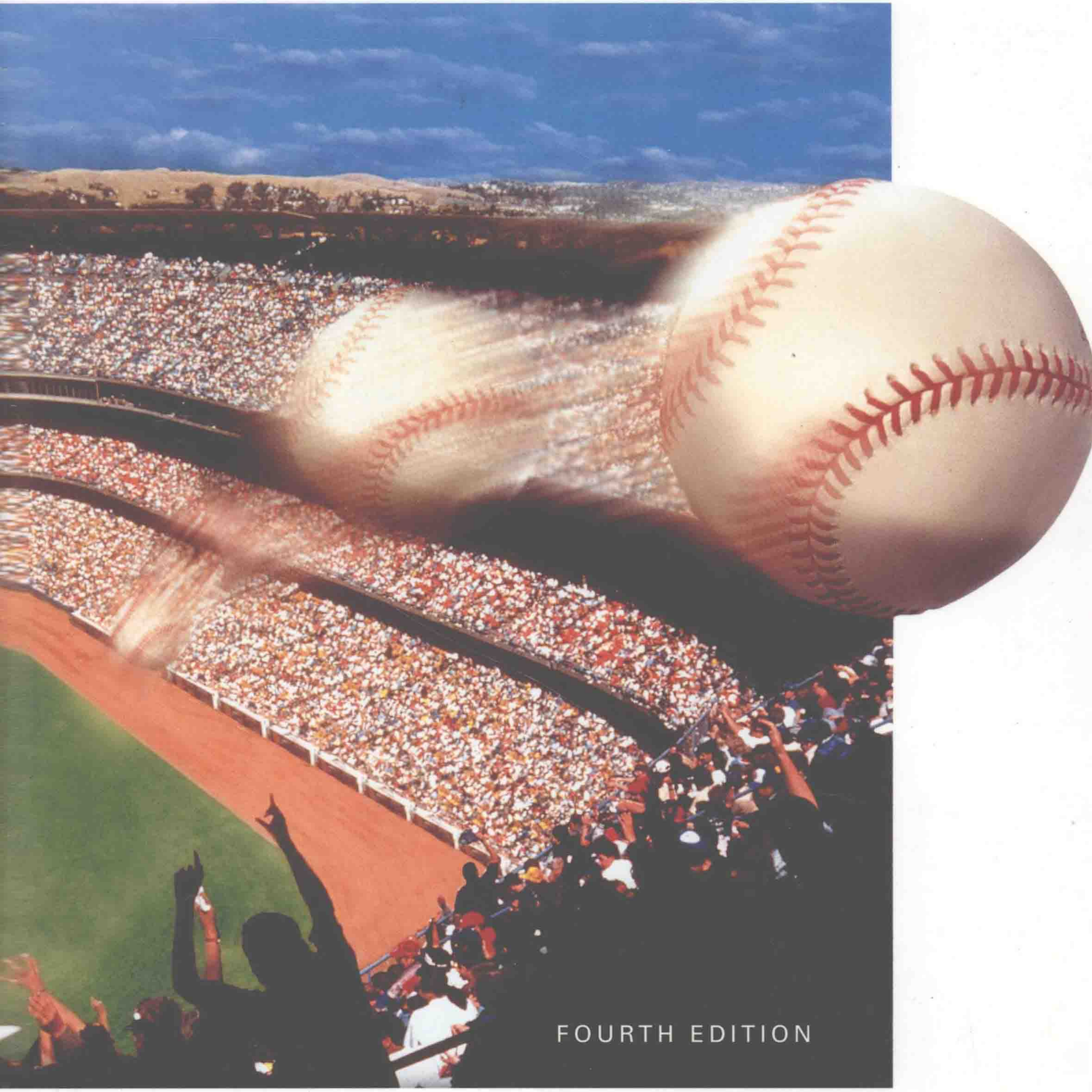
General Statistics



FOURTH EDITION

Warren Chase
Fred Bown

FOURTH EDITION

GENERAL STATISTICS

WARREN CHASE

FRED BOWN
Framingham State College

JOHN WILEY & SONS, INC.
New York / Chichester / Weinheim / Brisbane / Singapore / Toronto

ACQUISITIONS EDITOR *Brad Wiley II/Debbie Berridge*
MARKETING MANAGER *Peter Ryttel*
FULL SERVICE MANAGER *Jeanine Furino*
ILLUSTRATION EDITOR *Sigmund Malinowski*
PRODUCTION SERVICES *Publication Services*
COVER DESIGN *Carol C. Grobe*

This book was set in Times Roman by Publication Services and printed and bound by R.R. Donnelley & Sons, Inc. The cover was printed by Phoenix Color Corp.

