

ISBN: 978-1-119-28237-2 (PBK)
ISBN: 978-1-119-49668-7 (EVALC)

Library of Congress Cataloging-in-Publication Data

Names: Daniel, Wayne W., 1929– author. | Cross, Chad Lee, 1971– author.
Title: Biostatistics : a foundation for analysis in the health sciences /
Wayne W. Daniel, Chad L. Cross.
Description: Eleventh edition. | Hoboken, NJ : Wiley, [2019] | Includes
bibliographical references and index. |
Identifiers: LCCN 2018038950 (print) | LCCN 2018039731 (ebook) | ISBN
9781119496663 (Adobe PDF) | ISBN 9781119496571 (ePub) | ISBN 9781119282372
(pbk.) | ISBN 9781119496687 (evalc)
Subjects: | MESH: Biostatistics | Biometry
Classification: LCC RA409 (ebook) | LCC RA409 (print) | NLM WA 950 | DDC
610.72/7—dc23
LC record available at https://lccn.loc.gov/2018038950

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

Dr. Daniel

To my children, Jean, Carolyn, and John,
and to the memory of their mother, my wife, Mary.

Dr. Cross

To my wife Pamela and to my children,
Annabella Grace and Breanna Faith.
and
To Dr. Wayne Daniel, a trusted friend and colleague,
who has dedicated his life to providing
the best texts for statistics education.

Preface

The 11th edition of *Biostatistics: A Foundation for the Analysis in the Health Sciences* was prepared to meet the needs of students who may be using the book as a text in a course, and for professionals who may need a handy desk reference for basic, but widely used, statistical procedures in their applied work. For undergraduates, several chapters in this edition introduce concepts to students who are taking a first, generally junior-level or senior-level, course in statistics as part of their pre professional, nursing, or public health education. For beginning graduate students, both introductory chapters and more advanced topics in the text are suitable for master's students in health professions.

The breadth of coverage in the text is much more than may be generally covered in a one-semester course. This coverage, along with hundreds of practical and specific subject-matter exercises, allows instructors extensive flexibility in designing a course at various levels. We have developed some ideas on appropriate topical coverage based on our own use of this text in the classroom, and we present a matrix below in that regard.

As with previous editions of this book, the 11th edition requires little mathematical knowledge beyond college algebra. However, as many instructors will attest, it is not uncommon for students to lack solid proficiency in algebra prior to taking a statistics course. Our experience suggests that spending some time showing basic, algebraic manipulations of the formulas in the book goes a long way in quelling fears with mathematics that may easily undermine a statistics course. We have attempted to maintain an emphasis on practical and intuitive understanding of principles rather than on abstract concepts, and we therefore maintain a reliance on problem-solving utilizing examples and practice problems that are drawn largely from the health sciences literature instead of contrived problems, which makes the text more practical and less abstract. We believe that this makes the text more interesting for students, and more useful for health professionals who reference the text while performing their work duties.

There is no doubt that technological sophistication has changed how we teach and how we apply statistics professionally. The use of hand calculations can be a useful way to develop an understanding of how formulas work, and they also lead to an appreciation of underlying assumptions that need to be considered. However, once basic skills are learned, it is often useful to explore computer programs for dealing with large and/or real-world problem sets. Additionally, the reliance on statistical tables, once necessary for finding areas under curves, estimates of probability, and so on, has largely been replaced by efficient computer algorithms readily available to students and practitioners. To that end, you will find example outputs from MINITAB, SAS, SPSS, R, JASP, EXCEL, and others in the text. We do not endorse the use of any particular program, but simply note that many are available and both students and professionals will need to have some facility using the program of their choice. Additionally, we generally only provide outputs and explanation regarding programs, not instruction on their use, as there are many books dedicated to providing stepwise user guides for various programs.

Changes and Updates to This Edition

Many changes and updates have been made to this edition. We have attempted to incorporate corrections and clarifications that enhance the material presented in hopes of making the text more readable and accessible to the audience. We thank the reviewers of the many editions of this text for making useful comments and suggestions that have found their way into the new edition. Of course, there are always ways to improve and enhance, and we welcome comments and suggestions.

Specific changes to this edition include: (1) a newly rewritten introduction to the scientific method in Chapter 1; (2) a rearranged and rewritten Chapter 2 that now includes a section on data visualization and graphing; (3) an introduction to hypothesis testing and controversies surrounding p values in Chapter 7; (4) a brief introduction to Poisson regression in Chapter 11; (5) testing or dependent proportions using McNemar's Test in Chapter 12; and (6) the use of randomization procedures, including permutation-based p values and bootstrap confidence intervals, has been integrated throughout the text.

Other changes have occurred as well. Numerous changes to writing and phrasing have occurred to enhance clarity throughout the text. Also, by popular demand, we have integrated some R scripting ideas throughout many chapters for those using that particular software. Finally, for the benefit of instructors, we have provided some "Instructor-only" problems that will be made available to adopters of the text to use in their courses. Finally, the statistical tables are readily available through your instructor. Inasmuch as some professionals and professors still use tables, we believe it is important to retain access to them, and we continue to provide examples of their use in the current edition; however, we also show alternatives to tabled probabilities using computer programs.

Coverage Ideas

In the table below, we provide some suggestions for topical coverage in a variety of contexts, with "*X*" indicating those chapters we believe are most relevant for a variety of courses for which we believe this text is appropriate. As mentioned above, the text is designed to be flexible in order to accommodate various teaching styles and course presentations. Although the text is designed with progressive presentation of concepts in mind, certain topics may be skipped or briefly reviewed so that instructors may focus on concepts most useful for their courses.

	Chapters (X: Suggested coverage; O: Optional coverage)														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Undergraduate course for health sciences students	X	X	X	X	X	X	X	X	X	O	O	X	O	O	O
Graduate course for beginning health sciences master's students	X	X	X	X	X	X	X	X	X	X	O	X	X	X	O
Graduate course for graduate health sciences students who have completed an introductory statistics course	X	O	O	O	O	X	X	X	X	X	X	X	X	X	X

Supplements

Several supplements are available for the text on the instructor's website at www.wiley.com/go/Daniel/Biostatistics11e. These include:

- ***Instructor's Solution Manual***, available only to instructors who have adopted the text.

- **Data Sets**, over 200 data sets are available to be downloaded in CSV format for ready importing into any basic statistics program.

Acknowledgments

Many reviewers, students, and faculty have made contributions to this text through their careful review, inquisitive questions, and professional discussion of topics. In particular, we would like to thank:

- Dr. Sheniz Moonie, University of Nevada, Las Vegas
- Dr. Guogen Shan, University of Nevada, Las Vegas
- Dr. Gian Jhangri, University of Alberta
- Dr. Tina Cunningham, Eastern Virginal Medical School
- Dr. Shakhawat Hossain, University of Winnipeg
- Dr. Milind Phadnis, University of Kansas Medical Center
- Dr. David Anderson, Xavier University of Louisiana
- Dr. Derek Webb, Bemidji State University
- Dr. Keiji Oda, Loma Linda University
- Dr. David Zietler, Grand Valley State University
- Dr. Genady Grabarnik, St. John's University
- Dr. Al Bartolucci, University of Alabama at Birmingham
- Dr. Hwanseok Choi, University of Southern Mississippi
- Dr. Mark Kelley, University of Pittsburgh at Bradford
- Dr. Wan Tang, Tulane University
- Dr. Phil Gona, University of Massachusetts, Boston
- Dr. Jill Smith, University of California, Riverside
- Dr. Ronnie Brown, University of Baltimore
- Dr. Apoorv Goel, Indiana University-Purdue University Indianapolis
- Dr. Daniel Yorgov, Indiana University-Purdue University Fort Wayne

There are three additional acknowledgments that must be made to important contributors of the text. Dr. John. P. Holcomb of Cleveland State University updated many of the examples and exercises found in the text. Dr. Edward Danial of Morgan State University provided an extensive accuracy review of the 9th edition of the text, and his valuable comments added greatly to the later editions of the book. Dr. Jodi B. A. McKibben of the Uniformed Services University of the Health Sciences provided an extensive accuracy review of the 10th edition of the book, and we remain grateful for her contributions and comments.

We wish to acknowledge the cooperation of Minitab, Inc. for making available to the authors over many years and editions of the book the latest versions of their software.

Thanks are due to Professors Geoffrey Churchill and Brian Schott of Georgia State University, who wrote computer programs for generating some of the Appendix tables, and to Professor Lillian Lin, who read and commented on the logistic regression material in earlier editions of the book. Additionally, Dr. James T. Wassell provided useful assistance with some of the survival analysis methods presented in earlier editions of the text.

We are grateful to the many researchers in the health sciences field who publish their results and hence make available data that provide valuable practice to the students of biostatistics.

Final Note

I am eternally grateful that I have had the opportunity to work with Dr. Wayne Daniel on several editions of this text. I was invited by Wayne to work with him in various capacities beginning with the 8th edition. Since that time, I have had the pleasure to get to know Wayne and to appreciate his high standards and expectations. Unfortunately, Wayne was not able to participate in this edition. I am honored that he has entrusted me to carry forward his legacy.

Chad L. Cross
Las Vegas, Nevada

Brief Contents

The following supplements are available through your instructor

CONTENTS

4 Probability Distributions 80

5 Some Important Sampling Distributions 119

6 Estimation 143

7 Hypothesis Testing 189

The following supplements are available through your instructor

Introduction to Biostatistics

1

CHAPTER OVERVIEW

This chapter is intended to provide an overview of the basic statistical concepts and definitions used throughout the textbook. A course in statistics requires the student to learn new and specific terminology. Therefore, this chapter lays the foundation necessary for understanding basic statistical terms and concepts and the role that statisticians play in promoting scientific discovery.

TOPICS

LEARNING OUTCOMES

After studying this chapter, the student will

1. understand the basic concepts and terminology of biostatistics, including types of variables, measurement, and measurement scales.
2. be able to select a simple random sample and other scientific samples from a population of subjects.
3. understand the processes involved in the scientific method.
4. appreciate the advantages of using computers in the statistical analysis of data generated by studies and experiments conducted by researchers in the health sciences.

1.1 Introduction

We are frequently reminded of the fact that we are living in the information age. Appropriately, then, this book is about information—how it is obtained, how it is analyzed, and how it is interpreted. The information about which we are concerned we call data, and the data are available to us in the form of numbers or in other non numerical forms that can be analyzed.

The objectives of this book are twofold: (1) to teach the student to organize and summarize data and (2) to teach the student how to reach decisions about a large body of data by examining only a small part of it. The concepts and methods necessary for achieving the first objective are presented under the heading of *descriptive statistics*, and the second objective is reached through the study of what is called *inferential statistics*. This chapter discusses descriptive statistics. Chapters 2 through 5 discuss topics that form the foundation of statistical inference, and most of the remainder of the book deals with inferential statistics.

Because this volume is designed for persons preparing for or already pursuing a career in the health field, the illustrative material and exercises reflect the problems and activities that these persons are likely to encounter in the performance of their duties.

