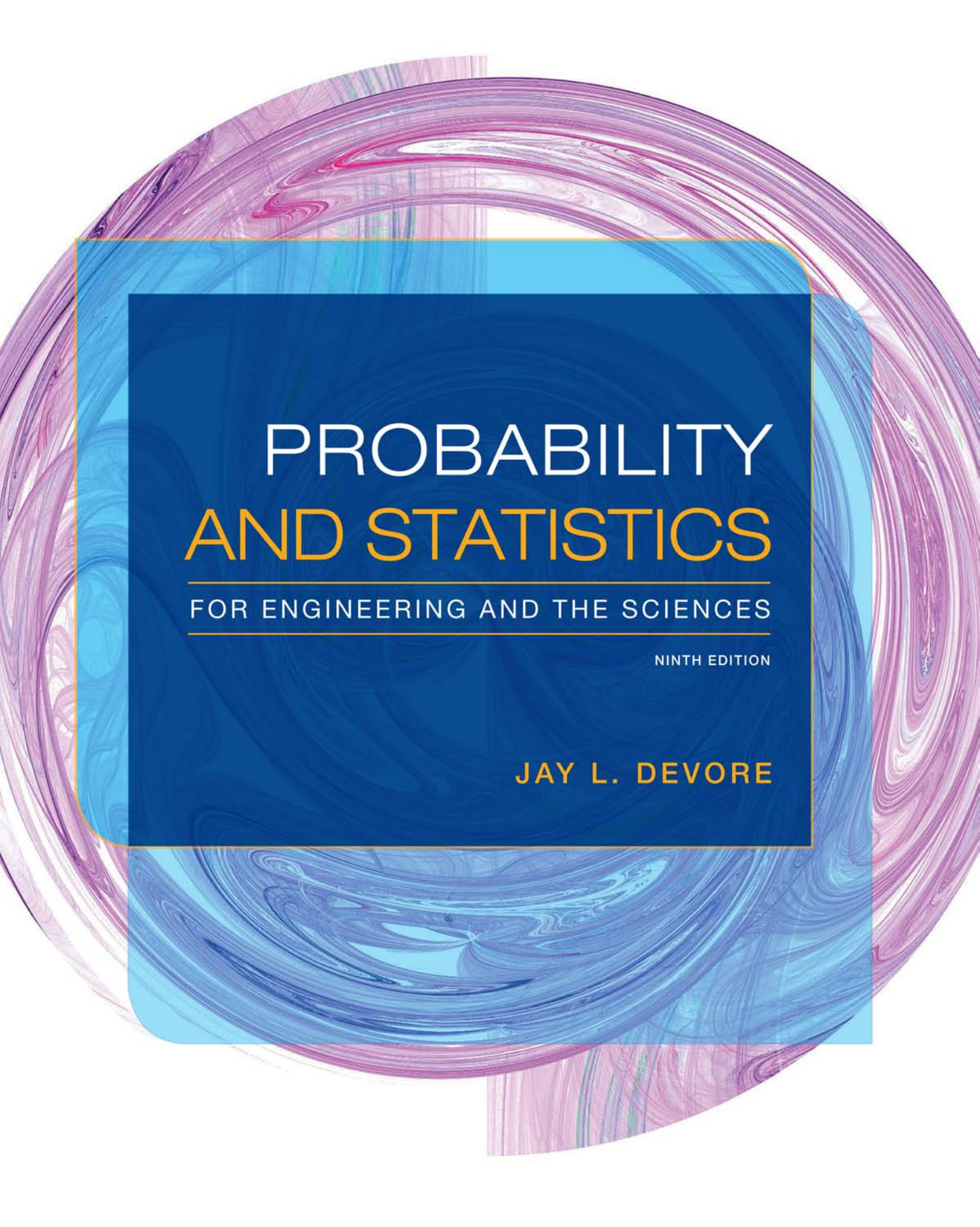

PROBABILITY
AND STATISTICS
FOR ENGINEERING AND THE SCIENCES
NINTH EDITION
JAY L. DEVORE

5 REASONS

to buy your textbooks and course materials at

CENGAGEbrain.com

1 SAVINGS:
Prices up to 75% off, daily coupons, and free shipping on orders over $25

2 CHOICE:
Multiple format options including textbook, eBook and eChapter rentals

3 CONVENIENCE:
Anytime, anywhere access of eBooks or eChapters via mobile devices

4 SERVICE:
Free eBook access while your text ships, and instant access to online homework products

5 STUDY TOOLS:
Study tools* for your text, plus writing, research, career and job search resources
**availability varies*

Find your course materials and start saving at:
www.cengagebrain.com

Source Code: 14M-AA0107

Engaged with you.
www.cengage.com

CENGAGE Learning®

NINTH EDITION

Probability and Statistics for Engineering and the Sciences

JAY DEVORE

California Polytechnic State University, San Luis Obispo

CENGAGE Learning®

Australia • Brazil • Mexico • Singapore • United Kingdom • United States

CENGAGE Learning®

***Probability and Statistics for Engineering and the Sciences*, Ninth Edition**
Jay L. Devore

Senior Product Team Manager: Richard Stratton
Senior Product Manager: Molly Taylor
Senior Content Developer: Jay Campbell
Product Assistant: Spencer Arritt
Media Developer: Andrew Coppola
Marketing Manager: Julie Schuster
Content Project Manager: Cathy Brooks
Art Director: Linda May
Manufacturing Planner: Sandee Milewski
IP Analyst: Christina Ciaramella
IP Project Manager: Farah Fard
Production Service and Compositor: MPS Limited
Text and Cover Designer: C Miller Design

Library of Congress Control Number: 2014946237

ISBN: 978-1-305-25180-9

Cengage Learning
20 Channel Center Street
Boston, MA 02210
USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil, and Japan. Locate your local office at **www.cengage.com/global**.

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

To learn more about Cengage Learning Solutions, **visit www.cengage.com**.

Purchase any of our products at your local college store or at our preferred online store **www.cengagebrain.com**.

Printed in the United States of America
Print Number: 01 Print Year: 2014

To my beloved grandsons
Philip and Elliot, who are
highly statistically significant.

Contents

Preface

Purpose

The use of probability models and statistical methods for analyzing data has become common practice in virtually all scientific disciplines. This book attempts to provide a comprehensive introduction to those models and methods most likely to be encountered and used by students in their careers in engineering and the natural sciences. Although the examples and exercises have been designed with scientists and engineers in mind, most of the methods covered are basic to statistical analyses in many other disciplines, so that students of business and the social sciences will also profit from reading the book.

Approach

Students in a statistics course designed to serve other majors may be initially skeptical of the value and relevance of the subject matter, but my experience is that students *can* be turned on to statistics by the use of good examples and exercises that blend their everyday experiences with their scientific interests. Consequently, I have worked hard to find examples of real, rather than artificial, data—data that someone thought was worth collecting and analyzing. Many of the methods presented, especially in the later chapters on statistical inference, are illustrated by analyzing data taken from published sources, and many of the exercises also involve working with such data. Sometimes the reader may be unfamiliar with the context of a particular problem (as indeed I often was), but I have found that students are more attracted by real problems with a somewhat strange context than by patently artificial problems in a familiar setting.

Mathematical Level

The exposition is relatively modest in terms of mathematical development. Substantial use of the calculus is made only in Chapter 4 and parts of Chapters 5 and 6. In particular, with the exception of an occasional remark or aside, calculus appears in the inference part of the book only—in the second section of Chapter 6. Matrix algebra is not used at all. Thus almost all the exposition should be accessible to those whose mathematical background includes one semester or two quarters of differential and integral calculus.

Content

Chapter 1 begins with some basic concepts and terminology—population, sample, descriptive and inferential statistics, enumerative versus analytic studies, and so on—and continues with a survey of important graphical and numerical descriptive methods. A rather traditional development of probability is given in Chapter 2, followed by probability distributions of discrete and continuous random variables in Chapters 3 and 4, respectively. Joint distributions and their properties are discussed in the first part of Chapter 5. The latter part of this chapter introduces statistics and their sampling distributions, which form the bridge between probability and inference. The next three

chapters cover point estimation, statistical intervals, and hypothesis testing based on a single sample. Methods of inference involving two independent samples and paired data are presented in Chapter 9. The analysis of variance is the subject of Chapters 10 and 11 (single-factor and multifactor, respectively). Regression makes its initial appearance in Chapter 12 (the simple linear regression model and correlation) and returns for an extensive encore in Chapter 13. The last three chapters develop chi-squared methods, distribution-free (nonparametric) procedures, and techniques from statistical quality control.

Helping Students Learn

Although the book's mathematical level should give most science and engineering students little difficulty, working toward an understanding of the concepts and gaining an appreciation for the logical development of the methodology may sometimes require substantial effort. To help students gain such an understanding and appreciation, I have provided numerous exercises ranging in difficulty from many that involve routine application of text material to some that ask the reader to extend concepts discussed in the text to somewhat new situations. There are many more exercises than most instructors would want to assign during any particular course, but I recommend that students be required to work a substantial number of them. In a problem-solving discipline, active involvement of this sort is the surest way to identify and close the gaps in understanding that inevitably arise. Answers to most odd-numbered exercises appear in the answer section at the back of the text. In addition, a Student Solutions Manual, consisting of worked-out solutions to virtually all the odd-numbered exercises, is available.

To access additional course materials and companion resources, please visit www.cengagebrain.com. At the CengageBrain.com home page, search for the ISBN of your title (from the back cover of your book) using the search box at the top of the page. This will take you to the product page where free companion resources can be found.

New for This Edition

- The major change for this edition is the elimination of the rejection region approach to hypothesis testing. Conclusions from a hypothesis-testing analysis are now based entirely on *P*-values. This has necessitated completely rewriting Section 8.1, which now introduces hypotheses and then test procedures based on *P*-values. Substantial revision of the remaining sections of Chapter 8 was then required, and this in turn has been propagated through the hypothesis-testing sections and subsections of Chapters 9–15.
- Many new examples and exercises, almost all based on real data or actual problems. Some of these scenarios are less technical or broader in scope than what has been included in previous editions—for example, investigating the nocebo effect (the inclination of those told about a drug's side effects to experience them), comparing sodium contents of cereals produced by three different manufacturers, predicting patient height from an easy-to-measure anatomical characteristic, modeling the relationship between an adolescent mother's age and the birth weight of her baby, assessing the effect of smokers' short-term abstinence on the accurate perception of elapsed time, and exploring the impact of phrasing in a quantitative literacy test.
- More examples and exercises in the probability material (Chapters 2–5) are based on information from published sources.

- The exposition has been polished whenever possible to help students gain a better intuitive understanding of various concepts.

Acknowledgments

My colleagues at Cal Poly have provided me with invaluable support and feedback over the years. I am also grateful to the many users of previous editions who have made suggestions for improvement (and on occasion identified errors). A special note of thanks goes to Jimmy Doi for his accuracy checking and to Matt Carlton for his work on the two solutions manuals, one for instructors and the other for students.

The generous feedback provided by the following reviewers of this and previous editions has been of great benefit in improving the book: Robert L. Armacost, University of Central Florida; Bill Bade, Lincoln Land Community College; Douglas M. Bates, University of Wisconsin–Madison; Michael Berry, West Virginia Wesleyan College; Brian Bowman, Auburn University; Linda Boyle, University of Iowa; Ralph Bravaco, Stonehill College; Linfield C. Brown, Tufts University; Karen M. Bursic, University of Pittsburgh; Lynne Butler, Haverford College; Troy Butler, Colorado State University; Barrett Caldwell, Purdue University; Kyle Caudle, South Dakota School of Mines & Technology; Raj S. Chhikara, University of Houston–Clear Lake; Edwin Chong, Colorado State University; David Clark, California State Polytechnic University at Pomona; Ken Constantine, Taylor University; Bradford Crain, Portland State University; David M. Cresap, University of Portland; Savas Dayanik, Princeton University; Don E. Deal, University of Houston; Annjanette M. Dodd, Humboldt State University; Jimmy Doi, California Polytechnic State University–San Luis Obispo; Charles E. Donaghey, University of Houston; Patrick J. Driscoll, U.S. Military Academy; Mark Duva, University of Virginia; Nassir Eltinay, Lincoln Land Community College; Thomas English, College of the Mainland; Nasser S. Fard, Northeastern University; Ronald Fricker, Naval Postgraduate School; Steven T. Garren, James Madison University; Mark Gebert, University of Kentucky; Harland Glaz, University of Maryland; Ken Grace, Anoka-Ramsey Community College; Celso Grebogi, University of Maryland; Veronica Webster Griffis, Michigan Technological University; Jose Guardiola, Texas A&M University–Corpus Christi; K. L. D. Gunawardena, University of Wisconsin–Oshkosh; James J. Halavin, Rochester Institute of Technology; James Hartman, Marymount University; Tyler Haynes, Saginaw Valley State University; Jennifer Hoeting, Colorado State University; Wei-Min Huang, Lehigh University; Aridaman Jain, New Jersey Institute of Technology; Roger W. Johnson, South Dakota School of Mines & Technology; Chihwa Kao, Syracuse University; Saleem A. Kassam, University of Pennsylvania; Mohammad T. Khasawneh, State University of NewYork–Binghamton; Kyungduk Ko, Boise State University; Stephen Kokoska, Colgate University; Hillel J. Kumin, University of Oklahoma; Sarah Lam, Binghamton University; M. Louise Lawson, Kennesaw State University; Jialiang Li, University of Wisconsin–Madison; Wooi K. Lim, William Paterson University; Aquila Lipscomb, The Citadel; Manuel Lladser, University of Colorado at Boulder; Graham Lord, University of California–Los Angeles; Joseph L. Macaluso, DeSales University; Ranjan Maitra, Iowa State University; David Mathiason, Rochester Institute of Technology; Arnold R. Miller, University of Denver; John J. Millson, University of Maryland; Pamela Kay Miltenberger, West Virginia Wesleyan College; Monica Molsee, Portland State University; Thomas Moore, Naval Postgraduate School; Robert M. Norton, College of Charleston; Steven Pilnick, Naval Postgraduate School; Robi Polikar, Rowan University; Justin Post, North Carolina State University; Ernest Pyle, Houston Baptist University;

Xianggui Qu, Oakland University; Kingsley Reeves, University of South Florida; Steve Rein, California Polytechnic State University–San Luis Obispo; Tony Richardson, University of Evansville; Don Ridgeway, North Carolina State University; Larry J. Ringer, Texas A&M University; Nabin Sapkota, University of Central Florida; Robert M. Schumacher, Cedarville University; Ron Schwartz, Florida Atlantic University; Kevan Shafizadeh, California State University–Sacramento; Mohammed Shayib, Prairie View A&M; Alice E. Smith, Auburn University; James MacGregor Smith, University of Massachusetts; Paul J. Smith, University of Maryland; Richard M. Soland, The George Washington University; Clifford Spiegelman, Texas A&M University; Jery Stedinger, Cornell University; David Steinberg, Tel Aviv University; William Thistleton, State University of New York Institute of Technology; J A Stephen Viggiano, Rochester Institute of Technology; G. Geoffrey Vining, University of Florida; Bhutan Wadhwa, Cleveland State University; Gary Wasserman, Wayne State University; Elaine Wenderholm, State University of New York–Oswego; Samuel P. Wilcock, Messiah College; Michael G. Zabetakis, University of Pittsburgh; and Maria Zack, Point Loma Nazarene University.