This book is printed on acid-free paper. ♾

ISBN 0-471-28310-X

Printed in the United States of America

10 9 8 7 6 5 4 3 2

PREFACE

The purpose of this book is to present a first course in statistics appropriate for students in a wide variety of disciplines, the only prerequisite being a knowledge of basic high school algebra. We feel that the objective of such a course should be to acquaint the student with the basic ideas of descriptive and inferential statistics. Great care has been taken in writing the book to make the subject matter understandable. All technical terms are defined in easy-to-grasp language; definitions, formulas, and summaries of statistical tests are set off in boxes for quick reference. Concepts are introduced and reinforced with examples and exercises from a range of fields from sports to medicine. All chapters begin with an introduction and end with a summary of the important ideas of the chapter. As in the third edition of the book, most exercises on hypothesis testing don't specify whether the traditional approach (finding the critical region) or the P-value approach should be used. We give answers in terms of both approaches. The instructor can then decided which approach to use. Also, in the fourth edition we continue to encourage the student to use exploratory data analysis techniques (especially with small samples) to determine whether various procedures are appropriate.

New Features of the Fourth Edition

- Mathematics educators generally agree that the topic students find most difficult is probability. In the third edition we addressed this impediment by making some of the sections in the probability chapter optional. In the fourth edition we've gone a step further: **the entire chapter on probability is optional.** For those who choose to skip this chapter, the next chapter (on discrete probability distributions) has a very brief and straightforward introductory section on probability. We have found that students have no difficulty with this section. For those who choose to cover the probability chapter, this introductory section may be skipped.
- In the third edition there was one rather large data set in the appendix: data on 1000 subjects of the Framingham Heart Study. Most chapters had a "Working with Data" section with problems involving this data set. At the request of users of the book, we have added more data sets to the appendix. There are now 11 data sets (which are also available on floppy disk). Problems in the "Working with Data" sections use these data sets.
- Structural changes include streamlining of the material where possible. For example, Chapter 2 (Organizing Data) and Chapter 3 (Describing Data) in the third edition have been combined in one chapter in the fourth edition. Also, the Minitab sections occurring at the end of most chapters have been moved to the instructor's manual. Appendix C in the third edition (which covered inference involving variances) has also been moved to the instructor's manual. Inference problems concerning variances that rely on chi-square or F tests require that the populations be normal. Since departures from normality can render the test inappropriate, most statisticians avoid using them.
- Over 100 new problems have been added. In some cases these problems have replaced old ones. All new problems involve real data. Also, many of the problems

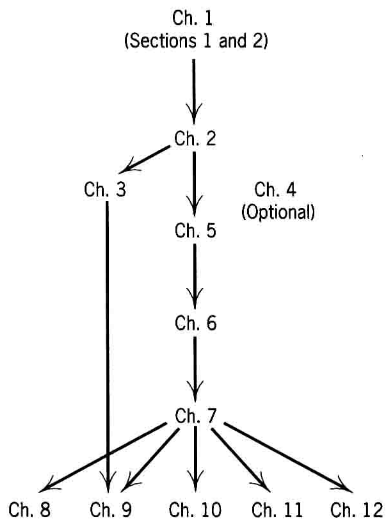


FIGURE 1 Prerequisite Chart: The Chapter at the Tail of an Arrow Is a Prerequisite for the Chapter at the Head of the Arrow

that were even-numbered in the third edition are odd-numbered in the fourth edition, and vice versa. Therefore, to users of the third edition who tend to assign odd-numbered problems, the problem sets in the fourth edition will seem quite new.