1.2 Basic Concepts and Definitions

Like all fields of learning, statistics has its own vocabulary. Some of the words and phrases encountered in the study of statistics will be new to those not previously exposed to the subject. Other terms, though appearing to be familiar, may have specialized meanings that are different from the meanings that we are accustomed to associating with these terms. The following are some common terms that we will use extensively in this book; others will be added as we progress through the material.

Data

The raw material of statistics is *data*. For our purposes, we may define data as *numbers*. The two kinds of numbers that we use in statistics are numbers that result from the taking—in the usual sense of the term—of a *measurement*, and those that result from the process of *counting*. For example, when a nurse weighs a patient or takes a patient's temperature, a measurement, consisting of a number such as 150 pounds or 100 degrees Fahrenheit, is obtained. Quite a different type of number is obtained when a hospital administrator counts the number of patients—perhaps 20—discharged from the hospital on a given day. Each of the three numbers is a *datum*, and the three taken together are data. Data can also be understood to be non numerical, and may include things such as text or other qualitative items. However, we will focus our interests in this text largely on numerical data and their associated analyses.

Statistics

The meaning of *statistics* is implicit in the previous section. More concretely, however, we may say that *statistics is a field of study concerned with* (1) *the collection, organization, summarization, and analysis of data and* (2) *the drawing of inferences about a body of data when only a part of the data is observed*.

The person who performs these statistical activities must be prepared to *interpret* and to *communicate* the results to someone else as the situation demands. Simply put, we may say that data are numbers, numbers contain information, and the purpose of statistics is to investigate and evaluate the nature and meaning of this information.

Sources of Data

The performance of statistical activities is motivated by the need to answer a question. For example, clinicians may want answers to questions regarding the relative merits of competing treatment procedures. Administrators may want answers to questions regarding such areas of concern as employee morale or facility utilization. When we determine that the appropriate approach to seeking an answer to a question will require the use of statistics, we begin to search for suitable data to serve as the raw material for our investigation. Such data are usually available from one or more of the following sources:

1. **Routinely kept records.** It is difficult to imagine any type of organization that does not keep records of day-to-day transactions of its activities. Hospital medical records, for example, contain immense amounts of information on patients, while hospital accounting records contain a wealth of data on the facility's business activities. When the need for data arises, we should look for them first among routinely kept records.

2. **Surveys.** If the data needed to answer a question are not available from routinely kept records, the logical source may be a survey. Suppose, for example, that the administrator of a clinic wishes to obtain information regarding the mode of transportation used by patients to visit the clinic. If admission forms do not contain a question on mode of transportation, we may conduct a survey among patients to obtain this information.

3. **Experiments.** Frequently, the data needed to answer a question are available only as the result of an experiment. A nurse may wish to know which of several strategies is best for maximizing patient compliance. The nurse might conduct an experiment in which the different strategies of motivating compliance are tried with different patients. Subsequent evaluation of the responses to the different strategies might enable the nurse to decide which is most effective.

4. **External sources.** The data needed to answer a question may already exist in the form of published reports, commercially available data banks, or the research literature. In other words, we may find that someone else has already asked the same question, and the answer obtained may be applicable to our present situation.

Biostatistics

The tools of statistics are employed in many fields—business, education, psychology, agriculture, and economics, to mention only a few. When the data analyzed are derived from the biological sciences and medicine, we use the term *biostatistics* to distinguish this particular application of statistical tools and concepts. This area of application is the concern of this book.

Variable

If, as we observe a characteristic, we find that it takes on different values in different persons, places, or things, we label the characteristic a *variable*. We do this for the simple reason that the characteristic is not the same when observed in different possessors of it. Some examples of variables include diastolic blood pressure, heart rate, the heights of adult males, the weights of preschool children, and the ages of patients seen in a dental clinic.

Quantitative Variables

A *quantitative variable* is one that can be measured in the usual sense. We can, for example, obtain measurements on the heights of adult males, the weights of preschool children, and the ages of patients seen in a dental clinic. These are examples of *quantitative variables*. Measurements made on quantitative variables convey information regarding amount.

Qualitative Variables

Some characteristics are not capable of being measured in the sense that height, weight, and age are measured. Many characteristics can be categorized only, as, for example, when an ill person is given a medical diagnosis, a person is designated as belonging to an ethnic group, or a person, place, or object is said to possess or not to possess some characteristic of interest. In such cases, measuring consists of categorizing. We refer to variables of this kind as *qualitative variables*. Measurements made on qualitative variables convey information regarding an attribute.

Although, in the case of qualitative variables, measurement in the usual sense of the word is not achieved, we can count the number of persons, places, or things belonging to various categories. A hospital administrator, for example, can count the number of patients admitted during a day under each of the various admitting diagnoses. These counts, or *frequencies* as they are called, are the numbers that we manipulate when our analysis involves qualitative variables.

Random Variable

Whenever we determine the height, weight, or age of an individual, the result is frequently referred to as a *value* of the respective variable. When the values obtained arise as a result of chance factors, so that they cannot be exactly predicted in advance, the variable is called a *random variable*. An example of a random variable is adult height. When a child is born, we cannot predict exactly his or her height at maturity. Attained adult height is the result of numerous genetic and environmental factors. Values resulting from measurement procedures are often referred to as *observations* or *measurements*.

Discrete Random Variable

Variables may be characterized further as to whether they are *discrete* or *continuous*. Since mathematically rigorous definitions of discrete and continuous variables are beyond the level of this book, we offer, instead, nonrigorous definitions and give an example of each.

A discrete variable is characterized by gaps or interruptions in the values that it can assume. These gaps or interruptions indicate the absence of values between particular values that the variable can assume. Some examples illustrate the point. The number of daily admissions to a general hospital is a discrete random variable since the number of admissions each day must be represented by a whole number, such as 0, 1, 2, or 3. The number of admissions on a given day cannot be a number such as 1.5, 2.997, or 3.333. The number of decayed, missing, or filled teeth per child in an elementary school is another example of a discrete variable.

Continuous Random Variable

A continuous random variable does not possess the gaps or interruptions characteristic of a discrete random variable. A continuous random variable can assume any value within a specified relevant interval of values assumed by the variable. Examples of continuous variables include the various measurements that can be made on individuals such as height, weight, and skull circumference. No matter how close together the observed heights of two people, for example, we can, theoretically, find another person whose height falls somewhere in between.

Because of the limitations of available measuring instruments, however, observations on variables that are inherently continuous are recorded as if they were discrete. Height, for example, is usually recorded to the nearest one-quarter, one-half, or whole inch, whereas, with a perfect measuring device, such a measurement could be made as precise as desired. Therefore, in a nontechnical sense, continuity is limited only by our ability to precisely measure it.

Population

The average person thinks of a population as a collection of entities, usually people. A population or collection of entities may, however, consist of animals, machines, places, or cells. For our purposes, we define a *population of entities as the largest collection of entities for which we have an interest at a particular time*. If we take a measurement of some variable on each of the entities in a population, we generate a population of values of that variable. We may, therefore, define a *population of values as the largest collection of values of a random variable for which we have an interest at a particular time*. If, for example, we are interested in the weights of all the children enrolled in a certain county elementary school system, our population consists of all these weights. If our interest lies only in the weights of first-grade students in the system, we have a different population—weights of first-grade students enrolled in the school system. Hence, populations are determined or defined by our sphere of interest. Populations may be *finite* or *infinite*. If a population of values consists of a fixed number of these values, the population is said to be *finite*. If, on the other hand, a population consists of an endless succession of values, the population is an *infinite* one. An exact value calculated from a population is referred to as a *parameter*.

Sample

A sample may be defined simply as *a part of a population*. Suppose our population consists of the weights of all the elementary school children enrolled in a certain county school system. If we collect for analysis the weights of only a fraction of these children, we have only a part of our population of weights, that is, we have a *sample*. An estimated value calculated from a sample is referred to as a *statistic*.

1.3 Measurement and Measurement Scales

In the preceding discussion, we used the word *measurement* several times in its usual sense, and presumably the reader clearly understood the intended meaning. The word *measurement*, however, may be given a more scientific definition. In fact, there is a whole body of scientific literature devoted to the subject of measurement. Part of this literature is concerned also with the nature of the numbers that result from measurements. Authorities on the subject of measurement speak of measurement scales that result in the categorization of measurements according to their nature. In this section, we define measurement and the four resulting measurement scales. A more detailed discussion of the subject is to be found in the writings of Stevens (1,2).

Measurement

This may be defined as the assignment of numbers to objects or events according to a set of rules. The various measurement scales result from the fact that measurement may be carried out under different sets of rules.

The Nominal Scale

The lowest measurement scale is the *nominal scale*. As the name implies it consists of "naming" observations or classifying them into various mutually exclusive and collectively exhaustive categories. The practice of using numbers to distinguish among the various medical diagnoses constitutes measurement on a nominal scale. Other examples include such dichotomies as positive–negative, well–sick, under 65 years of age–65 and over, child–adult, and married–not married.

The Ordinal Scale

Whenever observations are not only different from category to category but can be ranked according to some criterion, they are said to be measured on an ordinal scale. Convalescing patients may be characterized as unimproved, improved, and much improved. Individuals may be classified according to socioeconomic status as low, medium, or high. The intelligence of children may be above average, average, or below average. In each of these examples, the members of any one category are all considered equal, but the members of one category are considered lower, worse, or smaller than those in another category, which in turn bears a similar relationship to another category. For example, a much improved patient is in better health than one classified as improved, while a patient who has improved is in better condition than one who has not improved. It is usually impossible to infer that the difference between members of one category and the next adjacent category is equal to the difference between members of that category and the members of the next category adjacent to it. The degree of improvement between unimproved and improved is probably not the same as that between improved and much improved. The implication is that if a finer breakdown were made resulting in more categories, these, too, could be ordered in a similar manner. The function of numbers assigned to ordinal data is to order (or rank) the observations from lowest to highest and, hence, the term *ordinal*.

The Interval Scale

The *interval scale* is a more sophisticated scale than the nominal or ordinal in that with this scale not only is it possible to order measurements, but also the distance between any two measurements is known. We know, say, that the difference between a measurement of 20 and a measurement of 30 is equal to the difference between measurements of 30 and 40. The ability to do this implies the use of a unit distance and a zero point, both of which are arbitrary. The selected zero point is not necessarily a true zero in that it does not have to indicate a total absence of the quantity being measured. Perhaps the best example of an interval scale is provided by the way in which temperature is usually measured (degrees Fahrenheit or Celsius). The unit of measurement is the degree, and the point of comparison is the arbitrarily chosen "zero degrees," which does not indicate a lack of heat. The interval scale unlike the nominal and ordinal scales is a truly quantitative scale.

The Ratio Scale

The highest level of measurement is the *ratio scale*. This scale is characterized by the fact that equality of ratios as well as equality of intervals may be determined. Fundamental to the ratio scale is a true zero point. The measurement of such familiar traits as height, weight, and length makes use of the ratio scale.