Preeti Longia Sinha of MPS Limited has done a terrific job of supervising the book's production. Once again I am compelled to express my gratitude to all those people at Cengage who have made important contributions over the course of my textbook writing career. For this most recent edition, special thanks go to Jay Campbell (for his timely and informed feedback throughout the project), Molly Taylor, Ryan Ahern, Spencer Arritt, Cathy Brooks, and Andrew Coppola. I also greatly appreciate the stellar work of all those Cengage Learning sales representatives who have labored to make my books more visible to the statistical community. Last but by no means least, a heartfelt thanks to my wife Carol for her decades of support, and to my daughters for providing inspiration through their own achievements.

Jay Devore

Overview and Descriptive Statistics

1

"I took statistics at business school, and it was a transformative experience. Analytical training gives you a skill set that differentiates you from most people in the labor market."

—Laszlo Bock, Senior Vice President of People Operations (in charge of all hiring) at Google

April 20, 2014, *The New York Times*, interview with columnist Thomas Friedman

"I am not much given to regret, so I puzzled over this one a while. Should have taken much more statistics in college, I think."

—Max Levchin, Paypal Co-founder, Slide Founder

Quote of the week from the Web site of the American Statistical Association on November 23, 2010

"I keep saying that the sexy job in the next 10 years will be statisticians, and I'm not kidding."

—Hal Varian, Chief Economist at Google

August 6, 2009, *The New York Times*

INTRODUCTION

Statistical concepts and methods are not only useful but indeed often indispensable in understanding the world around us. They provide ways of gaining new insights into the behavior of many phenomena that you will encounter in your chosen field of specialization in engineering or science.

The discipline of statistics teaches us how to make intelligent judgments and informed decisions in the presence of uncertainty and variation. Without uncertainty or variation, there would be little need for statistical methods or statisticians. If every component of a particular type had exactly the same lifetime, if all resistors produced by a certain manufacturer had the same resistance value,

if pH determinations for soil specimens from a particular locale gave identical results, and so on, then a single observation would reveal all desired information.

An interesting manifestation of variation appeared in connection with determining the "greenest" way to travel. The article **"Carbon Conundrum" (*Consumer Reports*, 2008: 9)** identified organizations that help consumers calculate carbon output. The following results on output for a flight from New York to Los Angeles were reported:

Carbon Calculator	CO_2 (lb)
Terra Pass	1924
Conservation International	3000
Cool It	3049
World Resources Institute/Safe Climate	3163
National Wildlife Federation	3465
Sustainable Travel International	3577
Native Energy	3960
Environmental Defense	4000
Carbonfund.org	4820
The Climate Trust/CarbonCounter.org	5860
Bonneville Environmental Foundation	6732

There is clearly rather substantial disagreement among these calculators as to exactly how much carbon is emitted, characterized in the article as "from a ballerina's to Bigfoot's." A website address was provided where readers could learn more about how the various calculators work.

How can statistical techniques be used to gather information and draw conclusions? Suppose, for example, that a materials engineer has developed a coating for retarding corrosion in metal pipe under specified circumstances. If this coating is applied to different segments of pipe, variation in environmental conditions and in the segments themselves will result in more substantial corrosion on some segments than on others. Methods of statistical analysis could be used on data from such an experiment to decide whether the *average* amount of corrosion exceeds an upper specification limit of some sort or to predict how much corrosion will occur on a single piece of pipe.

Alternatively, suppose the engineer has developed the coating in the belief that it will be superior to the currently used coating. A comparative experiment could be carried out to investigate this issue by applying the current coating to some segments of pipe and the new coating to other segments. This must be done with care lest the wrong conclusion emerge. For example, perhaps the average amount of corrosion is identical for the two coatings. However, the new coating may be applied to segments that have superior ability to resist corrosion and under less stressful environmental conditions compared to the segments and conditions for the current coating. The investigator would then likely observe a difference

between the two coatings attributable not to the coatings themselves, but just to extraneous variation. Statistics offers not only methods for analyzing the results of experiments once they have been carried out but also suggestions for how experiments can be performed in an efficient manner to mitigate the effects of variation and have a better chance of producing correct conclusions.

1.1 Populations, Samples, and Processes

Engineers and scientists are constantly exposed to collections of facts, or **data**, both in their professional capacities and in everyday activities. The discipline of statistics provides methods for organizing and summarizing data and for drawing conclusions based on information contained in the data.

An investigation will typically focus on a well-defined collection of objects constituting a **population** of interest. In one study, the population might consist of all gelatin capsules of a particular type produced during a specified period. Another investigation might involve the population consisting of all individuals who received a B.S. in engineering during the most recent academic year. When desired information is available for all objects in the population, we have what is called a **census**. Constraints on time, money, and other scarce resources usually make a census impractical or infeasible. Instead, a subset of the population—**a sample**—is selected in some prescribed manner. Thus we might obtain a sample of bearings from a particular production run as a basis for investigating whether bearings are conforming to manufacturing specifications, or we might select a sample of last year's engineering graduates to obtain feedback about the quality of the engineering curricula.

We are usually interested only in certain characteristics of the objects in a population: the number of flaws on the surface of each casing, the thickness of each capsule wall, the gender of an engineering graduate, the age at which the individual graduated, and so on. A characteristic may be categorical, such as gender or type of malfunction, or it may be numerical in nature. In the former case, the *value* of the characteristic is a category (e.g., female or insufficient solder), whereas in the latter case, the value is a number (e.g., age = 23 years or diameter = .502 cm). A **variable** is any characteristic whose value may change from one object to another in the population. We shall initially denote variables by lowercase letters from the end of our alphabet. Examples include

x = brand of calculator owned by a student

y = number of visits to a particular Web site during a specified period

z = braking distance of an automobile under specified conditions

Data results from making observations either on a single variable or simultaneously on two or more variables. A **univariate** data set consists of observations on a single variable. For example, we might determine the type of transmission, automatic (A) or manual (M), on each of ten automobiles recently purchased at a certain dealership, resulting in the categorical data set

M A A A M A A M A A

The following sample of pulse rates (beats per minute) for patients recently admitted to an adult intensive care unit is a numerical univariate data set:

88 80 71 103 154 132 67 110 60 105

We have **bivariate** data when observations are made on each of two variables. Our data set might consist of a (height, weight) pair for each basketball player on a team, with the first observation as (72, 168), the second as (75, 212), and so on. If an engineer determines the value of both x = component lifetime and y = reason for component failure, the resulting data set is bivariate with one variable numerical and the other categorical. **Multivariate** data arises when observations are made on more than one variable (so bivariate is a special case of multivariate). For example, a research physician might determine the systolic blood pressure, diastolic blood pressure, and serum cholesterol level for each patient participating in a study. Each observation would be a triple of numbers, such as (120, 80, 146). In many multivariate data sets, some variables are numerical and others are categorical. Thus the annual automobile issue of *Consumer Reports* gives values of such variables as type of vehicle (small, sporty, compact, mid-size, large), city fuel efficiency (mpg), highway fuel efficiency (mpg), drivetrain type (rear wheel, front wheel, four wheel), and so on.

Branches of Statistics

An investigator who has collected data may wish simply to summarize and describe important features of the data. This entails using methods from **descriptive statistics**. Some of these methods are graphical in nature; the construction of histograms, boxplots, and scatter plots are primary examples. Other descriptive methods involve calculation of numerical summary measures, such as means, standard deviations, and correlation coefficients. The wide availability of statistical computer software packages has made these tasks much easier to carry out than they used to be. Computers are much more efficient than human beings at calculation and the creation of pictures (once they have received appropriate instructions from the user!). This means that the investigator doesn't have to expend much effort on "grunt work" and will have more time to study the data and extract important messages. Throughout this book, we will present output from various packages such as Minitab, SAS, JMP, and R. The R software can be downloaded without charge from the site **http://www.r-project.org**. It has achieved great popularity in the statistical community, and many books describing its various uses are available (it does entail programming as opposed to the pull-down menus of Minitab and JMP).

EXAMPLE 1.1 Charity is a big business in the United States. The Web site **charitynavigator.com** gives information on roughly 6000 charitable organizations, and there are many smaller charities that fly below the navigator's radar screen. Some charities operate very efficiently, with fundraising and administrative expenses that are only a small percentage of total expenses, whereas others spend a high percentage of what they take in on such activities. Here is data on fundraising expenses as a percentage of total expenditures for a random sample of 60 charities:

6.1	12.6	34.7	1.6	18.8	2.2	3.0	2.2	5.6	3.8
2.2	3.1	1.3	1.1	14.1	4.0	21.0	6.1	1.3	20.4
7.5	3.9	10.1	8.1	19.5	5.2	12.0	15.8	10.4	5.2
6.4	10.8	83.1	3.6	6.2	6.3	16.3	12.7	1.3	0.8
8.8	5.1	3.7	26.3	6.0	48.0	8.2	11.7	7.2	3.9
15.3	16.6	8.8	12.0	4.7	14.7	6.4	17.0	2.5	16.2

Without any organization, it is difficult to get a sense of the data's most prominent features—what a typical (i.e., representative) value might be, whether values are highly concentrated about a typical value or quite dispersed, whether there are any gaps in the data, what fraction of the values are less than 20%, and so on. Figure 1.1

```
Stem-and-leaf of FundRsng N = 60
Leaf Unit = 1.0
    0 | 0111112222333333344
    0 | 55556666666778888
    1 | 0001222244
    1 | 55666789
    2 | 01
    2 | 6
    3 | 4
    3 |
    4 |
    4 | 8
    5 |
    5 |
    6 |
    6 |
    7 |
    7 |
    8 | 3
```

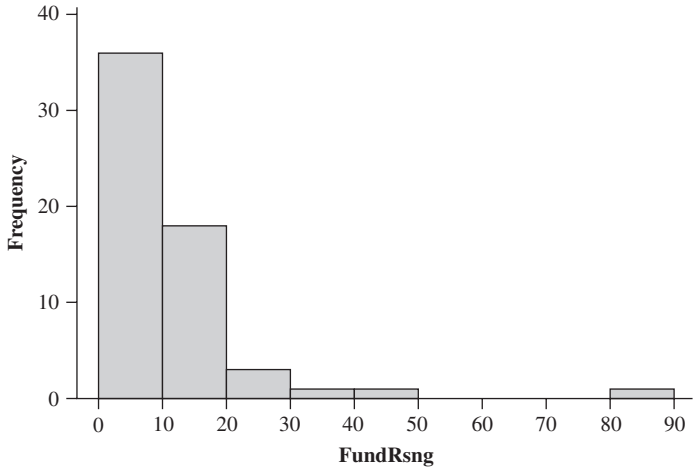


Figure 1.1 A Minitab stem-and-leaf display (tenths digit truncated) and histogram for the charity fundraising percentage data

shows what is called a *stem-and-leaf display* as well as a *histogram.* In Section 1.2 we will discuss construction and interpretation of these data summaries. For the moment, we hope you see how they begin to describe how the percentages are distributed over the range of possible values from 0 to 100. Clearly a substantial majority of the charities in the sample spend less than 20% on fundraising, and only a few percentages might be viewed as beyond the bounds of sensible practice. ■

Having obtained a sample from a population, an investigator would frequently like to use sample information to draw some type of conclusion (make an inference of some sort) about the population. That is, the sample is a means to an end rather than an end in itself. Techniques for generalizing from a sample to a population are gathered within the branch of our discipline called **inferential statistics**.

EXAMPLE 1.2 Material strength investigations provide a rich area of application for statistical methods. The article **"Effects of Aggregates and Microfillers on the Flexural Properties of Concrete" (*Magazine of Concrete Research*, 1997: 81–98)** reported on a study of strength properties of high-performance concrete obtained by using superplasticizers and certain binders. The compressive strength of such concrete had previously been investigated, but not much was known about flexural strength (a measure of ability to resist failure in bending). The accompanying data on flexural strength (in MegaPascal, MPa, where 1 Pa (Pascal) = 1.45×10^{-4} psi) appeared in the article cited:

5.9	7.2	7.3	6.3	8.1	6.8	7.0	7.6	6.8	6.5	7.0	6.3	7.9	9.0
8.2	8.7	7.8	9.7	7.4	7.7	9.7	7.8	7.7	11.6	11.3	11.8	10.7	

Suppose we want an *estimate* of the average value of flexural strength for all beams that could be made in this way (if we conceptualize a population of all such beams, we are trying to estimate the population mean). It can be shown that, with a high degree of confidence, the population mean strength is between 7.48 MPa and 8.80 MPa; we call this a *confidence interval* or *interval estimate.* Alternatively, this data could be used to predict the flexural strength of a *single* beam of this type. With a high degree of confidence, the strength of a single such beam will exceed 7.35 MPa; the number 7.35 is called a *lower prediction bound.* ■

The main focus of this book is on presenting and illustrating methods of inferential statistics that are useful in scientific work. The most important types of inferential procedures—point estimation, hypothesis testing, and estimation by confidence intervals—are introduced in Chapters 6–8 and then used in more complicated settings in Chapters 9–16. The remainder of this chapter presents methods from descriptive statistics that are most used in the development of inference.

Chapters 2–5 present material from the discipline of probability. This material ultimately forms a bridge between the descriptive and inferential techniques. Mastery of probability leads to a better understanding of how inferential procedures are developed and used, how statistical conclusions can be translated into everyday language and interpreted, and when and where pitfalls can occur in applying the methods. Probability and statistics both deal with questions involving populations and samples, but do so in an "inverse manner" to one another.

In a probability problem, properties of the population under study are assumed known (e.g., in a numerical population, some specified distribution of the population values may be assumed), and questions regarding a sample taken from the population are posed and answered. In a statistics problem, characteristics of a sample are available to the experimenter, and this information enables the experimenter to draw conclusions about the population. The relationship between the two disciplines can be summarized by saying that probability reasons from the population to the sample (deductive reasoning), whereas inferential statistics reasons from the sample to the population (inductive reasoning). This is illustrated in Figure 1.2.