- We have included more computer output in examples and exercises in the fourth edition. The objective is that the student will be able to understand computer output from a variety of statistical packages. Included is output from Minitab, SPSS, Data Desk, JMP, Resampling Stats, Maple V, and Mathematica.

Chapters 1–3 and 5–7 would be suitable for the core of a three-semester-hour introductory course, and Chapters 1–3 and 5–9 would form the core of a four-semester-hour course. In each case, additional material could be covered. After Chapter 7, the remaining chapters can be covered in any order. (See the prerequisite chart in Figure 1.)

Supplements

- An *Instructor's Manual,* which includes complete solutions for all exercises, chapters on the Poisson distribution, the hypergeometric distribution, multiple regression, inference concerning population variance(s), a brief Minitab manual, and a list of answers to the even-numbered exercises. Any portion of the manual may be reproduced for class use. At the request of the instructor, Wiley will shrink-wrap any portion of the *Instructor's Manual* with the book.
- A *Student Solutions Manual* with complete solutions to all odd-numbered exercises.
- A *Test Bank* with true-false, fill-in-the-blank, multiple-choice, and computational exercises. This is available in hard copy and on a floppy disk that may be edited to include the instructor's test items.
- An *Excel Manual and Workbook,* with Microsoft Excel instructions, simulations, and experiments corresponding to the chapters in the text.

- A Data Disk in ASCII file format, available from the publisher for each adopter of this text. The floppy disk contains 11 data sets, including a data set on 1000 randomly selected subjects from the Framingham Heart Study. These data sets are also printed in Appendix C of the book.
- A Web site, www.wiley.com/college/chase, from which the Data Disk can be downloaded.

Acknowledgments

The authors wish to thank Mack Hill of Worcester State College, who worked through all the examples and exercises and checked the manuscript of the fourth edition for accuracy. He made many valuable suggestions. Thanks are also in order to Brad Wiley II, Mary O'Sullivan, and Deborah Berridge of John Wiley & Sons, Inc. for their patience, encouragement, and valuable suggestions.

CONTENTS

CHAPTER 1 *INTRODUCTION* **1**

1.1 The Nature of Statistics **1**
1.2 Sampling **5**
1.3 Topics in the Design of Experiments **11**
1.4 Summary **16**
Sources **17**

PART 1

DESCRIPTIVE STATISTICS **19**

CHAPTER 2 *ORGANIZING AND DESCRIBING DATA* **21**

2.1 Introduction **21**
2.2 Summarizing Data **22**
2.3 Graphic Representations **30**
2.4 The Shape of a Distribution **39**
2.5 Stem-and-Leaf Plots **40**
2.6 Misleading Graphs (Optional) **46**
2.7 Measures of Central Tendency **49**
2.8 Measures of Dispersion **58**
2.9 Percentiles, Quartiles, and the Interquartile Range **67**
2.10 Boxplots **74**
2.11 Working with Data (Optional) **81**
2.12 Summary **83**
Review Exercises **84**
Sources **94**

CHAPTER 3 *DESCRIPTIVE METHODS FOR REGRESSION AND CORRELATION* **96**

3.1 Introduction **96**
3.2 The Least Squares Regression Line **97**
3.3 The Linear Correlation Coefficient **120**
3.4 Some Words of Caution Concerning Correlation and Regression **137**
3.5 Working with Data (Optional) **137**
3.6 Summary **139**
Review Exercises **139**
Sources **145**

PART 2
PROBABILITY DISTRIBUTIONS **147**

CHAPTER 4 *PROBABILITY (OPTIONAL)* **149**

4.1 Introduction **149**
4.2 Sample Spaces and Events **150**
4.3 The Probability of an Event **155**
4.4 Compound Events **160**
4.5 More on Compound Events **169**
4.6 Combinatorics **177**
4.7 Summary **182**
Review Exercises **183**
Sources **188**