1.4 Sampling and Statistical Inference

As noted earlier, one of the purposes of this book is to teach the concepts of statistical inference, which we may define as follows:

DEFINITION

Statistical inference is the procedure by which we reach a conclusion about a population on the basis of the information contained in a sample that has been drawn from that population.

There are many kinds of samples that may be drawn from a population. Not every kind of sample, however, can be used as a basis for making valid inferences about a population. In general, in order to make a valid inference about a population, we need a scientific sample from the population. There are also many kinds of scientific samples that may be drawn from a population. The simplest of these is the *simple random sample*. In this section, we define a simple random sample and show you how to draw one from a population.

If we use the letter N to designate the size of a finite population and the letter n to designate the size of a sample, we may define a simple random sample as follows:

DEFINITION

If a sample of size n is drawn from a population of size N in such a way that every possible sample of size n has the same chance of being selected, the sample is called a simple random sample.

The mechanics of drawing a sample to satisfy the definition of a simple random sample is called *simple random sampling*.

We will demonstrate the procedure of simple random sampling shortly, but first let us consider the problem of whether to sample *with replacement* or *without replacement*. When sampling with replacement is employed, every member of the population is available at each draw. For example, suppose that we are drawing a sample from a population of former hospital patients as part of a study of length of stay. Let us assume that the sampling involves selecting from the electronic health records a sample of charts of discharged patients. In sampling with replacement we would proceed as follows: select a chart to be in the sample, record the length of stay, and close the electronic chart. The chart is back in the "population" and may be selected again on some subsequent draw, in which case the length of stay will again be recorded. In sampling without replacement, we would not record a length of stay for a patient whose chart was already selected from the database. Following this procedure, a given chart could appear in the sample only once. In practice, sampling is almost always done without replacement. The significance and consequences of this will be explained later, but first let us see how one goes about selecting a simple random sample. To ensure true randomness of selection, we will need to follow some objective procedure. We certainly will want to avoid using our own judgment to decide which members of the population constitute a random sample. The following example illustrates one method of selecting a simple random sample from a population.

EXAMPLE 1.4.1

Gold et al. (A-1) studied the effectiveness on smoking cessation of bupropion SR, a nicotine patch, or both, when co administered with cognitive behavioral therapy. Consecutive consenting patients assigned themselves to one of the three conditions. For illustrative purposes, let us consider all these subjects to be a population of size $N = 189$. We wish to select a simple random sample of size 10 from this population whose ages are shown in Table 1.4.1.

Table 1.4.1 Ages of 189 Subjects Who Participated in a Study on Smoking Cessation

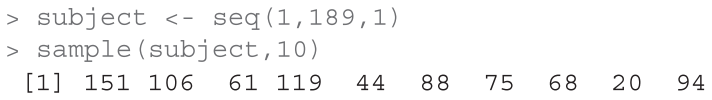
Subject No.	Age	Subject No.	Age	Subject No.	Age	Subject No.	Age
1	48	49	38	97	51	145	52
2	35	50	44	98	50	146	53
3	46	51	43	99	50	147	61
4	44	52	47	100	55	148	60
5	43	53	46	101	63	149	53
6	42	54	57	102	50	150	53
7	39	55	52	103	59	151	50
8	44	56	54	104	54	152	53
9	49	57	56	105	60	153	54
10	49	58	53	106	50	154	61
11	44	59	64	107	56	155	61
12	39	60	53	108	68	156	61
13	38	61	58	109	66	157	64
14	49	62	54	110	71	158	53
15	49	63	59	111	82	159	53
16	53	64	56	112	68	160	54
17	56	65	62	113	78	161	61
18	57	66	50	114	66	162	60
19	51	67	64	115	70	163	51
20	61	68	53	116	66	164	50
21	53	69	61	117	78	165	53
22	66	70	53	118	69	166	64
23	71	71	62	119	71	167	64
24	75	72	57	120	69	168	53
25	72	73	52	121	78	169	60
26	65	74	54	122	66	170	54
27	67	75	61	123	68	171	55
28	38	76	59	124	71	172	58
29	37	77	57	125	69	173	62
30	46	78	52	126	77	174	62
31	44	79	54	127	76	175	54
32	44	80	53	128	71	176	53
33	48	81	62	129	43	177	61
34	49	82	52	130	47	178	54
35	30	83	62	131	48	179	51
36	45	84	57	132	37	180	62

Table 1.4.1 **(Continued)**

Subject No.	Age	Subject No.	Age	Subject No.	Age	Subject No.	Age
37	47	85	59	133	40	181	57
38	45	86	59	134	42	182	50
39	48	87	56	135	38	183	64
40	47	88	57	136	49	184	63
41	47	89	53	137	43	185	65
42	44	90	59	138	46	186	71
43	48	91	61	139	34	187	71
44	43	92	55	140	46	188	73
45	45	93	61	141	46	189	66
46	40	94	56	142	48		
47	48	95	52	143	47		
48	49	96	54	144	43		

Source: Data provided courtesy of Paul B. Gold, Ph.D.

SOLUTION: One way of selecting a simple random sample is to use a table of random numbers, which was very commonly done historically, or to use an online random number generator. However, most statistical computer packages provide a way to select a sample of given size. For example, using program R, we can create a sequence of numbers from 1 to 189, and call them "subject" and then use the sample() function to randomly select 10 of them. One such output is shown in Figure 1.4.1, with the output summarized in Table 1.4.2.



```
> subject <- seq(1,189,1)
> sample(subject,10)
 [1] 151 106  61 119  44  88  75  68  20  94
```

FIGURE 1.4.1 Simple R program for selecting a sample of size 10 from 189 subjects.

Table 1.4.2 **Sample of 10 Ages Drawn from the Ages in Table 1.4.1**

Random Number	Sample Subject Number	Age
151	1	50
106	2	50
61	3	58
119	4	71
44	5	79
88	6	57
75	7	61
68	8	53
20	9	61
94	10	56

Thus we have drawn a simple random sample of size 10 from a population of size 189. In future discussions, whenever the term *simple random sample* is used, it will be understood that the sample has been drawn in this or an equivalent manner.

The preceding discussion of random sampling is presented because of the important role that the sampling process plays in designing *research studies* and *experiments*. The methodology and concepts employed in sampling processes will be described in more detail in Section 1.5.

DEFINITION

A research study is a scientific study of a phenomenon of interest. Research studies involve designing sampling protocols, collecting and analyzing data, and providing valid conclusions based on the results of the analyses.

DEFINITION

Experiments are a special type of research study in which observations are made after specific manipulations of conditions have been carried out; they provide the foundation for scientific research.

Despite the tremendous importance of random sampling in the design of research studies and experiments, there are some occasions when random sampling may not be the most appropriate method to use. Consequently, other sampling methods must be considered. The intention here is not to provide a comprehensive review of sampling methods, but rather to acquaint the student with two additional sampling methods that are often employed in the health sciences, *systematic sampling* and *stratified random sampling*. Interested readers are referred to the books by Thompson (3) and Levy and Lemeshow (4) for detailed overviews of various sampling methods and explanations of how sample statistics are calculated when these methods are applied in research studies and experiments.

Systematic Sampling

A sampling method that is widely used in health-care research is the systematic sample. Medical records, which contain raw data used in health-care research, are generally stored in a file system or on a computer and hence are easy to select in a systematic way. Using systematic sampling methodology, a researcher calculates the total number of records needed for the study or experiment at hand. A random number is then chosen to use as a starting point for initiating sampling. The record located at this starting point is called record x. A second number, determined by the number of records desired, is selected to define the sampling interval (call this interval k). Consequently, the data set would consist of records $x, x + k, x + 2k, x + 3k$, and so on, until the necessary number of records are obtained.

EXAMPLE 1.4.2

Continuing with the study of Gold et al. (A-1) illustrated in the previous example, imagine that we wanted a systematic sample of 10 subjects from those listed in Table 1.4.1.

SOLUTION: To obtain a starting point, we can employ the strategy from above using R and simply select only one subject. Let us assume that the sample() function returned subject number 4, which will serve as our starting point, *x*. Since we are starting at subject 4, this leaves 185 remaining subjects (i.e., 189 − 4) from which to choose. Since we wish to select 10 subjects, one method to define the sample interval, *k*, would be to take 185/10 = 18.5. To ensure that there will be enough subjects, it is customary to round this quotient down, and hence we will round the result to 18. In this scenario, the samples would be 4, 4 + 18 = 22, 4 + 18(2) = 40, 4 + 18(3) = 58, and so on. The resulting sample is shown in Table 1.4.3.

Table 1.4.3 **Sample of 10 Ages Selected Using a Systematic Sample from the Ages in Table 1.4.1**

Systematically Selected Subject Number	Age
4	44
22	66
40	47
58	53
76	59
94	56
112	68
130	47
148	60
166	64

Stratified Random Sampling

A common situation that may be encountered in a population under study is one in which the sample units occur together in a grouped fashion. On occasion, when the sample units are not inherently grouped, it may be possible and desirable to group them for sampling purposes. In other words, it may be desirable to partition a population of interest into groups, or *strata*, in which the sample units within a particular stratum are more similar to each other than they are to the sample units that compose the other strata. After the population is stratified, it is customary to take a random sample independently from each stratum. This technique is called *stratified random sampling*. The resulting sample is called a *stratified random sample*. Although the benefits of stratified random sampling may not be readily observable, it is most often the case that random samples taken within a stratum will have much less variability than a random sample taken across all strata. This is true because sample units within each stratum tend to have characteristics that are similar.

EXAMPLE 1.4.3

Hospital trauma centers are given ratings depending on their capabilities to treat various traumas. In this system, a level 1 trauma center is the highest level of available trauma care and a level 4 trauma center is the lowest level of available trauma care. Imagine that we are interested in estimating the survival rate of trauma victims treated at hospitals within a large metropolitan area. Suppose that the metropolitan area has a level 1, a level 2, and a level 3 trauma center. We wish to take samples of patients from these trauma centers in such a way that the total sample size is 30.

SOLUTION: We assume that the survival rates of patients may depend quite significantly on the trauma that they experienced and therefore on the level of care that they receive. As a result, a simple random sample of all trauma patients, without regard to the center at which they were treated, may not represent true survival rates, since patients receive different care at the various trauma centers. One way to better estimate the survival rate is to treat each trauma center as a stratum and then randomly select 10 patient files from each of the three centers. This procedure is based on the fact that we suspect that the survival rates within the trauma centers are less variable than the survival rates across trauma centers. Therefore, we believe that the stratified random sample provides a better representation of survival than would a sample taken without regard to differences within strata.

It should be noted that two slight modifications of the stratified sampling technique are frequently employed. To illustrate, consider again the trauma center example. In the first place, a systematic sample of patient files could have been selected from each trauma center (stratum). Such a sample is called a *stratified systematic sample*.

The second modification of stratified sampling involves selecting the sample from a given stratum in such a way that the number of sample units selected from that stratum is proportional to the size of the population of that stratum. Suppose, in our trauma center example that the level 1 trauma center treated 100 patients and the level 2 and level 3 trauma centers treated only 10 each. In that case, selecting a random sample of 10 from each trauma center overrepresents the trauma centers with smaller patient loads. To avoid this problem, we adjust the size of the sample taken from a stratum so that it is proportional to the size of the stratum's population. This type of sampling is called *stratified sampling proportional to size*. The within-stratum samples can be either random or systematic as described above.