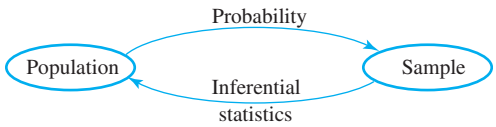


Figure 1.2 The relationship between probability and inferential statistics

Before we can understand what a particular sample can tell us about the population, we should first understand the uncertainty associated with taking a sample from a given population. This is why we study probability before statistics.

EXAMPLE 1.3 As an example of the contrasting focus of probability and inferential statistics, consider drivers' use of manual lap belts in cars equipped with automatic shoulder belt systems. (The article **"Automobile Seat Belts: Usage Patterns in Automatic Belt Systems,"** ***Human Factors*****, 1998: 126–135,** summarizes usage data.) In probability, we might assume that 50% of all drivers of cars equipped in this way in a certain metropolitan area regularly use their lap belt (an assumption about the population), so we might ask, "How likely is it that a sample of 100 such drivers will include at least 70 who regularly use their lap belt?" or "How many of the drivers in a sample of size 100 can we expect to regularly use their lap belt?" On the other hand, in inferential statistics, we have sample information available; for example, a sample of 100 drivers of such cars revealed that 65 regularly use their lap belt. We might then ask, "Does this provide substantial evidence for concluding that more than 50% of all such drivers in this area regularly use their lap belt?" In this latter scenario, we are attempting to use sample information to answer a question about the structure of the entire population from which the sample was selected. ■

In the foregoing lap belt example, the population is well defined and concrete: all drivers of cars equipped in a certain way in a particular metropolitan area. In Example 1.2, however, the strength measurements came from a sample of prototype beams that

had not been selected from an existing population. Instead, it is convenient to think of the population as consisting of all possible strength measurements that might be made under similar experimental conditions. Such a population is referred to as a **conceptual** or **hypothetical population**. There are a number of problem situations in which we fit questions into the framework of inferential statistics by conceptualizing a population.

The Scope of Modern Statistics

These days statistical methodology is employed by investigators in virtually all disciplines, including such areas as

- molecular biology (analysis of microarray data)
- ecology (describing quantitatively how individuals in various animal and plant populations are spatially distributed)
- materials engineering (studying properties of various treatments to retard corrosion)
- marketing (developing market surveys and strategies for marketing new products)
- public health (identifying sources of diseases and ways to treat them)
- civil engineering (assessing the effects of stress on structural elements and the impacts of traffic flows on communities)

As you progress through the book, you'll encounter a wide spectrum of different scenarios in the examples and exercises that illustrate the application of techniques from probability and statistics. Many of these scenarios involve data or other material extracted from articles in engineering and science journals. The methods presented herein have become established and trusted tools in the arsenal of those who work with data. Meanwhile, statisticians continue to develop new models for describing randomness, and uncertainty and new methodology for analyzing data. As evidence of the continuing creative efforts in the statistical community, here are titles and capsule descriptions of some articles that have recently appeared in statistics journals (*Journal of the American Statistical Association* is abbreviated *JASA*, and *AAS* is short for the *Annals of Applied Statistics,* two of the many prominent journals in the discipline):

- **"How Many People Do You Know? Efficiently Estimating Personal Network Size" (*JASA*, 2010: 59–70):** How many of the N individuals at your college do you know? You could select a random sample of students from the population and use an estimate based on the fraction of people in this sample that you know. Unfortunately this is very inefficient for large populations because the fraction of the population someone knows is typically very small. A "latent mixing model" was proposed that the authors asserted remedied deficiencies in previously used techniques. A simulation study of the method's effectiveness based on groups consisting of first names ("How many people named Michael do you know?") was included as well as an application of the method to actual survey data. The article concluded with some practical guidelines for the construction of future surveys designed to estimate social network size.
- **"Active Learning Through Sequential Design, with Applications to the Detection of Money Laundering" (*JASA*, 2009: 969–981):** Money laundering involves concealing the origin of funds obtained through illegal activities. The huge number of transactions occurring daily at financial institutions makes detection of money laundering difficult. The standard approach has been to extract various summary quantities from the transaction history and conduct a time-consuming investigation of suspicious activities. The article proposes a more efficient statistical method and illustrates its use in a case study.

- **"Robust Internal Benchmarking and False Discovery Rates for Detecting Racial Bias in Police Stops" (*JASA*, 2009: 661–668):** Allegations of police actions that are attributable at least in part to racial bias have become a contentious issue in many communities. This article proposes a new method that is designed to reduce the risk of flagging a substantial number of "false positives" (individuals falsely identified as manifesting bias). The method was applied to data on 500,000 pedestrian stops in New York City in 2006; of the 3000 officers regularly involved in pedestrian stops, 15 were identified as having stopped a substantially greater fraction of Black and Hispanic people than what would be predicted were bias absent.
- **"Records in Athletics Through Extreme Value Theory" (*JASA*, 2008: 1382–1391):** The focus here is on the modeling of extremes related to world records in athletics. The authors start by posing two questions: (1) What is the ultimate world record within a specific event (e.g., the high jump for women)? and (2) How "good" is the current world record, and how does the quality of current world records compare across different events? A total of 28 events (8 running, 3 throwing, and 3 jumping for both men and women) are considered. For example, one conclusion is that only about 20 seconds can be shaved off the men's marathon record, but that the current women's marathon record is almost 5 minutes longer than what can ultimately be achieved. The methodology also has applications to such issues as ensuring airport runways are long enough and that dikes in Holland are high enough.
- **"Self-Exciting Hurdle Models for Terrorist Activity" (*AAS*, 2012: 106–124):** The authors developed a predictive model of terrorist activity by considering the daily number of terrorist attacks in Indonesia from 1994 through 2007. The model estimates the chance of future attacks as a function of the times since past attacks. One feature of the model considers the excess of nonattack days coupled with the presence of multiple coordinated attacks on the same day. The article provides an interpretation of various model characteristics and assesses its predictive performance.
- **"Prediction of Remaining Life of Power Transformers Based on Left Truncated and Right Censored Lifetime Data" (*AAS*, 2009: 857–879):** There are roughly 150,000 high-voltage power transmission transformers in the United States. Unexpected failures can cause substantial economic losses, so it is important to have predictions for remaining lifetimes. Relevant data can be complicated because lifetimes of some transformers extend over several decades during which records were not necessarily complete. In particular, the authors of the article use data from a certain energy company that began keeping careful records in 1980. But some transformers had been installed before January 1, 1980, and were still in service after that date ("left truncated" data), whereas other units were still in service at the time of the investigation, so their complete lifetimes are not available ("right censored" data). The article describes various procedures for obtaining an interval of plausible values (a *prediction interval*) for a remaining lifetime and for the cumulative number of failures over a specified time period.
- **"The BARISTA: A Model for Bid Arrivals in Online Auctions" (*AAS*, 2007: 412–441):** Online auctions such as those on eBay and uBid often have characteristics that differentiate them from traditional auctions. One particularly important difference is that the number of bidders at the outset of many traditional auctions is fixed, whereas in online auctions this number and the number of resulting bids are not predetermined. The article proposes a new BARISTA (for Bid ARrivals In STAges) model for describing the way in which bids arrive online. The model allows for higher bidding intensity at the outset of the auction and also as the auction comes to a close. Various properties of the model are investigated and

then validated using data from eBay.com on auctions for Palm M515 personal assistants, Microsoft Xbox games, and Cartier watches.

- **"Statistical Challenges in the Analysis of Cosmic Microwave Background Radiation" (*AAS*, 2009: 61–95):** The cosmic microwave background (CMB) is a significant source of information about the early history of the universe. Its radiation level is uniform, so extremely delicate instruments have been developed to measure fluctuations. The authors provide a review of statistical issues with CMB data analysis; they also give many examples of the application of statistical procedures to data obtained from a recent NASA satellite mission, the *Wilkinson Microwave Anisotropy Probe*.

Statistical information now appears with increasing frequency in the popular media, and occasionally the spotlight is even turned on statisticians. For example, the **Nov. 23, 2009, *New York Times*** reported in an article "Behind Cancer Guidelines, Quest for Data" that the new science for cancer investigations and more sophisticated methods for data analysis spurred the U.S. Preventive Services task force to re-examine guidelines for how frequently middle-aged and older women should have mammograms. The panel commissioned six independent groups to do statistical modeling. The result was a new set of conclusions, including an assertion that mammograms every two years are nearly as beneficial to patients as annual mammograms, but confer only half the risk of harms. Donald Berry, a very prominent biostatistician, was quoted as saying he was pleasantly surprised that the task force took the new research to heart in making its recommendations. The task force's report has generated much controversy among cancer organizations, politicians, and women themselves.

It is our hope that you will become increasingly convinced of the importance and relevance of the discipline of statistics as you dig more deeply into the book and the subject. Hopefully you'll be turned on enough to want to continue your statistical education beyond your current course.

Enumerative Versus Analytic Studies

W. E. Deming, a very influential American statistician who was a moving force in Japan's quality revolution during the 1950s and 1960s, introduced the distinction between *enumerative studies* and *analytic studies.* In the former, interest is focused on a finite, identifiable, unchanging collection of individuals or objects that make up a population. A *sampling frame*—that is, a listing of the individuals or objects to be sampled—is either available to an investigator or else can be constructed. For example, the frame might consist of all signatures on a petition to qualify a certain initiative for the ballot in an upcoming election; a sample is usually selected to ascertain whether the number of *valid* signatures exceeds a specified value. As another example, the frame may contain serial numbers of all furnaces manufactured by a particular company during a certain time period; a sample may be selected to infer something about the average lifetime of these units. The use of inferential methods to be developed in this book is reasonably noncontroversial in such settings (though statisticians may still argue over which particular methods should be used).

An analytic study is broadly defined as one that is not enumerative in nature. Such studies are often carried out with the objective of improving a future product by taking action on a process of some sort (e.g., recalibrating equipment or adjusting the level of some input such as the amount of a catalyst). Data can often be obtained only on an existing process, one that may differ in important respects from the future process. There is thus no sampling frame listing the individuals or objects of interest. For example, a sample of five turbines with a new design may be experimentally manufactured and

tested to investigate efficiency. These five could be viewed as a sample from the conceptual population of all prototypes that could be manufactured under similar conditions, but *not* necessarily as representative of the population of units manufactured once regular production gets underway. Methods for using sample information to draw conclusions about future production units may be problematic. Someone with expertise in the area of turbine design and engineering (or whatever other subject area is relevant) should be called upon to judge whether such extrapolation is sensible. A good exposition of these issues is contained in the article **"Assumptions for Statistical Inference" by Gerald Hahn and William Meeker (*The American Statistician*, 1993: 1–11).**

Collecting Data

Statistics deals not only with the organization and analysis of data once it has been collected but also with the development of techniques for collecting the data. If data is not properly collected, an investigator may not be able to answer the questions under consideration with a reasonable degree of confidence. One common problem is that the target population—the one about which conclusions are to be drawn—may be different from the population actually sampled. For example, advertisers would like various kinds of information about the television-viewing habits of potential customers. The most systematic information of this sort comes from placing monitoring devices in a small number of homes across the United States. It has been conjectured that placement of such devices in and of itself alters viewing behavior, so that characteristics of the sample may be different from those of the target population.

When data collection entails selecting individuals or objects from a frame, the simplest method for ensuring a representative selection is to take a *simple random sample.* This is one for which any particular subset of the specified size (e.g., a sample of size 100) has the same chance of being selected. For example, if the frame consists of 1,000,000 serial numbers, the numbers 1, 2, …, up to 1,000,000 could be placed on identical slips of paper. After placing these slips in a box and thoroughly mixing, slips could be drawn one by one until the requisite sample size has been obtained. Alternatively (and much to be preferred), a table of random numbers or a software package's random number generator could be employed.

Sometimes alternative sampling methods can be used to make the selection process easier, to obtain extra information, or to increase the degree of confidence in conclusions. One such method, *stratified sampling,* entails separating the population units into nonoverlapping groups and taking a sample from each one. For example, a study of how physicians feel about the Affordable Care Act might proceed by stratifying according to specialty: select a sample of surgeons, another sample of radiologists, yet another sample of psychiatrists, and so on. This would result in information separately from each specialty and ensure that no one specialty is over- or underrepresented in the entire sample.

Frequently a "convenience" sample is obtained by selecting individuals or objects without systematic randomization. As an example, a collection of bricks may be stacked in such a way that it is extremely difficult for those in the center to be selected. If the bricks on the top and sides of the stack were somehow different from the others, resulting sample data would not be representative of the population. Often an investigator will assume that such a convenience sample approximates a random sample, in which case a statistician's repertoire of inferential methods can be used; however, this is a judgment call. Most of the methods discussed herein are based on a variation of simple random sampling described in Chapter 5.

Engineers and scientists often collect data by carrying out some sort of designed experiment. This may involve deciding how to allocate several different treatments (such as fertilizers or coatings for corrosion protection) to the various experimental units (plots

of land or pieces of pipe). Alternatively, an investigator may systematically vary the levels or categories of certain factors (e.g., pressure or type of insulating material) and observe the effect on some response variable (such as yield from a production process).

EXAMPLE 1.4 An article in the *New York Times* (Jan. 27, 1987) reported that heart attack risk could be reduced by taking aspirin. This conclusion was based on a designed experiment involving both a control group of individuals that took a placebo having the appearance of aspirin but known to be inert and a treatment group that took aspirin according to a specified regimen. Subjects were randomly assigned to the groups to protect against any biases and so that probability-based methods could be used to analyze the data. Of the 11,034 individuals in the control group, 189 subsequently experienced heart attacks, whereas only 104 of the 11,037 in the aspirin group had a heart attack. The incidence rate of heart attacks in the treatment group was only about half that in the control group. One possible explanation for this result is chance variation—that aspirin really doesn't have the desired effect and the observed difference is just typical variation in the same way that tossing two identical coins would usually produce different numbers of heads. However, in this case, inferential methods suggest that chance variation by itself cannot adequately explain the magnitude of the observed difference. ■

EXAMPLE 1.5 An engineer wishes to investigate the effects of both adhesive type and conductor material on bond strength when mounting an integrated circuit (IC) on a certain substrate. Two adhesive types and two conductor materials are under consideration. Two observations are made for each adhesive-type/conductor-material combination, resulting in the accompanying data:

Adhesive Type	Conductor Material	Observed Bond Strength	Average
1	1	82, 77	79.5
1	2	75, 87	81.0
2	1	84, 80	82.0
2	2	78, 90	84.0

The resulting average bond strengths are pictured in Figure 1.3. It appears that adhesive type 2 improves bond strength as compared with type 1 by about the same amount whichever one of the conducting materials is used, with the 2, 2 combination being best. Inferential methods can again be used to judge whether these effects are real or simply due to chance variation.

