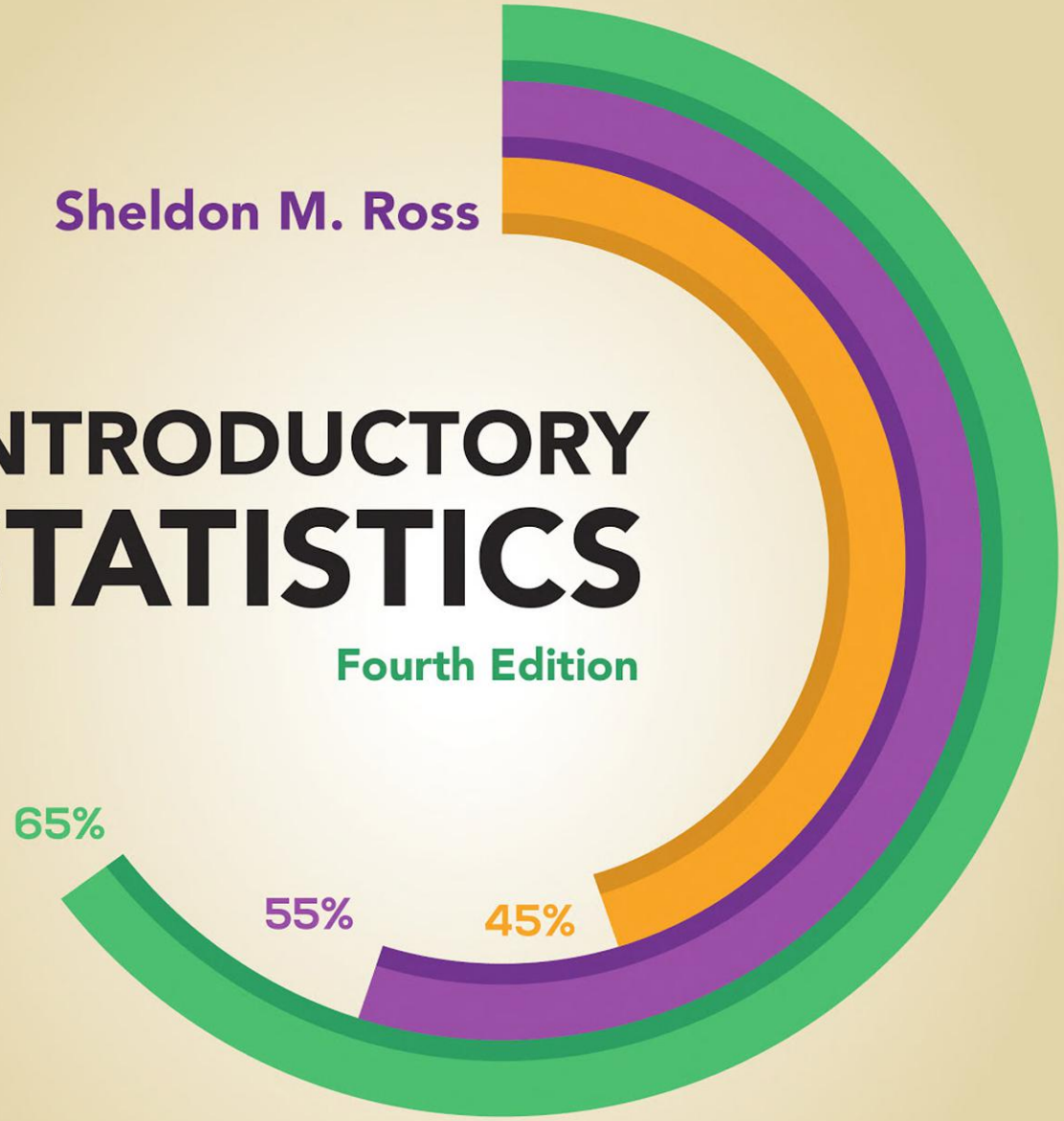
Sheldon M. Ross

INTRODUCTORY STATISTICS

Fourth Edition



Introductory Statistics

Introductory Statistics

Fourth Edition

Sheldon M. Ross

University of Southern California

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, United Kingdom
525 B Street, Suite 1800, San Diego, CA 92101-4495, United States
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-0-12-804317-2

Publisher: Nikki Levy
Acquisition Editor: Graham Nisbet
Editorial Project Manager: Susan Ikeda
Production Project Manager: Paul Prasad Chandramohan
Designer: Maria Ines Cruz

