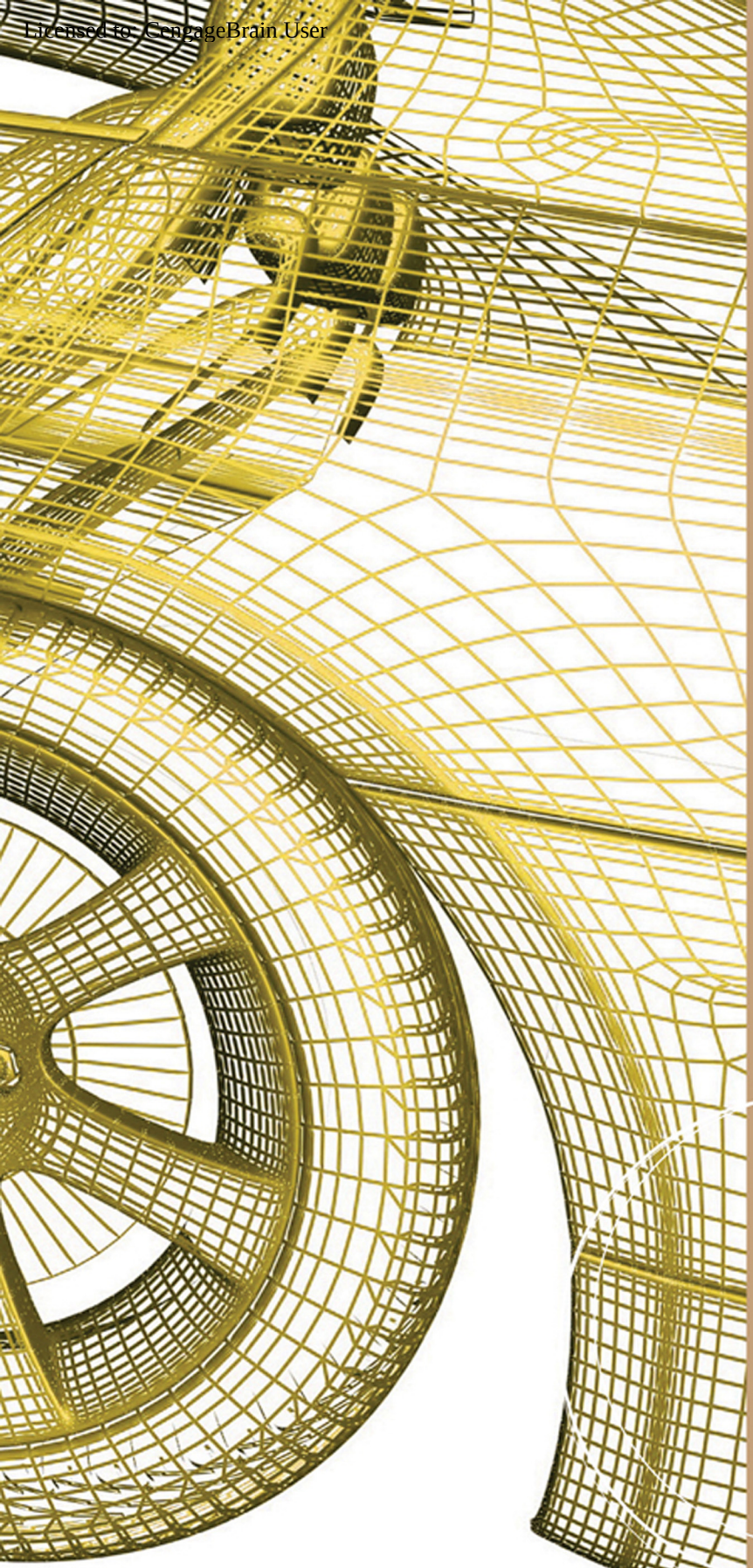


8th edition

# Probability and Statistics for Engineering and the Sciences

Jay L. Devore

International Edition

BROOKS/COLE
CENGAGE Learning™

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

Editor in Chief: Michelle Julet
Publisher: Richard Stratton
Senior Sponsoring Editor: Molly Taylor
Senior Development Editor: Jay Campbell
Assistant Editor: Shaylin Walsh
Media Editor: Andrew Coppola
Marketing Manager: Ashley Pickering
Marketing Communications Manager: Mary Anne Payumo
Content Project Manager: Cathy Brooks
Art Director: Linda Helcher
Print Buyer: Diane Gibbons
Rights Acquisitions Specialists: Image: Mandy Groszko; Text: Katie Huha
Production Service: Integra-Chicago
Text Designer: Diane Beasley
Cover Designer: Rokusek Design



International Student Edition:
ISBN-13: 978-0-8400-6827-9
ISBN-10: 0-8400-6827-1

Cengage Learning International Offices

**Asia**
cengageasia.com
tel: (65) 6410 1200

**Australia/New Zealand**
cengage.com.au
tel: (61) 3 9685 4111

**Brazil**
cengage.com.br
tel: (011) 3665 9900

**India**
cengage.co.in
tel: (91) 11 30484837/38

**Latin America**
cengage.com.mx
tel: +52 (55) 1500 6000

**UK/Europe/Middle East/Africa**
cengage.co.uk
tel: (44) 207 067 2500

**Represented in Canada by Nelson Education, Ltd.**
tel: (416) 752 9100 / (800) 668 0671
nelson.com

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.

For product information: **www.cengage.com/international**
Visit your local office: **www.cengage.com/global**
Visit our corporate website: **www.cengage.com**

Printed in Canada
2 3 4 5 14 13 12 11

# 1 Overview and Descriptive Statistics

"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 arises in the course of performing emissions testing on motor vehicles. The expense and time requirements of the Federal Test Procedure (FTP) preclude its widespread use in vehicle inspection programs. As a result, many agencies have developed less costly and quicker tests, which it is hoped replicate FTP results. According to the journal article "Motor

Vehicle Emissions Variability" (*J. of the Air and Waste Mgmt. Assoc.*, 1996: 667–675), the acceptance of the FTP as a gold standard has led to the widespread belief that repeated measurements on the same vehicle would yield identical (or nearly identical) results. The authors of the article applied the FTP to seven vehicles characterized as "high emitters." Here are the results for one such vehicle:

| | | | | |
|---|---|---|---|---|
| *HC (gm/mile)* | 13.8 | 18.3 | 32.2 | 32.5 |
| *CO (gm/mile)* | 118 | 149 | 232 | 236 |

The substantial variation in both the HC and CO measurements casts considerable doubt on conventional wisdom and makes it much more difficult to make precise assessments about emissions levels.

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 lifetimes (hours) of brand D 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 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, S-Plus, and R. The R software can be downloaded without charge from the site http://www.r-project.org.

**Example 1.1** Charity is a big business in the United States. The Web site charitynavigator.com gives information on roughly 5500 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

```
Stem-and-leaf of FundRsng N = 60
Leaf Unit = 1.0
    0 | 01111122223333333344
    0 | 55556666666778888
    1 | 0001222244
    1 | 55666789
    2 | 01
    2 | 6
    3 | 4
    3 |
    4 |
    4 | 8
    5 |
    5 |
    6 |
    6 |
    7 |
    7 |
    8 | 3
```

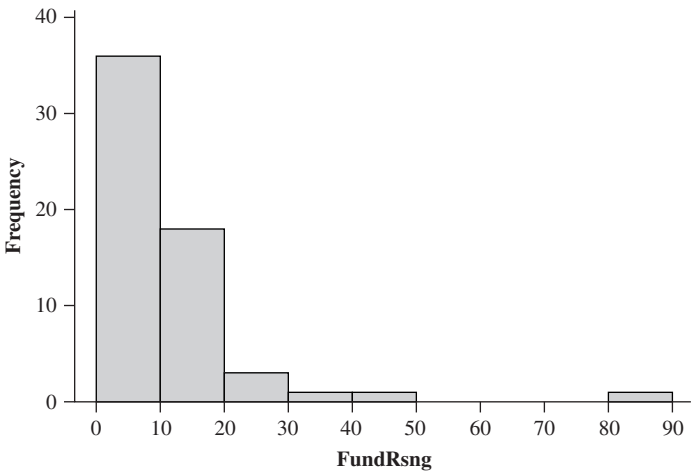



**Figure 1.1** A Minitab stem-and-leaf display (tenths digit truncated) and histogram for the charity fundraising percentage data

gaps in the data, what fraction of the values are less than 20%, and so on. 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. 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) = $1.45 \times 10^{-4}$ 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

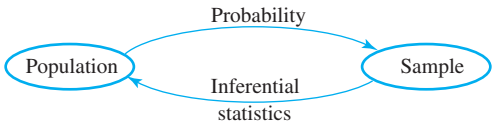



**Figure 1.2** The relationship between probability and inferential statistics

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.

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):

- "Modeling Spatiotemporal Forest Health Monitoring Data" (*JASA,* 2009: 899–911): Forest health monitoring systems were set up across Europe in the 1980s in response to concerns about air-pollution-related forest dieback, and have continued operation with a more recent focus on threats from climate change and increased ozone levels. The authors develop a quantitative description of tree crown defoliation, an indicator of tree health.
- "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.

- "Analysis of Episodic Data with Application to Recurrent Pulmonary Exacerbations in Cystic Fibrosis Patients" (*JASA,* 2008: 498–510): The analysis of recurrent medical events such as migraine headaches should take into account not only when such events first occur but also how long they last—length of episodes may contain important information about the severity of the disease or malady, associated medical costs, and the quality of life. The article proposes a technique that summarizes both episode frequency and length of episodes, and allows effects of characteristics that cause episode occurrence to vary over time. The technique is applied to data on cystic fibrosis patients (CF is a serious genetic disorder affecting sweat and other glands).
- "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 ARivals 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 computer'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 manufacturer of DVD players might want information about customer satisfaction for units produced during the previous year. If three different models were manufactured and sold, a separate sample could be selected from each of the three corresponding strata. This would result in information on all three models and ensure that no one model was 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.

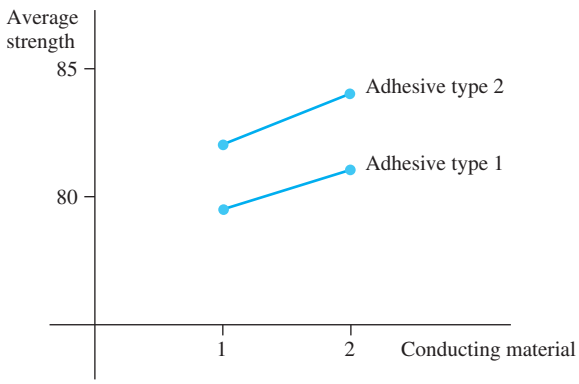



**Figure 1.3** Average bond strengths in Example 1.5

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.** 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.

**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.** 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

**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 techniques. 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. Thus if {29.7, 31.6, 30.9} and {28.7, 29.5, 29.4, 30.3} are thermal-efficiency measurements for two different types of diesel engines, then $m = 3$ 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 $i$th 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.

If the data set consists of exam scores, each between 0 and 100, the score of 83 would have a stem of 8 and a leaf of 3. For a data set of automobile fuel efficiencies (mpg), all between 8.1 and 47.8, we could use the tens digit as the stem, so 32.6 would then have a leaf of 2.6. In general, a display based on between 5 and 20 stems is recommended.

**Example 1.6** The use of alcohol by college students is of great concern not only to those in the academic community but also, because of potential health and safety consequences, to society at large. The article "Health and Behavioral Consequences of Binge Drinking in College" (*J. of the Amer. Med. Assoc.,* 1994: 1672–1677) reported on a comprehensive study of heavy drinking on campuses across the United States. A binge episode was defined as five or more drinks in a row for males and four or more for females. Figure 1.4 shows a stem-and-leaf display of 140 values of $x$ = the percentage of undergraduate students who are binge drinkers. (These values were not given in the cited article, but our display agrees with a picture of the data that did appear.)

```
0 | 4
1 | 1345678889
2 | 1223456666777889999          Stem: tens digit
3 | 011223334455566667777788899999   Leaf: ones digit
4 | 111222223344445566666677788888999
5 | 00111222233455666667777888899
6 | 01111244455666778
```

**Figure 1.4** Stem-and-leaf display for the percentage of binge drinkers at each of the 140 colleges

The first leaf on the stem 2 row is 1, which tells us that 21% of the students at one of the colleges in the sample were binge drinkers. Without the identification of stem digits and leaf digits on the display, we wouldn't know whether the stem 2, leaf 1 observation should be read as 21%, 2.1%, or .21%.