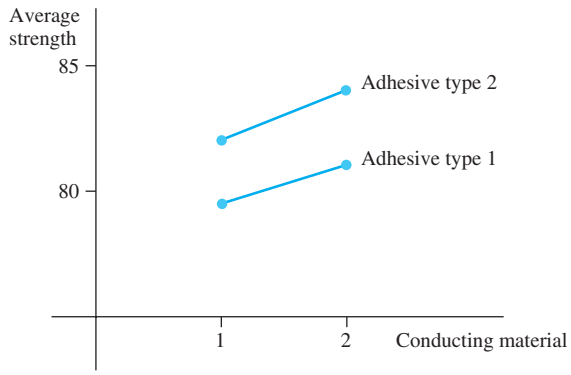


Figure 1.3 Average bond strengths in Example 1.5

Suppose additionally that there are two cure times under consideration and also two types of IC post coating. There are then $2 \cdot 2 \cdot 2 \cdot 2 = 16$ combinations of these four

factors, and our engineer may not have enough resources to make even a single observation for each of these combinations. In Chapter 11, we will see how the careful selection of a fraction of these possibilities will usually yield the desired information. ■

EXERCISES Section 1.1 (1–9)

1. Give one possible sample of size 4 from each of the following populations:
- **a.** All daily newspapers published in the United States
- **b.** All companies listed on the New York Stock Exchange
- **c.** All students at your college or university
- **d.** All grade point averages of students at your college or university

2. For each of the following hypothetical populations, give a plausible sample of size 4:
- **a.** All distances that might result when you throw a football
- **b.** Page lengths of books published 5 years from now
- **c.** All possible earthquake-strength measurements (Richter scale) that might be recorded in California during the next year
- **d.** All possible yields (in grams) from a certain chemical reaction carried out in a laboratory

3. Consider the population consisting of all computers of a certain brand and model, and focus on whether a computer needs service while under warranty.
- **a.** Pose several probability questions based on selecting a sample of 100 such computers.
- **b.** What inferential statistics question might be answered by determining the number of such computers in a sample of size 100 that need warranty service?

4. **a.** Give three different examples of concrete populations and three different examples of hypothetical populations.
- **b.** For one each of your concrete and your hypothetical populations, give an example of a probability question and an example of an inferential statistics question.

5. Many universities and colleges have instituted supplemental instruction (SI) programs, in which a student facilitator meets regularly with a small group of students enrolled in the course to promote discussion of course material and enhance subject mastery. Suppose that students in a large statistics course (what else?) are randomly divided into a control group that will not participate in SI and a treatment group that will participate. At the end of the term, each student's total score in the course is determined.
- **a.** Are the scores from the SI group a sample from an existing population? If so, what is it? If not, what is the relevant conceptual population?
- **b.** What do you think is the advantage of randomly dividing the students into the two groups rather than letting each student choose which group to join?
- **c.** Why didn't the investigators put all students in the treatment group? [*Note:* The article **"Supplemental Instruction: An Effective Component of Student Affairs Programming" (*J. of College Student Devel.*, 1997: 577–586)** discusses the analysis of data from several SI programs.]

6. The California State University (CSU) system consists of 23 campuses, from San Diego State in the south to Humboldt State near the Oregon border. A CSU administrator wishes to make an inference about the average distance between the hometowns of students and their campuses. Describe and discuss several different sampling methods that might be employed. Would this be an enumerative or an analytic study? Explain your reasoning.

7. A certain city divides naturally into ten district neighborhoods. How might a real estate appraiser select a sample of single-family homes that could be used as a basis for developing an equation to predict appraised value from characteristics such as age, size, number of bathrooms, distance to the nearest school, and so on? Is the study enumerative or analytic?

8. The amount of flow through a solenoid valve in an automobile's pollution-control system is an important characteristic. An experiment was carried out to study how flow rate depended on three factors: armature length, spring load, and bobbin depth. Two different levels (low and high) of each factor were chosen, and a single observation on flow was made for each combination of levels.
- **a.** The resulting data set consisted of how many observations?
- **b.** Is this an enumerative or analytic study? Explain your reasoning.

9. In a famous experiment carried out in 1882, Michelson and Newcomb obtained 66 observations on the time it took for light to travel between two locations in Washington, D.C. A few of the measurements (coded in a certain manner) were 31, 23, 32, 36, −2, 26, 27, and 31.
- **a.** Why are these measurements not identical?
- **b.** Is this an enumerative study? Why or why not?

1.2 Pictorial and Tabular Methods in Descriptive Statistics

Descriptive statistics can be divided into two general subject areas. In this section, we consider representing a data set using visual displays. In Sections 1.3 and 1.4, we will develop some numerical summary measures for data sets. Many visual techniques may already be familiar to you: frequency tables, tally sheets, histograms, pie charts, bar graphs, scatter diagrams, and the like. Here we focus on a selected few of these techniques that are most useful and relevant to probability and inferential statistics.

Notation

Some general notation will make it easier to apply our methods and formulas to a wide variety of practical problems. The number of observations in a single sample, that is, the *sample size,* will often be denoted by n, so that $n = 4$ for the sample of universities {Stanford, Iowa State, Wyoming, Rochester} and also for the sample of pH measurements {6.3, 6.2, 5.9, 6.5}. If two samples are simultaneously under consideration, either m and n or n_1 and n_2 can be used to denote the numbers of observations. An experiment to compare thermal efficiencies for two different types of diesel engines might result in samples {29.7, 31.6, 30.9} and {28.7, 29.5, 29.4, 30.3}, in which case $m = 3$ and $n = 4$.

Given a data set consisting of n observations on some variable x, the individual observations will be denoted by $x_1, x_2, x_3, \ldots, x_n$. The subscript bears no relation to the magnitude of a particular observation. Thus x_1 will not in general be the smallest observation in the set, nor will x_n typically be the largest. In many applications, x_1 will be the first observation gathered by the experimenter, x_2 the second, and so on. The ith observation in the data set will be denoted by x_i.

Stem-and-Leaf Displays

Consider a numerical data set $x_1, x_2, \ldots, x_n$ for which each x_i consists of at least two digits. A quick way to obtain an informative visual representation of the data set is to construct a *stem-and-leaf display.*

Constructing a Stem-and-Leaf Display

1. Select one or more leading digits for the stem values. The trailing digits become the leaves.
2. List possible stem values in a vertical column.
3. Record the leaf for each observation beside the corresponding stem value.
4. Indicate the units for stems and leaves someplace in the display.

For a data set consisting of exam scores, each between 0 and 100, the score of 83 would have a stem of 8 and a leaf of 3. If all exam scores are in the 90s, 80s, and 70s (an instructor's dream!), use of the tens digit as the stem would give a display

with only three rows. In this case, it is desirable to stretch the display by repeating each stem value twice—9H, 9L, 8H, . . . ,7L—once for high leaves 9, . . . , 5 and again for low leaves 4, . . . , 0. Then a score of 93 would have a stem of 9L and leaf of 3. In general, a display based on between 5 and 20 stems is recommended.

EXAMPLE 1.6 A common complaint among college students is that they are getting less sleep than they need. The article **"Class Start Times, Sleep, and Academic Performance in College: A Path Analysis" (*Chronobiology Intl*., 2012: 318–335)** investigated factors that impact sleep time. The stem-and-leaf display in Figure 1.4 shows the average number of hours of sleep per day over a two-week period for a sample of 253 students.

5L	00
5H	6889
6L	000111123444444
6H	55556778899999
7L	000011111112222223333333444444444
7H	55555555666666666666777777888888888999999999999999
8L	000000000000111111222222222222222233333333334444444444444
8H	555555556666666667777778888888889999999999
9L	00001111111222223334
9H	666678999
10L	00
10H	56

Stem: ones digit
Leaf: tenths digit

Figure 1.4 Stem-and-leaf display for average sleep time per day

The first observation in the top row of the display is 5.0, corresponding to a stem of 5 and leaf of 0, and the last observation at the bottom of the display is 10.6. Note that in the absence of a context, without the identification of stem and leaf digits in the display, we wouldn't know whether the observation with stem 7 and leaf 9 was .79, 7.9, or 79. The leaves in each row are ordered from smallest to largest; this is commonly done by software packages but is not necessary if a display is created by hand.

The display suggests that a typical or representative sleep time is in the stem 8L row, perhaps 8.1 or 8.2. The data is not highly concentrated about this typical value as would be the case if almost all students were getting between 7.5 and 9.5 hours of sleep on average. The display appears to rise rather smoothly to a peak in the 8L row and then decline smoothly (we conjecture that the minor peak in the 6L row would disappear if more data was available). The general shape of the display is rather symmetric, bearing strong resemblance to a bell-shaped curve; it does not stretch out more in one direction than the other. The two smallest and two largest values seem a bit separated from the remainder of the data—perhaps they are very mild, but certainly not extreme,"outliers". A reference in the cited article suggests that individuals in this age group need about 8.4 hours of sleep per day. So it appears that a substantial percentage of students in the sample are sleep deprived. ■

A stem-and-leaf display conveys information about the following aspects of the data:

- identification of a typical or representative value
- extent of spread about the typical value
- presence of any gaps in the data

- extent of symmetry in the distribution of values
- number and locations of peaks
- presence of any *outliers*—values far from the rest of the data

EXAMPLE 1.7 Figure 1.5 presents stem-and-leaf displays for a random sample of lengths of golf courses (yards) that have been designated by ***Golf Magazine*** as among the most challenging in the United States. Among the sample of 40 courses, the shortest is 6433 yards long, and the longest is 7280 yards. The lengths appear to be distributed in a roughly uniform fashion over the range of values in the sample. Notice that a stem choice here of either a single digit (6 or 7) or three digits (643, … , 728) would yield an uninformative display, the first because of too few stems and the latter because of too many.

64	35	64	33	70	Stem: Thousands and hundreds digits		
65	26	27	06	83	Leaf: Tens and ones digits		
66	05	94	14				
67	90	70	00	98	70	45	13
68	90	70	73	50			
69	00	27	36	04			
70	51	05	11	40	50	22	
71	31	69	68	05	13	65	
72	80	09					

(a)

```
Stem-and-leaf of yardage  N = 40
Leaf Unit = 10
     4          64   3367
     8          65   0228
    11          66   019
    18          67   0147799
   (4)          68   5779
    18          69   0023
    14          70   012455
     8          71   013666
     2          72   08
```

(b)

Figure 1.5 Stem-and-leaf displays of golf course lengths: (a) two-digit leaves; (b) display from Minitab with truncated one-digit leaves

Statistical software packages do not generally produce displays with multiple-digit stems. The Minitab display in Figure 1.5(b) results from *truncating* each observation by deleting the ones digit. ■

Dotplots

A dotplot is an attractive summary of numerical data when the data set is reasonably small or there are relatively few distinct data values. Each observation is represented by a dot above the corresponding location on a horizontal measurement scale. When a value occurs more than once, there is a dot for each occurrence, and these dots are stacked vertically. As with a stem-and-leaf display, a dotplot gives information about location, spread, extremes, and gaps.

EXAMPLE 1.8 There is growing concern in the U.S. that not enough students are graduating from college. America used to be number 1 in the world for the percentage of adults with college degrees, but it has recently dropped to 16th. Here is data on the percentage of 25- to 34-year-olds in each state who had some type of postsecondary degree as of 2010 (listed in alphabetical order, with the District of Columbia included):

31.5	32.9	33.0	28.6	37.9	43.3	45.9	37.2	68.8	36.2	35.5
40.5	37.2	45.3	36.1	45.5	42.3	33.3	30.3	37.2	45.5	54.3
37.2	49.8	32.1	39.3	40.3	44.2	28.4	46.0	47.2	28.7	49.6
37.6	50.8	38.0	30.8	37.6	43.9	42.5	35.2	42.2	32.8	32.2
38.5	44.5	44.6	40.9	29.5	41.3	35.4				

Figure 1.6 shows a dotplot of the data. Dots corresponding to some values close together (e.g., 28.6 and 28.7) have been vertically stacked to prevent crowding. There is clearly a great deal of state-to-state variability. The largest value, for D.C., is obviously an extreme outlier, and four other values on the upper end of the data are candidates for mild outliers (MA, MN, NY, and ND). There is also a cluster of states at the low end, primarily located in the South and Southwest. The overall percentage for the entire country is 39.3%; this is not a simple average of the 51 numbers but an average weighted by population sizes.

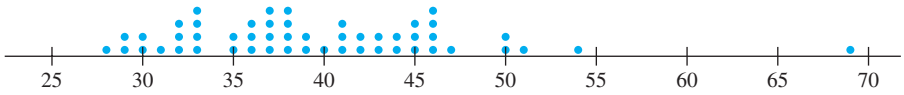


Figure 1.6 A dotplot of the data from Example 1.8 ■

A dotplot can be quite cumbersome to construct and look crowded when the number of observations is large. Our next technique is well suited to such situations.

Histograms

Some numerical data is obtained by counting to determine the value of a variable (the number of traffic citations a person received during the last year, the number of customers arriving for service during a particular period), whereas other data is obtained by taking measurements (weight of an individual, reaction time to a particular stimulus). The prescription for drawing a histogram is generally different for these two cases.

DEFINITION

A numerical variable is **discrete** if its set of possible values either is finite or else can be listed in an infinite sequence (one in which there is a first number, a second number, and so on). A numerical variable is **continuous** if its possible values consist of an entire interval on the number line.

A discrete variable x almost always results from counting, in which case possible values are 0, 1, 2, 3, … or some subset of these integers. Continuous variables arise from making measurements. For example, if x is the pH of a chemical substance, then in theory x could be any number between 0 and 14: 7.0, 7.03, 7.032, and so on. Of course, in practice there are limitations on the degree of accuracy of any measuring instrument, so we may not be able to determine pH, reaction time, height, and concentration to an arbitrarily large number of decimal places. However, from the point of view of creating mathematical models for distributions of data, it is helpful to imagine an entire continuum of possible values.