CHAPTER 5 *PROBABILITY DISTRIBUTIONS FOR DISCRETE RANDOM VARIABLES* **189**

5.1 Introduction to Probability (This section may be skipped if the first four sections of Chapter 4 have been covered.) **189**
5.2 Random Variables **194**
5.3 Discrete Probability Distributions **197**
5.4 Mean and Variance **202**
5.5 The Binomial Probability Distribution **210**
5.6 Summary **220**
Review Exercises **221**
Sources **224**

CHAPTER 6 *PROBABILITY DISTRIBUTIONS FOR CONTINUOUS RANDOM VARIABLES; THE NORMAL DISTRIBUTION* **226**

6.1 Introduction **226**
6.2 Continuous Probability Distributions **227**
6.3 The Normal Distribution **232**
6.4 The Standard Normal Distribution **233**
6.5 More on Normal Probability **241**
6.6 Normal Approximation to the Binomial Distribution **252**
6.7 The Central Limit Theorem **258**
6.8 Summary **268**
Review Exercises **269**
Sources **272**

PART 3
STATISTICAL INFERENCE **273**

CHAPTER 7 *STATISTICAL INFERENCE CONCERNING MEANS AND PROPORTIONS* **275**

7.1 Introduction **275**
7.2 Estimating a Population Mean (Large-Sample Case) **276**
7.3 Hypothesis Testing Concerning a Population Mean (Large-Sample Case) **287**
7.4 *P*-Values **301**

7.5 Inference Concerning a Population Mean (Small-Sample Case) **306**
7.6 Inference Concerning a Population Proportion **326**
7.7 Working with Data (Optional) **339**
7.8 Summary **340**
Review Exercises **342**
Sources **346**

CHAPTER 8 *INFERENCE CONCERNING TWO POPULATION PARAMETERS* **347**

8.1 Introduction **347**
8.2 Inference Concerning Two Population Means: Dependent Samples **348**
8.3 Inference Concerning Two Population Means Based on Independent Samples: Large-Sample Case **357**
8.4 Inference Concerning Two Population Means Based on Independent Samples: Small-Sample Case **363**
8.5 Inference Concerning Two Population Proportions **375**
8.6 Working with Data (Optional) **382**
8.7 Summary **385**
Review Exercises **385**
Sources **388**

CHAPTER 9 *INFERENCE CONCERNING REGRESSION AND CORRELATION* **390**

9.1 Introduction **390**
9.2 The Linear Regression Model **390**
9.3 Inference Concerning Slope **392**
9.4 Inference Concerning Correlation **401**
9.5 Checking Properties of the Regression Model **409**
9.6 Prediction Intervals and Confidence Intervals (Optional) **414**
9.7 Working with Data (Optional) **420**
9.8 Summary **421**
Review Exercises **422**
Sources **426**

CHAPTER 10 *ANALYSIS OF VARIANCE* **427**

10.1 Introduction **427**
10.2 The Idea Behind Analysis of Variance: The Case of Equal Sample Sizes **428**
10.3 Analysis of Variance When Sample Sizes Are Not Necessarily Equal. Conventional Terminology **441**
10.4 Alternative Formulas (Optional) **454**
10.5 Working with Data (Optional) **456**
10.6 Summary **457**
Review Exercises **458**
Sources **463**

CHAPTER 11 *ANALYSIS OF CATEGORICAL DATA* **464**

11.1 Introduction **464**
11.2 The Chi-Square Test for Goodness-of-Fit **465**

11.3 Tests of Independence; Contingency Tables **474**
11.4 Tests of Homogeneity **484**
11.5 Working with Data (Optional) **488**
11.6 Summary **489**
Review Exercises **489**
Sources **491**

CHAPTER 12 *NONPARAMETRIC STATISTICS* **492**

12.1 Introduction **492**
12.2 The Sign Test **493**
12.3 The Wilcoxon Signed-Rank test **499**
12.4 The Mann-Whitney *U* Test **511**
12.5 The Runs Test **520**
12.6 A Discussion of Parametric Versus Nonparametric Tests **524**
12.7 Working with Data (Optional) **525**
12.8 Summary **525**
Review Exercises **527**