When creating a display by hand, ordering the leaves from smallest to largest on each line can be time-consuming. This ordering usually contributes little if any extra information. Suppose the observations had been listed in alphabetical order by school name, as

16% 33% 64% 37% 31% . . .

Then placing these values on the display in this order would result in the stem 1 row having 6 as its first leaf, and the beginning of the stem 3 row would be

3 | 371 . . .

The display suggests that a typical or representative value is in the stem 4 row, perhaps in the mid-40% range. The observations are not highly concentrated about this typical value, as would be the case if all values were between 20% and 49%. The display rises to a single peak as we move downward, and then declines; there are no gaps in the display. The shape of the display is not perfectly symmetric, but instead appears to stretch out a bit more in the direction of low leaves than in the direction of high leaves. Lastly, there are no observations that are unusually far from the bulk of the data (no *outliers*), as would be the case if one of the 26% values had instead been 86%. The most surprising feature of this data is that, at most colleges in the sample, at least one-quarter of the students are binge drinkers. The problem of heavy drinking on campuses is much more pervasive than many had suspected. ■

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 location of peaks
- presence of any outlying values

**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.

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.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 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

## 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** Here is data on state-by-state appropriations for higher education as a percentage of state and local tax revenue for the fiscal year 2006–2007 (from the *Statistical Abstract of the United States*); values are listed in order of state abbreviations (AL first, WY last):

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 10.8 | 6.9 | 8.0 | 8.8 | 7.3 | 3.6 | 4.1 | 6.0 | 4.4 | 8.3 |
| 8.1 | 8.0 | 5.9 | 5.9 | 7.6 | 8.9 | 8.5 | 8.1 | 4.2 | 5.7 |
| 4.0 | 6.7 | 5.8 | 9.9 | 5.6 | 5.8 | 9.3 | 6.2 | 2.5 | 4.5 |
| 12.8 | 3.5 | 10.0 | 9.1 | 5.0 | 8.1 | 5.3 | 3.9 | 4.0 | 8.0 |
| 7.4 | 7.5 | 8.4 | 8.3 | 2.6 | 5.1 | 6.0 | 7.0 | 6.5 | 10.3 |

Figure 1.6 shows a dotplot of the data. The most striking feature is the substantial state-to-state variability. The largest value (for New Mexico) and the two smallest values (New Hampshire and Vermont) are somewhat separated from the bulk of the data, though not perhaps by enough to be considered outliers.

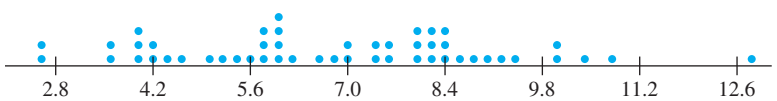



**Figure 1.6** A dotplot of the data from Example 1.8

If the number of compressive strength observations in Example 1.2 had been much larger than the $n = 27$ actually obtained, it would be quite cumbersome to construct a dotplot. 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.

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."

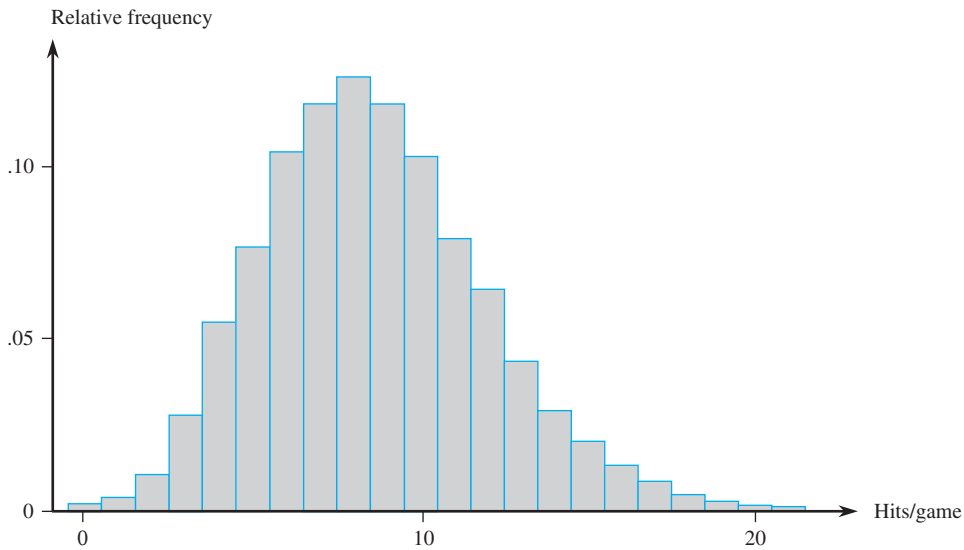



**Figure 1.7** Histogram of number of hits per nine-inning game

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

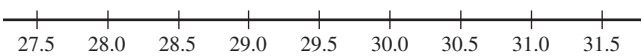
$$\begin{aligned}\text{proportion of games with at most two hits} &= \begin{matrix}\text{relative}\\ \text{frequency}\\ \text{for } x = 0\end{matrix} + \begin{matrix}\text{relative}\\ \text{frequency}\\ \text{for } x = 1\end{matrix} + \begin{matrix}\text{relative}\\ \text{frequency}\\ \text{for } x =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

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 |

We let Minitab select the class intervals. The most striking feature of the histogram in Figure 1.8 is its resemblance to a bell-shaped (and therefore symmetric) 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 |

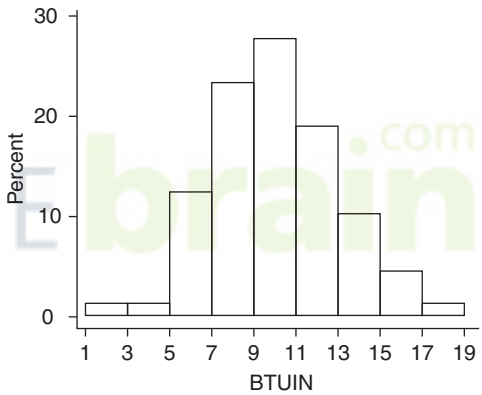



**Figure 1.8** Histogram of the energy consumption data from Example 1.10

From the histogram,

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

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

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

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.

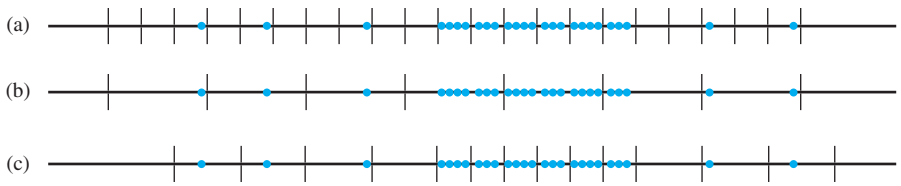



**Figure 1.9** Selecting class intervals for "varying density" data: (a) many short equal-width intervals; (b) a few wide equal-width intervals; (c) unequal-width intervals

### 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.

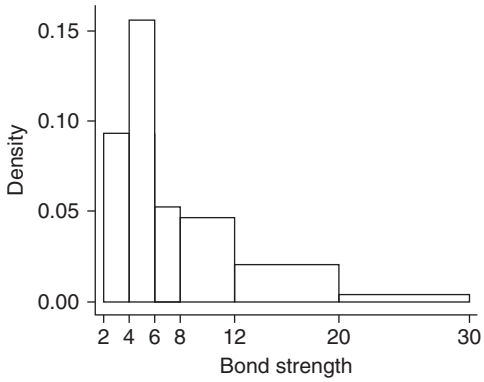



**Figure 1.10** A Minitab density histogram for the bond strength data of Example 1.11 ■

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

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

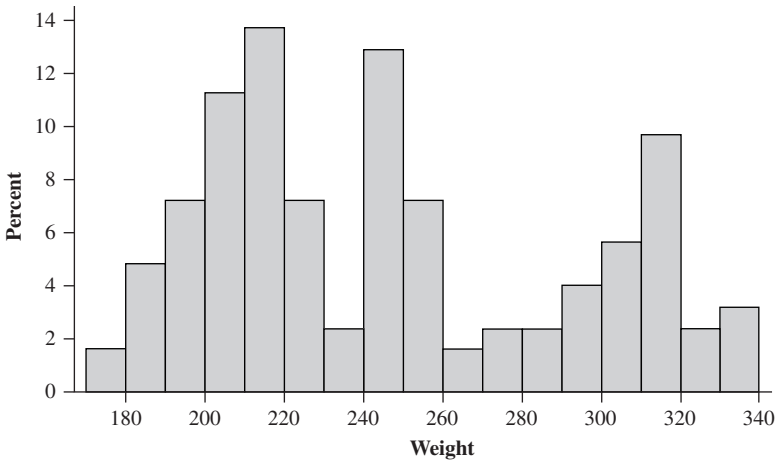
That is, *the area of each rectangle is the relative frequency of the corresponding class.* Furthermore, since the sum of relative frequencies should be 1, *the total area of all rectangles in a density histogram is* 1. It is always possible to draw a histogram so that the area equals the relative frequency (this is true also for a histogram of discrete data)—just use the density scale. This property will play an important role in creating models for distributions in Chapter 4.

## Histogram Shapes

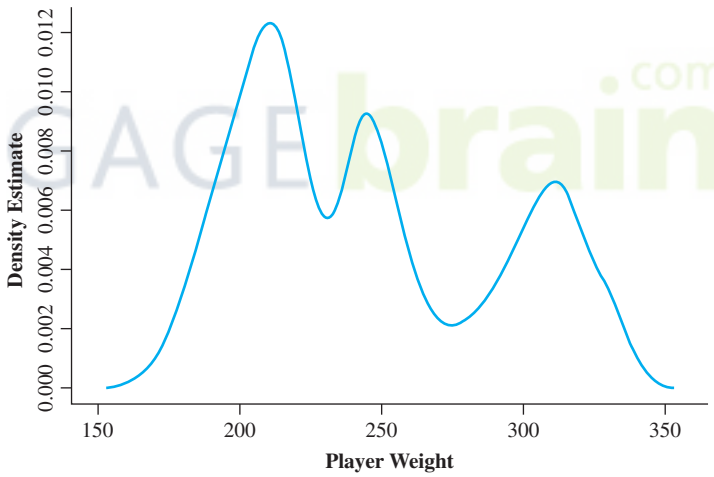
Histograms come in a variety of shapes. A **unimodal** histogram is one that rises to a single peak and then declines. A **bimodal** histogram has two different peaks. Bimodality can occur when the data set consists of observations on two quite different kinds of individuals or objects. For example, 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.).



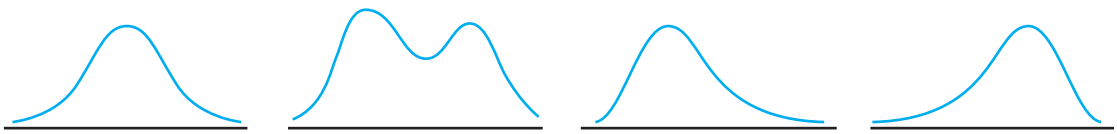

(a)




(b)

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

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.



(a) (b) (c) (d)

**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 |



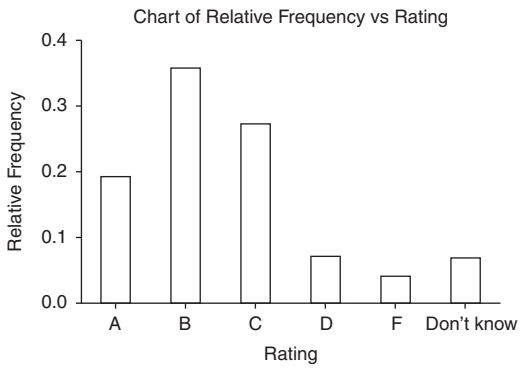



**Figure 1.13** Histogram of the school rating data from Minitab

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 scatter plots 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.** Every score in the following batch of exam scores is in the 60s, 70s, 80s, or 90s. A stem-and-leaf display with only the four stems 6, 7, 8, and 9 would not give a very detailed description of the distribution of scores. In such situations, it is desirable to use repeated stems. Here we could repeat the stem 6 twice, using 6L for scores in the low 60s (leaves 0, 1, 2, 3, and 4) and 6H for scores in the high 60s (leaves 5, 6, 7, 8, and 9). Similarly, the other stems can be repeated twice to obtain a display consisting of eight rows. Construct such a display for the given scores. What feature of the data is highlighted by this display?