Exercises

1.4.1 Using a table of random numbers or a computer program, select a new simple random sample of size 10 from the data in Table 1.4.1. Record the ages of the subjects in this new sample. Save your data for future use. What is the variable of interest in this exercise? What measurement scale was used to obtain the measurements?

1.4.2 Select another simple random sample of size 10 from the population represented in Table 1.4.1. Compare the subjects in this sample with those in the sample drawn in Exercise 1.4.1. Are there any subjects who showed up in both samples? How many? Compare the ages of the subjects in the two samples. How many ages in the first sample were duplicated in the second sample?

1.4.3 Using a table of random numbers or a computer program, select a random sample and a systematic sample, each of size 15, from the data in Table 1.4.1. Visually compare the distributions of the two samples. Do they appear similar? Which appears to be the best representation of the data?

1.4.4 Construct an example where it would be appropriate to use stratified sampling. Discuss how you would use stratified random sampling and stratified sampling proportional to size with this example. Which do you think would best represent the population that you described in your example? Why?

1.5 The Scientific Method

Data analyses using a broad range of statistical methods play a significant role in scientific studies. The previous section highlighted the importance of obtaining samples in a scientific manner. Appropriate sampling techniques enhance the likelihood that the results of statistical analyses of a data set will provide valid and scientifically defensible results. Because of the importance of the proper collection of data to support scientific discovery, it is necessary to consider the foundation of such discovery—the *scientific method*—and to explore the role of statistics in the context of this method.

DEFINITION

The scientific method is a process by which scientific information is collected, analyzed, and reported in order to produce unbiased and replicable results in an effort to provide an accurate representation of observable phenomena.

The scientific method is recognized universally as the only truly acceptable way to produce new scientific understanding of the world around us. It is based on an *empirical approach*, in that decisions and outcomes are based on data. There are several key elements associated with the scientific method, and the concepts and techniques of statistics play a prominent role in all these elements. The Scientific Method is illustrated in Figure 1.5.1.

Observations and Question Formation

First, an *observation* is made of a phenomenon or a group of phenomena. This observation leads to the formulation of questions or uncertainties that can be answered in a scientifically rigorous way. For example, it is readily observable that regular exercise reduces body weight in many people. It is also readily observable that changing diet may have a similar effect. In this case, there are two observable phenomena, regular exercise and diet change, that have the same endpoint. The nature of this endpoint can be determined by use of the scientific method.

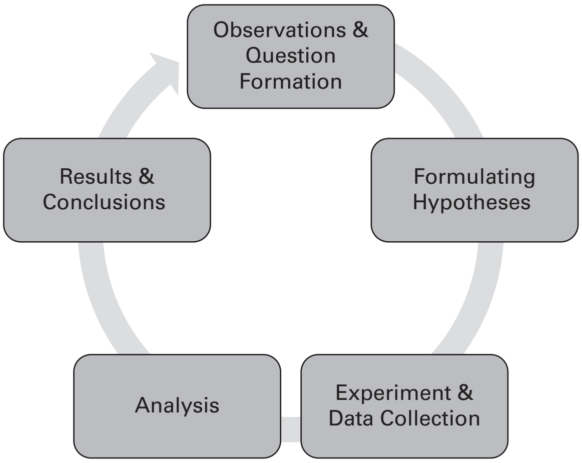

Observations & Question Formation
Formulating Hypotheses
Experiment & Data Collection
Analysis
Results & Conclusions

FIGURE 1.5.1 An illustration of the Scientific Method.

Formulating Hypotheses

In the second step of the scientific method, a *hypothesis* is formulated to explain the observation and to make quantitative *predictions* of new observations. Often hypotheses are generated as a result of extensive background research and literature reviews. The objective is to produce hypotheses that are scientifically sound. Hypotheses may be stated as either *research hypotheses* or *statistical hypotheses*. Explicit definitions of these terms are given in Chapter 7, which discusses the science of testing hypotheses. Suffice it to say for now that a research hypothesis from the weight-loss example would be a statement such as “Exercise appears to reduce body weight.” There is certainly nothing incorrect about this conjecture, but it lacks a truly quantitative basis for testing. A statistical hypothesis may be stated using quantitative terminology as follows: “The average (mean) loss of body weight of people who exercise is greater than the average (mean) loss of body weight of people who do not exercise.” In this statement a quantitative measure, the “average” or “mean” value, is hypothesized to be greater in the sample of patients who exercise. The role of the statistician in this step of the scientific method is to state the hypothesis in a way that valid conclusions may be drawn and to interpret correctly the results of such conclusions.

Experiment and Data Collection

The third step of the scientific method involves *designing an experiment* that will yield the data necessary to validly test an appropriate statistical hypothesis. This step of the scientific method, like that of data analysis, requires the expertise of a statistician. Improperly designed experiments are the leading cause of invalid results and unjustified conclusions. Further, most studies that are challenged by experts are challenged on the basis of the appropriateness or inappropriateness of the study’s research design, which can lead to a nonreproducible result.

Those who properly design research experiments make every effort to ensure that the measurement of the phenomenon of interest is both accurate and precise. *Accuracy* refers to the correctness of a measurement. *Precision*, on the other hand, refers to the consistency of a measurement. It should be noted that in the social sciences, the term *validity* is sometimes used to mean accuracy and that *reliability* is sometimes used to mean precision. In the context of the weight-loss example given earlier, the scale used to measure the weight of study participants would be accurate if the measurement is validated using a scale that is properly calibrated. If, however, the scale is off by +3 pounds, then each participant’s weight would be 3 pounds heavier; the measurements would be precise in that each would be wrong by +3 pounds, but the measurements would not be accurate. Measurements that are inaccurate or imprecise may invalidate research findings.

The design of an experiment depends on the type of data that need to be collected to test a specific hypothesis. As discussed in Section 1.2, data may be collected or made available through a variety of means. For much scientific research, however, the standard for data collection is experimentation. A true *experimental design* is one in which study subjects are randomly assigned to an *experimental group* (or *treatment group*) and a *control group* that is not directly exposed to a treatment. Continuing the weight-loss example, a sample of 100 participants could be randomly assigned to two conditions using the methods of Section 1.4. A sample of 50 of the participants would be assigned to a specific exercise program and the remaining 50 would be monitored, but asked not to exercise for a specific period of time. At the end of this experiment, the average (mean) weight losses of the two groups could be compared. The reason that experimental designs are desirable is that if all other potential factors are controlled, a *cause–effect relationship* may be tested; that is, all else being equal, we would be able to conclude or fail to conclude that the experimental group lost weight as a result of exercising.

The potential complexity of research designs requires statistical expertise, and Chapter 8 highlights some commonly used experimental designs. For a more in-depth discussion of research designs, the interested reader may wish to refer to texts by Kuehl (5), Keppel and Wickens (6), and Tabachnick and Fidell (7).

Analysis

Assuming that an appropriate experimental design is employed and data are correctly collected through a validated sampling protocol, these resulting data can be analyzed. Often *descriptive statistics*, as we will learn in Chapter 2, are used to understand characteristics of the data that have been collected. Additionally, based on the hypotheses that were developed, a researcher may have one or more analyses to complete using *inferential statistics*, as covered starting in Chapter 7, in order to make predictions or infer conclusions from the data.

Results and Conclusions

In the execution of a research study or experiment, one would hope to have collected the data necessary to draw conclusions, with some degree of confidence, about the hypotheses that were posed as part of the design. It is often the case that hypotheses need to be modified and retested with new data and a different design. Whatever the conclusions of the scientific process, however, results are rarely considered to be conclusive. That is, results need to be *replicated*, often a large number of times, before scientific credence is granted them.

Exercises

1.5.1 Using the example of weight loss as an endpoint, discuss how you would use the scientific method to test the observation that change in diet is related to weight loss. Include all of the steps, including the hypothesis to be tested and the design of your experiment.

1.5.2 Continuing with Exercise 1.5.1, consider how you would use the scientific method to test the observation that both exercise and change in diet are related to weight loss. Include all of the steps, paying particular attention to how you might design the experiment and which hypotheses would be testable given your design.

1.6 Computers and Technology

The widespread use of computers and related technology has had a tremendous impact on health sciences research in general and biostatistical analysis in particular. The necessity to perform long and tedious arithmetic computations as part of the statistical analysis of data lives only in the memory of those researchers and practitioners whose careers antedate the so-called computer revolution. Likewise, the use of statistical tables for finding standard comparative (i.e., critical) values in hypothesis testing largely has been replaced with statistical algorithms. Computers can perform more calculations faster and far more accurately than can human technicians. The use of computers makes it possible for investigators to devote more time to the improvement of the quality of raw data and the interpretation of the results.

The current prevalence of microcomputers and the abundance of available statistical software programs have further revolutionized statistical computing. The reader in search of a statistical software package may wish to consult *The American Statistician*, a quarterly publication of the American Statistical Association. Statistical software packages are regularly reviewed and advertised in the periodical.

As was illustrated in a previous example, an alternative to using printed tables of random numbers, investigators may use computers to generate the random numbers they need. Actually, the "random" numbers generated by most computers are in reality *pseudorandom numbers* because they are the result of a deterministic formula. However, as Fishman (8) points out, the numbers appear to serve satisfactorily for many practical purposes.

The usefulness of the computer in the health sciences is not limited to statistical analysis. The reader interested in learning more about the use of computers in the health sciences will find the books by Hersh (9), Johns (10), Miller et al. (11), and Saba and McCormick (12) helpful; though some of these are now dated, the perspectives discussed in these volumes is still relevant. Those who wish to derive maximum benefit from the Internet may wish to consult the books *Physicians' Guide to the Internet* (13) and *Computers in Nursing's Nurses' Guide to the Internet* (14). Current developments in the use of computers in biology, medicine, and related fields are reported in several periodicals devoted to the subject. A few such periodicals are *Computers in Biology and Medicine*, *Computers and Biomedical Research*, *International Journal of Bio-Medical Computing*, *Computer Methods and Programs in Biomedicine*, *Computer Applications in the Biosciences*, and *Computers in Nursing*.

As the reader makes their way through this text, it will be apparent just how important computers and technology have become to our current use of statistics for applied problems. We will generally forgo lengthy hand calculations and tabled values in favor of a more contemporary view of statistical science. In that regard, Computer printouts are used throughout this book to illustrate the use of computers in biostatistical analysis. The MINITAB, SPSS, R, JASP, and SAS® statistical software packages for the personal computer have been used for this purpose. Additionally, we will sometimes present some useful Microsoft Excel™ outputs as well.

1.7 SUMMARY

In this chapter, we introduced the reader to the basic concepts of statistics. We defined statistics as an area of study concerned with collecting and describing data and with making statistical inferences. We defined statistical inference as the procedure by which we reach a conclusion about a population on the basis of information contained in a sample drawn from that population. We learned that a basic type of sample that will allow us to make valid inferences is the simple random sample. We learned how to use a table of random numbers to draw a simple random sample from a population.