Consider data consisting of observations on a discrete variable x. The **frequency** of any particular x value is the number of times that value occurs in the data set. The **relative frequency** of a value is the fraction or proportion of times the value occurs:

$$\text{relative frequency of a value} = \frac{\text{number of times the value occurs}}{\text{number of observations in the data set}}$$

Suppose, for example, that our data set consists of 200 observations on x = the number of courses a college student is taking this term. If 70 of these x values are 3, then

$$\text{frequency of the } x \text{ value 3:} \quad 70$$

$$\text{relative frequency of the } x \text{ value 3:} \quad \frac{70}{200} = .35$$

Multiplying a relative frequency by 100 gives a percentage; in the college-course example, 35% of the students in the sample are taking three courses. The relative frequencies, or percentages, are usually of more interest than the frequencies themselves. In theory, the relative frequencies should sum to 1, but in practice the sum may differ slightly from 1 because of rounding. A **frequency distribution** is a tabulation of the frequencies and/or relative frequencies.

Constructing a Histogram for Discrete Data

First, determine the frequency and relative frequency of each x value. Then mark possible x values on a horizontal scale. Above each value, draw a rectangle whose height is the relative frequency (or alternatively, the frequency) of that value; the rectangles should have equal widths.

This construction ensures that the *area* of each rectangle is proportional to the relative frequency of the value. Thus if the relative frequencies of $x = 1$ and $x = 5$ are .35 and .07, respectively, then the area of the rectangle above 1 is five times the area of the rectangle above 5.

EXAMPLE 1.9 How unusual is a no-hitter or a one-hitter in a major league baseball game, and how frequently does a team get more than 10, 15, or even 20 hits? Table 1.1 is a frequency distribution for the number of hits per team per game for all nine-inning games that were played between 1989 and 1993.

Table 1.1 Frequency Distribution for Hits in Nine-Inning Games

Hits/Game	Number of Games	Relative Frequency	Hits/Game	Number of Games	Relative Frequency
0	20	.0010	14	569	.0294
1	72	.0037	15	393	.0203
2	209	.0108	16	253	.0131
3	527	.0272	17	171	.0088
4	1048	.0541	18	97	.0050
5	1457	.0752	19	53	.0027
6	1988	.1026	20	31	.0016
7	2256	.1164	21	19	.0010
8	2403	.1240	22	13	.0007
9	2256	.1164	23	5	.0003
10	1967	.1015	24	1	.0001
11	1509	.0779	25	0	.0000
12	1230	.0635	26	1	.0001
13	834	.0430	27	1	.0001
				19,383	1.0005

The corresponding histogram in Figure 1.7 rises rather smoothly to a single peak and then declines. The histogram extends a bit more on the right (toward large values) than it does on the left—a slight "positive skew."

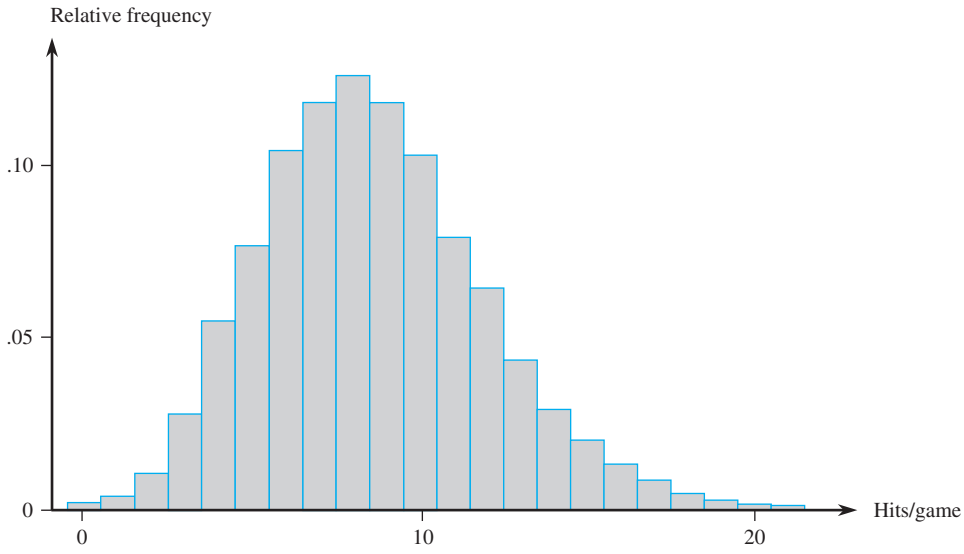


Figure 1.7 Histogram of number of hits per nine-inning game

Either from the tabulated information or from the histogram itself, we can determine the following:

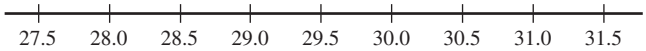
$$\begin{aligned}\text{proportion of games with at most two hits} &= \begin{matrix}\text{relative}\\ \text{frequency}\\ \text{for } x=0\end{matrix} + \begin{matrix}\text{relative}\\ \text{frequency}\\ \text{for } x=1\end{matrix} + \begin{matrix}\text{relative}\\ \text{frequency}\\ \text{for } x=2\end{matrix}\\ &= .0010 + .0037 + .0108 = .0155\end{aligned}$$

Similarly,

$$\text{proportion of games with between 5 and 10 hits (inclusive)} = .0752 + .1026 + \cdots + .1015 = .6361$$

That is, roughly 64% of all these games resulted in between 5 and 10 (inclusive) hits. ■

Constructing a histogram for continuous data (measurements) entails subdividing the measurement axis into a suitable number of **class intervals** or **classes**, such that each observation is contained in exactly one class. Suppose, for example, that we have 50 observations on x = fuel efficiency of an automobile (mpg), the smallest of which is 27.8 and the largest of which is 31.4. Then we could use the class boundaries 27.5, 28.0, 28.5, … , and 31.5 as shown here:



One potential difficulty is that occasionally an observation lies on a class boundary so therefore does not fall in exactly one interval, for example, 29.0. One way to deal with this problem is to use boundaries like 27.55, 28.05, … , 31.55. Adding a hundredths digit to the class boundaries prevents observations from falling on the resulting boundaries. Another approach is to use the classes $27.5 - < 28.0$, $28.0 - < 28.5, \ldots, 31.0 - < 31.5$. Then 29.0 falls in the class $29.0 - < 29.5$ rather than in the class $28.5 - < 29.0$. In other words, with this convention, an observation on a boundary is placed in the interval to the *right* of the boundary. This is how Minitab constructs a histogram.

Constructing a Histogram for Continuous Data: Equal Class Widths

Determine the frequency and relative frequency for each class. Mark the class boundaries on a horizontal measurement axis. Above each class interval, draw a rectangle whose height is the corresponding relative frequency (or frequency).

EXAMPLE 1.10 Power companies need information about customer usage to obtain accurate forecasts of demands. Investigators from Wisconsin Power and Light determined energy consumption (BTUs) during a particular period for a sample of 90 gas-heated homes. An adjusted consumption value was calculated as follows:

$$\text{adjusted consumption} = \frac{\text{consumption}}{(\text{weather, in degree days})(\text{house area})}$$

This resulted in the accompanying data (part of the stored data set FURNACE.MTW available in Minitab), which we have ordered from smallest to largest.

2.97	4.00	5.20	5.56	5.94	5.98	6.35	6.62	6.72	6.78
6.80	6.85	6.94	7.15	7.16	7.23	7.29	7.62	7.62	7.69
7.73	7.87	7.93	8.00	8.26	8.29	8.37	8.47	8.54	8.58
8.61	8.67	8.69	8.81	9.07	9.27	9.37	9.43	9.52	9.58
9.60	9.76	9.82	9.83	9.83	9.84	9.96	10.04	10.21	10.28
10.28	10.30	10.35	10.36	10.40	10.49	10.50	10.64	10.95	11.09
11.12	11.21	11.29	11.43	11.62	11.70	11.70	12.16	12.19	12.28
12.31	12.62	12.69	12.71	12.91	12.92	13.11	13.38	13.42	13.43
13.47	13.60	13.96	14.24	14.35	15.12	15.24	16.06	16.90	18.26

The most striking feature of the histogram in Figure 1.8 is its resemblance to a bell-shaped curve, with the point of symmetry roughly at 10.

Class	1−<3	3−<5	5−<7	7−<9	9−<11	11−<13	13−<15	15−<17	17−<19
Frequency	1	1	11	21	25	17	9	4	1
Relative frequency	.011	.011	.122	.233	.278	.189	.100	.044	.011

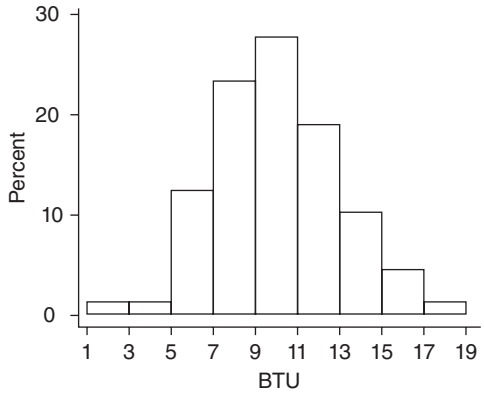


Figure 1.8 Histogram of the energy consumption data from Example 1.10

From the histogram,

$$\begin{pmatrix}\text{proportion of}\\ \text{observations}\\ \text{less than 9}\end{pmatrix} \approx .01 + .01 + .12 + .23 = .37 \quad \left(\text{exact value} = \frac{34}{90} = .378\right)$$

The relative frequency for the $9-<11$ class is about .27, so we estimate that roughly half of this, or .135, is between 9 and 10. Thus

$$\begin{pmatrix}\text{proportion of observations}\\ \text{less than 10}\end{pmatrix} \approx .37 + .135 = .505 \text{ (slightly more than 50\%)}$$

The exact value of this proportion is $47/90 = .522$. ■

There are no hard-and-fast rules concerning either the number of classes or the choice of classes themselves. Between 5 and 20 classes will be satisfactory for most data sets. Generally, the larger the number of observations in a data set, the more classes should be used. A reasonable rule of thumb is

$$\text{number of classes} \approx \sqrt{\text{number of observations}}$$

Equal-width classes may not be a sensible choice if there are some regions of the measurement scale that have a high concentration of data values and other parts where data is quite sparse. Figure 1.9 shows a dotplot of such a data set; there is high concentration in the middle, and relatively few observations stretched out to either side. Using a small number of equal-width classes results in almost all observations falling in just one or two of the classes. If a large number of equal-width classes are used, many classes will have zero frequency. A sound choice is to use a few wider intervals near extreme observations and narrower intervals in the region of high concentration.

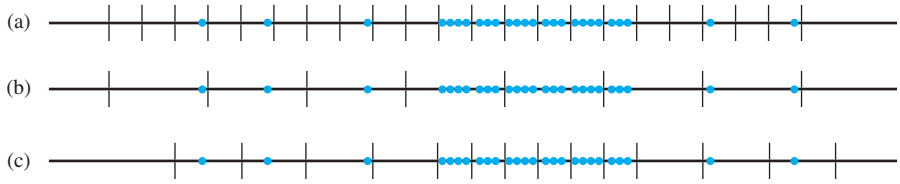


Figure 1.9 Selecting class intervals for "varying density" data: (a) many short equal-width intervals; (b) a few wide equal-width intervals; (c) unequal-width intervals

Constructing a Histogram for Continuous Data: Unequal Class Widths

After determining frequencies and relative frequencies, calculate the height of each rectangle using the formula

$$\text{rectangle height} = \frac{\text{relative frequency of the class}}{\text{class width}}$$

The resulting rectangle heights are usually called *densities,* and the vertical scale is the **density scale.** This prescription will also work when class widths are equal.

EXAMPLE 1.11 Corrosion of reinforcing steel is a serious problem in concrete structures located in environments affected by severe weather conditions. For this reason, researchers have been investigating the use of reinforcing bars made of composite material. One study was carried out to develop guidelines for bonding glass-fiber-reinforced plastic rebars to concrete (**"Design Recommendations for Bond of GFRP Rebars to Concrete," *J. of Structural Engr.*, 1996: 247–254).** Consider the following 48 observations on measured bond strength:

11.5	12.1	9.9	9.3	7.8	6.2	6.6	7.0	13.4	17.1	9.3	5.6
5.7	5.4	5.2	5.1	4.9	10.7	15.2	8.5	4.2	4.0	3.9	3.8
3.6	3.4	20.6	25.5	13.8	12.6	13.1	8.9	8.2	10.7	14.2	7.6
5.2	5.5	5.1	5.0	5.2	4.8	4.1	3.8	3.7	3.6	3.6	3.6

Class	2−<4	4−<6	6−<8	8−<12	12−<20	20−<30
Frequency	9	15	5	9	8	2
Relative frequency	.1875	.3125	.1042	.1875	.1667	.0417
Density	.094	.156	.052	.047	.021	.004

The resulting histogram appears in Figure 1.10. The right or upper tail stretches out much farther than does the left or lower tail—a substantial departure from symmetry.

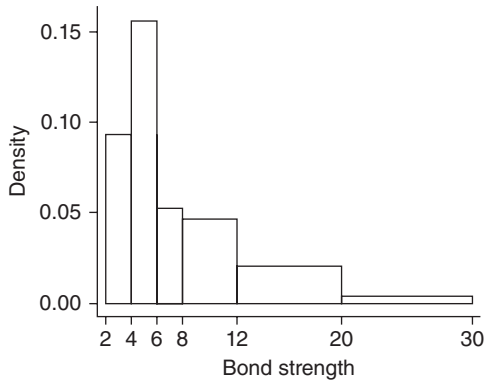


Figure 1.10 A Minitab density histogram for the bond strength data of Example 1.11 ■

When class widths are unequal, not using a density scale will give a picture with distorted areas. For equal-class widths, the divisor is the same in each density calculation, and the extra arithmetic simply results in a rescaling of the vertical axis (i.e., the histogram using relative frequency and the one using density will have exactly the same appearance). A density histogram does have one interesting property. Multiplying both sides of the formula for density by the class width gives

$$\begin{aligned}\text{relative frequency} &= (\text{class width})(\text{density}) = (\text{rectangle width})(\text{rectangle height}) \\ &= \text{rectangle area}\end{aligned}$$

That is, *the area of each rectangle is the relative frequency of the corresponding class.* Furthermore, since the sum of relative frequencies should be 1, *the total area of all rectangles in a density histogram is* 1. It is always possible to draw a histogram

so that the area equals the relative frequency (this is true also for a histogram of discrete data)—just use the density scale. This property will play an important role in motivating models for distributions in Chapter 4.