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 74 | 89 | 80 | 93 | 64 | 67 | 72 | 70 | 66 | 85 | 89 | 81 | 81 |
| 71 | 74 | 82 | 85 | 63 | 72 | 81 | 81 | 95 | 84 | 81 | 80 | 70 |
| 69 | 66 | 60 | 83 | 85 | 98 | 84 | 68 | 90 | 82 | 69 | 72 | 87 |
| 88 | | | | | | | | | | | | |

**12.** 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 (see the previous exercise), and comment on any interesting features of the display.

**13.** 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 $10-<20, 20-<30, \ldots$ to construct a histogram of the original data. Use intervals $1.1-<1.2, 1.2-<1.3, \ldots$ to do the same for the transformed data. What is the effect of the transformation?

**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.** 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.

**18.** In a study of author productivity ("Lotka's Test," *Collection Mgmt.,* 1982: 111–118), a large number of authors were classified according to the number of articles they had published during a certain period. The results were presented in the accompanying frequency distribution:

| *Number of papers* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| *Frequency* | 784 | 204 | 127 | 50 | 33 | 28 | 19 | 19 |

| *Number of papers* | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|
| *Frequency* | 6 | 7 | 6 | 7 | 4 | 4 | 5 | 3 | 3 |

**a.** Construct a histogram corresponding to this frequency distribution. What is the most interesting feature of the shape of the distribution?

**b.** What proportion of these authors published at least five papers? At least ten papers? More than ten papers?

**c.** Suppose the five 15s, three 16s, and three 17s had been lumped into a single category displayed as "≥15." Would you be able to draw a histogram? Explain.

**d.** Suppose that instead of the values 15, 16, and 17 being listed separately, they had been combined into a 15–17 category with frequency 11. Would you be able to draw a histogram? Explain.

**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).

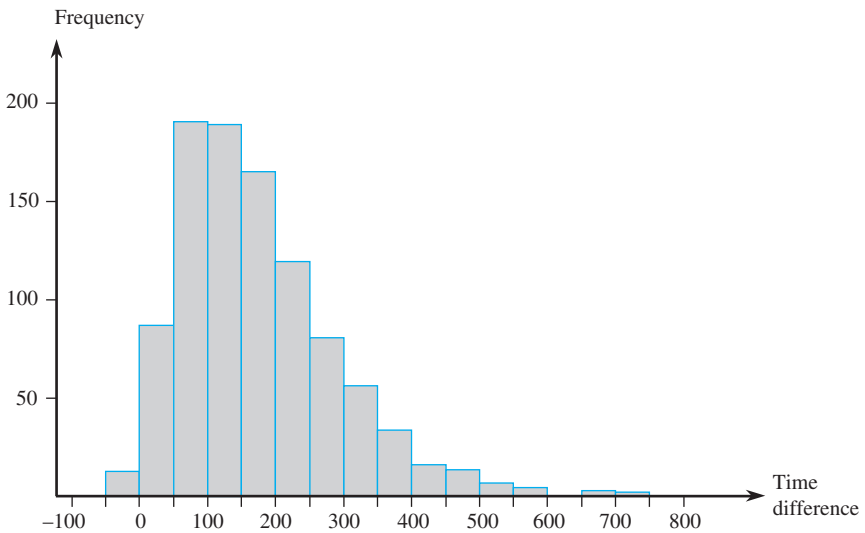
What are some interesting features of this histogram? What is a typical difference value? Roughly what proportion of the runners ran the late distance more quickly than the early distance?

**23.** Consider the following data on types of health complaint (J = joint swelling, F = fatigue, B = back pain, M = muscle weakness, C = coughing, N = nose running/irritation, O = other) made by tree planters. Obtain frequencies and relative frequencies for the various categories, and draw a histogram. (The data is consistent with percentages given in the article "Physiological Effects of Work Stress and Pesticide Exposure in Tree Planting by British Columbia Silviculture Workers," *Ergonomics,* 1993: 951–961.)

O O N J C F B B F O J O O M
O F F O O N O N J F J B O C
J O J J F N O B M O J M O B
O F J O O B N C O O O M B F
J O F N

**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.

Histogram for Exercise 22



| | | | | | | |
|---|---|---|---|---|---|---|
| 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.** Temperature transducers of a certain type are shipped in batches of 50. A sample of 60 batches was selected, and the number of transducers in each batch not conforming to design specifications was determined, resulting in the following data:

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 1 | 2 | 4 | 0 | 1 | 3 | 2 | 0 | 5 | 3 | 3 | 1 | 3 | 2 | 4 | 7 | 0 | 2 | 3 |
| 0 | 4 | 2 | 1 | 3 | 1 | 1 | 3 | 4 | 1 | 2 | 3 | 2 | 2 | 8 | 4 | 5 | 1 | 3 | 1 |
| 5 | 0 | 2 | 3 | 2 | 1 | 0 | 6 | 4 | 2 | 1 | 6 | 0 | 3 | 3 | 3 | 6 | 1 | 2 | 3 |

**a.** Determine frequencies and relative frequencies for the observed values of $x$ = number of nonconforming transducers in a batch.

**b.** What proportion of batches in the sample have at most five nonconforming transducers? What proportion have fewer than five? What proportion have at least five nonconforming units?

**c.** Draw a histogram of the data using relative frequency on the vertical scale, and comment on its features.

**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 paper "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.** Human measurements provide a rich area of application for statistical methods. The article "A Longitudinal Study of the Development of Elementary School Children's Private Speech" (*Merrill-Palmer Q.*, 1990: 443–463) reported on a study of children talking to themselves (private speech). It was thought that private speech would be related to IQ, because IQ is supposed to measure mental maturity, and it was known that private speech decreases as students progress through the primary grades. The study included 33 students whose first-grade IQ scores are given here:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 82 | 96 | 99 | 102 | 103 | 103 | 106 | 107 | 108 | 108 | 108 | 108 |
| 109 | 110 | 110 | 111 | 113 | 113 | 113 | 113 | 115 | 115 | 118 | 118 |
| 119 | 121 | 122 | 122 | 127 | 132 | 136 | 140 | 146 | | | |

Describe the data and comment on any interesting features.

**29.** 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.

**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 numbers—numbers that might serve to characterize the data set and convey some of its salient 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 measuring 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 $\sum x_i$, 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** Caustic stress corrosion cracking of iron and steel has been studied because of failures around rivets in steel boilers and failures of steam rotors. Consider the accompanying observations on $x$ = crack length ($\mu$m) as a result of constant load stress corrosion tests on smooth bar tensile specimens for a fixed length of time. (The data is consistent with a histogram and summary quantities from the article "On the Role of Phosphorus in the Caustic Stress Corrosion Cracking of Low Alloy Steels," *Corrosion Science,* 1989: 53–68.)

$$\begin{array}{lllllll} x_1 = 16.1 & x_2 = 9.6 & x_3 = 24.9 & x_4 = 20.4 & x_5 = 12.7 & x_6 = 21.2 & x_7 = 30.2 \\ x_8 = 25.8 & x_9 = 18.5 & x_{10} = 10.3 & x_{11} = 25.3 & x_{12} = 14.0 & x_{13} = 27.1 & x_{14} = 45.0 \\ x_{15} = 23.3 & x_{16} = 24.2 & x_{17} = 14.6 & x_{18} = 8.9 & x_{19} = 32.4 & x_{20} = 11.8 & x_{21} = 28.5 \end{array}$$

Figure 1.14 shows a stem-and-leaf display of the data; a crack length in the low 20s appears to be "typical."

| | | | | | | |
|---|---|---|---|---|---|---|
| 0H | 96 | 89 | | | | |
| 1L | 27 | 03 | 40 | 46 | 18 | |
| 1H | 61 | 85 | | | | |
| 2L | 49 | 04 | 12 | 33 | 42 | Stem: tens digit |
| 2H | 58 | 53 | 71 | 85 | | Leaf: one and tenths digit |
| 3L | 02 | 24 | | | | |
| 3H | | | | | | |
| 4L | | | | | | |
| 4H | 50 | | | | | |

**Figure 1.14** A stem-and-leaf display of the crack-length data



With $\sum x_i = 444.8$, the sample mean is

$$\bar{x} = \frac{444.8}{21} = 21.18$$

a value consistent with information conveyed by the stem-and-leaf display. ■

A physical interpretation of $\bar{x}$ demonstrates how it measures the location (center) of a sample. Think of drawing and scaling a horizontal measurement axis, and then represent each sample observation by a 1-lb weight placed at the corresponding point on the axis. The only point at which a fulcrum can be placed to balance the system of weights is the point corresponding to the value of $\bar{x}$ (see Figure 1.15).

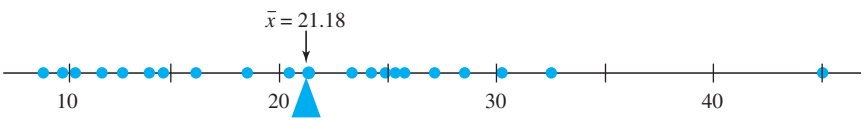



**Figure 1.15** The mean as the balance point for a system of weights

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 $\mu$. When there are $N$ values in the population (a finite population), then $\mu$ = (sum of the $N$ population values)/$N$. In Chapters 3 and 4, we will give a more general definition for $\mu$ that applies to both finite and (conceptually) infinite populations. Just as $\bar{x}$ is an interesting and important measure of sample location, $\mu$ is an interesting and important (often the most important) characteristic of a population. In the chapters on statistical

inference, we will present methods based on the sample mean for drawing conclusions about a population mean. For example, we might use the sample mean $\bar{x} = 21.18$ computed in Example 1.14 as a *point estimate* (a single number that is our "best" guess) of $\mu$ = the true average crack length 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). In Example 1.14, the value $x_{14} = 45.0$ is obviously an outlier. Without this observation, $\bar{x} = 399.8/20 = 19.99$; the outlier increases the mean by more than 1 $\mu$m. If the 45.0 $\mu$m observation were replaced by the catastrophic value 295.0 $\mu$m, a really extreme outlier, then $\bar{x} = 694.8/21 = 33.09$, which is larger than all but one of the observations!

A sample of incomes often produces such outlying values (those lucky few who earn astronomical amounts), and the use of average income as a measure of location will often be misleading. Such examples suggest that we look for 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 #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:

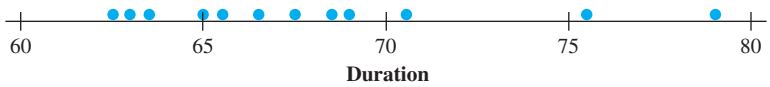



**Figure 1.16** Dotplot of the data from Example 1.14

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:

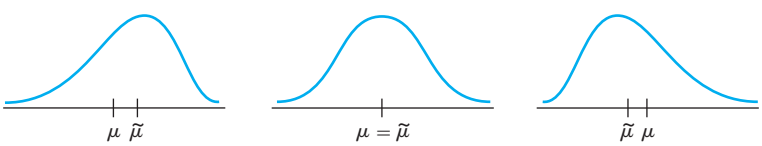
$$\tilde{x} = \frac{66.4 + 67.4}{2} = 66.90$$

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 $\bar{x} = \sum x_i = 816.1/12 = 68.01$, a bit more than a full minute larger than the median. The mean is pulled out a bit 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, we increased the two largest $x_i$s 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 $\mu$, 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. A median is often used to describe income or salary data (because it is not greatly influenced by a few large salaries). If the median salary for a sample of engineers were $\tilde{x} = \$66{,}416$, we might use this as a basis for concluding that the median salary for all engineers exceeds $60,000.

The population mean $\mu$ and median $\tilde{\mu}$ will not generally be identical. If the population distribution is positively or negatively skewed, as pictured in Figure 1.17, 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 1.17** 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 use percentiles in Chapter 4 in connection with certain models for infinite populations and so postpone discussion until that point.

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 1.18 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.