APPENDIX A *USE OF THE RANDOM NUMBER TABLE* **A-1**

APPENDIX B *TABLES* **A-3**

APPENDIX C *DATA SETS* **A-27**

APPENDIX D *BIBLIOGRAPHY* **A-49**

ANSWERS **ANS-1**

INDEX **I-1**

CHAPTER 1

INTRODUCTION

1.1 THE NATURE OF STATISTICS

1.2 SAMPLING

1.3 TOPICS IN THE DESIGN OF EXPERIMENTS

1.4 SUMMARY

SOURCES

1.1 THE NATURE OF STATISTICS

To get an idea of what statistics is all about, we consider a few concrete illustrations of the kinds of problems that involve statistics.

- A pollster interested in the outcome of an upcoming election interviews a certain number of voters and, based on the results obtained, makes a prediction as to who will probably win the election.
- The Environmental Protection Agency conducts tests on a certain number of cars of the same make and model to estimate the average gas mileage for all cars of that make and model.
- Each year the Federal Bureau of Investigation publishes the *Uniform Crime Reports.* Among other things, this document reports the violent crime rate (number per 100,000 people) for each of the Metropolitan Statistical Areas in the United States.
- To get an idea of the economic status of a town, an economist obtains the salaries of all wage earners in the town and then computes the average.
- To estimate the average age of all adult Americans, a sociologist selects 1000 adults, computes their average age, and then uses this figure as an estimate.

All of these illustrations make use of the raw material of statistics, namely, **data** (or **data values**), also called **observations.** We use these terms in a very broad sense. That is, a data value is simply a piece of information that might be numerical, such as the annual snowfall in Boston, or a person's weight or age. Or the information might be nonnumerical, such as the color of a car, a person's ethnic status, or whom you favor in the next presidential election. We give the following definition of **statistics:**

Definition *Statistics* is the science of collecting, simplifying, and describing data, as well as making inferences (drawing conclusions) based on the analysis of data.

As the definition suggests, there are two branches of statistics. The branch that deals with collecting, simplifying, and giving properties of data is called **descriptive statistics.** An important objective of descriptive statistics is to organize, summarize, or describe the data to make it more comprehensible. For example, suppose that we obtained a list of the salaries of all the wage earners in Boston. This list would be so long that it would be incomprehensible. But if we were to find the average of all these salaries, then we would understand something about the economic status of the residents of Boston.

The other branch of statistics, which involves drawing conclusions based on the analysis of data, is called **inferential statistics.** The pollster who predicts the outcome of an election based on a knowledge of only *some* of the votes, and the sociologist who estimates the average age of *all* adult Americans based on a knowledge of the average of the ages of *some* of these adults, are both using inferential statistics. Apparently, it was impractical for these researchers to obtain all the data they were interested in (i.e., all the votes or all the ages); therefore, in both cases a judgment was made about the larger body of data that was being studied by means of information obtained from only some of these data values. This leads to the following definition:

Definition The entire collection of all the elements we are interested in is called a *population.* (These elements might be people, automobiles, data values, etc.) A collection of some of the elements obtained from the population is called a *sample* from the population.

In an investigation such as a voter preference study, we may think of the population as consisting of all the voters or all the votes they will cast. The votes are the **data values** of interest, and it is these values that we are really investigating. In general, we are ultimately interested in data values in statistics. For this reason, in this book we will often think of populations as consisting of data values.

Now that we have introduced the terms *population* and *sample,* we can give a more precise definition of inferential statistics.

Definition *Inferential statistics* is concerned with making judgments (or inferences) about a population based on the properties of some sample obtained from the population.

We said that trying to estimate the average age of all adult Americans by using the average of the ages of some of these adults is a typical problem in inferential statistics. The average age for the population of all adult Americans is an example of what is called a **parameter.** The average of the ages for a sample of adult Americans is an example of a **statistic.**

Definition A numerical property of a population is called a *parameter.* A numerical property of a sample is called a *statistic.* (By numerical property we mean a property that can be expressed as a number.)