Histogram Shapes

Histograms come in a variety of shapes. A **unimodal** histogram is one that rises to a single peak and then declines. A **bimodal** histogram has two different peaks. Bimodality can occur when the data set consists of observations on two quite different kinds of individuals or objects. For example, consider a large data set consisting of driving times for automobiles traveling between San Luis Obispo, California, and Monterey, California (exclusive of stopping time for sightseeing, eating, etc.). This histogram would show two peaks: one for those cars that took the inland route (roughly 2.5 hours) and another for those cars traveling up the coast (3.5–4 hours). However, bimodality does not automatically follow in such situations. Only if the two separate histograms are "far apart" relative to their spreads will bimodality occur in the histogram of combined data. Thus a large data set consisting of heights of college students should not result in a bimodal histogram because the typical male height of about 69 inches is not far enough above the typical female height of about 64–65 inches. A histogram with more than two peaks is said to be **multimodal**. Of course, the number of peaks may well depend on the choice of class intervals, particularly with a small number of observations. The larger the number of classes, the more likely it is that bimodality or multimodality will manifest itself.

EXAMPLE 1.12 Figure 1.11(a) shows a Minitab histogram of the weights (lb) of the 124 players listed on the rosters of the San Francisco 49ers and the New England Patriots (teams the author would like to see meet in the Super Bowl) as of Nov. 20, 2009. Figure 1.11(b) is a smoothed histogram (actually what is called a *density estimate*) of the data from the R software package. Both the histogram and the smoothed histogram show three distinct peaks; the one on the right is for linemen, the middle peak corresponds to linebacker weights, and the peak on the left is for all other players (wide receivers, quarterbacks, etc.).

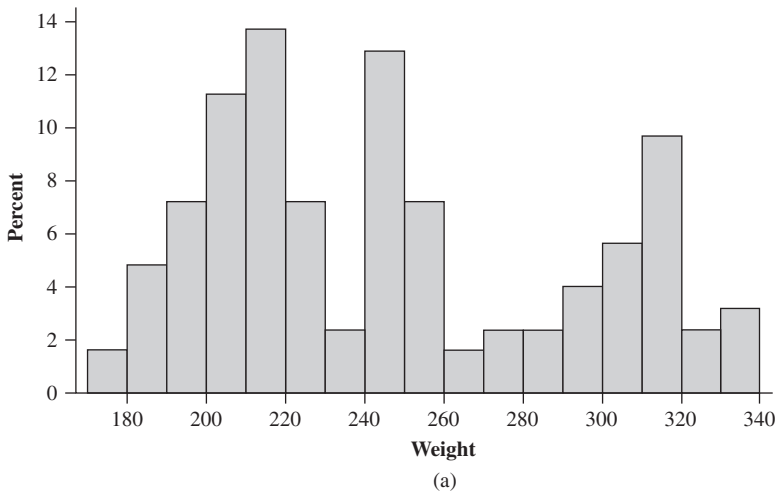


Figure 1.11 NFL player weights (a) Histogram (b) Smoothed histogram

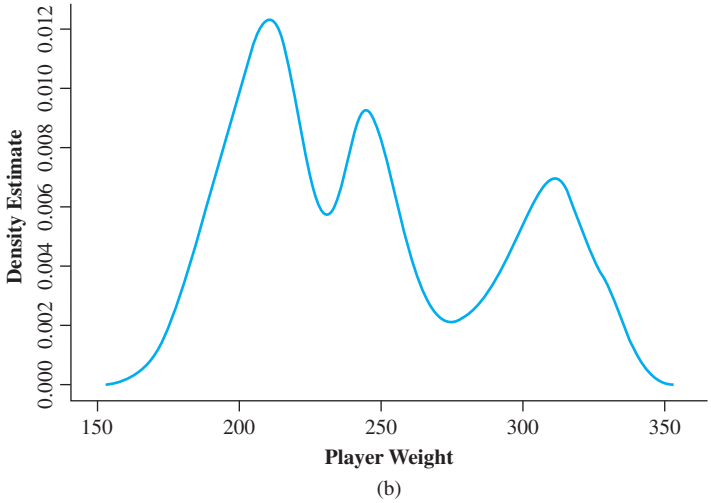


Figure 1.11 (continued)

A histogram is **symmetric** if the left half is a mirror image of the right half. A unimodal histogram is **positively skewed** if the right or upper tail is stretched out compared with the left or lower tail and **negatively skewed** if the stretching is to the left. Figure 1.12 shows "smoothed" histograms, obtained by superimposing a smooth curve on the rectangles, that illustrate the various possibilities.

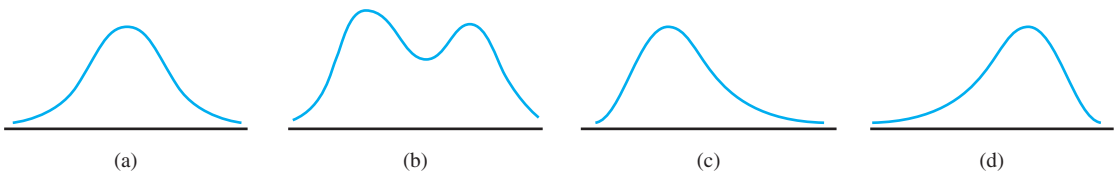


Figure 1.12 Smoothed histograms: (a) symmetric unimodal; (b) bimodal; (c) positively skewed; and (d) negatively skewed

Qualitative Data

Both a frequency distribution and a histogram can be constructed when the data set is *qualitative* (categorical) in nature. In some cases, there will be a natural ordering of classes—for example, freshmen, sophomores, juniors, seniors, graduate students—whereas in other cases the order will be arbitrary—for example, Catholic, Jewish, Protestant, and the like. With such categorical data, the intervals above which rectangles are constructed should have equal width.

EXAMPLE 1.13 **The Public Policy Institute of California** carried out a telephone survey of 2501 California adult residents during April 2006 to ascertain how they felt about various aspects of K–12 public education. One question asked was "Overall, how would you rate the quality of public schools in your neighborhood today?" Table 1.2 displays the frequencies and relative frequencies, and Figure 1.13 shows the corresponding histogram (bar chart).

Table 1.2 Frequency Distribution for the School Rating Data

Rating	Frequency	Relative Frequency
A	478	.191
B	893	.357
C	680	.272
D	178	.071
F	100	.040
Don't know	172	.069
	2501	1.000

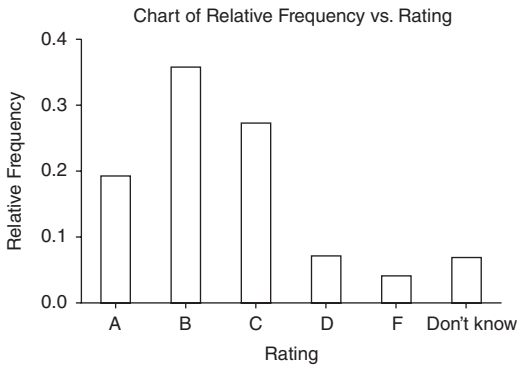


Figure 1.13 Histogram of the school rating data from Minitab

More than half the respondents gave an A or B rating, and only slightly more than 10% gave a D or F rating. The percentages for parents of public school children were somewhat more favorable to schools: 24%, 40%, 24%, 6%, 4%, and 2%. ■

Multivariate Data

Multivariate data is generally rather difficult to describe visually. Several methods for doing so appear later in the book, notably scatterplots for bivariate numerical data.

EXERCISES Section 1.2 (10–32)

10. Consider the strength data for beams given in Example 1.2.

- **a.** Construct a stem-and-leaf display of the data. What appears to be a representative strength value? Do the observations appear to be highly concentrated about the representative value or rather spread out?
- **b.** Does the display appear to be reasonably symmetric about a representative value, or would you describe its shape in some other way?
- **c.** Do there appear to be any outlying strength values?
- **d.** What proportion of strength observations in this sample exceed 10 MPa?

11. The accompanying specific gravity values for various wood types used in construction appeared in the article **"Bolted Connection Design Values Based on European Yield Model" (*J. of Structural Engr.*, 1993: 2169–2186):**

.31	.35	.36	.36	.37	.38	.40	.40	.40
.41	.41	.42	.42	.42	.42	.42	.43	.44
.45	.46	.46	.47	.48	.48	.48	.51	.54
.54	.55	.58	.62	.66	.66	.67	.68	.75

Construct a stem-and-leaf display using repeated stems, and comment on any interesting features of the display.

12. The accompanying summary data on CeO_2 particle sizes (nm) under certain experimental conditions was read from a graph in the article **"Nanoceria—Energetics of Surfaces, Interfaces and Water Adsorption" (*J. of the Amer. Ceramic Soc.*, 2011: 3992–3999):**

3.0–<3.5	3.5–<4.0	4.0–<4.5	4.5–<5.0	5.0–<5.5
5	15	27	34	22
5.5–<6.0	6.0–<6.5	6.5–<7.0	7.0–<7.5	7.5–<8.0
14	7	2	4	1

a. What proportion of the observations are less than 5?
b. What proportion of the observations are at least 6?
c. Construct a histogram with relative frequency on the vertical axis and comment on interesting features. In particular, does the distribution of particle sizes appear to be reasonably symmetric or somewhat skewed? [*Note:* The investigators fit a lognormal distribution to the data; this is discussed in Chapter 4.]
d. Construct a histogram with density on the vertical axis and compare to the histogram in (c).

13. Allowable mechanical properties for structural design of metallic aerospace vehicles requires an approved method for statistically analyzing empirical test data. The article **"Establishing Mechanical Property Allowables for Metals" (*J. of Testing and Evaluation*, 1998: 293–299)** used the accompanying data on tensile ultimate strength (ksi) as a basis for addressing the difficulties in developing such a method.

122.2	124.2	124.3	125.6	126.3	126.5	126.5	127.2	127.3
127.5	127.9	128.6	128.8	129.0	129.2	129.4	129.6	130.2
130.4	130.8	131.3	131.4	131.4	131.5	131.6	131.6	131.8
131.8	132.3	132.4	132.4	132.5	132.5	132.5	132.5	132.6
132.7	132.9	133.0	133.1	133.1	133.1	133.1	133.2	133.2
133.2	133.3	133.3	133.5	133.5	133.5	133.8	133.9	134.0
134.0	134.0	134.0	134.1	134.2	134.3	134.4	134.4	134.6
134.7	134.7	134.7	134.8	134.8	134.8	134.9	134.9	135.2
135.2	135.2	135.3	135.3	135.4	135.5	135.5	135.6	135.6
135.7	135.8	135.8	135.8	135.8	135.8	135.9	135.9	135.9
135.9	136.0	136.0	136.1	136.2	136.2	136.3	136.4	136.4
136.6	136.8	136.9	136.9	137.0	137.1	137.2	137.6	137.6
137.8	137.8	137.8	137.9	137.9	138.2	138.2	138.3	138.3
138.4	138.4	138.4	138.5	138.5	138.6	138.7	138.7	139.0
139.1	139.5	139.6	139.8	139.8	140.0	140.0	140.7	140.7
140.9	140.9	141.2	141.4	141.5	141.6	142.9	143.4	143.5
143.6	143.8	143.8	143.9	144.1	144.5	144.5	147.7	147.7

a. Construct a stem-and-leaf display of the data by first deleting (truncating) the tenths digit and then repeating each stem value five times (once for leaves 1 and 2, a second time for leaves 3 and 4, etc.). Why is it relatively easy to identify a representative strength value?
b. Construct a histogram using equal-width classes with the first class having a lower limit of 122 and an upper limit of 124. Then comment on any interesting features of the histogram.

14. The accompanying data set consists of observations on shower-flow rate (L/min) for a sample of $n = 129$ houses in Perth, Australia **("An Application of Bayes Methodology to the Analysis of Diary Records in a Water Use Study," *J. Amer. Stat. Assoc.*, 1987: 705–711)**:

4.6	12.3	7.1	7.0	4.0	9.2	6.7	6.9	11.5	5.1
11.2	10.5	14.3	8.0	8.8	6.4	5.1	5.6	9.6	7.5
7.5	6.2	5.8	2.3	3.4	10.4	9.8	6.6	3.7	6.4
8.3	6.5	7.6	9.3	9.2	7.3	5.0	6.3	13.8	6.2
5.4	4.8	7.5	6.0	6.9	10.8	7.5	6.6	5.0	3.3
7.6	3.9	11.9	2.2	15.0	7.2	6.1	15.3	18.9	7.2
5.4	5.5	4.3	9.0	12.7	11.3	7.4	5.0	3.5	8.2
8.4	7.3	10.3	11.9	6.0	5.6	9.5	9.3	10.4	9.7
5.1	6.7	10.2	6.2	8.4	7.0	4.8	5.6	10.5	14.6
10.8	15.5	7.5	6.4	3.4	5.5	6.6	5.9	15.0	9.6
7.8	7.0	6.9	4.1	3.6	11.9	3.7	5.7	6.8	11.3
9.3	9.6	10.4	9.3	6.9	9.8	9.1	10.6	4.5	6.2
8.3	3.2	4.9	5.0	6.0	8.2	6.3	3.8	6.0	

a. Construct a stem-and-leaf display of the data.
b. What is a typical, or representative, flow rate?
c. Does the display appear to be highly concentrated or spread out?
d. Does the distribution of values appear to be reasonably symmetric? If not, how would you describe the departure from symmetry?
e. Would you describe any observation as being far from the rest of the data (an outlier)?

15. Do running times of American movies differ somehow from running times of French movies? The author investigated this question by randomly selecting 25 recent movies of each type, resulting in the following running times:

Am:	94	90	95	93	128	95	125	91	104	116	162	102	90
	110	92	113	116	90	97	103	95	120	109	91	138	
Fr:	123	116	90	158	122	119	125	90	96	94	137	102	
	105	106	95	125	122	103	96	111	81	113	128	93	92

Construct a *comparative* stem-and-leaf display by listing stems in the middle of your paper and then placing the Am leaves out to the left and the Fr leaves out to

the right. Then comment on interesting features of the display.

16. The article cited in Example 1.2 also gave the accompanying strength observations for cylinders:

6.1	5.8	7.8	7.1	7.2	9.2	6.6	8.3	7.0	8.3
7.8	8.1	7.4	8.5	8.9	9.8	9.7	14.1	12.6	11.2

a. Construct a comparative stem-and-leaf display (see the previous exercise) of the beam and cylinder data, and then answer the questions in parts (b)–(d) of Exercise 10 for the observations on cylinders.

b. In what ways are the two sides of the display similar? Are there any obvious differences between the beam observations and the cylinder observations?

c. Construct a dotplot of the cylinder data.

17. The accompanying data came from a study of collusion in bidding within the construction industry **("Detection of Collusive Behavior," *J. of Construction Engr. and Mgmnt*, 2012: 1251–1258).**

No. Bidders	No. Contracts
2	7
3	20
4	26
5	16
6	11
7	9
8	6
9	8
10	3
11	2

a. What proportion of the contracts involved at most five bidders? At least five bidders?

b. What proportion of the contracts involved between five and 10 bidders, inclusive? Strictly between five and 10 bidders?

c. Construct a histogram and comment on interesting features.