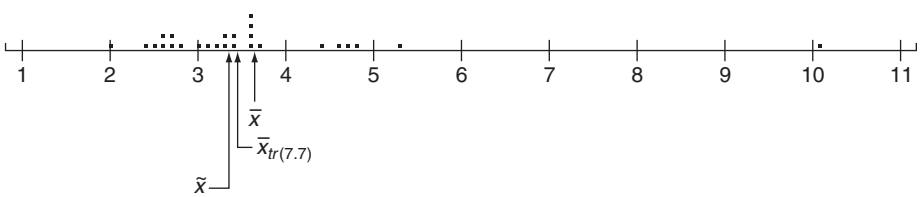



**Figure 1.18** Dotplot of copper contents from Example 1.16 ■

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 $\mu$. In particular, we will subsequently use $x/n$ to make inferences about $p$. If, for example, a sample of 100 car owners reveals that 22 owned their car at least 5 years, then we might use $22/100 = .22$ as a point estimate of the proportion of all owners who have owned their car at least 5 years. 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. The article "Dust Sampling Methods for Endotoxin—An Essential, But Underestimated Issue" (*Indoor Air,* 2006: 20–27) considered various issues associated with determining endotoxin concentration. The following data on concentration (EU/mg) in settled dust for one sample of urban homes and another of farm homes was kindly supplied by the authors of the cited article.

U: 6.0 5.0 11.0 33.0 4.0 5.0 80.0 18.0 35.0 17.0 23.0
F: 4.0 14.0 11.0 9.0 9.0 8.0 4.0 20.0 5.0 8.9 21.0
9.2 3.0 2.0 0.3

**a.** Determine the sample mean for each sample. How do they compare?
**b.** Determine the sample median for each sample. How do they compare? Why is the median for the urban sample so different from the mean for that sample?
**c.** Calculate the trimmed mean for each sample by deleting the smallest and largest observation. What are the corresponding trimming percentages? How do the values of these trimmed means compare to the corresponding means and medians?

**35.** An experiment to study the lifetime (in hours) for a certain type of component involved putting ten components into operation and observing them for 100 hours. Eight of the components failed during that period, and those lifetimes were recorded. Denote the lifetimes of the two components still functioning after 100 hours by 100+ . The resulting sample observations were

48 79 100+ 35 92 86 57 100+ 17 29

Which of the measures of center discussed in this section can be calculated, and what are the values of those measures? [*Note:* The data from this experiment is said to be "censored on the right."]

**36.** A sample of 26 offshore oil workers took part in a simulated escape exercise, resulting in the accompanying data on time (sec) to complete the escape ("Oxygen Consumption and Ventilation During Escape from an Offshore Platform," *Ergonomics,* 1997: 281–292):

389 356 359 363 375 424 325 394 402
373 373 370 364 366 364 325 339 393
392 369 374 359 356 403 334 397

**a.** Construct a stem-and-leaf display of the data. How does it suggest that the sample mean and median will compare?
**b.** Calculate the values of the sample mean and median. [*Hint:* $\sum x_i = 9638$.]
**c.** By how much could the largest time, currently 424, be increased without affecting the value of the sample median? By how much could this value be decreased without affecting the value of the sample median?
**d.** What are the values of $\bar{x}$ and $\tilde{x}$ when the observations are reexpressed in minutes?

**37.** The article "Snow Cover and Temperature Relationships in North America and Eurasia" (*J. Climate and Applied Meteorology,* 1983: 460–469) used statistical techniques to relate the amount of snow cover on each continent to average continental temperature. Data presented there included the following ten observations on October snow cover for Eurasia during the years 1970–1979 (in million $km^2$):

6.5 12.0 14.9 10.0 10.7 7.9 21.9 12.5 14.5 9.2

What would you report as a representative, or typical, value of October snow cover for this period, and what prompted your choice?

**38.** Blood pressure values are often reported to the nearest 5 mmHg (100, 105, 110, etc.). Suppose the actual blood pressure values for nine randomly selected individuals are

118.6 127.4 138.4 130.0 113.7 122.0 108.3
131.5 133.2

**a.** What is the median of the *reported* blood pressure values?
**b.** Suppose the blood pressure of the second individual is 127.6 rather than 127.4 (a small change in a single value). How does this affect the median of the reported values? What does this say about the sensitivity of the median to rounding or grouping in the data?

**39.** The minimum injection pressure (psi) for injection molding specimens of high amylose corn was determined for eight different specimens (higher pressure corresponds to greater processing difficulty), resulting in the following observations (from "Thermoplastic Starch Blends with a Polyethylene-Co-Vinyl Alcohol: Processability and Physical Properties," *Polymer Engr. and Science,* 1994: 17–23):

15.0 13.0 18.0 14.5 12.0 11.0 8.9 8.0

**a.** Determine the values of the sample mean, sample median, and 12.5% trimmed mean, and compare these values.
**b.** By how much could the smallest sample observation, currently 8.0, be increased without affecting the value of the sample median?

**c.** Suppose we want the values of the sample mean and median when the observations are expressed in kilograms per square inch (ksi) rather than psi. Is it necessary to reexpress each observation in ksi, or can the values calculated in part (a) be used directly? [*Hint:* 1 kg = 2.2 lb.]

**40.** Compute the sample median, 25% trimmed mean, 10% trimmed mean, and sample mean for the lifetime data given in Exercise 27, and compare these measures.

**41.** A sample of $n = 10$ automobiles was selected, and each was subjected to a 5-mph crash test. Denoting a car with no visible damage by S (for success) and a car with such damage by F, results were as follows:

S S F S S S F F S S

**a.** What is the value of the sample proportion of successes $x/n$?

**b.** Replace each S with a 1 and each F with a 0. Then calculate $\bar{x}$ for this numerically coded sample. How does $\bar{x}$ compare to $x/n$?

**c.** Suppose it is decided to include 15 more cars in the experiment. How many of these would have to be S's to give $x/n = .80$ for the entire sample of 25 cars?

**42. a.** If a constant $c$ is added to each $x_i$ in a sample, yielding $y_i = x_i + c$, how do the sample mean and median of the $y_i$s relate to the mean and median of the $x_i$s? Verify your conjectures.

**b.** If each $x_i$ is multiplied by a constant $c$, yielding $y_i = cx_i$, answer the question of part (a). Again, verify your conjectures.

**43.** The propagation of fatigue cracks in various aircraft parts has been the subject of extensive study in recent years. The accompanying data consists of propagation lives (flight hours/$10^4$) to reach a given crack size in fastener holes intended for use in military aircraft ("Statistical Crack Propagation in Fastener Holes Under Spectrum Loading," *J. Aircraft,* 1983: 1028–1032):

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .736 | .863 | .865 | .913 | .915 | .937 | .983 | 1.007 |
| 1.011 | 1.064 | 1.109 | 1.132 | 1.140 | 1.153 | 1.253 | 1.394 |

**a.** Compute and compare the values of the sample mean and median.

**b.** By how much could the largest sample observation be decreased without affecting the value of the median?

## 1.4 Measures of Variability

Reporting a measure of center gives only partial information about a data set or distribution. Different samples or populations may have identical measures of center yet differ from one another in other important ways. Figure 1.19 shows dotplots of three samples with the same mean and median, yet the extent of spread about the center is different for all three samples. The first sample has the largest amount of variability, the third has the smallest amount, and the second is intermediate to the other two in this respect.

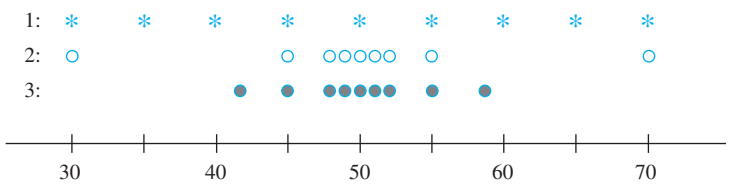



**Figure 1.19** Samples with identical measures of center but different amounts of variability

### Measures of Variability for Sample Data

The simplest measure of variability in a sample is the **range,** which is the difference between the largest and smallest sample values. The value of the range for sample 1 in Figure 1.19 is much larger than it is for sample 3, reflecting more variability in the first sample than in the third. A defect of the range, though, is that it depends on only the two most extreme observations and disregards the positions of the remaining

$n - 2$ values. Samples 1 and 2 in Figure 1.19 have identical ranges, yet when we take into account the observations between the two extremes, there is much less variability or dispersion in the second sample than in the first.

Our primary measures of variability involve the **deviations from the mean,** $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$. That is, the deviations from the mean are obtained by subtracting $\bar{x}$ from each of the $n$ sample observations. A deviation will be positive if the observation is larger than the mean (to the right of the mean on the measurement axis) and negative if the observation is smaller than the mean. If all the deviations are small in magnitude, then all $x_i$s are close to the mean and there is little variability. Alternatively, if some of the deviations are large in magnitude, then some $x_i$s lie far from $\bar{x}$, suggesting a greater amount of variability. A simple way to combine the deviations into a single quantity is to average them. Unfortunately, this is a bad idea:

$$\text{sum of deviations} = \sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

so that the average deviation is always zero. The verification uses several standard rules of summation and the fact that $\sum\bar{x} = \bar{x} + \bar{x} + \cdots + \bar{x} = n\bar{x}$:

$$\sum(x_i - \bar{x}) = \sum x_i - \sum\bar{x} = \sum x_i - n\bar{x} = \sum x_i - n\left(\frac{1}{n}\sum x_i\right) = 0$$

How can we prevent negative and positive deviations from counteracting one another when they are combined? One possibility is to work with the absolute values of the deviations and calculate the average absolute deviation $\sum|x_i - \bar{x}|/n$. Because the absolute value operation leads to a number of theoretical difficulties, consider instead the squared deviations $(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \ldots, (x_n - \bar{x})^2$. Rather than use the average squared deviation $\sum(x_i - \bar{x})^2/n$, for several reasons we divide the sum of squared deviations by $n - 1$ rather than $n$.



DEFINITION

The **sample variance,** denoted by $s^2$, is given by

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1} = \frac{S_{xx}}{n - 1}$$

The **sample standard deviation,** denoted by $s$, is the (positive) square root of the variance:

$$s = \sqrt{s^2}$$

Note that $s^2$ and $s$ are both nonnegative. The unit for $s$ is the same as the unit for each of the $x_i$s. If, for example, the observations are fuel efficiencies in miles per gallon, then we might have $s = 2.0$ mpg. A rough interpretation of the sample standard deviation is that it is the size of a typical or representative deviation from the sample mean within the given sample. Thus if $s = 2.0$ mpg, then some $x_i$'s in the sample are closer than 2.0 to $\bar{x}$, whereas others are farther away; 2.0 is a representative (or "standard") deviation from the mean fuel efficiency. If $s = 3.0$ for a second sample of cars of another type, a typical deviation in this sample is roughly 1.5 times what it is in the first sample, an indication of more variability in the second sample.

**Example 1.17** The Web site www.fueleconomy.gov contains a wealth of information about fuel characteristics of various vehicles. In addition to EPA mileage ratings, there are

many vehicles for which users have reported their own values of fuel efficiency (mpg). Consider the following sample of $n = 11$ efficiencies for the 2009 Ford Focus equipped with an automatic transmission (for this model, EPA reports an overall rating of 27 mpg–24 mpg for city driving and 33 mpg for highway driving):

| Car | $x_i$ | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ | |
|---|---|---|---|---|
| 1 | 27.3 | −5.96 | 35.522 | |
| 2 | 27.9 | −5.36 | 28.730 | |
| 3 | 32.9 | −0.36 | 0.130 | |
| 4 | 35.2 | 1.94 | 3.764 | |
| 5 | 44.9 | 11.64 | 135.490 | |
| 6 | 39.9 | 6.64 | 44.090 | |
| 7 | 30.0 | −3.26 | 10.628 | |
| 8 | 29.7 | −3.56 | 12.674 | |
| 9 | 28.5 | −4.76 | 22.658 | |
| 10 | 32.0 | −1.26 | 1.588 | |
| 11 | 37.6 | 4.34 | 18.836 | |
| | $\sum x_i = 365.9$ | $\sum (x_i - \bar{x}) = .04$ | $\sum (x_i - \bar{x})^2 = 314.106$ | $\bar{x} = 33.26$ |

Effects of rounding account for the sum of deviations not being exactly zero. The numerator of $s^2$ is $S_{xx} = 314.106$, from which

$$s^2 = \frac{S_{xx}}{n-1} = \frac{314.106}{11-1} = 31.41, \qquad s = 5.60$$

The size of a representative deviation from the sample mean 33.26 is roughly 5.6 mpg. Note: Of the nine people who also reported driving behavior, only three did more than 80% of their driving in highway mode; we bet you can guess which cars they drove. We haven't a clue why all 11 reported values exceed the EPA figure—maybe only drivers with really good fuel efficiencies communicate their results. ■

## Motivation for $s^2$

To explain the rationale for the divisor $n - 1$ in $s^2$, note first that whereas $s^2$ measures sample variability, there is a measure of variability in the population called the *population variance*. We will use $\sigma^2$ (the square of the lowercase Greek letter sigma) to denote the population variance and $\sigma$ to denote the population standard deviation (the square root of $\sigma^2$). When the population is finite and consists of $N$ values,

$$\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2 / N$$

which is the average of all squared deviations from the population mean (for the population, the divisor is $N$ and not $N - 1$). More general definitions of $\sigma^2$ appear in Chapters 3 and 4.

