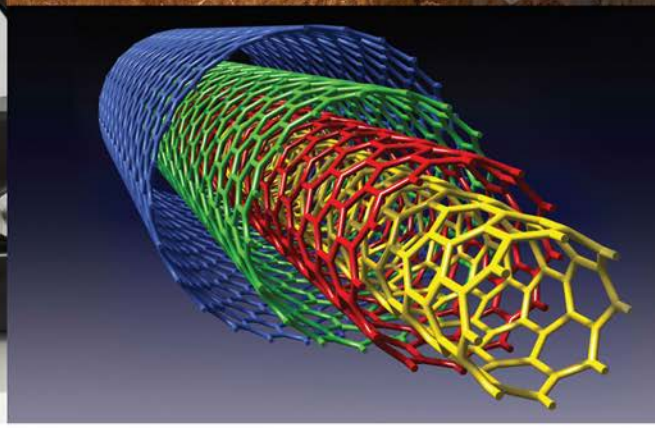
Applied Statistics

FOR ENGINEERS AND SCIENTISTS

THIRD EDITION



Devore - Farnum - Doi

Applied Statistics for Engineers and Scientists

THIRD EDITION

Applied Statistics for Engineers and Scientists

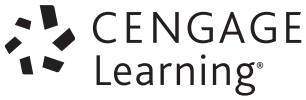
Jay Devore
California Polytechnic State University, San Luis Obispo

Nicholas Farnum
California State University, Fullerton

Jimmy Doi
California Polytechnic State University, San Luis Obispo

CENGAGE Learning®

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



Applied Statistics for Engineers and Scientists, Third Edition
Jay Devore, Nicholas Farnum, Jimmy Doi

Publisher: Richard Stratton

Senior Sponsoring Editor: Molly Taylor

Development Editor: Laura Wheel

Editorial Assistant: Danielle Hallock

Associate Media Editor: Andrew Coppola

Brand Manager: Gordon Lee

Content Project Manager: Jill Quinn

Senior Art Director: Linda May

Manufacturing Planner: Sandee Milewski

Rights Acquisition Specialist: Shalice Shah-Caldwell

Production Service: Prashant Kumar Das, MPS Limited

Text and Cover Designer: Jenny Willingham

Cover Image: Female Scientist: wavebreakmedia/Shutterstock.com; Solar Panels: portumen/Shutterstock.com; Nanotubes: PASIEKA/SPL/Getty images

Compositor: MPS Limited

Library of Congress Control Number: 2013944181

ISBN-13: 978-1-133-11136-8

ISBN-10: 1-133-11136-X

Cengage Learning
200 First Stamford Place, 4th Floor
Stamford, CT 06902
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 **international.cengage.com/region**

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

For your course and 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**

Instructors: Please visit **login.cengage.com** and log in to access instructor-specific resources.

Printed in the United States of America
1 2 3 4 5 6 7 17 16 15 14 13

This book is dedicated to

My grandsons, Philip and Elliot
J.L.D.

My grandchildren, Ava and Leo
N.R.F.

My wife and daughter, Midori and Alicia
J.A.D.

Contents

Preface

PURPOSE

The use of statistical models and methods for describing and analyzing data has become common practice in virtually all scientific disciplines. This book provides 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. It is appropriate for courses of one term (semester or quarter) in duration.

APPROACH

Students in a statistics course designed to serve other majors are too often initially skeptical of the value and relevance of the subject matter. Our experience, however, is that students *can* be turned on to the subject by the use of good examples and exercises that blend their everyday experiences with their scientific interests. We have worked hard to find examples involving real, rather than artificial, data—data that someone thought was worth collecting and analyzing. Many of the methods presented throughout the book are illustrated by analyzing data taken from a published source.