18. Every corporation has a governing board of directors. The number of individuals on a board varies from one corporation to another. One of the authors of the article **"Does Optimal Corporate Board Size Exist? An Empirical Analysis" (*J. of Applied Finance*, 2010: 57–69)** provided the accompanying data on the number of directors on each board in a random sample of 204 corporations.

No. directors:	4	5	6	7	8	9
Frequency:	3	12	13	25	24	42
No. directors:	10	11	12	13	14	15
Frequency:	23	19	16	11	5	4
No. directors:	16	17	21	24	32	
Frequency:	1	3	1	1	1	

a. Construct a histogram of the data based on relative frequencies and comment on any interesting features.

b. Construct a frequency distribution in which the last row includes all boards with at least 18 directors. If this distribution had appeared in the cited article, would you be able to draw a histogram? Explain.

c. What proportion of these corporations have at most 10 directors?

d. What proportion of these corporations have more than 15 directors?

19. The number of contaminating particles on a silicon wafer prior to a certain rinsing process was determined for each wafer in a sample of size 100, resulting in the following frequencies:

Number of particles	0	1	2	3	4	5	6	7
Frequency	1	2	3	12	11	15	18	10
Number of particles	8	9	10	11	12	13	14	
Frequency	12	4	5	3	1	2	1	

a. What proportion of the sampled wafers had at least one particle? At least five particles?

b. What proportion of the sampled wafers had between five and ten particles, inclusive? Strictly between five and ten particles?

c. Draw a histogram using relative frequency on the vertical axis. How would you describe the shape of the histogram?

20. The article **"Determination of Most Representative Subdivision" (*J. of Energy Engr.*, 1993: 43–55)** gave data on various characteristics of subdivisions that could be used in deciding whether to provide electrical power using overhead lines or underground lines. Here are the values of the variable x = total length of streets within a subdivision:

1280	5320	4390	2100	1240	3060	4770
1050	360	3330	3380	340	1000	960
1320	530	3350	540	3870	1250	2400
960	1120	2120	450	2250	2320	2400
3150	5700	5220	500	1850	2460	5850
2700	2730	1670	100	5770	3150	1890
510	240	396	1419	2109		

a. Construct a stem-and-leaf display using the thousands digit as the stem and the hundreds digit as the leaf, and comment on the various features of the display.

b. Construct a histogram using class boundaries 0, 1000, 2000, 3000, 4000, 5000, and 6000. What proportion

of subdivisions have total length less than 2000? Between 2000 and 4000? How would you describe the shape of the histogram?

21. The article cited in Exercise 20 also gave the following values of the variables y = number of culs-de-sac and z = number of intersections:

y	1	0	1	0	0	2	0	1	1	1	2	1	0	0	1	1	0	1	1
z	1	8	6	1	1	5	3	0	0	4	4	0	0	1	2	1	4	0	4
y	1	1	0	0	0	1	1	2	0	1	2	2	1	1	0	2	1	1	0
z	0	3	0	1	1	0	1	3	2	4	6	6	0	1	1	8	3	3	5
y	1	5	0	3	0	1	1	0	0										
z	0	5	2	3	1	0	0	0	3										

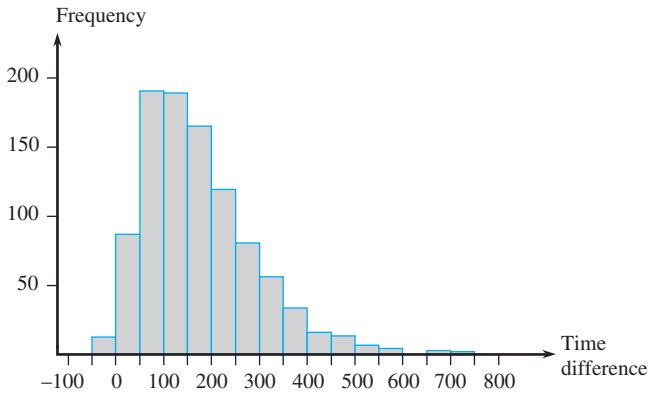
a. Construct a histogram for the y data. What proportion of these subdivisions had no culs-de-sac? At least one cul-de-sac?

b. Construct a histogram for the z data. What proportion of these subdivisions had at most five intersections? Fewer than five intersections?

22. How does the speed of a runner vary over the course of a marathon (a distance of 42.195 km)? Consider determining both the time to run the first 5 km and the time to run between the 35-km and 40-km points, and then subtracting the former time from the latter time. A positive value of this difference corresponds to a runner slowing down toward the end of the race. The accompanying histogram is based on times of runners who participated in several different Japanese marathons **("Factors Affecting Runners' Marathon Performance," *Chance, Fall*, 1993: 24–30).**

What are some interesting features of this histogram? What is a typical difference value? Roughly what proportion of the runners ran the late distance more quickly than the early distance?

Histogram for Exercise 22



23. The article **"Statistical Modeling of the Time Course of Tantrum Anger" (*Annals of Applied Stats*, 2009: 1013–1034)** discussed how anger intensity in children's tantrums could be related to tantrum duration as well as behavioral indicators such as shouting, stamping, and pushing or pulling. The following frequency distribution was given (and also the corresponding histogram):

0–<2 :	136	2–<4:	92	4–<11:	71
11–<20:	26	20–<30:	7	30–<40:	3

Draw the histogram and then comment on any interesting features.

24. The accompanying data set consists of observations on shear strength (lb) of ultrasonic spot welds made on a certain type of alclad sheet. Construct a relative frequency histogram based on ten equal-width classes with boundaries 4000, 4200, …. [The histogram will agree with the one in **"Comparison of Properties of Joints Prepared by Ultrasonic Welding and Other Means" (*J. of Aircraft*, 1983: 552–556).**] Comment on its features.

5434	4948	4521	4570	4990	5702	5241
5112	5015	4659	4806	4637	5670	4381
4820	5043	4886	4599	5288	5299	4848
5378	5260	5055	5828	5218	4859	4780
5027	5008	4609	4772	5133	5095	4618
4848	5089	5518	5333	5164	5342	5069
4755	4925	5001	4803	4951	5679	5256
5207	5621	4918	5138	4786	4500	5461
5049	4974	4592	4173	5296	4965	5170
4740	5173	4568	5653	5078	4900	4968
5248	5245	4723	5275	5419	5205	4452
5227	5555	5388	5498	4681	5076	4774
4931	4493	5309	5582	4308	4823	4417
5364	5640	5069	5188	5764	5273	5042
5189	4986					

25. A transformation of data values by means of some mathematical function, such as $\sqrt{x}$ or $1/x$, can often yield a set of numbers that has "nicer" statistical properties than the original data. In particular, it may be possible to find a function for which the histogram of transformed values is more symmetric (or, even better, more like a bell-shaped curve) than the original data. As an example, the article **"Time Lapse Cinematographic Analysis of Beryllium–Lung Fibroblast Interactions" (*Environ. Research*, 1983: 34–43)** reported the results of experiments designed to study the behavior of certain individual cells that had been exposed to beryllium. An important characteristic of such an individual cell is its interdivision time (IDT). IDTs were determined for a large number of cells, both in exposed (treatment) and unexposed (control) conditions. The authors of the article used a logarithmic transformation, that is,

transformed value = log(original value). Consider the following representative IDT data:

IDT	$\log_{10}$(IDT)	IDT	$\log_{10}$(IDT)	IDT	$\log_{10}$(IDT)
28.1	1.45	60.1	1.78	21.0	1.32
31.2	1.49	23.7	1.37	22.3	1.35
13.7	1.14	18.6	1.27	15.5	1.19
46.0	1.66	21.4	1.33	36.3	1.56
25.8	1.41	26.6	1.42	19.1	1.28
16.8	1.23	26.2	1.42	38.4	1.58
34.8	1.54	32.0	1.51	72.8	1.86
62.3	1.79	43.5	1.64	48.9	1.69
28.0	1.45	17.4	1.24	21.4	1.33
17.9	1.25	38.8	1.59	20.7	1.32
19.5	1.29	30.6	1.49	57.3	1.76
21.1	1.32	55.6	1.75	40.9	1.61
31.9	1.50	25.5	1.41		
28.9	1.46	52.1	1.72		

Use class intervals 10−<20, 20−<30,… to construct a histogram of the original data. Use intervals 1.1−<1.2, 1.2−<1.3,… to do the same for the transformed data. What is the effect of the transformation?

26. Automated electron backscattered diffraction is now being used in the study of fracture phenomena. The following information on misorientation angle (degrees) was extracted from the article **"Observations on the Faceted Initiation Site in the Dwell-Fatigue Tested Ti-6242 Alloy: Crystallographic Orientation and Size Effects" (*Metallurgical and Materials Trans.*, 2006: 1507–1518).**

Class:	0−<5	5−<10	10−<15	15−<20
Rel freq:	.177	.166	.175	.136
Class:	20−<30	30−<40	40−<60	60−<90
Rel freq:	.194	.078	.044	.030

a. Is it true that more than 50% of the sampled angles are smaller than 15°, as asserted in the paper?

b. What proportion of the sampled angles are at least 30°?

c. Roughly what proportion of angles are between 10° and 25°?

d. Construct a histogram and comment on any interesting features.

27. The article **"Study on the Life Distribution of Microdrills" (*J. of Engr. Manufacture*, 2002: 301–305)** reported the following observations, listed in increasing order, on drill lifetime (number of holes that a drill machines before it breaks) when holes were drilled in a certain brass alloy.

11	14	20	23	31	36	39	44	47	50
59	61	65	67	68	71	74	76	78	79
81	84	85	89	91	93	96	99	101	104
105	105	112	118	123	136	139	141	148	158
161	168	184	206	248	263	289	322	388	513

a. Why can a frequency distribution not be based on the class intervals 0–50, 50–100, 100–150, and so on?

b. Construct a frequency distribution and histogram of the data using class boundaries 0, 50, 100, …, and then comment on interesting characteristics.

c. Construct a frequency distribution and histogram of the natural logarithms of the lifetime observations, and comment on interesting characteristics.

d. What proportion of the lifetime observations in this sample are less than 100? What proportion of the observations are at least 200?

28. The accompanying frequency distribution on deposited energy (mJ) was extracted from the article **"Experimental Analysis of Laser-Induced Spark Ignition of Lean Turbulent Premixed Flames" (*Combustion and Flame*, 2013: 1414–1427).**

1.0–<2.0	5	2.0–<2.4	11
2.4–<2.6	13	2.6–<2.8	30
2.8–<3.0	46	3.0–<3.2	66
3.2–<3.4	133	3.4–<3.6	141
3.6–<3.8	126	3.8–<4.0	92
4.0–<4.2	73	4.2–<4.4	38
4.4–<4.6	19	4.6–<5.0	11

a. What proportion of these ignition trials resulted in a deposited energy of less than 3 mJ?

b. What proportion of these ignition trials resulted in a deposited energy of at least 4 mJ?

c. Roughly what proportion of the trials resulted in a deposited energy of at least 3.5 mJ?

d. Construct a histogram and comment on its shape.

29. The following categories for type of physical activity involved when an industrial accident occurred appeared in the article **"Finding Occupational Accident Patterns in the Extractive Industry Using a Systematic Data Mining Approach" (*Reliability Engr. and System Safety*, 2012: 108–122):**

A. Working with handheld tools
B. Movement
C. Carrying by hand
D. Handling of objects
E. Operating a machine
F. Other

Construct a frequency distribution, including relative frequencies, and histogram for the accompanying data from 100 accidents (the percentages agree with those in the cited article):

```
A B D A A F C A C B E B A C
F D B C D A A C B E B C E A
B A A A B C C D F D B B A F
C B A C B E E D A B C E A A
F C B D D D B D C A F A A B
D E A E D B C A F A C D D A
A B A F D C A C B F D A E A
C D
```

30. A **Pareto diagram** is a variation of a histogram for categorical data resulting from a quality control study. Each category represents a different type of product nonconformity or production problem. The categories are ordered so that the one with the largest frequency appears on the far left, then the category with the second largest frequency, and so on. Suppose the following information on nonconformities in circuit packs is obtained: failed component, 126; incorrect component, 210; insufficient solder, 67; excess solder, 54; missing component, 131. Construct a Pareto diagram.

31. The **cumulative frequency** and cumulative relative frequency for a particular class interval are the sum of frequencies and relative frequencies, respectively, for that interval and all intervals lying below it. If, for example, there are four intervals with frequencies 9, 16, 13, and 12, then the cumulative frequencies are 9, 25, 38, and 50, and the cumulative relative frequencies are .18, .50, .76, and 1.00. Compute the cumulative frequencies and cumulative relative frequencies for the data of Exercise 24.

32. Fire load (MJ/m^2) is the heat energy that could be released per square meter of floor area by combustion of contents and the structure itself. The article **"Fire Loads in Office Buildings" (*J. of Structural Engr.*, 1997: 365–368)** gave the following cumulative percentages (read from a graph) for fire loads in a sample of 388 rooms:

Value	0	150	300	450	600
Cumulative %	0	19.3	37.6	62.7	77.5
Value	750	900	1050	1200	1350
Cumulative %	87.2	93.8	95.7	98.6	99.1
Value	1500	1650	1800	1950	
Cumulative %	99.5	99.6	99.8	100.0	

a. Construct a relative frequency histogram and comment on interesting features.

b. What proportion of fire loads are less than 600? At least 1200?

c. What proportion of the loads are between 600 and 1200?

1.3 Measures of Location

Visual summaries of data are excellent tools for obtaining preliminary impressions and insights. More formal data analysis often requires the calculation and interpretation of numerical summary measures. That is, from the data we try to extract several summarizing quantities that might serve to characterize the data set and convey some of its prominent features. Our primary concern will be with numerical data; some comments regarding categorical data appear at the end of the section.

Suppose, then, that our data set is of the form $x_1, x_2, \ldots, x_n$, where each x_i is a number. What features of such a set of numbers are of most interest and deserve emphasis? One important characteristic of a set of numbers is its location, and in particular its center. This section presents methods for describing the location of a data set; in Section 1.4 we will turn to methods for assessing variability in a set of numbers.

The Mean

For a given set of numbers $x_1, x_2, \ldots, x_n$, the most familiar and useful measure of the center is the *mean,* or arithmetic average of the set. Because we will almost always think of the x_i's as constituting a sample, we will often refer to the arithmetic average as the *sample mean* and denote it by $\bar{x}$.