Just as $\bar{x}$ will be used to make inferences about the population mean $\mu$, we should define the sample variance so that it can be used to make inferences about $\sigma^2$. Now note that $\sigma^2$ involves squared deviations about the population mean $\mu$. If we actually knew the value of $\mu$, then we could define the sample variance as the average squared deviation of the sample $x_i$s about $\mu$. However, the value of $\mu$ is almost never known, so the sum of squared deviations about $\bar{x}$ must be used. But *the $x_i$s tend to be closer to their average $\bar{x}$ than to the population average $\mu$, so to compensate for this*

*the divisor $n - 1$ is used rather than $n$.* In other words, if we used a divisor $n$ in the sample variance, then the resulting quantity would tend to underestimate $\sigma^2$ (produce estimated values that are too small on the average), whereas dividing by the slightly smaller $n - 1$ corrects this underestimating.

It is customary to refer to $s^2$ as being based on $n - 1$ **degrees of freedom** (df). This terminology reflects the fact that although $s^2$ is based on the $n$ quantities $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$, these sum to 0, so specifying the values of any $n - 1$ of the quantities determines the remaining value. For example, if $n = 4$ and $x_1 - \bar{x} = 8$, $x_2 - \bar{x} = -6$, and $x_4 - \bar{x} = -4$, then automatically $x_3 - \bar{x} = 2$, so only three of the four values of $x_i - \bar{x}$ are freely determined (3 df).

## A Computing Formula for $s^2$

It is best to obtain $s^2$ from statistical software or else use a calculator that allows you to enter data into memory and then view $s^2$ with a single keystroke. If your calculator does not have this capability, there is an alternative formula for $S_{xx}$ that avoids calculating the deviations. The formula involves both $(\sum x_i)^2$, summing and then squaring, and $\sum x_i^2$, squaring and then summing.

An alternative expression for the numerator of $s^2$ is

$$S_{xx} = \sum(x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$



**Proof** Because $\bar{x} = \sum x_i/n$, $n\bar{x}^2 = (\sum x_i)^2/n$. Then,

$$\begin{aligned}\sum(x_i - \bar{x})^2 &= \sum(x_i^2 - 2\bar{x} \cdot x_i + \bar{x}^2) = \sum x_i^2 - 2\bar{x}\sum x_i + \sum(\bar{x})^2 \\ &= \sum x_i^2 - 2\bar{x} \cdot n\bar{x} + n(\bar{x})^2 = \sum x_i^2 - n(\bar{x})^2\end{aligned}$$

**Example 1.18** Traumatic knee dislocation often requires surgery to repair ruptured ligaments. One measure of recovery is range of motion (measured as the angle formed when, starting with the leg straight, the knee is bent as far as possible). The given data on post-surgical range of motion appeared in the article "Reconstruction of the Anterior and Posterior Cruciate Ligaments After Knee Dislocation" (*Amer. J. Sports Med.,* 1999: 189–197):

154 142 137 133 122 126 135 135 108 120 127 134 122

The sum of these 13 sample observations is $\sum x_i = 1695$, and the sum of their squares is

$$\sum x_i^2 = (154)^2 + (142)^2 + \cdots + (122)^2 = 222{,}581$$

Thus the numerator of the sample variance is

$$S_{xx} = \sum x_i^2 - [(\sum x_i)^2]/n = 222{,}581 - (1695)^2/13 = 1579.0769$$

from which $s^2 = 1579.0769/12 = 131.59$ and $s = 11.47$. ■

Both the defining formula and the computational formula for $s^2$ can be sensitive to rounding, so as much decimal accuracy as possible should be used in intermediate calculations.

Several other properties of $s^2$ can enhance understanding and facilitate computation.

PROPOSITION

Let $x_1, x_2, \ldots, x_n$ be a sample and $c$ be any nonzero constant.

**1.** If $y_1 = x_1 + c, y_2 = x_2 + c, \ldots, y_n = x_n + c$, then $s_y^2 = s_x^2$, and

**2.** If $y_1 = cx_1, \ldots, y_n = cx_n$, then $s_y^2 = c^2 s_x^2$, $s_y = |c| s_x$

where $s_x^2$ is the sample variance of the $x$'s and $s_y^2$ is the sample variance of the $y$'s.

In words, Result 1 says that if a constant $c$ is added to (or subtracted from) each data value, the variance is unchanged. This is intuitive, since adding or subtracting $c$ shifts the location of the data set but leaves distances between data values unchanged. According to Result 2, multiplication of each $x_i$ by $c$ results in $s^2$ being multiplied by a factor of $c^2$. These properties can be proved by noting in Result 1 that $\bar{y} = \bar{x} + c$ and in Result 2 that $\bar{y} = c\bar{x}$.

## Boxplots

Stem-and-leaf displays and histograms convey rather general impressions about a data set, whereas a single summary such as the mean or standard deviation focuses on just one aspect of the data. In recent years, a pictorial summary called a *boxplot* has been used successfully to describe several of a data set's most prominent features. These features include (1) center, (2) spread, (3) the extent and nature of any departure from symmetry, and (4) identification of "outliers," observations that lie unusually far from the main body of the data. Because even a single outlier can drastically affect the values of $\bar{x}$ and $s$, a boxplot is based on measures that are "resistant" to the presence of a few outliers—the median and a measure of variability called the *fourth spread.*



DEFINITION

Order the $n$ observations from smallest to largest and separate the smallest half from the largest half; the median $\tilde{x}$ is included in both halves if $n$ is odd. Then the **lower fourth** is the median of the smallest half and the **upper fourth** is the median of the largest half. A measure of spread that is resistant to outliers is the **fourth spread** $f_s$, given by

$$f_s = \text{upper fourth} - \text{lower fourth}$$

Roughly speaking, the fourth spread is unaffected by the positions of those observations in the smallest 25% or the largest 25% of the data. Hence it is resistant to outliers.

The simplest boxplot is based on the following five-number summary:

$$\text{smallest } x_i \quad \text{lower fourth} \quad \text{median} \quad \text{upper fourth} \quad \text{largest } x_i$$

First, draw a horizontal measurement scale. Then place a rectangle above this axis; the left edge of the rectangle is at the lower fourth, and the right edge is at the upper fourth (so box width $= f_s$). Place a vertical line segment or some other symbol inside the rectangle at the location of the median; the position of the median symbol relative to the two edges conveys information about skewness in the middle 50% of the data. Finally, draw "whiskers" out from either end of the rectangle to the smallest and largest observations. A boxplot with a vertical orientation can also be drawn by making obvious modifications in the construction process.

**Example 1.19** Ultrasound was used to gather the accompanying corrosion data on the thickness of the floor plate of an aboveground tank used to store crude oil ("Statistical Analysis of UT Corrosion Data from Floor Plates of a Crude Oil Aboveground Storage Tank," *Materials Eval.,* 1994: 846–849); each observation is the largest pit depth in the plate, expressed in milli-in.

40 52 55 60 70 75 85 85 90 90 92 94 94 95 98 100 115 125 125

The five-number summary is as follows:

smallest $x_i = 40$ lower fourth $= 72.5$ $\tilde{x} = 90$ upper fourth $= 96.5$
largest $x_i = 125$

Figure 1.20 shows the resulting boxplot. The right edge of the box is much closer to the median than is the left edge, indicating a very substantial skew in the middle half of the data. The box width ($f_s$) is also reasonably large relative to the range of the data (distance between the tips of the whiskers).

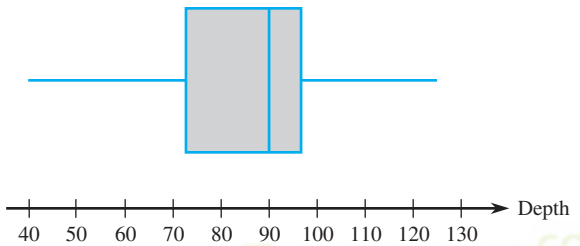



**Figure 1.20** A boxplot of the corrosion data



Figure 1.21 shows Minitab output from a request to describe the corrosion data. Q1 and Q3 are the lower and upper quartiles; these are similar to the fourths but are calculated in a slightly different manner. SE Mean is $s/\sqrt{n}$; this will be an important quantity in our subsequent work concerning inferences about $\mu$.

```
Variable        N      Mean   Median   TrMean   StDev   SE Mean
   depth       19     86.32    90.00    86.76   23.32      5.35

Variable   Minimum   Maximum       Q1       Q3
   depth     40.00    125.00    70.00    98.00
```

**Figure 1.21** Minitab description of the pit-depth data ■

## Boxplots That Show Outliers

A boxplot can be embellished to indicate explicitly the presence of outliers. Many inferential procedures are based on the assumption that the population distribution is normal (a certain type of bell curve). Even a single extreme outlier in the sample warns the investigator that such procedures may be unreliable, and the presence of several mild outliers conveys the same message.

DEFINITION

Any observation farther than $1.5f_s$ from the closest fourth is an **outlier.** An outlier is **extreme** if it is more than $3f_s$ from the nearest fourth, and it is **mild** otherwise.

Let's now modify our previous construction of a boxplot by drawing a whisker out from each end of the box to the smallest and largest observations that are *not* outliers. Each mild outlier is represented by a closed circle and each extreme outlier by an open circle. Some statistical computer packages do not distinguish between mild and extreme outliers.

**Example 1.20** The Clean Water Act and subsequent amendments require that all waters in the United States meet specific pollution reduction goals to ensure that water is "fishable and swimmable." The article "Spurious Correlation in the USEPA Rating Curve Method for Estimating Pollutant Loads" (*J. of Environ. Engr.,* 2008: 610–618) investigated various techniques for estimating pollutant loads in watersheds; the authors "discuss the imperative need to use sound statistical methods" for this purpose. Among the data considered is the following sample of TN (total nitrogen) loads (kg N/day) from a particular Chesapeake Bay location, displayed here in increasing order.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 9.69 | 13.16 | 17.09 | 18.12 | 23.70 | 24.07 | 24.29 | 26.43 |
| 30.75 | 31.54 | 35.07 | 36.99 | 40.32 | 42.51 | 45.64 | 48.22 |
| 49.98 | 50.06 | 55.02 | 57.00 | 58.41 | 61.31 | 64.25 | 65.24 |
| 66.14 | 67.68 | 81.40 | 90.80 | 92.17 | 92.42 | 100.82 | 101.94 |
| 103.61 | 106.28 | 106.80 | 108.69 | 114.61 | 120.86 | 124.54 | 143.27 |
| 143.75 | 149.64 | 167.79 | 182.50 | 192.55 | 193.53 | 271.57 | 292.61 |
| 312.45 | 352.09 | 371.47 | 444.68 | 460.86 | 563.92 | 690.11 | 826.54 |
| 1529.35 | | | | | | | |

Relevant summary quantities are

$$\tilde{x} = 92.17 \qquad \text{lower } 4^{\text{th}} = 45.64 \qquad \text{upper } 4^{\text{th}} = 167.79$$

$$f_s = 122.15 \qquad 1.5f_s = 183.225 \qquad 3f_s = 366.45$$

Subtracting $1.5f_s$ from the lower $4^{\text{th}}$ gives a negative number, and none of the observations are negative, so there are no outliers on the lower end of the data. However,

$$\text{upper } 4^{\text{th}} + 1.5f_s = 351.015 \qquad \text{upper } 4^{\text{th}} + 3f_s = 534.24$$

Thus the four largest observations—563.92, 690.11, 826.54, and 1529.35—are extreme outliers, and 352.09, 371.47, 444.68, and 460.86 are mild outliers.

The whiskers in the boxplot in Figure 1.22 extend out to the smallest observation, 9.69, on the low end and 312.45, the largest observation that is not an outlier, on the upper end. There is some positive skewness in the middle half of the data (the median line is somewhat closer to the left edge of the box than to the right edge) and a great deal of positive skewness overall.