Typeset by VTeX

About the Author

Sheldon M. Ross

Sheldon M. Ross received his Ph.D. in Statistics at Stanford University in 1968 and then joined the Department of Industrial Engineering and Operations Research at the University of California at Berkeley. He remained at Berkeley until Fall 2004, when he became the Daniel J. Epstein Professor of Industrial and Systems Engineering in the Daniel J. Epstein Department of Industrial and Systems Engineering at the University of Southern California. He has published many technical articles and textbooks in the areas of statistics and applied probability. Among his texts are *A First Course in Probability* (ninth edition), *Introduction to Probability Models* (eleventh edition), *Simulation* (fifth edition), and *Introduction to Probability and Statistics for Engineers and Scientists* (fifth edition).

Professor Ross is the founding and continuing editor of the journal *Probability in the Engineering and Informational Sciences*. He is a fellow of the Institute of Mathematical Statistics, the Institute for Operations Research and Management Sciences, and a recipient of the Humboldt U.S. Senior Scientist Award.

For Rebecca and Elise

Contents

Preface

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.

H. G. Wells (1866–1946)

In today's complicated world, very few issues are clear-cut and without controversy. In order to understand and form an opinion about an issue, one must usually gather information, or data. To learn from data, one must know something about statistics, which is the art of learning from data.

This introductory statistics text is written for college-level students in any field of study. It can be used in a quarter, semester, or full-year course. Its only prerequisite is high school algebra. Our goal in writing it is to present statistical concepts and techniques in a manner that will teach students not only how and when to utilize the statistical procedures developed, but also to understand why these procedures should be used. As a result we have made a great effort to explain the ideas behind the statistical concepts and techniques presented. Concepts are motivated, illustrated, and explained in a way that attempts to increase one's intuition. It is only when a student develops a feel or intuition for statistics that she or he is really on the path toward making sense of data.

To illustrate the diverse applications of statistics and to offer students different perspectives about the use of statistics, we have provided a wide variety of text examples and problems to be worked by students. Most refer to real-world issues, such as gun control, stock price models, health issues, driving age limits, school admission ages, public policy issues, gender issues, use of helmets, sports, disputed authorship, scientific fraud, and Vitamin C, among many others. Many of them use data that not only are real but are themselves of interest. The examples have been posed in a clear and concise manner and include many thought-provoking problems that emphasize thinking and problem-solving skills. In addition, some of the problems are designed to be open-ended and can be used as starting points for term projects.

SOME SPECIAL FEATURES OF THE TEXT

Introduction The first numbered section of each chapter is an introduction that poses a realistic statistical situation to help students gain perspective on what they will encounter in the chapter.

Statistics in Perspective Statistics in Perspective highlights are placed throughout the book to illustrate real-world application of statistical techniques and concepts. These perspectives are designed to help students analyze and interpret data while utilizing proper statistical techniques and methodology.

Real Data Throughout the text discussions, examples, perspective highlights, and problems, real data sets are used to enhance the students' understanding of the material. These data sets provide information for the study of current issues in a variety of disciplines, such as health, medicine, sports, business, and education.

Historical Perspectives These enrichment sections profile prominent statisticians and historical events, giving students an understanding of how the discipline of statistics has evolved.

Problems/Review Problems This text includes hundreds of exercises placed at the end of each section within a chapter, as well as more comprehensive review problems at the end of each chapter. Many of these problems utilize real data and are designed to assess the students' conceptual as well as computational understanding of the material. Selected problems are open-ended and offer excellent opportunity for extended discussion, group activities, or student projects.

Summary/Key Terms An end-of-chapter summary provides a detailed review of important concepts and formulas covered in the chapter. Key terms and their definitions are listed that serve as a working glossary within each chapter.

Formula Summary Important tables and formulas that students often refer to and utilize are included on the inside front and back covers of the book. These can serve as a quick reference when doing homework or studying for an exam.

Program CD-ROM A CD-ROM is provided with each volume that includes programs that can be used to solve basic statistical computation problems. Please refer to Appendix E for a listing of these programs.

THE TEXT

In Chap. 1 we introduce the subject matter of statistics and present its two branches. The first of these, called descriptive statistics, is concerned with the

collection, description, and summarization of data. The second branch, called inferential statistics, deals with the drawing of conclusions from data.

Chapters 2 and 3 are concerned with descriptive statistics. In Chap. 2 we discuss tabular and graphical methods of presenting a set of data. We see that an effective presentation of a data set can often reveal certain of its essential features. Chap. 3 shows how to summarize certain features of a data set.