Consider the population of all voters in the 1936 presidential election. The percentage of the voters that were for Roosevelt is a parameter. Viewing the voters in Peoria, Illinois, as a sample from this population, we see that the percentage in this sample that were for Roosevelt is a statistic.

Many problems in inferential statistics involve estimating the value of, or making some decision concerning, a parameter based on the value of a statistic.

The methods used in making inferences in statistics are probabilistic. For example, suppose that a pollster interviews 100 voters selected by chance and finds that 96 of them favor candidate A in an upcoming election. Then the pollster would say that the evidence points to candidate A winning the election, because it would be highly unlikely or improbable that so many voters (in the sample of 100 voters) would be in favor of candidate A if candidate A were not going to win the election. **Probability, which deals with the laws of chance, plays an important role in statistics.** Therefore, we will study this topic in some detail in this book.

EXERCISES

In Exercises 1.1 and 1.2, discuss similarities and differences between the following terms.

1.1 Population and sample

1.2 Parameter and statistic

In Exercises 1.3–1.5, which are true?

1.3 The value of a statistic always remains unchanged under repeated sampling.

1.4 The value of a parameter always remains unchanged under repeated sampling.

1.5 Often, the value of a parameter is unknown.

In Exercises 1.6–1.15, determine whether the results given are examples of descriptive or inferential statistics.

1.6 In the 1996 presidential election, voters in Massachusetts cast 1,571,763 votes for Bill Clinton, 718,107 for Bob Dole, and 227,217 for H. Ross Perot (*Source* [1], p. 98).

1.7 As of January, 1997, Nielsen Media Research estimated various percentages for United States homes (*Source* [1], p. 746). Percentages with

Color TV sets	85%	Black and white only	1%
Two or more sets	73%	One set	27%
Cable	67.2%		

1.8 The 1997 population of the United States was estimated to be 267,368,000, and the 1997 median age was estimated to be 34.9 years (*Source* [1], p. 823). (*Note:* A median age of 34.9 years means that 50% of the population is estimated to be younger than 34.9 years of age.)

1.9 In the 1996 presidential election, Bill Clinton, Bob Dole, and H. Ross Perot received 379, 159, and 0 electoral votes, respectively (*Source* [1], p. 98).

1.10 Massachusetts Institute of Technology Professor Richard Larson studies the physics and psychology of queues. He estimates that people spend an average of 30 minutes a day in lines (*Source* [2]).

1.11 A Boston-based nurses health study of 80,000 nurses concluded that the key to reducing the risk of a heart attack depends more on the type of fat than the amount of fat that you eat. The study suggests that by halving the intake of transfats from the current average of 4% to 2% of total calories consumed, the risk of heart attack and coronary deaths could be slashed by 42% (*Source* [3], p. A1).

1.12 Researchers from the University of Texas Southwestern Medical Center in Dallas conducted a study on youngsters with symptoms of severe depression. The 8-week study of 96 children aged 7–17 years showed that 74% of those given Prozac had improved compared with 58% of those given a placebo (*Source* [3], p. A3).

1.13 Researchers at the Center for Employment Futures (CEF) at Drexel University in Philadelphia conducted a study of 2001 15- to 17-year-olds, and 91% of the teenagers envisioned themselves working in full-time jobs with a five-day work week (*Source* [4]).

1.14 It was reported that gold mutual funds are down an average of 44% this year (*Source* [2]).

1.15 A postmaster reported that his office had delivered 8453 packages during the month of January. He also estimated that his office would deliver 125,400 packages during the year.

1.16 A linguist was interested in the population of words in James Joyce's *Ulysses*. The word *the* occurs 14,887 times (*Source* [5], p. 143). Is 14,887 the value of a parameter or a statistic?

1.17 In the latest national survey of more than 348,000 college freshmen by the American Council on Education and the University of California at Los Angeles, 36% said they were frequently bored in class during the senior year in high school (*Source* [6], p. A4). Is 36 the value of a parameter or a statistic?

1.18 A national organization of personnel managers has estimated that about 25% of all résumés contain a major fabrication (*Source* [7], p. A14). Is 25 the value of a parameter or a statistic?