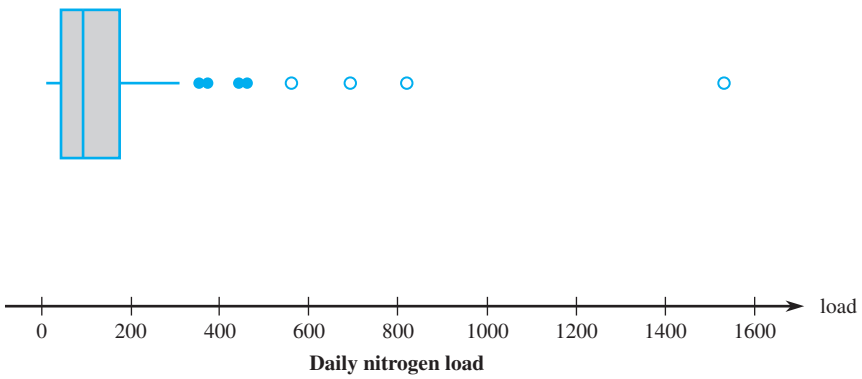



**Figure 1.22** A boxplot of the nitrogen load data showing mild and extreme outliers ■

## Comparative Boxplots

A comparative or side-by-side boxplot is a very effective way of revealing similarities and differences between two or more data sets consisting of observations on the same variable—fuel efficiency observations for four different types of automobiles, crop yields for three different varieties, and so on.

**Example 1.21** In recent years, some evidence suggests that high indoor radon concentration may be linked to the development of childhood cancers, but many health professionals remain unconvinced. A recent article ("Indoor Radon and Childhood Cancer," *The Lancet,* 1991: 1537–1538) presented the accompanying data on radon concentration ($Bq/m^3$) in two different samples of houses. The first sample consisted of houses in which a child diagnosed with cancer had been residing. Houses in the second sample had no recorded cases of childhood cancer. Figure 1.23 presents a stem-and-leaf display of the data.

| 1. Cancer | | 2. No cancer |
|---|---|---|
| 9683795 | 0 | 95768397678993 |
| 86071815066815233150 | 1 | 12271713114 |
| 12302731 | 2 | 99494191 |
| 8349 | 3 | 839 |
| 5 | 4 | |
| 7 | 5 | 55 |
| | 6 | |
| | 7 | Stem: Tens digit |
| HI: 210 | 8 | 5 Leaf: Ones digit |

**Figure 1.23** Stem-and-leaf display for Example 1.21

Numerical summary quantities are as follows:

| | $\bar{x}$ | $\tilde{x}$ | $s$ | $f_s$ |
|---|---|---|---|---|
| Cancer | 22.8 | 16.0 | 31.7 | 11.0 |
| No cancer | 19.2 | 12.0 | 17.0 | 18.0 |



The values of both the mean and median suggest that the cancer sample is centered somewhat to the right of the no-cancer sample on the measurement scale. The mean, however, exaggerates the magnitude of this shift, largely because of the observation 210 in the cancer sample. The values of $s$ suggest more variability in the cancer sample than in the no-cancer sample, but this impression is contradicted by the fourth spreads. Again, the observation 210, an extreme outlier, is the culprit. Figure 1.24 shows a comparative boxplot from the S-Plus computer package. The no-cancer box

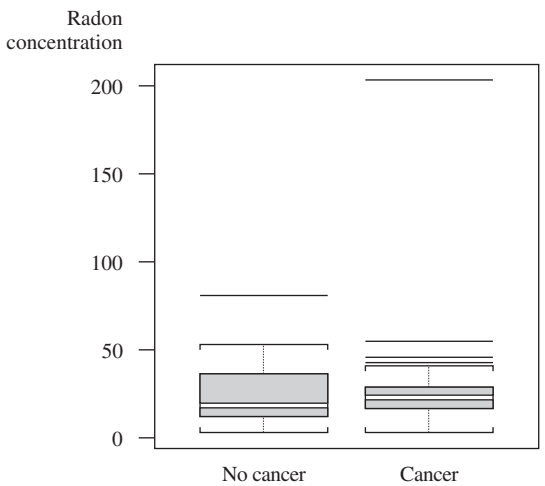



**Figure 1.24** A boxplot of the data in Example 1.21, from S-Plus

is stretched out compared with the cancer box ($f_s = 18$ vs. $f_s = 11$), and the positions of the median lines in the two boxes show much more skewness in the middle half of the no-cancer sample than the cancer sample. Outliers are represented by horizontal line segments, and there is no distinction between mild and extreme outliers. ■

## EXERCISES Section 1.4 (44–61)

**44.** The article "Oxygen Consumption During Fire Suppression: Error of Heart Rate Estimation" (*Ergonomics,* 1991: 1469–1474) reported the following data on oxygen consumption (mL/kg/min) for a sample of ten firefighters performing a fire-suppression simulation:

29.5 49.3 30.6 28.2 28.0 26.3 33.9 29.4 23.5 31.6

Compute the following:

**a.** The sample range
**b.** The sample variance $s^2$ from the definition (i.e., by first computing deviations, then squaring them, etc.)
**c.** The sample standard deviation
**d.** $s^2$ using the shortcut method

**45.** The article "A Thin-Film Oxygen Uptake Test for the Evaluation of Automotive Crankcase Lubricants" (*Lubric. Engr.,* 1984: 75–83) reported the following data on oxidation-induction time (min) for various commercial oils:

87 103 130 160 180 195 132 145 211 105 145
153 152 138 87 99 93 119 129

**a.** Calculate the sample variance and standard deviation.
**b.** If the observations were reexpressed in hours, what would be the resulting values of the sample variance and sample standard deviation? Answer without actually performing the reexpression.

**46.** The accompanying observations on stabilized viscosity (cP) for specimens of a certain grade of asphalt with 18% rubber added are from the article "Viscosity Characteristics of Rubber-Modified Asphalts" (*J. of Materials in Civil Engr.,* 1996: 153–156):

2781 2900 3013 2856 2888

**a.** What are the values of the sample mean and sample median?
**b.** Calculate the sample variance using the computational formula. [*Hint:* First subtract a convenient number from each observation.]

**47.** Calculate and interpret the values of the sample median, sample mean, and sample standard deviation for the following observations on fracture strength (MPa, read from a graph in "Heat-Resistant Active Brazing of Silicon Nitride: Mechanical Evaluation of Braze Joints," *Welding J.,* August, 1997):

87 93 96 98 105 114 128 131 142 168

**48.** Exercise 34 presented the following data on endotoxin concentration in settled dust both for a sample of urban homes and for a sample of farm homes:

U: 6.0 5.0 11.0 33.0 4.0 5.0 80.0 18.0 35.0 17.0 23.0
F: 4.0 14.0 11.0 9.0 9.0 8.0 4.0 20.0 5.0 8.9 21.0
9.2 3.0 2.0 0.3

**a.** Determine the value of the sample standard deviation for each sample, interpret these values, and then contrast variability in the two samples. [*Hint:* $\sum x_i = 237.0$ for the urban sample and $= 128.4$ for the farm sample, and $\sum x_i^2 = 10{,}079$ for the urban sample and 1617.94 for the farm sample.]
**b.** Compute the fourth spread for each sample and compare. Do the fourth spreads convey the same message about variability that the standard deviations do? Explain.
**c.** The authors of the cited article also provided endotoxin concentrations in dust bag dust:

U: 34.0 49.0 13.0 33.0 24.0 24.0 35.0 104.0 34.0 40.0 38.0 1.0
F: 2.0 64.0 6.0 17.0 35.0 11.0 17.0 13.0 5.0 27.0 23.0
28.0 10.0 13.0 0.2

Construct a comparative boxplot (as did the cited paper) and compare and contrast the four samples.

**49.** A study of the relationship between age and various visual functions (such as acuity and depth perception) reported the following observations on the area of scleral lamina ($mm^2$) from human optic nerve heads ("Morphometry of Nerve Fiber Bundle Pores in the Optic Nerve Head of the Human," *Experimental Eye Research,* 1988: 559–568):

2.75 2.62 2.74 3.85 2.34 2.74 3.93 4.21 3.88
4.33 3.46 4.52 2.43 3.65 2.78 3.56 3.01

**a.** Calculate $\sum x_i$ and $\sum x_i^2$.
**b.** Use the values calculated in part (a) to compute the sample variance $s^2$ and then the sample standard deviation $s$.

**50.** In 1997 a woman sued a computer keyboard manufacturer, charging that her repetitive stress injuries were caused by the keyboard (*Genessy v. Digital Equipment Corp.*). The injury awarded about $3.5 million for pain and suffering, but the court then set aside that award as being unreasonable compensation. In making this determination, the court identified a "normative" group of 27 similar cases and specified a reasonable award as one within two standard deviations of the mean of the awards in the 27 cases. The 27 awards were

(in $1000s) 37, 60, 75, 115, 135, 140, 149, 150, 238, 290, 340, 410, 600, 750, 750, 750, 1050, 1100, 1139, 1150, 1200, 1200, 1250, 1576, 1700, 1825, and 2000, from which $\sum x_i = 20{,}179$, $\sum x_i^2 = 24{,}657{,}511$. What is the maximum possible amount that could be awarded under the two-standard-deviation rule?

**51.** A sample of 20 glass bottles of a particular type was selected, and the internal pressure strength of each bottle was determined. Consider the following partial sample information:

median = 202.2 lower fourth = 196.0
upper fourth = 216.8

| | | | |
|---|---|---|---|
| *Three smallest observations* | 125.8 | 188.1 | 193.7 |
| *Three largest observations* | 221.3 | 230.5 | 250.2 |

**a.** Are there any outliers in the sample? Any extreme outliers?
**b.** Construct a boxplot that shows outliers, and comment on any interesting features.

**52.** The first four deviations from the mean in a sample of $n = 5$ reaction times were .3, .9, 1.0, and 1.3. What is the fifth deviation from the mean? Give a sample for which these are the five deviations from the mean.

**53.** A **mutual fund** is a professionally managed investment scheme that pools money from many investors and invests in a variety of securities. Growth funds focus primarily on increasing the value of investments, whereas blended funds seek a balance between current income and growth. Here is data on the expense ratio (expenses as a % of assets, from www.morningstar.com) for samples of 20 large-cap balanced funds and 20 large-cap growth funds ("large-cap" refers to the sizes of companies in which the funds invest; the population sizes are 825 and 762, respectively):

| | | | | | |
|---|---|---|---|---|---|
| Bl | 1.03 | 1.23 | 1.10 | 1.64 | 1.30 |
| | 1.27 | 1.25 | 0.78 | 1.05 | 0.64 |
| | 0.94 | 2.86 | 1.05 | 0.75 | 0.09 |
| | 0.79 | 1.61 | 1.26 | 0.93 | 0.84 |
| Gr | 0.52 | 1.06 | 1.26 | 2.17 | 1.55 |
| | 0.99 | 1.10 | 1.07 | 1.81 | 2.05 |
| | 0.91 | 0.79 | 1.39 | 0.62 | 1.52 |
| | 1.02 | 1.10 | 1.78 | 1.01 | 1.15 |

**a.** Calculate and compare the values of $\bar{x}$, $\tilde{x}$, and $s$ for the two types of funds.
**b.** Construct a comparative boxplot for the two types of funds, and comment on interesting features.

**54.** Grip is applied to produce normal surface forces that compress the object being gripped. Examples include two people shaking hands, or a nurse squeezing a patient's forearm to stop bleeding. The article "Investigation of Grip Force, Normal Force, Contact Area, Hand Size, and Handle Size for Cylindrical Handles" (*Human Factors,* 2008: 734–744) included the following data on grip strength (N) for a sample of 42 individuals:

16 18 18 26 33 41 54 56 66 68 87 91 95
98 106 109 111 118 127 127 135 145 147 149 151 168
172 183 189 190 200 210 220 229 230 233 238 244 259
294 329 403

**a.** Construct a stem-and-leaf display based on repeating each stem value twice, and comment on interesting features.
**b.** Determine the values of the fourths and the fourthspread.
**c.** Construct a boxplot based on the five-number summary, and comment on its features.
**d.** How large or small does an observation have to be to qualify as an outlier? An extreme outlier? Are there any outliers?
**e.** By how much could the observation 403, currently the largest, be decreased without affecting $f_s$?