The exercises form a very important component of the book. A really good lecturer can deceive students into thinking they have an excellent mastery of the subject, only to discover otherwise when they start working problems. We have therefore provided a rich assortment of exercises designed to reinforce understanding of the material. A substantial majority of these are based on real data, and we have tried as much as possible to avoid mathematical manipulation for its own sake. Someone who attempts a good portion of the exercises will gain a greater appreciation of the scope and applicability of the subject than would be gleaned simply by reading the text.

Sometimes the reader may be unfamiliar with the context of a particular problem situation (as indeed we often were), but we believe that students will find scenarios,

such as the one below, more appealing than they would in patently artificial situations dealing with widgets or brand A versus brand B.

64. The use of microorganisms to dissolve metals from ores has offered an ecologically friendly and less expensive alternative to traditional methods. The dissolution of metals by this method can be done in a two-stage bioleaching process: (1) microorganisms are grown in culture to produce metabolites (e.g. organic acids) and (2) ore is added to the culture medium to initiate leaching. The article "Two-Stage Fungal Leaching of Vanadium from Uranium Ore Residue of the Leaching Stage using Statistical Experimental Design" (*Annals of Nuclear Energy*, 2013: 48–52) reported on a two-stage bioleaching process of vanadium by using the fungus *Aspergillus niger*. In one study, the authors examined the impact of the variables x_1 = pH, x_2 = sucrose concentration (g/L), and x_3 = spore population (10^6 cells/ml) on y = oxalic acid production (mg/L). The accompanying SAS output resulted from a request to fit the model with predictors x_1, x_2, and x_3 only.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5861301	1953767	7.53	0.0052
Error	11	2855951	259632		
Corrected Total	14	8717252			

Fitting the complete second-order model resulted in SSResid = 541,632. Carry out a test at significance level .01 to decide whether at least one of the second-order predictors provides useful information about oxalic acid production.

MATHEMATICAL AND COMPUTING LEVEL

The exposition is relatively modest in terms of mathematical development. Limited use of univariate calculus is made in the first two chapters, and a bit of univariate and multivariate calculus is employed later on. Matrix algebra appears nowhere in the book. Thus virtually all of the exposition should be accessible to those whose mathematical background includes one semester or two quarters of differential and integral calculus.

The computer is an indispensable tool these days for organizing, displaying, and analyzing data. We have included many examples, as illustrated on the next page, of output from the most widely used statistical computer packages, including Minitab, SAS, R, and JMP, both to convince students that the statistical methods discussed herein are available in these packages and to expose them to format and contents of typical output. Because availability of packages and nature of platforms vary widely from institution to institution, we decided not to include instructions for obtaining output from any particular package. Based on our experience, it should be straightforward to supplement the text by independently introducing students to any one of the aforementioned packages. They can then be asked to use the computer in working the many problems that contain raw data.

Example 10.2 Over the past decade researchers and consumers have shown increased interest in renewable fuels such as biodiesel, a form of diesel fuel derived from vegetable oils and animal fats. According to www.fueleconomy.gov, compared to petroleum diesel, the advantages of using biodiesel include its nontoxicity, biodegradability,

and lower greenhouse gas emissions. One popular biodiesel fuel is fatty acid ethyl ester (FAEE). The authors of "Application of the Full Factorial Design to Optimization of Base-Catalyzed Sunflower Oil Ethanolysis" (*Fuel*, 2013: 433–442) performed an experiment to determine optimal process conditions for producing FAEE from the ethanolysis of sunflower oils. In one study, the effects of three process factors on FAEE purity (%) were investigated.

Factor	Factor name	Factor levels
A	Reaction Temperature	25°C, 50°C, 75°C
B	Ethanol-to-oil molar ratio	6:1, 9:1, 12:1
C	Catalyst loading	.75 wt.%, 1.00 wt.%, 1.25 wt.%

(See Page 467 for the complete data)

Plots of all two-factor interactions are shown in Figure 10.18, along with the main effects Plots for the three factors. Suppose we are interested in maximizing the value of the response variable, FAEE purity. Looking at the interaction plots, the combination of factor levels that best accomplishes this objective is $A = 75°C$, $B = 12{:}1$, and $C = 1.25\%$. In this example, the conclusions from the interaction plots agree with the conclusions that we would have drawn from inspecting the main effects plots.

