WAYNE W. DANIEL • CHAD L. CROSS

BIOSTATISTICS
A Foundation for Analysis in the Health Sciences

Eleventh Edition

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 Quad Graphics.

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

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

Contents

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