**55.** Here is a stem-and-leaf display of the escape time data introduced in Exercise 36 of this chapter.

| | |
|---|---|
| 32 | 55 |
| 33 | 49 |
| 34 | |
| 35 | 6699 |
| 36 | 34469 |
| 37 | 03345 |
| 38 | 9 |
| 39 | 2347 |
| 40 | 23 |
| 41 | |
| 42 | 4 |

**a.** Determine the value of the fourth spread.
**b.** Are there any outliers in the sample? Any extreme outliers?
**c.** Construct a boxplot and comment on its features.
**d.** By how much could the largest observation, currently 424, be decreased without affecting the value of the fourth spread?

**56.** The following data on distilled alcohol content (%) for a sample of 35 port wines was extracted from the article "A Method for the Estimation of Alcohol in Fortified Wines Using Hydrometer Baumé and Refractometer Brix" (*Amer. J. Enol. Vitic.,* 2006: 486–490). Each value is an average of two duplicate measurements.

16.35 18.85 16.20 17.75 19.58 17.73 22.75 23.78 23.25
19.08 19.62 19.20 20.05 17.85 19.17 19.48 20.00 19.97
17.48 17.15 19.07 19.90 18.68 18.82 19.03 19.45 19.37
19.20 18.00 19.60 19.33 21.22 19.50 15.30 22.25

Use methods from this chapter, including a boxplot that shows outliers, to describe and summarize the data.

**57.** The value of Young's modulus (GPa) was determined for cast plates consisting of certain intermetallic substrates, resulting in the following sample observations ("Strength and Modulus of a Molybdenum-Coated Ti-25Al-10Nb-3U-1Mo Intermetallic," *J. of Materials Engr. and Performance,* 1997: 46–50):

116.4 115.9 114.6 115.2 115.8

**a.** Calculate $\bar{x}$ and the deviations from the mean.
**b.** Use the deviations calculated in part (a) to obtain the sample variance and the sample standard deviation.
**c.** Calculate $s^2$ by using the computational formula for the numerator $S_{xx}$.
**d.** Subtract 100 from each observation to obtain a sample of transformed values. Now calculate the sample variance of these transformed values, and compare it to $s^2$ for the original data.

**58.** A company utilizes two different machines to manufacture parts of a certain type. During a single shift, a sample of $n = 20$ parts produced by each machine is obtained, and the value of a particular critical dimension for each part is determined. The comparative boxplot at the bottom of this page is constructed from the resulting data. Compare and contrast the two samples.

**59.** Blood cocaine concentration (mg/L) was determined both for a sample of individuals who had died from cocaine-induced excited delirium (ED) and for a sample of those who had died from a cocaine overdose without excited delirium; survival time for people in both groups was at most 6 hours. The accompanying data was read from a comparative boxplot in the article "Fatal Excited Delirium Following Cocaine Use" (*J. of Forensic Sciences,* 1997: 25–31).

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *ED* | 0 | 0 | 0 | 0 | .1 | .1 | .1 | .1 | .2 | .2 | .3 | .3 |
| | .3 | .4 | .5 | .7 | .8 | 1.0 | 1.5 | 2.7 | 2.8 | | | |
| | 3.5 | 4.0 | 8.9 | 9.2 | 11.7 | 21.0 | | | | | | |
| *Non-ED* | 0 | 0 | 0 | 0 | 0 | .1 | .1 | .1 | .1 | .2 | .2 | .2 |
| | .3 | .3 | .3 | .4 | .5 | .5 | .6 | .8 | .9 | 1.0 | | |
| | 1.2 | 1.4 | 1.5 | 1.7 | 2.0 | 3.2 | 3.5 | 4.1 | | | | |
| | 4.3 | 4.8 | 5.0 | 5.6 | 5.9 | 6.0 | 6.4 | 7.9 | | | | |
| | 8.3 | 8.7 | 9.1 | 9.6 | 9.9 | 11.0 | 11.5 | | | | | |
| | 12.2 | 12.7 | 14.0 | 16.6 | 17.8 | | | | | | | |

**a.** Determine the medians, fourths, and fourth spreads for the two samples.
**b.** Are there any outliers in either sample? Any extreme outliers?
**c.** Construct a comparative boxplot, and use it as a basis for comparing and contrasting the ED and non-ED samples.

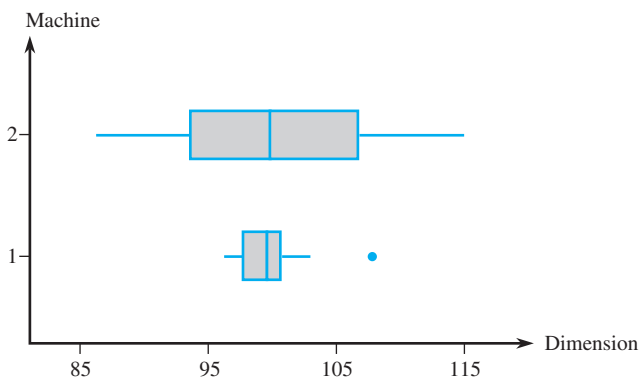
**60.** Observations on burst strength (lb/in$^2$) were obtained both for test nozzle closure welds and for production cannister nozzle welds ("Proper Procedures Are the Key to Welding Radioactive Waste Cannisters," *Welding J.,* Aug. 1997: 61–67).

| | | | | | | |
|---|---|---|---|---|---|---|
| *Test* | 7200 | 6100 | 7300 | 7300 | 8000 | 7400 |
| | 7300 | 7300 | 8000 | 6700 | 8300 | |
| *Cannister* | 5250 | 5625 | 5900 | 5900 | 5700 | 6050 |
| | 5800 | 6000 | 5875 | 6100 | 5850 | 6600 |

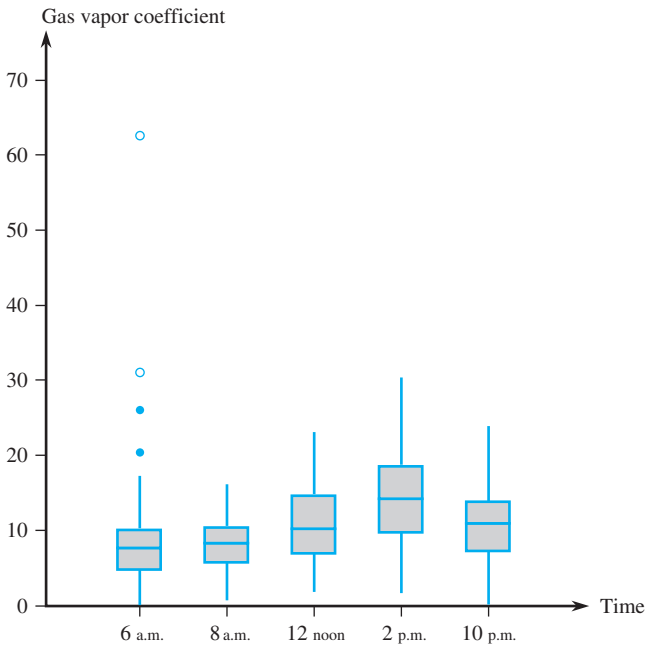
Construct a comparative boxplot and comment on interesting features (the cited article did not include such a picture, but the authors commented that they had looked at one).

**61.** The accompanying comparative boxplot of gasoline vapor coefficients for vehicles in Detroit appeared in the article "Receptor Modeling Approach to VOC Emission Inventory Validation" (*J. of Envir. Engr.,* 1995: 483–490). Discuss any interesting features.

Comparative boxplot for Exercise 58




Comparative boxplot for Exercise 61



## SUPPLEMENTARY EXERCISES (62–83)

**62.** Consider the following information on ultimate tensile strength (lb/in) for a sample of $n = 4$ hard zirconium copper wire specimens (from "Characterization Methods for Fine Copper Wire," *Wire J. Intl.,* Aug., 1997: 74–80):

$\bar{x} = 76{,}831 \quad s = 180$ smallest $x_i = 76{,}683$
largest $x_i = 77{,}048$

Determine the values of the two middle sample observations (and don't do it by successive guessing!).

**63.** A sample of 77 individuals working at a particular office was selected and the noise level (dBA) experienced by each individual was determined, yielding the following data ("Acceptable Noise Levels for Construction Site Offices," *Building Serv. Engr. Research and Technology*, 2009: 87–94).

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 55.3 | 55.3 | 55.3 | 55.9 | 55.9 | 55.9 | 55.9 | 56.1 | 56.1 | 56.1 | 56.1 |
| 56.1 | 56.1 | 56.8 | 56.8 | 57.0 | 57.0 | 57.0 | 57.8 | 57.8 | 57.8 | 57.9 |
| 57.9 | 57.9 | 58.8 | 58.8 | 58.8 | 59.8 | 59.8 | 59.8 | 62.2 | 62.2 | 63.8 |
| 63.8 | 63.8 | 63.9 | 63.9 | 63.9 | 64.7 | 64.7 | 64.7 | 65.1 | 65.1 | 65.1 |
| 65.3 | 65.3 | 65.3 | 65.3 | 67.4 | 67.4 | 67.4 | 67.4 | 68.7 | 68.7 | 68.7 |
| 68.7 | 69.0 | 70.4 | 70.4 | 71.2 | 71.2 | 71.2 | 73.0 | 73.0 | 73.1 | 73.1 |
| 74.6 | 74.6 | 74.6 | 74.6 | 79.3 | 79.3 | 79.3 | 79.3 | 83.0 | 83.0 | 83.0 |

Use various techniques discussed in this chapter to organize, summarize, and describe the data.

**64.** Fretting is a wear process that results from tangential oscillatory movements of small amplitude in machine parts. The article "Grease Effect on Fretting Wear of Mild Steel" (*Industrial Lubrication and Tribology*, 2008: 67–78) included the following data on volume wear ($10^{-4}$mm$^3$) for base oils having four different viscosities.

| Viscosity | Wear | | | | | |
|---|---|---|---|---|---|---|
| 20.4 | 58.8 | 30.8 | 27.3 | 29.9 | 17.7 | 76.5 |
| 30.2 | 44.5 | 47.1 | 48.7 | 41.6 | 32.8 | 18.3 |
| 89.4 | 73.3 | 57.1 | 66.0 | 93.8 | 133.2 | 81.1 |
| 252.6 | 30.6 | 24.2 | 16.6 | 38.9 | 28.7 | 23.6 |

**a.** The *sample coefficient of variation* $100s/\bar{x}$ assesses the extent of variability relative to the mean (specifically, the standard deviation as a percentage of the mean). Calculate the coefficient of variation for the sample at each viscosity. Then compare the results and comment.

**b.** Construct a comparative boxplot of the data and comment on interesting features.

**65.** Let $\bar{x}_n$ and $s_n^2$ denote the sample mean and variance for the sample $x_1, \ldots, x_n$ and let $\bar{x}_{n+1}$ and $s_{n+1}^2$ denote these quantities when an additional observation $x_{n+1}$ is added to the sample.

**a.** Show how $\bar{x}_{n+1}$ can be computed from $\bar{x}_n$ and $x_{n+1}$.

**b.** Show that

$$ns_{n+1}^2 = (n-1)s_n^2 + \frac{n}{n+1}(x_{n+1} - \bar{x}_n)^2$$

so that $s_{n+1}^2$ can be computed from $x_{n+1}$, $\bar{x}_n$, and $s_n^2$.

**c.** Suppose that a sample of 15 strands of drapery yarn has resulted in a sample mean thread elongation of 12.58 mm and a sample standard deviation of .512 mm. A 16th strand results in an elongation value of 11.8. What are the values of the sample mean and sample standard deviation for all 16 elongation observations?

**66.** A deficiency of the trace element selenium in the diet can negatively impact growth, immunity, muscle and neuromuscular function, and fertility. The introduction of selenium supplements to dairy cows is justified when pastures have low selenium levels. Authors of the paper "Effects of Short-Term Supplementation with Selenised Yeast on Milk Production and Composition of Lactating Cows" (*Australian J. of Dairy Tech.*, 2004: 199–203) supplied the following data on milk selenium concentration (mg/L) for a sample of cows given a selenium supplement and a control sample given no supplement, both initially and after a 9-day period.

| Obs | Init Se | Init Cont | Final Se | Final Cont |
|---|---|---|---|---|
| 1 | 11.4 | 9.1 | 138.3 | 9.3 |
| 2 | 9.6 | 8.7 | 104.0 | 8.8 |
| 3 | 10.1 | 9.7 | 96.4 | 8.8 |
| 4 | 8.5 | 10.8 | 89.0 | 10.1 |
| 5 | 10.3 | 10.9 | 88.0 | 9.6 |
| 6 | 10.6 | 10.6 | 103.8 | 8.6 |
| 7 | 11.8 | 10.1 | 147.3 | 10.4 |
| 8 | 9.8 | 12.3 | 97.1 | 12.4 |
| 9 | 10.9 | 8.8 | 172.6 | 9.3 |
| 10 | 10.3 | 10.4 | 146.3 | 9.5 |
| 11 | 10.2 | 10.9 | 99.0 | 8.4 |
| 12 | 11.4 | 10.4 | 122.3 | 8.7 |
| 13 | 9.2 | 11.6 | 103.0 | 12.5 |
| 14 | 10.6 | 10.9 | 117.8 | 9.1 |
| 15 | 10.8 | | 121.5 | |
| 16 | 8.2 | | 93.0 | |

**a.** Do the initial Se concentrations for the supplement and control samples appear to be similar? Use various techniques from this chapter to summarize the data and answer the question posed.