In order to be able to draw conclusions from data it is necessary to have some understanding of what they represent. For instance, it is often assumed that the data constitute a "random sample from some population." In order to understand exactly what this and similar phrases signify, it is necessary to have some understanding of probability, and that is the subject of Chap. 4. The study of probability is often a troublesome issue in an introductory statistics class because many students find it a difficult subject. As a result, certain textbooks have chosen to downplay its importance and present it in a rather cursory style. We have chosen a different approach and attempted to concentrate on its essential features and to present them in a clear and easily understood manner. Thus, we have briefly but carefully dealt with the concept of the events of an experiment, the properties of the probabilities that are assigned to the events, and the idea of conditional probability and independence. Our study of probability is continued in Chap. 5, where discrete random variables are introduced, and in Chap. 6, which deals with the normal and other continuous random variables.

Chapter 7 is concerned with the probability distributions of sampling statistics. In this chapter we learn why the normal distribution is of such importance in statistics.

Chapter 8 deals with the problem of using data to estimate certain parameters of interest. For instance, we might want to estimate the proportion of people who are presently in favor of congressional term limits. Two types of estimators are studied. The first of these estimates the quantity of interest with a single number (for instance, it might estimate that 52 percent of the voting population favors term limits). The second type provides an estimator in the form of an interval of values (for instance, it might estimate that between 49 and 55 percent of the voting population favors term limits).

Chapter 9 introduces the important topic of statistical hypothesis testing, which is concerned with using data to test the plausibility of a specified hypothesis. For instance, such a test might reject the hypothesis that over 60 percent of the voting population favors term limits. The concept of p value, which measures the degree of plausibility of the hypothesis after the data have been observed, is introduced.

Whereas the tests in Chap. 9 deal with a single population, the ones in Chap. 10 relate to two separate populations. For instance, we might be inter-

ested in testing whether the proportions of men and of women that favor term limits are the same.

Probably the most widely used statistical inference technique is that of the analysis of variance; this is introduced in Chap. 11. This technique allows us to test inferences about parameters that are affected by many different factors. Both one- and two-factor analysis of variance problems are considered in this chapter.

In Chap. 12 we learn about linear regression and how it can be used to relate the value of one variable (say, the height of a man) to that of another (the height of his father). The concept of regression to the mean is discussed, and the regression fallacy is introduced and carefully explained. We also learn about the relation between regression and correlation. Also, in an optional section, we use regression to the mean along with the central limit theorem to present a simple, original argument to explain why biological data sets often appear to be normally distributed.

In Chap. 13 we present goodness-of-fit tests, which can be used to test whether a proposed model is consistent with data. This chapter also considers populations classified according to two characteristics and shows how to test whether the characteristics of a randomly chosen member of the population are independent.

Chapter 14 deals with nonparametric hypothesis tests, which are tests that can be used in situations where the ones of earlier chapters are inappropriate.

Chapter 15 introduces the subject matter of quality control, a key statistical technique in manufacturing and production processes.

Chapter 16 deals with the topics of machine learning and big data. The techniques described have become popular in recent years due to the preponderance of large amounts of data. A general problem considered is to determine the probability that a cross country flight will be late, with the flight defined by a characterizing vector giving such information as the airline, the departure airport, the arrival airport, the time of departure, and the weather conditions. We consider a variety of estimation procedures, with names like naive Bayes and distance based approaches. We then consider what are known as bandit problems, and which can be applied, among other things, to sequentially choosing among different medications for treating a particular medical condition.

NEW TO THIS EDITION

The fourth edition has many new and updated examples and exercises. In addition, are the following:

1. A new section (Section 3.8) on Lorenz Curves and the Gini Index. Lorenz curves are plots, as p ranges from 0 to 1, of the fraction of the total in-

come earned by all members of a population that is earned by the $100p$ percent lowest paid workers. We show that the more this curve is below the straight line connecting $(0, 0)$ to $(1, 1)$ the greater is the inequality in the incomes of the population (where equality is said to occur when all workers earn the same amount). The Gini index, which is commonly used to measure that inequality, is presented.