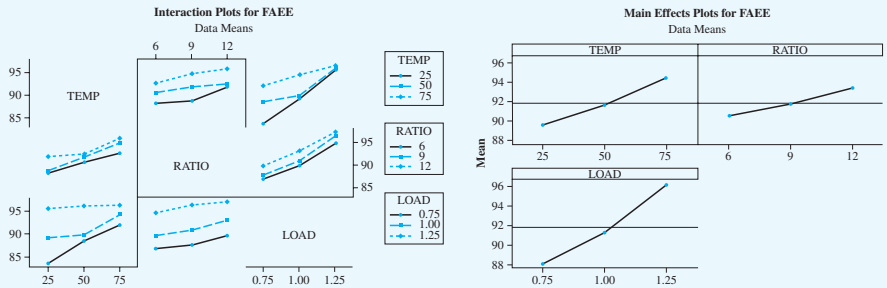


Figure 10.18 Two-factor interaction plots and main effects plots for Example 10.2

FOCUS AND CONTENT

We have written this book for an audience whose primary interest is in statistical methodology and the analysis of data. The ordering of topics herein is rather different from what is found in virtually all competing texts. The usual approach is to inject a heavy dose of probability at the outset, then develop probability distributions and use these as a basis for inferential methods (drawing conclusions from data). Unfortunately, an introductory one-term course rarely allows sufficient time for comprehensive treatments of both probability and statistical inference. If probability is emphasized, statistics gets short shrift. An additional problem is that many students find probability to be a difficult and

intimidating subject, so starting out in this way creates an aura of mathematical formalism that makes it all too easy to lose sight of the applied and practical aspects of statistics.

Certainly descriptive statistical methods can be developed in detail with virtually no probability background, and even an understanding of the most commonly used inferential techniques requires familiarity with only the most basic of probability properties. So we decided to proceed along a path first blazed by David Moore and George McCabe in their book *Introduction to the Practice of Statistics*, written for a nonscience audience. In their Chapter 1, the normal distribution is introduced and employed to address many interesting questions, whereas probability does not surface until much later in the book. Our Chapter 1 first presents some basic concepts and terminology, continues with an introduction to some descriptive techniques, and then extends the notion of a histogram for sample data to a distribution of values for an entire population or process. This allows us to develop and use not only the family of normal distributions but also other continuous and discrete distributions such as the lognormal, Weibull, Poisson, and binomial. Chapter 2 covers numerical summary measures for sample data (e.g., the sample mean $\bar{x}$ and sample standard deviation s) in tandem with analogous measures for populations and processes (e.g., the population or process mean μ and standard deviation σ).

The focus of the first two chapters is on univariate data (observations on or values of a single variable, such as tensile strength). In the third chapter we consider descriptive methods for bivariate data (e.g., measuring both thickness and strength for wire specimens) and then multivariate data, emphasizing in particular correlation and regression. This chapter should be especially useful for courses in which there is insufficient time to cover regression models from a probabilistic viewpoint (such models and inferences based on them are the subject of Chapter 11).

Most other books intended for our target audience say rather little about how data is obtained. Yet statistics has much to say not only about how to analyze data once it is available but also about sensible and efficient techniques for collecting data. Several lower-level texts, notably the one by Moore and McCabe cited earlier, successfully and entertainingly covered this territory prior to probability and inference, and we follow their lead with our Chapter 4. Sampling and experimental design are discussed, and the last section contains an introduction to various aspects of measurement.