The reader is provided with the definitions of some basic terms, such as variable and sample, that are used in the study of statistics. We also discussed measurement and defined four measurement scales—nominal, ordinal, interval, and ratio. The reader is also introduced to the scientific method and the role of statistics and the statistician in this process.

Finally, we discussed the importance of computers in the performance of the activities involved in statistics.

REVIEW QUESTIONS AND EXERCISES

1. Explain what is meant by descriptive statistics.

2. Explain what is meant by inferential statistics.

3. Define:

(a) Statistics

(b) Biostatistics

(c) Variable

(d) Quantitative variable

(e) Qualitative variable

(f) Random variable

(g) Population

(h) Finite population

(i) Infinite population

(j) Sample

(k) Discrete variable

(l) Continuous variable

(m) Simple random sample

(n) Sampling with replacement

(o) Sampling without replacement

4. Define the word *measurement*.

5. List, describe, and compare the four measurement scales.

6. For each of the following variables, indicate whether it is quantitative or qualitative and specify the measurement scale that is employed when taking measurements on each:

(a) Class standing of the members of this class relative to each other

(b) Admitting diagnosis of patients admitted to a mental health clinic

(c) Weights of babies born in a hospital during a year

(d) Gender of babies born in a hospital during a year

(e) Range of motion of elbow joint of students enrolled in a university health sciences curriculum

(f) Under-arm temperature of day-old infants born in a hospital

7. For each of the following situations, answer questions a through e:

(a) What is the sample in the study?

(b) What is the population?

(c) What is the variable of interest?

(d) How many measurements were used in calculating the reported results?

(e) What measurement scale was used?

Situation A. A study of 300 households in a small southern town revealed that 20 percent had at least one school-age child present.

Situation B. A study of 250 patients admitted to a hospital during the past year revealed that, on the average, the patients lived 15 miles from the hospital.

8. Consider the two situations given in Exercise 7. For Situation A, describe how you would use a stratified random sample to collect the data. For Situation B, describe how you would use systematic sampling of patient records to collect the data.

REFERENCES

Methodology References

1. Stanley S. Stevens, "On the Theory of Scales of Measurement," *Science*, 103 (1946), 677–680.
2. Stanley S. Stevens, "Mathematics, Measurement and Psychophysics," in Stanley S. Stevens (ed.), *Handbook of Experimental Psychology*, Wiley, New York, 1951.
3. Steven K. Thompson, *Sampling* (3rd ed.), Wiley, New York, 2012.
4. Paul S. Levy and Stanley Lemeshow, *Sampling of Populations: Methods and Applications* (4th ed.), Wiley, New York, 2008.
5. Robert O. Kuehl, *Statistical Principles of Research Design and Analysis* (2nd ed.), Duxbury Press, Belmont, CA, 1999.
6. Geoffrey Keppel and Thomas D. Wickens, *Design and Analysis: A Researcher's Handbook* (4th ed.), Prentice Hall, Upper Saddle River, NJ, 2004.
7. Barbara G. Tabachnick and Linda S. Fidell, *Experimental Designs using ANOVA*, Thomson, Belmont, CA, 2007.
8. George S. Fishman, *Concepts and Methods in Discrete Event Digital Simulation*, Wiley, New York, 1973.
9. William R. Hersh, *Information Retrieval: A Health Care Perspective* (3rd ed.), Springer, New York, 2010.
10. Merida L. Johns, *Information Management for Health Professions*, Delmar Publishers, Albany, NY, 1997.
11. Marvin J. Miller, Kenric W. Hammond, and Matthew G. Hile (eds.), *Mental Health Computing*, Springer, New York, 1996.
12. Virginia K. Saba and Kathleen A. McCormick, *Essentials of Computers for Nurses*, McGraw-Hill, New York, 1996.
13. Lee Hancock, *Physicians' Guide to the Internet*, Lippincott Williams & Wilkins Publishers, Philadelphia, 1996.
14. Leslie H. Nicoll and Teena H. Ouellette, *Computers in Nursing's Nurses' Guide to the Internet* (3rd ed.), Lippincott Williams & Wilkins Publishers, Philadelphia, 2001.

Applications References

A-1. Paul B. Gold, Robert N. Rubey, and Richard T. Harvey, "Naturalistic, Self-Assignment Comparative Trial of Bupropion SR, a Nicotine Patch, or Both for Smoking Cessation Treatment in Primary Care," *American Journal on Addictions*, 11 (2002), 315–331.

2 Descriptive Statistics

CHAPTER OVERVIEW

This chapter introduces a set of basic procedures and statistical measures for describing data. Data generally consist of an extensive number of measurements or observations that are too numerous or complicated to be understood through simple observation. Therefore, this chapter introduces several techniques including the construction of tables, graphical displays, and basic statistical computations that provide ways to condense and organize information into a set of descriptive measures and visual devices that enhance the understanding of complex data.

TOPICS

LEARNING OUTCOMES

After studying this chapter, the student will

1. understand how data can be appropriately organized and displayed.
2. understand how to reduce data sets into a few useful, descriptive measures.
3. be able to calculate and interpret measures of central tendency, such as the mean, median, and mode.
4. be able to calculate and interpret measures of dispersion, such as the range, variance, and standard deviation.

2.1 Introduction

In Chapter 1, we stated that the taking of a measurement and the process of counting yield numbers that contain information. The objective of the person applying the tools of statistics to these numbers is to determine the underlying nature of this information. This task is made much easier if the numbers are organized and summarized. When measurements of a random variable are taken on the entities of a population or sample, the resulting values are made available to the researcher or statistician as a mass of unordered data. Measurements that have not been organized, summarized, or otherwise manipulated are called *raw data*. Unless the number of observations is extremely small, it will be unlikely that these raw data will impart much information until they have been organized in some way.

In this chapter, we learn several techniques for organizing and summarizing data so that we may more easily determine what information they contain. The ultimate in summarization of data is the calculation of a single number that in some way conveys important information about the data from which it was calculated. Such single numbers that are used to describe data are called *descriptive measures*. After studying this chapter, you will be able to compute several descriptive measures for both populations and samples of data.

The purpose of this chapter is to equip you with skills that will enable you to manipulate the information—in the form of numbers—that you encounter as a health sciences professional. The better able you are to manipulate such information, the better understanding you will have of the environment and forces that generate the information.

2.2 The Ordered Array

An easy first step in organizing data is simply to order, or rank, the data. An *ordered array* is a listing of the values of a collection (either population or sample) in order of magnitude from the smallest value to the largest value.

An ordered array enables one to determine quickly the value of the smallest measurement, the value of the largest measurement, and other facts about the arrayed data that might be immediately useful. We illustrate the construction of an ordered array with the data discussed in Example 1.4.1.

EXAMPLE 2.2.1

Table 1.4.1 contains a list of the ages of subjects who participated in the study on smoking cessation discussed in Example 1.4.1. As can be seen, this unordered table requires considerable searching for us to ascertain such elementary information as the age of the youngest and oldest subjects.

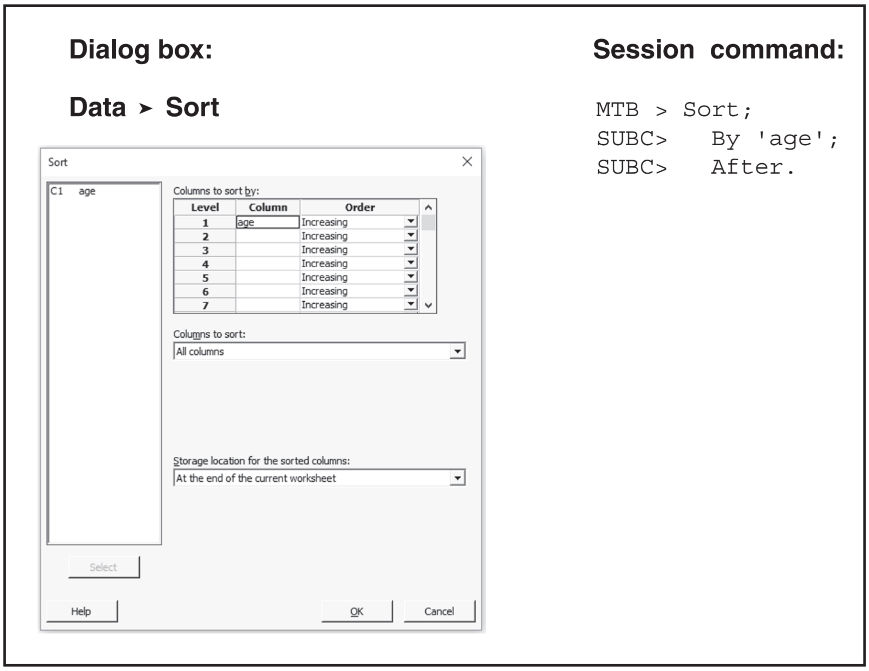
SOLUTION: Table 2.2.1 presents the data of Table 1.4.1 in the form of an ordered array. By referring to Table 2.2.1 we are able to determine quickly the age of the youngest subject (30) and the age of the oldest subject (82). We also readily note that about one-third of the subjects are 50 years of age or younger.

Table 2.2.1 **Ordered Array of Ages of Subjects from Table 1.4.1**

30	34	35	37	37	38	38	38	38	39	39	40	40	42	42
43	43	43	43	43	43	44	44	44	44	44	44	44	45	45
45	46	46	46	46	46	46	47	47	47	47	47	47	48	48
48	48	48	48	48	49	49	49	49	49	49	49	50	50	50
50	50	50	50	50	51	51	51	51	52	52	52	52	52	52
53	53	53	53	53	53	53	53	53	53	53	53	53	53	53
53	53	54	54	54	54	54	54	54	54	54	54	54	55	55
55	56	56	56	56	56	56	57	57	57	57	57	57	57	58
58	59	59	59	59	59	59	60	60	60	60	61	61	61	61
61	61	61	61	61	61	61	62	62	62	62	62	62	62	63
63	64	64	64	64	64	64	65	65	66	66	66	66	66	66
67	68	68	68	69	69	69	70	71	71	71	71	71	71	71
72	73	75	76	77	78	78	78	82						

Computer Analysis

If additional computations and organization of a data set have to be done by hand, the work may be facilitated by working from an ordered array. If the data are to be analyzed by a computer, it may be undesirable to prepare an ordered array, unless one is needed for reference purposes or for some other use. A computer does not need for its user to first construct an ordered array before entering data for the construction of frequency distributions and the performance of other analyses. However, almost all computer statistical packages and spreadsheet programs contain a routine for sorting data in either an ascending or descending order. See Figure 2.2.1, for example.



Dialog box:

Data ➤ Sort

Sort

C1 age

Columns to sort by:

Level	Column	Order
1	age	Increasing
2		Increasing
3		Increasing
4		Increasing
5		Increasing
6		Increasing
7		Increasing

Columns to sort: All columns

Storage location for the sorted columns: At the end of the current worksheet

Session command:

```
MTB > Sort;
SUBC>   By 'age';
SUBC>   After.
```

FIGURE 2.2.1 MINITAB dialog box for Example 2.2.1.