2. A new optional Subsection 3.8.1 on the 80–20 and Pareto rules.
3. Material on Benford's Law for first digits.
4. A new Subsection 5.4.2 dealing with finding the expectation of a function of a random variable, with an example (Example 5.18) on the friendship paradox.
5. Section 12.12 on Logistic Regression.
6. Chapter 16 on Machine Learning and Big Data.
7. Illustration throughout of how to use the statistical package R to do the necessary computations. Although the text's Program CD Rom can be used to solve the problems in the text, we highly recommend that students download the free statistical software package R. To install R on your computer

   Go to http://cran.cnr.berkeley.edu
   Hit the button Download R for your type of computer
   Install the PKG file that came in the download

   R is very simple to use and, in the latter parts of relevant chapters, we illustrate its use as it relates to that chapter.

Acknowledgments

We would like to thank the reviewers of the fourth edition that asked to remain anonymous. In addition, we wish to thank the following reviewers of earlier editions for their many helpful comments: William H. Beyer, University of Akron; Patricia Buchanan, Pennsylvania State University; Michael Eurgubian, Santa Rosa Junior College; Larry Griffey, Florida Community College, Jacksonville; Katherine T. Halvorsen, Smith College, James E. Holstein, University of Missouri; James Householder, Humboldt State University; Robert Lacher, South Dakota State University; Margaret Lin, University of California Berkeley; Jacinta Mann, Seton Hill College; C. J. Park, San Diego State University; Liam O'Brien, Colby College; Erol Pekoz, Boston University; Ronald Pierce, Eastern Kentucky University; Lawrence Riddle, Agnes Scott College; Gaspard T. Rizzuto, University of Southwestern Louisiana; Jim Robison-Cox, Montana State University; Walter Rosenkrantz, University of Massachusetts, Amherst; Bruce Sisko, Belleville Area College; Glen Swindle, University of California, Santa Barbara; Paul Vetrano, Santa Rose Junior College; Joseph J. Walker, Georgia State University; Deborah White, College of the Redwoods; and Cathleen Zucco, LeMoyne College.

Sheldon M. Ross

CHAPTER 1

Introduction to Statistics

> Statisticians have already overrun every branch of science with a rapidity of conquest rivalled only by Attila, Mohammed, and the Colorado beetle.
>
> **Maurice Kendall (British statistician)**

This chapter introduces the subject matter of statistics, the art of learning from data. It describes the two branches of statistics, descriptive and inferential. The idea of learning about a population by sampling and studying certain of its members is discussed. Some history is presented.

CONTENTS

1.1 INTRODUCTION

Is it better for children to start school at a younger or older age? This is certainly a question of interest to many parents as well as to people who set public policy. How can we answer it?

It is reasonable to start by thinking about this question, relating it to your own experiences, and talking it over with friends. However, if you want to convince others and obtain a consensus, it is then necessary to gather some objective information. For instance, in many states, achievement tests are given to children at the end of their first year in school. The children's results on these tests can be obtained and then analyzed to see whether there appears to be a connection between children's ages at school entrance and their scores on the test. In fact, such studies have been done, and they have generally concluded that older student entrants have, as a group, fared better than younger entrants. However, it has also been noted that the reason for this may just be that those students who entered at an older age would be older at the time of the examination, and this by itself may be what is responsible for their higher scores. For instance, suppose parents did not send their 6-year-olds to school but rather waited an additional year. Then, since these children will probably learn a great deal at home in that year, they will probably score higher when they take the test at the end of their first year of school than they would have if they had started school at age 6.

Introductory Statistics. DOI:10.1016/B978-0-12-804317-2.00001-1

Table 1.1 Total Years in School Related to Starting Age

Year	Younger half of children		Older half of children	
	Average age on starting school	Average number of years completed	Average age on starting school	Average number of years completed
1946	6.38	13.84	6.62	13.67
1947	6.34	13.80	6.59	13.86
1948	6.31	13.78	6.56	13.79
1949	6.29	13.77	6.54	13.78
1950	6.24	13.68	6.53	13.68
1951	6.18	13.63	6.45	13.65
1952	6.08	13.49	6.37	13.53

Source: J. Angrist and A. Krueger, "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association*, vol. 87, no. 18, 1992, pp. 328–336.

A recent study (Table 1.1) has attempted to improve upon earlier work by examining the effect of children's age upon entering school on the eventual number of years of school completed. These authors argue that the total number of years spent in school is a better measure of school success than is a score on an achievement test taken in an early grade. Using 1960 and 1980 census data, they concluded that the age at which a child enters school has very little effect on the total number of years that a child spends in school. Table 1.1 is an abridgment of one presented in their work. The table indicates that for children beginning school in 1949, the younger half (whose average entrance age was 6.29 years) spent an average of 13.77 years, and the older half an average of 13.78 years, in school.