At last probability makes its appearance in Chapter 5. Our minimalist treatment of this subject is intended to move readers expeditiously into the inferential part of the book. Since only the notion of probability as limiting or long-run relative frequency is needed to understand the basis for most of the usual inferential procedures, little time is spent on topics such as addition and multiplication rules and conditional probability, and no material on counting techniques is included here (combinations enter briefly in Chapter 1 in connection with the binomial distribution). The concept of a random variable and its probability distribution is then introduced and related to the distributional material in Chapter 1. Finally, the notion of a statistic and its sampling distribution is discussed and illustrated.

The remaining six chapters focus on the most widely used methods from statistical inference. Descriptive techniques from earlier chapters, such as boxplots and quantile plots, are employed in many of our examples. Chapter 6 covers topics from quality control and reliability. Estimation and various statistical intervals—confidence, prediction,

and tolerance—are introduced in Chapter 7. Hypothesis testing is discussed in Chapter 8. Chapter 9 covers the analysis of variance for comparing more than two populations or treatments, and these ideas are extended in Chapter 10 to the analysis of data from designed multifactor experiments. Finally, regression models and associated inferential procedures are covered in Chapter 11.

SOME SUGGESTIONS CONCERNING COVERAGE

It should be possible to cover virtually all the material in the book in a semester-long course that meets four hours per week. For a course of this duration that meets only three times per week or for a one-quarter course, some pruning will have to be done (perhaps combined with reading assignments on topics not discussed in lecture). The first four sections of Chapter 1 are essential, but Section 5 on other (than the normal) continuous distributions and Section 6 on the binomial and Poisson distributions can be covered very lightly or even omitted altogether. The first two sections of Chapter 2, on measures of center and spread, are also required. The material on more detailed summary measures (e.g., boxplots) in Section 3 can be just touched on or skipped, and quantile plots from Section 4 can be presented very quickly.

When time does not allow for coverage of inferences in regression, we strongly recommend that at least a bit of bivariate descriptive methods from Chapter 3 be covered. At minimum, this could consume just two or three one-hour lectures in which scatterplots, correlation, and fitting a line by least squares are discussed. More time would provide the opportunity to introduce r^2 as an assessment of fit, nonlinear relationships, and even multiple regression. If inference in regression is to be covered, this chapter can be skipped over for the moment and then combined with Chapter 11 at the end of the course.

Chapter 4, on obtaining data, can be covered next or postponed until later. There is no mathematics here, only some definitions and examples, so this is one place where a minimal amount of lecture time can be expended along with a request that students read on their own. Most of Chapter 5 is crucial; inferential methods cannot be understood without a modest exposure to probability and sampling distributions of various statistics. The quality control and reliability techniques of Chapter 6 are attractive applications of sampling distribution and probability properties. When time is limited, as few as two lectures might be devoted to some general concepts and a single type of control chart. Another possibility is to postpone this material until after hypothesis testing has been introduced.

From this point on, it is local option as to what is covered and in how much detail. We certainly believe that students deserve at least minimal exposure to point estimation, confidence intervals, and hypothesis testing. Time may permit presentation of just some selected one-sample procedures (Sections 7.1, 7.2, 8.1, and perhaps a bit of Sections 7.4 and 8.2). A longer course would accommodate topics from among prediction and tolerance intervals, two-sample situations, chi-squared tests, testing the plausibility of some particular type of distribution (e.g., testing the assumption that the data came from a normal distribution), analysis of variance and experimental design, and more on regression.

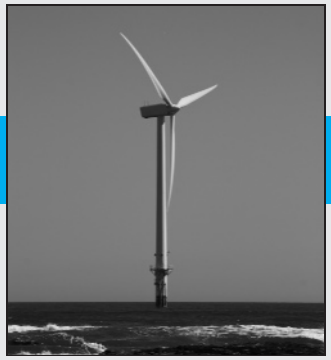
Changes for the Third Edition

- There are nearly 200 new exercises and 40 new examples, most of which include real data or other information from published sources.
- Chapter 1 contains a new subsection on "The Scope of Modern Statistics" to illustrate how statisticians continue to develop new methodology while working on problems in a wide spectrum of disciplines.
- Section 8.3, on hypothesis testing based on categorical data, now contains a subsection on Fisher's Exact Test that is a useful alternative when assumptions for the standard chi-squared test fail.
- Section 11.6, on regression, now contains a subsection on the multiple logistic regression model that accommodates multiple predictor variables for a dichotomous response.
- In general, the exposition has been polished, tightened, and improved.