2.3 Frequency Tables

Although a set of observations can be made more comprehensible and meaningful by means of an ordered array, further useful summarization may be achieved by grouping the data. Before the days of computers one of the main objectives in grouping large data sets was to facilitate the calculation of various descriptive measures such as percentages and averages. Because computers can perform these calculations on large data sets without first grouping the data, the main purpose in grouping data now is summarization, and all statistical packages provide some basic set of algorithms for developing group summaries and visualizations. One must bear in mind that data contain information and that summarization is a way of making it easier to determine the nature of this information. One must also be aware that reducing a large quantity of information in order to summarize the data succinctly carries with it the potential to inadvertently lose some amount of specificity with regard to the underlying data set. Therefore, it is important to group the data sufficiently such that the vast amounts of information are reduced into understandable summaries. At the same time data should be summarized only to the extent that useful intricacies in the data are not obfuscated.

To group a set of observations, we select a set of contiguous, nonoverlapping intervals such that each value in the set of observations can be placed in one, and only one, of the intervals. These intervals are usually referred to as *class intervals*.

One of the first considerations when data are to be grouped is how many intervals to include. Too few intervals are undesirable because of the resulting loss of information. On the other hand, if too many intervals are used, the objective of summarization will not be met. The best guide to this, as well as to other decisions to be made in grouping data, is your knowledge of the data. It may be that class intervals have been determined by precedent, as in the case of annual tabulations, when the class intervals of previous years are maintained for comparative purposes. A commonly followed rule of thumb states that there should be no fewer than 5 intervals and no more than 15. If there are fewer than five intervals, the data have been summarized too much and the information they contain has been lost. If there are more than 15 intervals, the data have not been summarized enough.

Those who need more specific guidance in the matter of deciding how many class intervals to employ may use a formula given by Sturges (1). This formula gives $k = 1 + 3.322(\log_{10} n)$, where k stands for the number of class intervals and n is the number of values in the data set under consideration. The answer obtained by applying *Sturges's rule* should not be regarded as final, but should be considered as a guide only. The number of class intervals specified by the rule should be increased or decreased for convenience and clear presentation. This rule is one among many, and the literature is replete with various algorithms. We provide Sturges's rule here as an historic example and note that computer programs use various algorithms for determining the number of classes.

Suppose, for example, that we have a sample of 275 observations that we want to group. The logarithm to the base 10 of 275 is 2.4393. Applying Sturges's formula gives $k = 1 + 3.322(2.4393) \simeq 9$. In practice, other considerations might cause us to use eight or fewer or perhaps 10 or more class intervals.

Another question that must be decided regards the width of the class intervals. Class intervals generally should be of the same width, although this is sometimes impossible to accomplish. This width may be determined by dividing the range by k, the number of class intervals. Symbolically, the class interval width is given by

$$w = \frac{R}{k} \tag{2.3.1}$$

where R (the range; see Section 2.5) is the difference between the smallest and the largest observation in the data set, and k is defined as above. As a rule this procedure yields a width that is inconvenient for use. Again, we may exercise our good judgment and select a width (usually close to one given by Equation 2.3.1) that is more convenient.

There are other rules of thumb that are helpful in setting up useful class intervals. When the nature of the data makes them appropriate, class interval widths of 5 units, 10 units, and widths that are multiples of 10 tend to make the summarization more comprehensible. When these widths are employed, it is generally good practice to have the lower limit of each interval end in a 0 or 5. Usually class intervals are ordered from smallest to largest; that is, the first class interval contains the smaller measurements and the last class interval contains the larger measurements. When this is the case, the lower limit of the first class interval should be equal to or smaller than the smallest measurement in the data set, and the upper limit of the last class interval should be equal to or greater than the largest measurement.

Most statistical packages allow users to interactively change the number of class intervals and/or the class widths, so that several visualizations of the data can be obtained quickly. This feature allows users to exercise their judgment in deciding which data display is most appropriate for a given purpose. Let us use the 189 ages shown in Table 1.4.1 and arrayed in Table 2.2.1 to illustrate the construction of a frequency distribution.

EXAMPLE 2.3.1

We wish to know how many class intervals to have in the frequency distribution of the data. We also want to know how wide the intervals should be.

SOLUTION: To get an idea as to the number of class intervals to use, we can apply Sturges's rule to obtain

$$\begin{aligned} k &= 1 + 3.322(\log 189) \\ &= 1 + 3.322(2.2764618) \\ &\approx 9 \end{aligned}$$

Now let us divide the range by 9 to get some idea about the class interval width. We have

$$\frac{R}{k} = \frac{82 - 30}{9} = \frac{52}{9} = 5.778$$

It is apparent that a class interval width of 5 or 10 will be more convenient to use, as well as more meaningful to the reader. Suppose we decide on 10. We may now construct our intervals. Since the smallest value in Table 2.2.1 is 30 and the largest value is 82, we may begin our intervals with 30 and end with 89. This gives the following intervals:

30–39

40–49

50–59

60–69

70–79

80–89

We see that there are six of these intervals, three fewer than the number suggested by Sturges's rule.

It is sometimes useful to refer to the center, called the *midpoint*, of a class interval. The midpoint of a class interval is determined by obtaining the sum of the upper and lower limits of the class interval and dividing by 2. Thus, for example, the midpoint of the class interval 30–39 is found to be $(30 + 39)/2 = 34.5$.

When we group data manually, determining the number of values falling into each class interval is merely a matter of looking at the ordered array and counting the number of observations falling in the various intervals. When we do this for our example, we have Table 2.3.1.

A table such as Table 2.3.1 is called a *frequency distribution*. This table shows the way in which the values of the variable are distributed among the specified class intervals. By consulting it, we can determine the frequency of occurrence of values within any one of the class intervals shown.

Relative Frequencies

It may be useful at times to know the proportion, rather than the number, of values falling within a particular class interval. We obtain this information by dividing the number of values in the particular class interval by the total number of values. If, in our example, we wish to know the proportion of values between 50 and 59, inclusive, we divide 70 by 189, obtaining .3704. Thus we say that 70 out of 189, or 70/189ths, or .3704, of the values are between 50 and 59. Multiplying .3704 by 100 gives us the percentage of values between 50 and 59. We can say, then, that 37.04% of the subjects are between 50 and 59 years of age. We may refer to the proportion of values falling within a class interval as the *relative frequency of occurrence* of values in that interval. In Section 3.2, we shall see that a relative frequency may be interpreted also as the probability of occurrence within the given interval. This probability of occurrence is also called the *experimental probability* or the *empirical probability*.

In determining the frequency of values falling within two or more class intervals, we obtain the sum of the number of values falling within the class intervals of interest. Similarly, if we want to know the relative frequency of occurrence of values falling within two or more class intervals, we add the respective relative frequencies. We may sum, or *cumulate*, the frequencies and relative frequencies to facilitate obtaining information regarding the frequency or relative frequency of values within two or more contiguous class intervals. Table 2.3.2 shows the data of Table 2.3.1 along with the *cumulative frequencies*, the *relative frequencies*, and *cumulative relative frequencies*.

Table 2.3.1 Frequency Distribution of Ages of 189 Subjects Shown in Tables 1.4.1 and 2.2.1

Class Interval	Frequency
30–39	11
40–49	46
50–59	70
60–69	45
70–79	16
80–89	1
Total	189

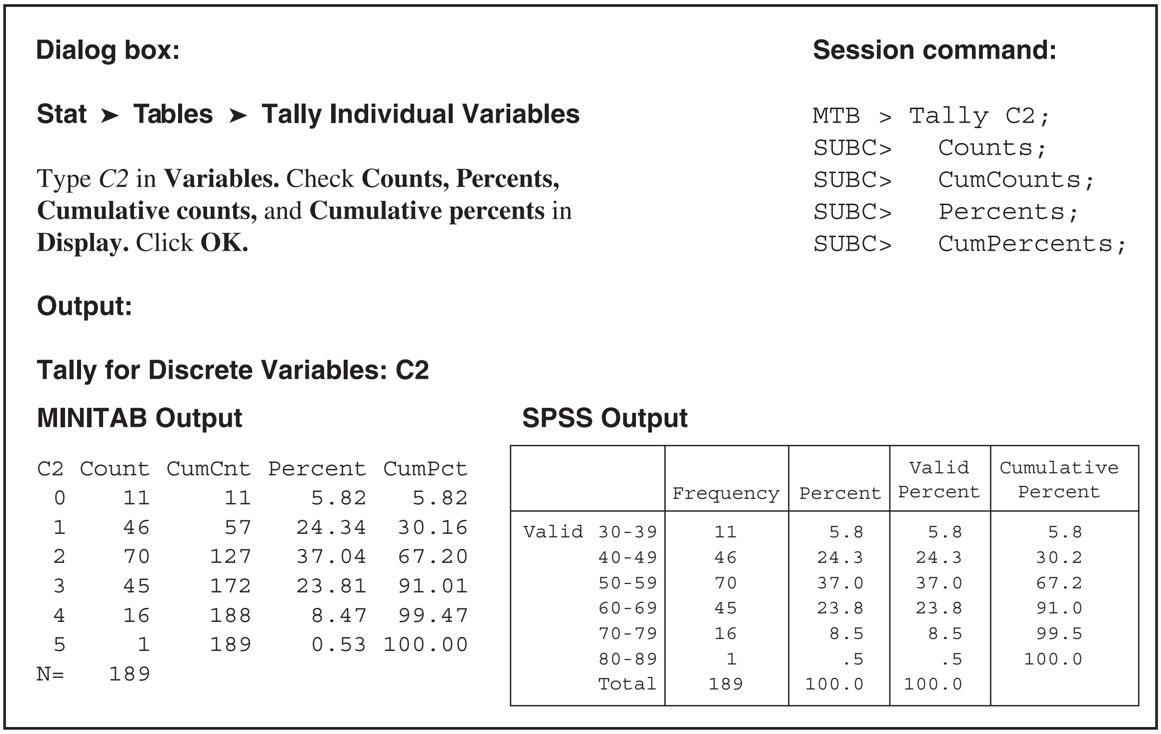
Table 2.3.2 Frequency, Cumulative Frequency, Relative Frequency, and Cumulative Relative Frequency Distributions of the Ages of Subjects Described in Example 1.4.1

Class Interval	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
30–39	11	**11**	(11/189 =) **.0582**	(11/189 =) **.0582**
40–49	46	(11 + 46 =) **57**	(46/189 =) **.2434**	(57/189 = .0582 + .2434 =) **.3016**
50–59	70	(57 + 70 =) **127**	(70/189 =) **.3704**	(127/189 = .3016 + .3704 =) **.6720**
60–69	45	(127 + 45 =) **172**	(45/189 =) **.2381**	(172/189 = .6720 + .2381 =) **.9101**
70–79	16	(172 + 16 =) **188**	(16/189 =) **.0847**	(188/189 = .9101 + .0847 =) **.9948**
80–89	1	(188 + 1 =) **189**	(1/189 =) **.0053**	(189/189 ≈ .9948 + .0053 ≈) **1.0001**
Total	189		1.0001	

Note: Frequencies do not add to 1.0000 exactly because of rounding.