Note that we have not presented the preceding in order to make the case that the ages at which children enter school do not affect their performance in school. Rather we are using it to indicate the modern approach to learning about a complicated question. Namely, one must collect relevant information, or *data*, and these data must then be described and analyzed. Such is the subject matter of statistics.

1.2 THE NATURE OF STATISTICS

It has become a truism in today's world that in order to learn about something, you must first collect data. For instance, the first step in learning about such things as

1. The present state of the economy
2. The percentage of the voting public who favors a certain proposition
3. The average miles per gallon of a newly developed automobile
4. The efficacy of a new drug

5. The usefulness of a new way of teaching reading to children in elementary school

is to collect relevant data.

Definition. Statistics *is the art of learning from data. It is concerned with the collection of data, their subsequent description, and their analysis, which often leads to the drawing of conclusions.*

1.2.1 Data Collection

Sometimes a statistical analysis begins with a given set of data; for instance, the government regularly collects and publicizes data about such quantities as the unemployment rate and the gross domestic product. Statistics would then be used to describe, summarize, and analyze these data.

In other situations, data are not yet available, and statistics can be utilized to design an appropriate experiment to generate data. The experiment chosen should depend on the use that one wants to make of the data. For instance, if a cholesterol-lowering drug has just been developed and its efficacy needs to be determined, volunteers will be recruited and their cholesterol levels noted. They will then be given the drug for some period, and their levels will be measured again. However, it would be an ineffective experiment if *all* the volunteers were given the drug. For if this were so, then even if the cholesterol levels of all the volunteers were significantly reduced, we would not be justified in concluding that the improvements were due to the drug used and not to some other possibility. For instance, it is a well-documented fact that any medication received by a patient, whether or not it is directly related to that patient's suffering, will often lead to an improvement in the patient's condition. This is the *placebo effect*, which is not as surprising as it might seem at first, since a patient's belief that she or he is being effectively treated often leads to a reduction in stress, which can result in an improved state of health. In addition, there might have been other—usually unknown—factors that played a role in the reduction of cholesterol levels. Perhaps the weather was unusually warm (or cold), causing the volunteers to spend more or less time outdoors than usual, and this was a factor. Thus, we see that the experiment that calls for giving the drug to all the volunteers is not well designed for generating data from which we can learn about the efficacy of that drug.

A better experiment is one that tries to neutralize all other possible causes of the change of cholesterol level except the drug. The accepted way of accomplishing this is to divide the volunteers into two groups; then one group receives the drug, and the other group receives a tablet (known as a *placebo*) that looks and tastes like the drug but has no physiological effect. The volunteers should not know whether they are receiving the true drug or the placebo, and indeed it is best if the medical people overseeing the experiment also do not know, so their own biases will not play a role. In addition, we want the division of the volunteers into the two groups to be done such that neither of the

groups is favored in that it tends to have the "better" patients. The accepted best approach for arranging this is to break up the volunteers "at random," where by this term we mean that the breakup is done in such a manner that all possible choices of people in the group receiving the drug are equally likely. The group that does not receive any treatment (that is, the volunteers that receive a placebo) is called the *control* group.

At the end of the experiment, the data should be described. For instance, the before and after cholesterol levels of each volunteer should be presented, and the experimenter should note whether the volunteer received the drug or the placebo. In addition, summary measures such as the average reduction in cholesterol of members of the control group and members of the drug group should be determined.

Definition. *The part of statistics concerned with the description and summarization of data is called* descriptive statistics.

1.2.2 Inferential Statistics and Probability Models

When the experiment is completed and the data are described and summarized, we hope to be able to draw a conclusion about the efficacy of the drug. For instance, can we conclude that it is effective in reducing blood cholesterol levels?

Definition. *The part of statistics concerned with the drawing of conclusions from data is called* inferential statistics.

To be able to draw a conclusion from the data, we must take into account the possibility of chance. For instance, suppose that the average reduction in cholesterol is lower for the group receiving the drug than for the control group. Can we conclude that this result is due to the drug? Or is it possible that the drug is really ineffective and that the improvement was just a chance occurrence? For instance, the fact that a coin comes up heads 7 times in 10 flips does not necessarily mean that the coin is more likely to come up heads than tails in future flips. Indeed, it could be a perfectly ordinary coin that, by chance, just happened to land heads 7 times out of the total of 10 flips. (On the other hand, if the coin had landed heads 47 times out of 50 flips, then we would be quite certain that it was not an ordinary coin.)