ACKNOWLEDGMENTS

We greatly appreciate the feedback and useful advice from the many individuals who reviewed various parts of our manuscript: Christine Anderson-Cook, Virginia Tech; Olcay Arslan, St. Cloud State; Peyton Cook, The University of Tulsa; Jean-Yves "Pip" Courbois, University of Washington; Charles Donaghey, University of Houston; Dale O. Everson, University of Idaho; William P. Fox, United States Military Academy; William Fulkerson, Deere & Company; Roger Hoerl, General Electric Company; Marianne Huebner, Michigan State University; Alan M. Johnson, University of Arkansas, Little Rock; Steven L. Johnson, University of Arkansas; Janusz Kawczak, University of North Carolina, Charlotte; Mohammed Kazemi, University of North Carolina, Charlotte; David P. Kessler, Purdue University; Barbara McKinney, Western Michigan University; Jang W. Ra, University of Alaska, Anchorage; John Ramberg, University of Arizona; Stephen E. Rigdon, Southern Illinois University at Edwardsville; Amy L. Rocha, San Jose State University; Joe Romano, Stanford University; Lewis H. Shoemaker, Millersville University; and Paul Wilson, Rochester Institute of Technology.

The editorial and production services provided by numerous people from Cengage Learning are greatly appreciated, especially the support of Shaylin Walsh, Laura Wheel, and Jill Quinn. It was indeed a great pleasure to have Prashant Kumar Das overseeing production of the book; his attention to detail, timely feedback, and willingness to tolerate the authors' idiosyncrasies made our work during production much more tolerable than would otherwise have been the case. A special thanks goes to Soma Roy for her accuracy checking and work on the solutions manuals. Finally, the continuing support of family, colleagues, and friends has helped smooth out the bumps in the road. We are truly grateful to all of you.



Crepesoles/Shutterstock.com

1 Data and Distributions

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 an effort to determine the "greenest" way to travel. The article titled "Carbon Conundrum" (*Consumer Reports*, 2008: 9) described websites that help consumers calculate carbon output. The results for carbon output for a flight from New York to Los Angeles appear in the accompanying table.

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

Substantial disagreement clearly exists among these online calculators as to exactly how much carbon is emitted, characterized in the article as "from a ballerina's to Bigfoot's." A website also 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.

In Chapters 1–3, we concentrate on describing and summarizing statistical information obtained from populations or processes under investigation. Chapter 4 discusses how information can be collected either by the mechanism of sampling or by designing

and carrying out an experiment. Chapter 5 formalizes the notion of randomness and uncertainty by introducing the language of probability. The remainder of the book focuses on the development of inferential methods for drawing interesting conclusions from data in a wide variety of situations. We hope you will find the subject matter and our presentation to be as interesting, relevant, and exciting as we do.

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 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 generally denote variables by lowercase letters from the end of our alphabet. Examples include

x = gender of a graduating engineer
y = number of major defects on a newly manufactured automobile
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 lifetimes (hours) of brand X batteries put to a certain use is a numerical univariate data set:

5.6 5.1 6.2 6.0 5.8 6.5 5.8 5.5

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 two variables. 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 scatterplots 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, and R. The R software can be downloaded without charge from www.r-project.org.

