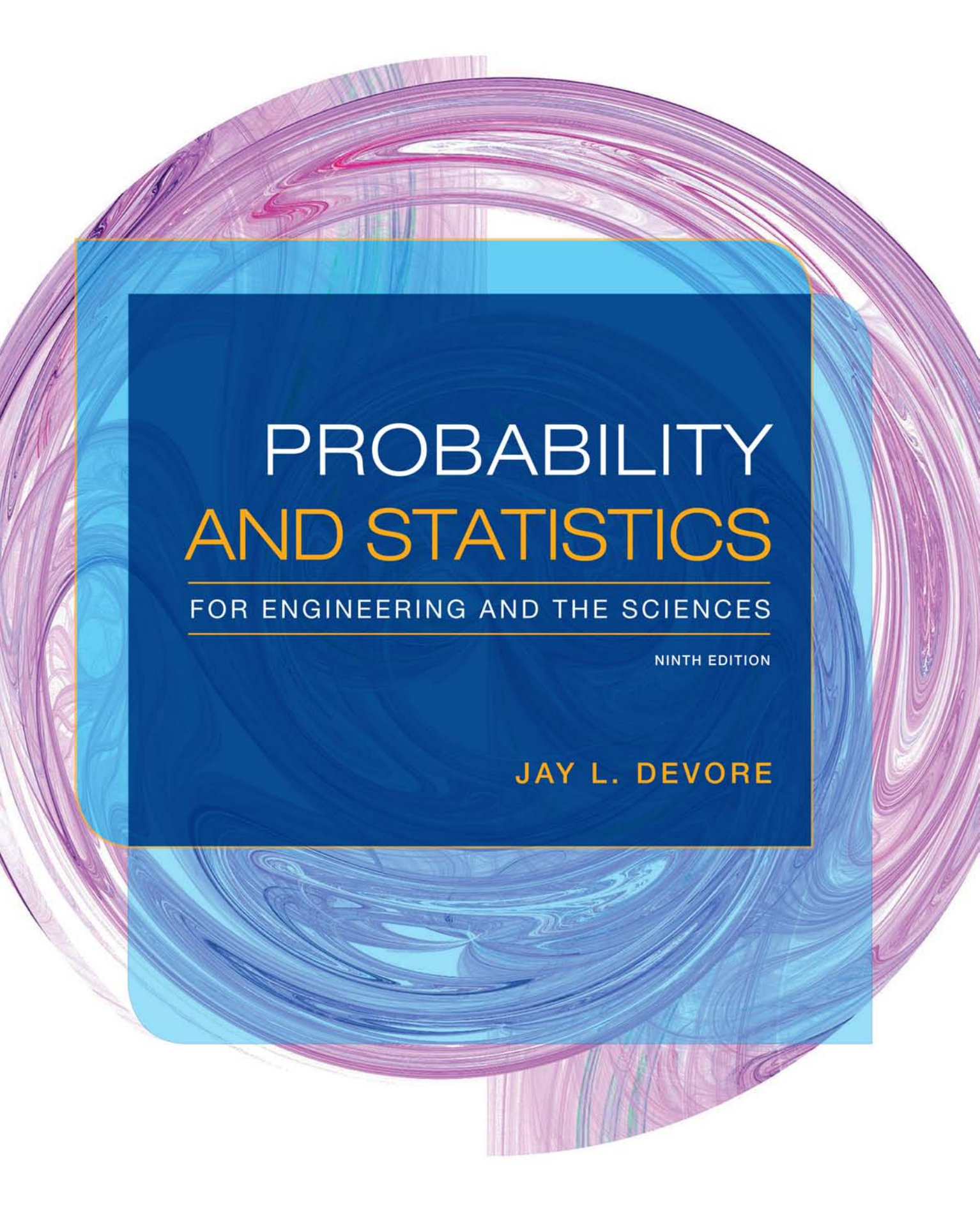
PROBABILITY
AND STATISTICS
FOR ENGINEERING AND THE SCIENCES
NINTH EDITION
JAY L. DEVORE

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
```

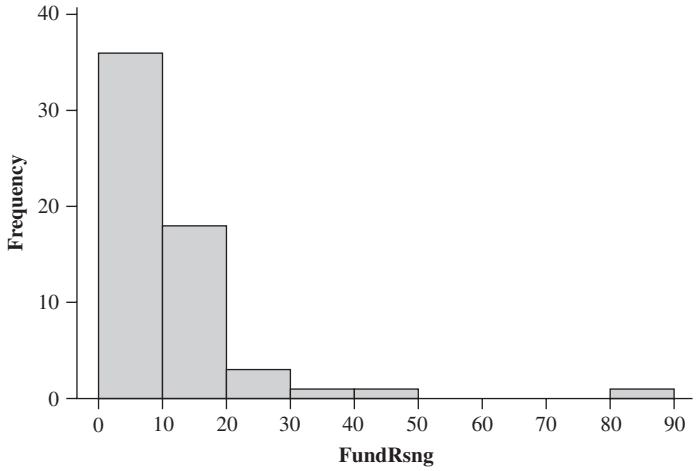

Figure 1.1 A Minitab stem-and-leaf display (tenths digit truncated) and histogram for the charity fundraising percentage data

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.

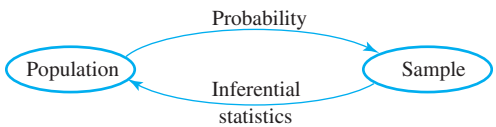

Figure 1.2 The relationship between probability and inferential statistics

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