To be able to draw logical conclusions from data, it is usually necessary to make some assumptions about the chances (or *probabilities*) of obtaining the different data values. The totality of these assumptions is referred to as a *probability model* for the data.

Sometimes the nature of the data suggests the form of the probability model that is assumed. For instance, suppose the data consist of the responses of a selected group of individuals to a question about whether they are in favor of a senator's welfare reform proposal. Provided that this group was *randomly*

selected, it is reasonable to suppose that each individual queried was in favor of the proposal with probability p, where p represents the unknown proportion of all citizens in favor of the proposal. The resultant data can then be used to make inferences about p.

In other situations, the appropriate probability model for a given data set will not be readily apparent. However, a careful description and presentation of the data sometimes enable us to infer a reasonable model, which we can then try to verify with the use of additional data.

Since the basis of statistical inference is the formulation of a probability model to describe the data, an understanding of statistical inference requires some knowledge of the theory of probability. In other words, statistical inference starts with the assumption that important aspects of the phenomenon under study can be described in terms of probabilities, and then it draws conclusions by using data to make inferences about these probabilities.

1.3 POPULATIONS AND SAMPLES

In statistics, we are interested in obtaining information about a total collection of elements, which we will refer to as the *population*. The population is often too large for us to examine each of its members. For instance, we might have all the residents of a given state, or all the television sets produced in the last year by a particular manufacturer, or all the households in a given community. In such cases, we try to learn about the population by choosing and then examining a subgroup of its elements. This subgroup of a population is called a *sample*.

Definition. *The total collection of all the elements that we are interested in is called a* population.

A subgroup of the population that will be studied in detail is called a sample.

In order for the sample to be informative about the total population, it must be, in some sense, representative of that population. For instance, suppose that we are interested in learning about the age distribution of people residing in a given city, and we obtain the ages of the first 100 people to enter the town library. If the average age of these 100 people is 46.2 years, are we justified in concluding that this is approximately the average age of the entire population? Probably not, for we could certainly argue that the sample chosen in this case is not representative of the total population because usually more young students and senior citizens use the library than do working-age citizens. Note that *representative* does not mean that the age distribution of people in the sample is exactly that of the total population, but rather that the sample was chosen in such a way that all parts of the population had an equal chance to be included in the sample.

In certain situations, such as the library illustration, we are presented with a sample and must then decide whether this sample is reasonably representative of the entire population. In practice, a given sample generally cannot be considered to be representative of a population unless that sample has been chosen in a random manner. This is because any specific nonrandom rule for selecting a sample often results in one that is inherently biased toward some data values as opposed to others.

Definition. *A sample of k members of a population is said to be a* random sample, *sometimes called a* simple random sample, *if the members are chosen in such a way that all possible choices of the k members are equally likely.*

Thus, although it may seem paradoxical, we are most likely to obtain a representative sample by choosing its members in a totally random fashion without any prior considerations of the elements that will be chosen. In other words, we need not attempt to deliberately choose the sample so that it contains, for instance, the same gender percentage and the same percentage of people in each profession as found in the general population. Rather, we should just leave it up to "chance" to obtain roughly the correct percentages. The actual mechanics of choosing a random sample involve the use of random numbers and will be presented in App. C.

Once a random sample is chosen, we can use statistical inference to draw conclusions about the entire population by studying the elements of the sample.

*1.3.1 Stratified Random Sampling[1]

A more sophisticated approach to sampling than simple random sampling is the *stratified random sampling* approach. This approach, which requires more initial information about the population than does simple random sampling, can be explained as follows. Consider a high school that contains 300 students in the first-year class, 500 in the second-year class, and 600 each in the third- and fourth-year classes. Suppose that in order to learn about the students' feelings concerning a military draft for 18-year-olds, an in-depth interview of 100 students will be done. Rather than randomly choosing 100 people from the 2000 students, in a stratified sample one calculates how many to choose from each class. Since the proportion of students who are first-year is $300/2000 = 0.15$, in a stratified sample the percentage is the same and thus there are $100 \times 0.15 = 15$ first-year students in the sample. Similarly, one selects $100 \times 0.25 = 25$ second-year students and $100 \times 0.30 = 30$ third-year and 30 fourth-year students. Then one selects students from each class at random.

In other words, in this type of sample, first the population is *stratified* into subpopulations, and then the correct number of elements is randomly chosen from each of the subpopulations. As a result, the proportions of the sample

[1] The asterisk * signifies optional material not used in the sequel.