1.19 Unaware that 35% of the 10,000 voters in his district still support him, a politician decides to estimate his political strength. A sample of 200 voters shows that 40% support him.

(a) What is the population?

(b) What is the value of the parameter of interest?

(c) What is the value of the statistic of interest?

(d) Compare your answers in (b) and (c). Is it surprising that they are different? If the politician were to sample another 200 voters, which of the two numbers would most likely change? Explain.

1.20 Consider the problem of estimating the average grade point average (GPA) of the 750 seniors at a college. (The average, which is unknown, is the sum of 750 GPAs divided by 750.)

(a) What is the population? How many data values are in the population?

(b) What is the parameter of interest?

(c) Suppose that a sample of 10 seniors is selected, and their GPAs are 2.72, 2.81, 2.65, 2.69, 3.17, 2.74, 2.57, 2.17, 3.48, 3.10. Calculate a statistic that you would use to estimate the parameter.

(d) Suppose that another sample of 10 seniors was selected. Would it be likely that the value of the statistic is the same as in part (c)? Why or why not? Would the value of the parameter remain the same?

1.21 A sociologist was interested in estimating some aspects of family life in a town. Information about the entire town (unknown to the sociologist) and the results of a sample obtained by the sociologist follow.

Number of children per family	Number of families in town	Number of families in sample
0	120	6
1	180	10
2	270	12
3	300	8
4	80	6
5	50	8

(a) Identify the elements of the population and give the population size.

(b) Identify the elements of the sample and give the sample size.

(c) Suppose that the sociologist were interested in the *number* of families with more than two children.

i. Calculate a statistic to estimate the parameter of interest. (*Hint:* Assume that the sociologist knows that the population is 20 times the size of the sample.)

ii. What is the value of the parameter?

(d) Suppose that the sociologist were interested in the *proportion* of families with less than four children.

i. What is the value of the statistic?

ii. What is the value of the parameter?

1.22 In the Massachusetts State Lottery Megabucks game, six numbers are selected from the set of 42 numbers $\{1, 2, 3, 4, \ldots, 39, 40, 41, 42\}$. A player thought that the percentage of single-digit numbers being selected was too high. He obtained a partial list of past drawings and observed that 26.4% of the numbers were single digit.

(a) If the numbers were generated randomly, what would the value of the parameter of interest be?

(b) What is the value of the statistic?

1.2 SAMPLING

We said that in inferential statistics we use samples to make judgments about populations. We want the samples we obtain to be **representative** of the population, that is, to resemble the population. There are many ways of obtaining samples; we mention only a few here.

Random Samples

Random sampling is one of the most important types of sampling in statistics.

Definition A *random sample* is a sample obtained from the population in such a manner that all samples of the same size have equal likelihood of being selected. Any method of obtaining random samples is called *random sampling.*

For example, one method of random sampling is the **lottery method.** With this method, elements in the population are identified by a name or number written on a tag. The tags are placed in a container and are then well mixed. A tag is drawn by chance from the container, and this process is repeated until the desired number of tags is obtained.

When the elements in the population are identified by numbers (such as employee identification numbers for employees of a large corporation), we may use the **random number method** to obtain random samples. Random numbers are found in Appendix Table B.1. Instructions on how to use the table are found in Appendix A. Suppose there are 9000 employees with ID numbers 0001, 0002, 0003, and so on, up to 9000. We can use the random number table to obtain a random sample of, say, five employees, by choosing a sequence of five ID numbers that occurred in a purely random manner. When the sampling is done in such a way that there are no repetitions in the sample (for example, by ignoring ID numbers that repeat), we call the resulting sample a **simple random sample** (SRS).

Stratified Samples

Another type of sample frequently encountered is a **stratified sample.**

Definition If the population is divided into subgroups, and we take a random sample from each, the resulting sample is a *stratified sample.* In stratified sampling, these subgroups are called *strata*. The members of a stratum usually share some characteristic, such as race or income level.