Suppose that we are interested in the relative frequency of values between 50 and 79. We use the cumulative relative frequency column of Table 2.3.2 and subtract .3016 from .9948, obtaining .6932.

We may use a statistical package to obtain a table similar to that shown in Table 2.3.2. Tables obtained from both MINITAB and SPSS software are shown in Figure 2.3.1. Additionally, we may plot our tabled values into various types of graphical displays as shown in Section 2.6.



Dialog box:

Stat ➤ Tables ➤ Tally Individual Variables

Type *C2* in **Variables.** Check **Counts, Percents, Cumulative counts,** and **Cumulative percents** in **Display.** Click **OK.**

Session command:

```
MTB > Tally C2;
SUBC>   Counts;
SUBC>   CumCounts;
SUBC>   Percents;
SUBC>   CumPercents;
```

Output:

Tally for Discrete Variables: C2

MINITAB Output

```
C2  Count  CumCnt  Percent  CumPct
 0     11      11     5.82    5.82
 1     46      57    24.34   30.16
 2     70     127    37.04   67.20
 3     45     172    23.81   91.01
 4     16     188     8.47   99.47
 5      1     189     0.53  100.00
N=    189
```

SPSS Output

```
                                       Valid    Cumulative
                  Frequency  Percent  Percent    Percent
Valid  30-39          11        5.8      5.8        5.8
       40-49          46       24.3     24.3       30.2
       50-59          70       37.0     37.0       67.2
       60-69          45       23.8     23.8       91.0
       70-79          16        8.5      8.5       99.5
       80-89           1         .5       .5      100.0
       Total         189      100.0    100.0
```

FIGURE 2.3.1 Frequency, cumulative frequencies, percent, and cumulative percent distribution of the ages of subjects described in Example 1.4.1 as constructed by MINITAB and SPSS.

Exercises

2.3.1 In a study of the oral home care practice and reasons for seeking dental care among individuals on renal dialysis, Atassi (A-1) studied 90 subjects on renal dialysis. The oral hygiene status of all subjects was examined using a plaque index with a range of 0 to 3 (0 = no soft plaque deposits, 3 = an abundance of soft plaque deposits). The following table shows the plaque index scores for all 90 subjects.

1.17	2.50	2.00	2.33	1.67	1.33
1.17	2.17	2.17	1.33	2.17	2.00
2.17	1.17	2.50	2.00	1.50	1.50
1.00	2.17	2.17	1.67	2.00	2.00
1.33	2.17	2.83	1.50	2.50	2.33
0.33	2.17	1.83	2.00	2.17	2.00
1.00	2.17	2.17	1.33	2.17	2.50
0.83	1.17	2.17	2.50	2.00	2.50
0.50	1.50	2.00	2.00	2.00	2.00
1.17	1.33	1.67	2.17	1.50	2.00
1.67	0.33	1.50	2.17	2.33	2.33
1.17	0.00	1.50	2.33	1.83	2.67
0.83	1.17	1.50	2.17	2.67	1.50
2.00	2.17	1.33	2.00	2.33	2.00
2.17	2.17	2.00	2.17	2.00	2.17

Source: Data provided courtesy of Farhad Atassi, DDS, MSc, FICOI.

(a) Use these data to prepare
A frequency distribution
A relative frequency distribution
A cumulative frequency distribution
A cumulative relative frequency distribution

(b) What percentage of the measurements are less than 2.00?

(c) What proportion of the subjects have measurements greater than or equal to 1.50?

(d) What percentage of the measurements are between 1.50 and 1.99 inclusive?

(e) How many of the measurements are greater than 2.49?

(f) What proportion of the measurements are either less than 1.0 or greater than 2.49?

(g) Someone picks a measurement at random from this data set and asks you to guess the value. What would be your answer? Why?

2.3.2 Janardhan et al. (A-2) conducted a study in which they measured incidental intracranial aneurysms (IIAs) in 125 patients. The researchers examined postprocedural complications and concluded that IIAs can be safely treated without causing mortality and with a lower complications rate than previously reported. The following are the sizes (in millimeters) of the 159 IIAs in the sample:

8.1	10.0	5.0	7.0	10.0	3.0
20.0	4.0	4.0	6.0	6.0	7.0
10.0	4.0	3.0	5.0	6.0	6.0
6.0	6.0	6.0	5.0	4.0	5.0
6.0	25.0	10.0	14.0	6.0	6.0
4.0	15.0	5.0	5.0	8.0	19.0
21.0	8.3	7.0	8.0	5.0	8.0
5.0	7.5	7.0	10.0	15.0	8.0
10.0	3.0	15.0	6.0	10.0	8.0
7.0	5.0	10.0	3.0	7.0	3.3
15.0	5.0	5.0	3.0	7.0	8.0
3.0	6.0	6.0	10.0	15.0	6.0
3.0	3.0	7.0	5.0	4.0	9.2
16.0	7.0	8.0	5.0	10.0	10.0
9.0	5.0	5.0	4.0	8.0	4.0
3.0	4.0	5.0	8.0	30.0	14.0
15.0	2.0	8.0	7.0	12.0	4.0
3.8	10.0	25.0	8.0	9.0	14.0
30.0	2.0	10.0	5.0	5.0	10.0
22.0	5.0	5.0	3.0	4.0	8.0
7.5	5.0	8.0	3.0	5.0	7.0
8.0	5.0	9.0	11.0	2.0	10.0
6.0	5.0	5.0	12.0	9.0	8.0
15.0	18.0	10.0	9.0	5.0	6.0
6.0	8.0	12.0	10.0	5.0	
5.0	16.0	8.0	5.0	8.0	
4.0	16.0	3.0	7.0	13.0	

Source: Data provided courtesy of Vallabh Janardhan, M.D.

(a) Use these data to prepare:
A frequency distribution
A relative frequency distribution
A cumulative frequency distribution
A cumulative relative frequency distribution

(b) What percentage of the measurements are between 10 and 14.9 inclusive?

(c) How many observations are less than 20?

(d) What proportion of the measurements are greater than or equal to 25?

(e) What percentage of the measurements are either less than 10.0 or greater than 19.95?

2.3.3 Hoekema et al. (A-3) studied the craniofacial morphology of patients diagnosed with obstructive sleep apnea syndrome (OSAS) in healthy male subjects. One of the demographic variables the researchers collected for all subjects was the Body Mass Index (calculated by dividing weight in kg by the square of the patient's height in cm). The following are the BMI values of 29 OSAS subjects.

33.57	27.78	40.81
38.34	29.01	47.78
26.86	54.33	28.99
25.21	30.49	27.38
36.42	41.50	29.39
24.54	41.75	44.68
24.49	33.23	47.09
29.07	28.21	42.10
26.54	27.74	33.48
31.44	30.08	

Source: Data provided courtesy of A. Hoekema, D.D.S.

(a) Use these data to construct:
A frequency distribution
A relative frequency distribution
A cumulative frequency distribution
A cumulative relative frequency distribution

(b) What percentage of the measurements are less than 30?

(c) What percentage of the measurements are between 40.0 and 49.99 inclusive?

(d) What percentage of the measurements are greater than 34.99?

(e) How many of the measurements are less than 40?

2.3.4 David Holben (A-4) studied selenium levels in beef raised in a low selenium region of the United States. The goal of the study was to compare selenium levels in the region-raised beef to selenium levels in cooked venison, squirrel, and beef from other regions of the United States. The data below are the selenium levels calculated on a dry weight basis in μg/100 g for a sample of 53 region-raised cattle:

11.23	15.82
29.63	27.74
20.42	22.35
10.12	34.78
39.91	35.09
32.66	32.60
38.38	37.03
36.21	27.00
16.39	44.20
27.44	13.09
17.29	33.03
56.20	9.69
28.94	32.45

20.11	37.38
25.35	34.91
21.77	27.99
31.62	22.36
32.63	22.68
30.31	26.52
46.16	46.01
56.61	38.04
24.47	30.88
29.39	30.04
40.71	25.91
18.52	18.54
27.80	25.51
19.49	

Source: Data provided courtesy of David Holben, Ph.D.

(a) Use these data to construct:
A frequency distribution
A relative frequency distribution
A cumulative frequency distribution
A cumulative relative frequency distribution
(b) How many of the measurements are greater than 40?
(c) What percentage of the measurements are less than 25?

2.3.5 The following table shows the number of hours 45 hospital patients slept following the administration of a certain anesthetic:

7	10	12	4	8	7	3	8	5
12	11	3	8	1	1	13	10	4
4	5	5	8	7	7	3	2	3
8	13	1	7	17	3	4	5	5
3	1	17	10	4	7	7	11	8

From these data construct:

A frequency distribution
A relative frequency distribution

2.3.6 The following are the number of babies born during a year in 60 community hospitals:

30	55	27	45	56	48	45	49	32	57	47	56
37	55	52	34	54	42	32	59	35	46	24	57
32	26	40	28	53	54	29	42	42	54	53	59
39	56	59	58	49	53	30	53	21	34	28	50
52	57	43	46	54	31	22	31	24	24	57	29

From these data construct:

A frequency distribution
A relative frequency distribution

2.3.7 In a study of physical endurance levels of male college freshman, the following composite endurance scores based on several exercise routines were collected:

254	281	192	260	212	179	225	179	181	149
182	210	235	239	258	166	159	223	186	190
180	188	135	233	220	204	219	211	245	151
198	190	151	157	204	238	205	229	191	200
222	187	134	193	264	312	214	227	190	212
165	194	206	193	218	198	241	149	164	225
265	222	264	249	175	205	252	210	178	159
220	201	203	172	234	198	173	187	189	237
272	195	227	230	168	232	217	249	196	223
232	191	175	236	152	258	155	215	197	210
214	278	252	283	205	184	172	228	193	130
218	213	172	159	203	212	117	197	206	198
169	187	204	180	261	236	217	205	212	218
191	124	199	235	139	231	116	182	243	217
251	206	173	236	215	228	183	204	186	134
188	195	240	163	208					

From these data construct:

A frequency distribution
A relative frequency distribution

2.4 Measures of Central Tendency

Although frequency distributions serve useful purposes, there are many situations that require other types of data summarization. What we need in many instances is the ability to summarize the data by means of a single number called a *descriptive measure*. Descriptive measures may be computed from the data of a sample or the data of a population. To distinguish between them, we have the following definitions:

DEFINITIONS

1. A descriptive measure computed from the data of a sample is called a *statistic*. Most often statistics are shown using the standard alphabet (e.g., $\bar{x}$ or s).
2. A descriptive measure computed from the data of a population is called a *parameter*. Most often parameters are shown using the Greek alphabet (e.g., μ or σ).

Several types of descriptive measures can be computed from a set of data. In this chapter, however, we limit discussion to *measures of central tendency* and *measures of dispersion*. We consider measures of central tendency in this section and measures of dispersion in the following one.