Example 1.1

Charity is a big business in the United States. The website charitynavigator.com gives information on approximately 5500 charitable organizations, and many smaller charities fly below the navigator's radar screen. Some charities operate very efficiently, with fund-raising and administrative expenses only a small percentage of total expenses, whereas others spend a high percentage of what they take in to perform the same activities. Here is data on fund-raising 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, making sense of the data's most prominent features is difficult: What is a typical (i.e., representative) value? Are values highly concentrated

about a typical value or are they quite dispersed? Are there any gaps in the data? What fraction of the values are less than 20%? Figure 1.1 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. A substantial majority of the charities in the sample obviously spend less than 20% on fund-raising, and only a few percentages might be viewed as beyond the bounds of sensible practice.

```
Stem-and-leaf of FundRsng N = 60
Leaf Unit = 1.0
   0  011111222233333344
   0  55556666666778888
   1  0001222244
   1  55666789
   2  01
   2  6
   3  4
   3
   4
   4  8
   5
   5
   6
   6
   7
   7
   8  3
```

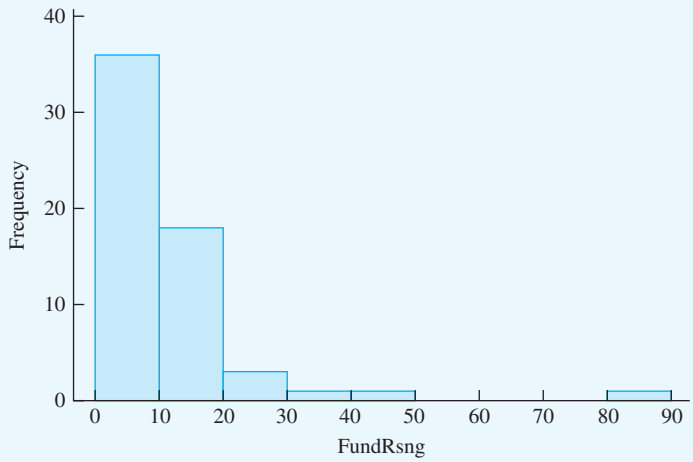


Figure 1.1 A Minitab stem-and-leaf display (10ths digit truncated) and histogram for the charity fund-raising percentage data

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 megapascals, 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 Scope of Modern Statistics

Statistical methodology is commonly employed by investigators in virtually every discipline, 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 here have become established and trusted tools in the arsenal of those who work with data. Meanwhile, statisticians continue to develop new models to describe

randomness and uncertainty and new methodology to analyze 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 *APS* is short for the *Annals of Applied Statistics*, just two of the many prominent journals in the discipline):

- "Application of Branching Models in the Study of Invasive Species" (*JASA*, 2012: 467–476): Seismologists often predict earthquake occurrences using what is known as *epidemic-type aftershock sequence* (ETAS) models. The name stems from the model feature that allows earthquakes to cause aftershocks, which in turn may induce subsequent aftershocks, and so on, thereby generating a cascading effect. The authors propose the use of ETAS models in studying invasive plant and animal species. In particular, the article considers the spread of an invasive species in Costa Rica (*Musa velutina*, or red banana). The authors determine the estimated spatial–temporal rate of spread of red banana plants using a space–time ETAS model.
- "Spatio-Spectral Mixed-Effects Model for Functional Magnetic Resonance Imaging Data" (*JASA*, 2012: 568–577): For many years, scientists have attempted to model cognitive control-related activation among specific regions of the human brain. Researchers measure this brain activity through *functional magnetic resonance imaging* (fMRI). fMRI data often exhibit spatial and temporal correlations (i.e., observations made at nearby locations or time points are often strongly related). Standard approaches to fMRI analysis, however, fail to incorporate these relationships. The article proposes a statistical model to study activation in specific regions in the prefrontal cortex while also incorporating the underlying spatio–temporal correlations. The authors provide a simulation study that shows that significant errors can occur by ignoring the correlation structure in the network.
- "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 at least partly attributable 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 from New York City in 2006; 15 officers from the pool of 3000 regularly involved in pedestrian stops were identified as having stopped a substantially greater fraction of black and Hispanic people than what would be predicted if bias were absent.