For example, the students at a college could be divided into strata according to class: freshmen, sophomores, juniors, and seniors. To obtain a stratified sample, we could take a random sample from each class. In stratified sampling, the size of the sample from each stratum is often proportional to the size of the stratum in the population. For example, suppose 40% of the student body are freshmen, 25% sophomores, 20% juniors, and 15% seniors. To obtain a stratified sample of size 100, we could sample 40 freshmen, 25 sophomores, 20 juniors, and 15 seniors. The result would be a **proportional stratified sample.**

Stratified samples are often easier to obtain than true random samples. For example, if we wanted a sample of Massachusetts voters, it would be inconvenient to compile a list or computer file of all voters from which to select a random sample. However, we could sample voters from each voting precinct, thereby obtaining a stratified sample.

EXAMPLE 1.1

An engineering firm has a professional staff of 20 employees with ID numbers 01, 02, ..., 20, and a support staff of 30 employees with ID numbers 21, 22, ..., 50. Using the first column of Appendix Table B.1 (i.e., 10, 37, 08, ...),

(a) Give a simple random sample of five ID numbers.
(b) Using the professional staff and support staff as two strata, give a proportional stratified sample of five ID numbers.

Solution

(a) Moving down the first column of Appendix Table B.1, we list the first five two-digit numbers that are legitimate ID numbers (i.e., falling in the range 01 to 50). Those are 10, 37, 08, 12, 31.
(b) Twenty out of 50, or 40%, are professional staff, and 30 out of 50, or 60%, are support staff. Thus, we should have two professionals (40% of 5) and three support staff (60% of 5) in our sample. Starting at the top of the first column, pick two IDs in the range 01 to 20. These are 10, 08. Start at the top again and pick three IDs in the range 21 to 50. These are 37, 31, 44. So our sample is 10, 08, 37, 31, 44. ■

Cluster Samples

Definition Suppose the population is divided into subgroups. If we select some of these subgroups and take a sample from each, the resulting combined sample is a *cluster sample.* Sometimes all the members of the selected subgroups are included in the sample.

In stratified sampling, the strata are frequently quite different from one another but are homogeneous internally. In cluster sampling, however, the subgroups are often defined in such a way that they are as internally diverse as possible, yet easy to access by the interviewer. This could be accomplished by defining the subgroups as compact geographical regions, saving the interviewer's travel time.

In practice, most large surveys, such as the Current Population Survey (CPS) of the Census Bureau, use **multistage cluster sampling.** To obtain a multistage cluster sample of U.S. households, we could start with a list of all U.S. counties. These are called **primary sampling units** (PSUs). The procedure is as follows:

- Take a sample of PSUs.
- Take a sample of towns in each PSU.
- Take a sample of streets in each town.
- Take a sample of households on each street.

At each stage, simple random sampling could be used.

The multistage cluster sample used for the Current Population Survey differs from this scheme in a number of respects. For example, the PSUs are placed in strata based on various criteria, such as size. A stratified sample of the PSUs is then obtained. Also, instead of streets, the CPS uses enumeration districts (EDs) selected from a list by a procedure called **systematic sampling,** to be discussed next.

Systematic Samples

If we have a list of the elements in the population, an easy way to obtain a sample is by **systematic sampling.** For example, from a list of all employees of a corporation, we could choose by chance some starting point on the list and then select perhaps every fifth name on the list. The starting point could be chosen by selecting an ID number from the random number table.

> **Definition** From a list of members of a population, choose a starting point by chance and then select every nth element on the list (for some appropriate value of n). The result is a *systematic sample.*

For example, suppose you have a list of 100 ID numbers and you want a sample of size 12. Notice that $100/12 \doteq 8.33$ (where $\doteq$ means "is approximately equal to"). Round to get 8. Randomly select a starting point. Then select every eighth ID number until you have 12 numbers. You could use the random number table to select a starting ID number. If the starting point was 004, the next number would be 012.

The samples discussed thus far involve a chance process and fall under the general heading of **probability samples.** However, occasionally we encounter samples that are not probability samples. An example of this is a **sample of convenience.**

Samples of Convenience

An instructor who was teaching two sections of statistics wanted to compare two books, so one book was used in one section and the other book was used in the other section. The two statistics sections are examples of **samples of convenience.**