In each of the measures of central tendency, of which we discuss three, we have a single value that is considered to be typical of the set of data as a whole. Measures of central tendency convey information regarding the average value of a set of values. As we will see, the word *average* can be defined in different ways.

The three most commonly used measures of central tendency are the *mean*, the *median*, and the *mode*.

Arithmetic Mean

The most familiar measure of central tendency is the arithmetic mean. It is the descriptive measure most people have in mind when they speak of the "average." The adjective *arithmetic* distinguishes this mean from other means that can be computed. Since we are not covering these other means in this book, we shall refer to the arithmetic mean simply as the *mean*. The mean is obtained by adding all the values in a population or sample and dividing by the number of values that are added.

EXAMPLE 2.4.1

We wish to obtain the mean age of the population of 189 subjects represented in Table 1.4.1.

SOLUTION: We proceed as follows:

$$\text{mean age} = \frac{48 + 35 + 46 + \cdots + 73 + 66}{189} = 55.032$$

The three dots in the numerator represent the values we did not show in order to save space.

General Formula for the Mean

It will be convenient if we can generalize the procedure for obtaining the mean and, also, represent the procedure in a more compact notational form. Let us begin by designating the random variable of interest by the capital letter X. In our present illustration we let X represent the random variable, age. Specific values of a random variable will be designated by the lowercase letter x. To distinguish one value from another, we attach a subscript to the x and let the subscript refer to the first, the second, the third value, and so on. For example, from Table 1.4.1 we have

$$x_1 = 48, x_2 = 35, \ldots, x_{189} = 66$$

In general, a typical value of a random variable will be designated by x_i and the final value, in a finite population of values, by x_N, where N is the number of values in the population. Finally, we will use the Greek letter μ to stand for the population mean. We may now write the general formula for a finite population mean as follows:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \tag{2.4.1}$$

The symbol $\sum_{i=1}^{N}$ instructs us to add all values of the variable from the first to the last. This symbol $\sum$, called the *summation sign*, will be used extensively in this book. When from the context it is obvious which values are to be added, the symbols above and below $\sum$ will be omitted.

The Sample Mean

When we compute the mean for a sample of values, the procedure just outlined is followed with some modifications in notation. We use $\bar{x}$ to designate the sample mean and n to indicate the number of values in the sample. The sample mean then is expressed as

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.4.2}$$

EXAMPLE 2.4.2

The Centers for Disease Control and Prevention collect influenza vaccination records for health-care workers. According to their 2016–2017 reports, the percentage of vaccinated health-care workers for select western states were: AZ (88.9), CO (97.1), NV (79.0), ID (92.0), CA (83.9), OR (81.2), NM (86.2) and WA (89.8). We wish to calculate the sample mean percentage of vaccinated health-care workers in this region.

SOLUTION: We label our data points as $x_1 = 88.9$, $x_2 = 97.1$, … , $x_7 = 89.8$ and apply Formula 2.4.2.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{88.9 + 97.1 + \cdots + 89.8}{8} = 87.3\%$$

Properties of the Mean

The arithmetic mean possesses certain properties, some desirable and some not so desirable. These properties include the following:

1. Uniqueness. For a given set of data, there is one and only one arithmetic mean.
2. Simplicity. The arithmetic mean is easily understood and easy to compute.
3. Since each and every value in a set of data enters into the computation of the mean, it is affected by each value. Extreme values, therefore, have an influence on the mean and, in some cases, can so distort it that it becomes undesirable as a measure of central tendency.

As an example of how extreme values may affect the mean, consider the following situation. Suppose the five physicians who practice in an area are surveyed to determine their charges for a certain procedure. Assume that they report these charges: \$75, \$75, \$80, \$80, and \$280. The mean charge for the five physicians is found to be \$118, a value that is not very representative of the set of data as a whole. The single atypical value had the effect of inflating the mean.

Median

The median of a finite set of values is that value which divides the set into two equal parts such that the number of values equal to or greater than the median is equal to the number of values equal to or less than the median. If the number of values is odd, the median will be the middle value when all values have been arranged in order of magnitude. When the number of values is even, there is no single middle value. Instead there are two middle values. In this case, the median is taken to be the mean of these two middle values, when all values have been arranged in the order of their magnitudes. In other words, the median observation of a data set is the $(n + 1)/2$th one when the observation has been ordered. If, for example, we have 11 observations, the median is the $(11 + 1)/2 = 6$th ordered observation. If we have 12 observations, the median is the $(12 + 1)/2 = 6.5$th ordered observation and is a value halfway between the 6th and 7th ordered observations.

EXAMPLE 2.4.3

Let us illustrate by finding the median of the data in Table 2.2.1.

SOLUTION: The values are already ordered so we need only to find the two middle values. The middle value is the $(n + 1)/2 = (189 + 1)/2 = 190/2 = 95$th one. Counting from the smallest up to the 95th value, we see that it is 54. Thus the median age of the 189 subjects is 54 years.

EXAMPLE 2.4.4

We wish to find the median age of the subjects represented in the sample described in Example 2.4.2.

SOLUTION: Arraying the 8 states in order of magnitude from smallest to largest gives 79.0, 81.2, 83.9, 86.2, 88.9, 89.8, 92.0, 97.1. Since we have an even number of states, there is no middle value. The two middle values, however, are 86.2 and 88.9. The median, then, is $(86.2 + 88.9)/2 = 87.6$.

Properties of the Median

Properties of the median include the following:

1. Uniqueness. As is true with the mean, there is only one median for a given set of data.
2. Simplicity. The median is easy to calculate.
3. It is not as drastically affected by extreme values as is the mean.

The Mode

The mode of a set of values is that value which occurs most frequently. If all the values are different, there is no mode; on the other hand, a set of values may have more than one mode.

EXAMPLE 2.4.5

Find the modal age of the subjects whose ages are given in Table 2.2.1.

SOLUTION: A count of the ages in Table 2.2.1 reveals that the age 53 occurs most frequently (17 times). The mode for this population of ages is 53.

For an example of a set of values that has more than one mode, let us consider a laboratory with 10 employees whose ages are 20, 21, 20, 20, 34, 22, 24, 27, 27, and 27. We could say that these data have two modes, 20 and 27. The sample consisting of the values 10, 21, 33, 53, and 54 has no mode since all the values are different.

The mode may be used also for describing qualitative data. For example, suppose the patients seen in a mental health clinic during a given year received one of the following diagnoses: mental retardation, organic brain syndrome, psychosis, neurosis, and personality disorder. The diagnosis occurring most frequently in the group of patients would be called the modal diagnosis.

An attractive property of a data distribution occurs when the mean, median, and mode are all equal. The well-known "bell-shaped curve" is a graphical representation of a distribution for which the mean, median, and mode are all equal. Much statistical inference is based on this distribution, the most common of which is the normal distribution. The normal distribution is introduced in Section 4.6 and discussed further in subsequent chapters. Another common distribution of this type is the *t*-distribution, which is introduced in Section 6.3.

Skewness

Data distributions may be classified on the basis of whether they are symmetric or asymmetric. If a distribution is symmetric, the left half of its graph (see Section 2.6 for graphing details) will be a mirror image of its right half. When the left half and right half of the graph of a distribution are not mirror images of each other, the distribution is asymmetric. Skewness can also be visualized by examining frequency tables as described in Section 2.3.

DEFINITION

If the graph (histogram or frequency polygon) of a distribution is asymmetric, the distribution is said to be *skewed*. If a distribution is not symmetric because its graph extends further to the right than to the left, that is, if it has a long tail to the right, we say that the distribution is *skewed to the right* or is *positively skewed*. If a distribution is not symmetric because its graph extends further to the left than to the right, that is, if it has a long tail to the left, we say that the distribution is *skewed to the left* or is *negatively skewed*.

A distribution will be skewed to the right, or positively skewed, if its mean is greater than its mode. A distribution will be skewed to the left, or negatively skewed, if its mean is less than its mode. Skewness can be expressed as follows:

$$\mathit{Skewness} = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{3/2}} = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{(n-1)\sqrt{n-1}s^3} \tag{2.4.3}$$

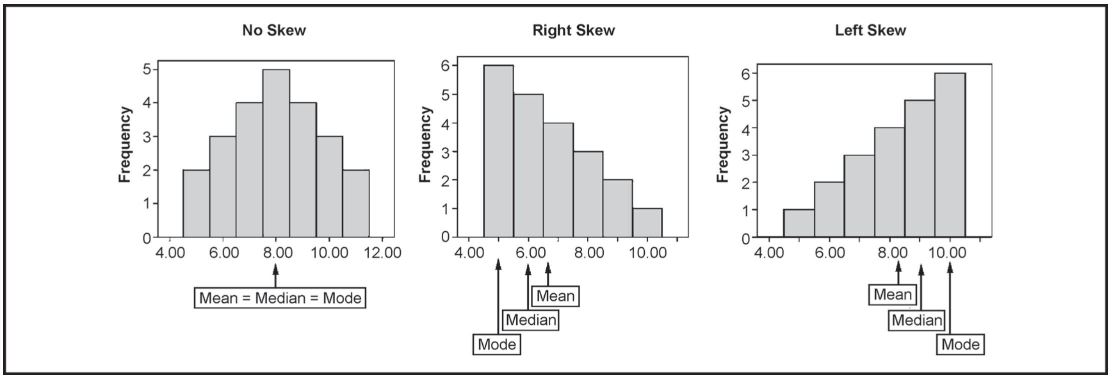


FIGURE 2.4.1 Three histograms illustrating skewness.

In Equation 2.4.3, s is the standard deviation of a sample as defined in Equation 2.5.4. Most computer statistical packages include this statistic as part of a standard printout. A value of skewness > 0 indicates positive skewness and a value of skewness < 0 indicates negative skewness. An illustration of skewness is shown in Figure 2.4.1.

EXAMPLE 2.4.6

Consider the three distributions shown in Figure 2.4.1. These were produced using a statistical package and are further described in Section 2.6. For example, observation of the "No Skew" distribution would yield the following data: 5, 5, 6, 6, 6, 7, 7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 9, 10, 10, 10, 11, 11. Values can be obtained from the skewed distributions in a similar fashion. Using SPSS software, the following descriptive statistics were obtained for these three distributions:

	No Skew	Right Skew	Left Skew
Mean	8.0000	6.6667	8.3333
Median	8.0000	6.0000	9.0000
Mode	8.00	5.00	10.00
Skewness	.000	.627	−.627

2.5 Measures of Dispersion

The *dispersion* of a set of observations refers to the variety that they exhibit. A measure of dispersion conveys information regarding the amount of variability present in a set of data. If all the values are the same, there is no dispersion; if they are not all the same, dispersion is present in the data. The amount of dispersion may be small when the values, though different, are close together. Figure 2.5.1 shows the frequency polygons for two populations that have equal means but different amounts of variability. Population B, which is more variable than population A, is more spread out. If the values are widely scattered, the dispersion is greater. Other terms used synonymously with dispersion include *variation, spread*, and *scatter*.