DEFINITION

The **sample mean** $\bar{x}$ of observations $x_1, x_2, \ldots, x_n$ is given by

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The numerator of $\bar{x}$ can be written more informally as Σx_i, where the summation is over all sample observations.

For reporting $\bar{x}$, we recommend using decimal accuracy of one digit more than the accuracy of the x_i's. Thus if observations are stopping distances with $x_1 = 125$, $x_2 = 131$, and so on, we might have $\bar{x} = 127.3$ ft.

EXAMPLE 1.14 Recent years have seen growing commercial interest in the use of what is known as *internally cured concrete*. This concrete contains porous inclusions most commonly in the form of lightweight aggregate (LWA). The article **"Characterizing Lightweight Aggregate Desorption at High Relative Humidities Using a Pressure Plate Apparatus" (*J. of Materials in Civil Engr*, 2012: 961–969)** reported on a study in which researchers examined various physical properties of 14 LWA specimens. Here are the 24-hour water-absorption percentages for the specimens:

$x_1 = 16.0$	$x_2 = 30.5$	$x_3 = 17.7$	$x_4 = 17.5$	$x_5 = 14.1$
$x_6 = 10.0$	$x_7 = 15.6$	$x_8 = 15.0$	$x_9 = 19.1$	$x_{10} = 17.9$
$x_{11} = 18.9$	$x_{12} = 18.5$	$x_{13} = 12.2$	$x_{14} = 6.0$	

Figure 1.14 shows a dotplot of the data; a water-absorption percentage in the mid-teens appears to be "typical." With $\Sigma x_i = 229.0$, the sample mean is

$$\bar{x} = \frac{229.0}{14} = 16.36$$

A physical interpretation of the sample mean demonstrates how it assesses the center of a sample. Think of each dot in the dotplot as representing a 1-lb weight. Then a fulcrum placed with its tip on the horizontal axis will balance precisely when it is located at $\bar{x}$. So the sample mean can be regarded as the balance point of the distribution of observations.

Figure 1.14 Dotplot of the data from Example 1.14 ■

Just as $\bar{x}$ represents the average value of the observations in a sample, the average of all values in the population can be calculated. This average is called the **population mean** and is denoted by the Greek letter μ. When there are N values in the population (a finite population), then μ = (sum of the N population values)/N. We will give a more general definition for μ in Chapters 3 and 4 that applies to both finite and (conceptually) infinite populations. Just as $\bar{x}$ is an interesting and important measure of sample location, μ is an interesting and important (often the most important) characteristic of a population. One of our first tasks in statistical inference will be to present methods based on the sample mean for drawing

conclusions about a population mean. For example, we might use the sample mean $\bar{x} = 16.36$ computed in Example 1.14 as a *point estimate* (a single number that is our "best" guess) of μ = the true average water-absorption percentage for all specimens treated as described.

The mean suffers from one deficiency that makes it an inappropriate measure of center under some circumstances: Its value can be greatly affected by the presence of even a single outlier (unusually large or small observation). For example, if a sample of employees contains nine who earn \$50,000 per year and one whose yearly salary is \$150,000, the sample mean salary is \$60,000; this value certainly does not seem representative of the data. In such situations, it is desirable to employ a measure that is less sensitive to outlying values than $\bar{x}$, and we will momentarily propose one. However, although $\bar{x}$ does have this potential defect, it is still the most widely used measure, largely because there are many populations for which an extreme outlier in the sample would be highly unlikely. When sampling from such a population (a normal or bell-shaped population being the most important example), the sample mean will tend to be stable and quite representative of the sample.

The Median

The word *median* is synonymous with "middle," and the sample median is indeed the middle value once the observations are ordered from smallest to largest. When the observations are denoted by $x_1, \ldots, x_n$, we will use the symbol $\tilde{x}$ to represent the sample median.

DEFINITION

The **sample median** is obtained by first ordering the n observations from smallest to largest (with any repeated values included so that every sample observation appears in the ordered list). Then,

$$\tilde{x} = \begin{cases} \text{The single middle value if } n \text{ is odd} & = \left(\frac{n+1}{2}\right)^{\text{th}} \text{ ordered value} \\ \text{The average of the two middle values if } n \text{ is even} & = \text{average of } \left(\frac{n}{2}\right)^{\text{th}} \text{ and } \left(\frac{n}{2}+1\right)^{\text{th}} \text{ ordered values} \end{cases}$$

EXAMPLE 1.15

People not familiar with classical music might tend to believe that a composer's instructions for playing a particular piece are so specific that the duration would not depend at all on the performer(s). However, there is typically plenty of room for interpretation, and orchestral conductors and musicians take full advantage of this. The author went to the Web site **ArkivMusic.com** and selected a sample of 12 recordings of Beethoven's Symphony No. 9 (the "Choral," a stunningly beautiful work), yielding the following durations (min) listed in increasing order:

62.3 62.8 63.6 65.2 65.7 66.4 67.4 68.4 68.8 70.8 75.7 79.0

Here is a dotplot of the data:

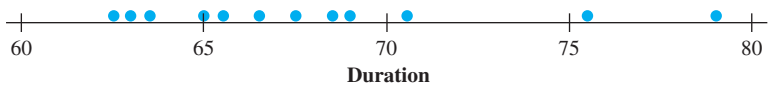


Figure 1.15 Dotplot of the data from Example 1.15

Since $n = 12$ is even, the sample median is the average of the $n/2 = 6^{th}$ and $(n/2 + 1) = 7^{th}$ values from the ordered list:

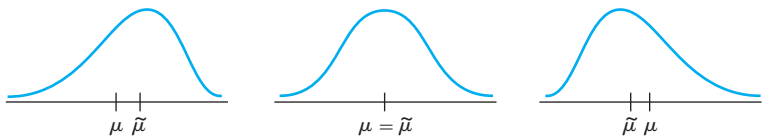
$$\tilde{x} = \frac{66.4 + 67.4}{2} = 66.90$$

Note that if the largest observation 79.0 had not been included in the sample, the resulting sample median for the $n = 11$ remaining observations would have been the single middle value 66.4 (the $[n + 1]/2 = 6^{th}$ ordered value, i.e., the 6^{th} value in from either end of the ordered list). The sample mean is $\bar{x} = \Sigma x_i = 816.1/12 = 68.01$, roughly a minute larger than the median. The mean is pulled out relative to the median because the sample "stretches out" somewhat more on the upper end than on the lower end. ■

The data in Example 1.15 illustrates an important property of $\tilde{x}$ in contrast to $\bar{x}$: The sample median is very insensitive to outliers. If, for example, the two largest x_is are increased from 75.7 and 79.0 to 85.7 and 89.0, respectively, $\tilde{x}$ would be unaffected. Thus, in the treatment of outlying data values, $\bar{x}$ and $\tilde{x}$ are at opposite ends of a spectrum. Both quantities describe where the data is centered, but they will not in general be equal because they focus on different aspects of the sample.

Analogous to $\tilde{x}$ as the middle value in the sample is a middle value in the population, the **population median**, denoted by $\tilde{\mu}$. As with $\bar{x}$ and μ, we can think of using the sample median $\tilde{x}$ to make an inference about $\tilde{\mu}$. In Example 1.15, we might use $\tilde{x} = 66.90$ as an estimate of the median time for the population of all recordings.

The population mean μ and median $\tilde{\mu}$ will not generally be identical. If the population distribution is positively or negatively skewed, as pictured in Figure 1.16, then $\mu \neq \tilde{\mu}$. When this is the case, in making inferences we must first decide which of the two population characteristics is of greater interest and then proceed accordingly.



(a) Negative skew (b) Symmetric (c) Positive skew

Figure 1.16 Three different shapes for a population distribution

Other Measures of Location: Quartiles, Percentiles, and Trimmed Means

The median (population or sample) divides the data set into two parts of equal size. To obtain finer measures of location, we could divide the data into more than two such parts. Roughly speaking, quartiles divide the data set into four equal parts, with the observations above the third quartile constituting the upper quarter of the data set, the second quartile being identical to the median, and the first quartile separating the lower quarter from the upper three-quarters.

Similarly, a data set (sample or population) can be even more finely divided using percentiles; the 99th percentile separates the highest 1% from the bottom 99%, and so on. Unless the number of observations is a multiple of 100, care must be exercised in obtaining percentiles. We will revisit percentiles in Chapter 4 in connection with certain models for infinite populations.

The mean is quite sensitive to a single outlier, whereas the median is impervious to many outliers. Since extreme behavior of either type might be undesirable, we briefly consider alternative measures that are neither as sensitive as $\bar{x}$ nor as insensitive as $\tilde{x}$. To motivate these alternatives, note that $\bar{x}$ and $\tilde{x}$ are at opposite extremes of the same "family" of measures. The mean is the average of all the data, whereas the median results from eliminating all but the middle one or two values and then averaging. To paraphrase, the mean involves trimming 0% from each end of the sample, whereas for the median the maximum possible amount is trimmed from each end. A **trimmed mean** is a compromise between $\bar{x}$ and $\tilde{x}$. A 10% trimmed mean, for example, would be computed by eliminating the smallest 10% and the largest 10% of the sample and then averaging what remains.

EXAMPLE 1.16 The production of Bidri is a traditional craft of India. Bidri wares (bowls, vessels, and so on) are cast from an alloy containing primarily zinc along with some copper. Consider the following observations on copper content (%) for a sample of Bidri artifacts in London's Victoria and Albert Museum **("Enigmas of Bidri," *Surface Engr.*, 2005: 333–339),** listed in increasing order:

2.0 2.4 2.5 2.6 2.6 2.7 2.7 2.8 3.0 3.1 3.2 3.3 3.3
3.4 3.4 3.6 3.6 3.6 3.6 3.7 4.4 4.6 4.7 4.8 5.3 10.1

Figure 1.17 is a dotplot of the data. A prominent feature is the single outlier at the upper end; the distribution is somewhat sparser in the region of larger values than is the case for smaller values. The sample mean and median are 3.65 and 3.35, respectively. A trimmed mean with a trimming percentage of $100(2/26) = 7.7\%$ results from eliminating the two smallest and two largest observations; this gives $\bar{x}_{tr(7.7)} = 3.42$. Trimming here eliminates the larger outlier and so pulls the trimmed mean toward the median.

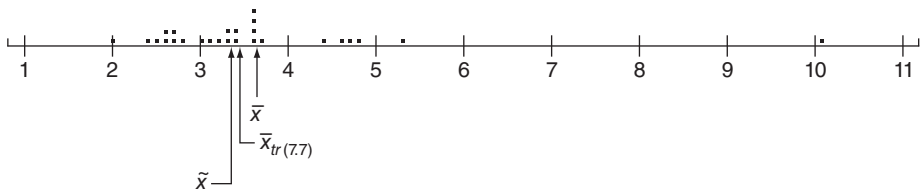


Figure 1.17 Dotplot of copper contents from Example 1.16 ■

A trimmed mean with a moderate trimming percentage—someplace between 5% and 25%—will yield a measure of center that is neither as sensitive to outliers as is the mean nor as insensitive as the median. If the desired trimming percentage is $100\alpha\%$ and $n\alpha$ is not an integer, the trimmed mean must be calculated by interpolation. For example, consider $\alpha = .10$ for a 10% trimming percentage and $n = 26$ as in Example 1.16. Then $\bar{x}_{tr(10)}$ would be the appropriate weighted average of the 7.7% trimmed mean calculated there and the 11.5% trimmed mean resulting from trimming three observations from each end.

Categorical Data and Sample Proportions

When the data is categorical, a frequency distribution or relative frequency distribution provides an effective tabular summary of the data. The natural numerical summary quantities in this situation are the individual frequencies and the relative frequencies. For example, if a survey of individuals who own digital cameras is undertaken to study brand preference, then each individual in the sample would identify the brand of camera that he or she owned, from which we could count the number owning Canon, Sony, Kodak, and so on. Consider sampling a dichotomous population—one that consists of only two categories (such as voted or did not vote in the last election, does or does not own a digital camera, etc.). If we let x denote the number in the sample falling in category 1, then the number in category 2 is $n - x$. The relative frequency or *sample proportion* in category 1 is x/n and the sample proportion in category 2 is $1 - x/n$. Let's denote a response that falls in category 1 by a 1 and a response that falls in category 2 by a 0. A sample size of $n = 10$ might then yield the responses 1, 1, 0, 1, 1, 1, 0, 0, 1, 1. The sample mean for this numerical sample is (since number of 1s $= x = 7$)

$$\frac{x_1 + \cdots + x_n}{n} = \frac{1 + 1 + 0 + \cdots + 1 + 1}{10} = \frac{7}{10} = \frac{x}{n} = \text{sample proportion}$$

More generally, *focus attention on a particular category and code the sample results so that a* 1 *is recorded for an observation in the category and a* 0 *for an observation not in the category. Then the sample proportion of observations in the category is the sample mean of the sequence of* 1*'s and* 0*'s.* Thus a sample mean can be used to summarize the results of a categorical sample. These remarks also apply to situations in which categories are defined by grouping values in a numerical sample or population (e.g., we might be interested in knowing whether individuals have owned their present automobile for at least 5 years, rather than studying the exact length of ownership).

Analogous to the sample proportion x/n of individuals or objects falling in a particular category, let p represent the proportion of those in the entire population falling in the category. As with x/n, p is a quantity between 0 and 1, and while x/n is a sample characteristic, p is a characteristic of the population. The relationship between the two parallels the relationship between $\tilde{x}$ and $\tilde{\mu}$ and between $\bar{x}$ and μ. In particular, we will subsequently use x/n to make inferences about p. If a sample of 100 students from a large university reveals that 38 have Macintosh computers, then we could use $38/100 = .38$ as a point estimate of the proportion of all students at the university who have Macs. Or we might ask whether this sample provides strong evidence for concluding that at least 1/3 of all students are Mac owners. With k categories ($k > 2$), we can use the k sample proportions to answer questions about the population proportions $p_1, \ldots, p_k$.

EXERCISES Section 1.3 (33–43)

33. The **May 1, 2009, issue of *The Montclarian*** reported the following home sale amounts for a sample of homes in Alameda, CA that were sold the previous month (1000s of $):

590 815 575 608 350 1285 408 540 555 679

a. Calculate and interpret the sample mean and median.

b. Suppose the 6th observation had been 985 rather than 1285. How would the mean and median change?

c. Calculate a 20% trimmed mean by first trimming the two smallest and two largest observations.

d. Calculate a 15% trimmed mean.

34. Exposure to microbial products, especially endotoxin, may have an impact on vulnerability to allergic diseases.