- "Measuring the Vulnerability of the Uruguayan Population to Vector-Borne Diseases via Spatially Hierarchical Factor Models" (*APS*, 2012: 284–303): Vector-borne diseases are illnesses caused by infections transmitted to people by organisms such as insects and spiders. According to the World Health Organization, the most deadly vector-borne disease is malaria, which kills more than 1 million people annually, mostly African children under age five. The authors develop a statistical index to model the vulnerability of Uruguayans to vector-borne diseases by accounting for variation attributable to factors such as different census tracts within cities and different cities in the country.
- "Self-Exciting Hurdle Models for Terrorist Activity" (*APS*, 2012: 106–124): The authors develop 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.
- "The BARISTA: A Model for Bid Arrivals in Online Auctions" (*APS*, 2007: 412–441): Online auctions such as those on eBay and uBid often have characteristics that differentiate them from traditional auctions. One particularly important such property 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 ARivals In STAges) model for describing the way in which bids arrive that allows for higher bidding intensity not only at the outset of the auction but 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 information now appears with increasing frequency in the popular media, and occasionally the spotlight is even turned on statisticians. For example, "Behind Cancer Guidelines, Quest for Data," a *New York Times* article from November 23, 2009, reported that the new science for cancer investigations and more sophisticated methods for data analysis spurred the U.S. Preventive Services task force to reexamine 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, in particular one that mammograms every two years give nearly the same benefit as annual ones and confer only half the risk of harm. Donald Berry, a 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.

We hope you will become increasingly convinced of the importance and relevance of the discipline of statistics as you dig more deeply into the book and subject. We also anticipate you'll be intrigued enough to want to continue your statistical education beyond your current course.

Enumerative Versus Analytic Studies

W. E. Deming was a very influential American statistician whose ideas concerning the use of statistical methods in industrial production found great favor with Japanese companies in the years after World War II. He used the phrase **enumerative study** to describe investigations involving a finite collection of identifiable, unchanging objects that make up a population. In such studies, a *sampling frame*—that is, a listing of the objects to be sampled—is available or can be created. One example of such a frame is the collection of all signatures on petitions to qualify an initiative for inclusion on the ballot for an upcoming election. A sample is usually selected to ascertain whether the number of *valid* signatures exceeds a specified value. The variable on which observations are made is dichotomous, the two possible values being *valid* (S, for success) and *not valid* (F, for failure). As another example, the frame may contain serial numbers of all ovens manufactured by a particular company during a particular period. A sample may be selected to infer something about the average actual temperature of these units when the temperature control is set to 400°F (an inference about the population mean temperature).

Many problem situations faced by engineers involve some sort of ongoing **process**—a group of interrelated activities undertaken to accomplish some objective—rather than a specified, unchanging population. An investigator wants to learn something about how the process is operating so that the process can then be modified to better achieve the desired goal. Deming described such scenarios as **analytic studies.**

Example 1.3

The process of making ignition keys for automobiles consists of trimming and pressing raw key blanks, cutting grooves and notches, and then plating the keys. Dimensions associated with groove and notch cutting are crucial to proper key functioning. There will always be "normal" variation in dimensions because of fluctuations in materials, worker behavior, and environmental conditions. It is important, though, to monitor production to ensure that there are no unusual sources of variation, such as incorrect machine settings or contaminated material, which might result in nonconforming units or substantial changes in product characteristics. For this purpose, a sample (subgroup) of five keys is selected every 20 minutes, and critical dimensions are measured. Here are a few of the resulting observations for one particular dimension (in thousandths of an inch):

Subgroup 1:	6.1	8.4	7.6	7.5	4.4
Subgroup 2:	8.8	8.3	5.9	7.4	7.6
Subgroup 3:	8.0	7.5	7.0	6.8	9.3

This is indeed sample data, which can be used as a basis for drawing conclusions. However, the conclusions are about production process behavior rather than about a particular population of keys.

Analytic studies sometimes involve figuring out what actions to take to improve the performance of a future product.