**b.** Again use methods from this chapter to summarize the data and then describe how the final Se concentration values in the treatment group differ from those in the control group.

**67.** *Aortic stenosis* refers to a narrowing of the aortic valve in the heart. The paper "Correlation Analysis of Stenotic Aortic Valve Flow Patterns Using Phase Contrast MRI" (*Annals of Biomed. Engr.*, 2005: 878–887) gave the following data on

aortic root diameter (cm) and gender for a sample of patients having various degrees of aortic stenosis:

M: 3.7 3.4 3.7 4.0 3.9 3.8 3.4 3.6 3.1 4.0 3.4 3.8 3.5
F: 3.8 2.6 3.2 3.0 4.3 3.5 3.1 3.1 3.2 3.0

**a.** Compare and contrast the diameter observations for the two genders.
**b.** Calculate a 10% trimmed mean for each of the two samples, and compare to other measures of center (for the male sample, the interpolation method mentioned in Section 1.3 must be used).

**68.** **a.** For what value of $c$ is the quantity $\sum(x_i - c)^2$ minimized? [*Hint:* Take the derivative with respect to $c$, set equal to 0, and solve.]
**b.** Using the result of part (a), which of the two quantities $\sum(x_i - \bar{x})^2$ and $\sum(x_i - \mu)^2$ will be smaller than the other (assuming that $\bar{x} \neq \mu$)?

**69.** **a.** Let $a$ and $b$ be constants and let $y_i = ax_i + b$ for $i = 1, 2, \ldots, n$. What are the relationships between $\bar{x}$ and $\bar{y}$ and between $s_x^2$ and $s_y^2$?
**b.** A sample of temperatures for initiating a certain chemical reaction yielded a sample average (°C) of 87.3 and a sample standard deviation of 1.04. What are the sample average and standard deviation measured in °F? [*Hint:* $F = \frac{9}{5}C + 32$.]

**70.** Elevated energy consumption during exercise continues after the workout ends. Because calories burned after exercise contribute to weight loss and have other consequences, it is important to understand this process. The paper "Effect of Weight Training Exercise and Treadmill Exercise on Post-Exercise Oxygen Consumption" (*Medicine and Science in Sports and Exercise,* 1998: 518–522) reported the accompanying data from a study in which oxygen consumption (liters) was measured continuously for 30 minutes for each of 15 subjects both after a weight training exercise and after a treadmill exercise.

| *Subject* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| *Weight* ($x$) | 14.6 | 14.4 | 19.5 | 24.3 | 16.3 | 22.1 | 23.0 | |
| *Treadmill* ($y$) | 11.3 | 5.3 | 9.1 | 15.2 | 10.1 | 19.6 | 20.8 | |
| *Subject* | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| *Weight* ($x$) | 18.7 | 19.0 | 17.0 | 19.1 | 19.6 | 23.2 | 18.5 | 15.9 |
| *Treadmill* ($y$) | 10.3 | 10.3 | 2.6 | 16.6 | 22.4 | 23.6 | 12.6 | 4.4 |

**a.** Construct a comparative boxplot of the weight and treadmill observations, and comment on what you see.
**b.** Because the data is in the form of $(x, y)$ pairs, with $x$ and $y$ measurements on the same variable under two different conditions, it is natural to focus on the differences within pairs: $d_1 = x_1 - y_1, \ldots, d_n = x_n - y_n$. Construct a boxplot of the sample differences. What does it suggest?

**71.** The accompanying frequency distribution of fracture strength (MPa) observations for ceramic bars fired in a particular kiln appeared in the article "Evaluating Tunnel Kiln Performance" (*Amer. Ceramic Soc. Bull.,* Aug. 1997: 59–63).

| *Class* | 81−<83 | 83−<85 | 85−<87 | 87−<89 | 89−<91 |
|---|---|---|---|---|---|
| *Frequency* | 6 | 7 | 17 | 30 | 43 |
| *Class* | 91−<93 | 93−<95 | 95−<97 | 97−<99 | |
| *Frequency* | 28 | 22 | 13 | 3 | |

**a.** Construct a histogram based on relative frequencies, and comment on any interesting features.
**b.** What proportion of the strength observations are at least 85? Less than 95?
**c.** Roughly what proportion of the observations are less than 90?

**72.** Anxiety disorders and symptoms can often be effectively treated with benzodiazepine medications. It is known that animals exposed to stress exhibit a decrease in benzodiazepine receptor binding in the frontal cortex. The paper "Decreased Benzodiazepine Receptor Binding in Prefrontal Cortex in Combat-Related Posttraumatic Stress Disorder" (*Amer. J. of Psychiatry,* 2000: 1120–1126) described the first study of benzodiazepine receptor binding in individuals suffering from PTSD. The accompanying data on a receptor binding measure (adjusted distribution volume) was read from a graph in the paper.

*PTSD:* 10, 20, 25, 28, 31, 35, 37, 38, 38, 39, 39, 42, 46
*Healthy:* 23, 39, 40, 41, 43, 47, 51, 58, 63, 66, 67, 69, 72

Use various methods from this chapter to describe and summarize the data.

**73.** The article "Can We Really Walk Straight?" (*Amer. J. of Physical Anthropology,* 1992: 19–27) reported on an experiment in which each of 20 healthy men was asked to walk as straight as possible to a target 60 m away at normal speed. Consider the following observations on cadence (number of strides per second):

.95 .85 .92 .95 .93 .86 1.00 .92 .85 .81
.78 .93 .93 1.05 .93 1.06 1.06 .96 .81 .96

Use the methods developed in this chapter to summarize the data; include an interpretation or discussion wherever appropriate. [*Note:* The author of the article used a rather sophisticated statistical analysis to conclude that people cannot walk in a straight line and suggested several explanations for this.]

**74.** The **mode** of a numerical data set is the value that occurs most frequently in the set.
**a.** Determine the mode for the cadence data given in Exercise 73.
**b.** For a categorical sample, how would you define the modal category?

**75.** Specimens of three different types of rope wire were selected, and the fatigue limit (MPa) was determined for each specimen, resulting in the accompanying data.

*Type 1* 350 350 350 358 370 370 370 371
371 372 372 384 391 391 392

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Type 2* | 350 | 354 | 359 | 363 | 365 | 368 | 369 | 371 |
| | 373 | 374 | 376 | 380 | 383 | 388 | 392 | |
| *Type 3* | 350 | 361 | 362 | 364 | 364 | 365 | 366 | 371 |
| | 377 | 377 | 377 | 379 | 380 | 380 | 392 | |

**a.** Construct a comparative boxplot, and comment on similarities and differences.

**b.** Construct a comparative dotplot (a dotplot for each sample with a common scale). Comment on similarities and differences.

**c.** Does the comparative boxplot of part (a) give an informative assessment of similarities and differences? Explain your reasoning.

**76.** The three measures of center introduced in this chapter are the mean, median, and trimmed mean. Two additional measures of center that are occasionally used are the *midrange,* which is the average of the smallest and largest observations, and the *midfourth,* which is the average of the two fourths.Which of these five measures of center are resistant to the effects of outliers and which are not? Explain your reasoning.

**77.** The authors of the article "Predictive Model for Pitting Corrosion in Buried Oil and Gas Pipelines" (*Corrosion*, 2009: 332–342) provided the data on which their investigation was based.

**a.** Consider the following sample of 61 observations on maximum pitting depth (mm) of pipeline specimens buried in clay loam soil.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0.41 | 0.41 | 0.41 | 0.41 | 0.43 | 0.43 | 0.43 | 0.48 | 0.48 |
| 0.58 | 0.79 | 0.79 | 0.81 | 0.81 | 0.81 | 0.91 | 0.94 | 0.94 |
| 1.02 | 1.04 | 1.04 | 1.17 | 1.17 | 1.17 | 1.17 | 1.17 | 1.17 |
| 1.17 | 1.19 | 1.19 | 1.27 | 1.40 | 1.40 | 1.59 | 1.59 | 1.60 |
| 1.68 | 1.91 | 1.96 | 1.96 | 1.96 | 2.10 | 2.21 | 2.31 | 2.46 |
| 2.49 | 2.57 | 2.74 | 3.10 | 3.18 | 3.30 | 3.58 | 3.58 | 4.15 |
| 4.75 | 5.33 | 7.65 | 7.70 | 8.13 | 10.41 | 13.44 | | |

Construct a stem-and-leaf display in which the two largest values are shown in a last row labeled HI.

**b.** Refer back to (a), and create a histogram based on eight classes with 0 as the lower limit of the first class and class widths of .5, .5, .5, .5, 1, 2, 5, and 5, respectively.

**c.** The accompanying comparative boxplot from Minitab shows plots of pitting depth for four different types of soils. Describe its important features.

**78.** Consider a sample $x_1, x_2, \ldots, x_n$ and suppose that the values of $\bar{x}$, $s^2$, and $s$ have been calculated.

**a.** Let $y_i = x_i - \bar{x}$ for $i = 1, \ldots, n$. How do the values of $s^2$ and $s$ for the $y_i$'s compare to the corresponding values for the $x_i$'s? Explain.

**b.** Let $z_i = (x_i - \bar{x})/s$ for $i = 1, \ldots, n$. What are the values of the sample variance and sample standard deviation for the $z_i$'s?

**79.** Here is a description from Minitab of the strength data given in Exercise 17.

```
Variable   N    Mean Median TrMean StDev SE Mean
strength 153 135.39 135.40 135.41  4.59    0.37

Variable    Minimum    Maximum        Q1        Q3
strength     122.20     147.70    132.95    138.25
```

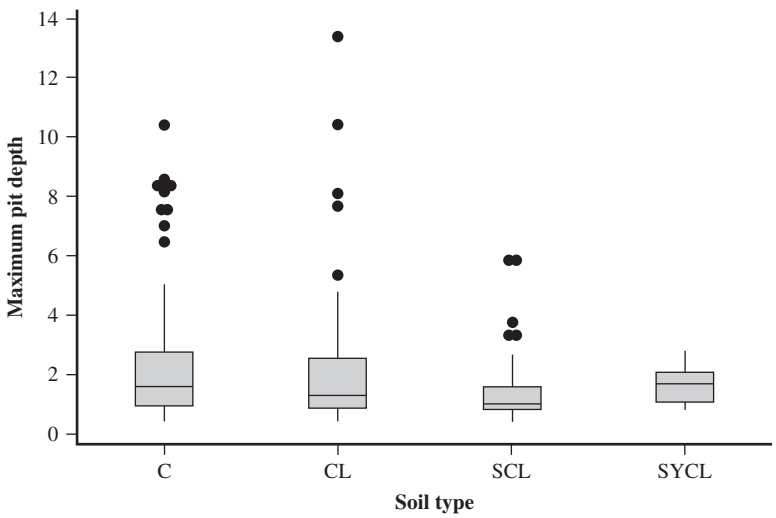
**a.** Comment on any interesting features (the quartiles and fourths are virtually identical here).

**b.** Construct a boxplot of the data based on the quartiles, and comment on what you see.

**80.** Lengths of bus routes for any particular transit system will typically vary from one route to another. The article "Planning of City Bus Routes" (*J. of the Institution of Engineers,* 1995: 211–215) gives the following information on lengths (km) for one particular system:

| | | | | | |
|---|---|---|---|---|---|
| *Length* | 6−<8 | 8−<10 | 10−<12 | 12−<14 | 14−<16 |
| *Frequency* | 6 | 23 | 30 | 35 | 32 |
| *Length* | 16−<18 | 18−<20 | 20−<22 | 22−<24 | 24−<26 |
| *Frequency* | 48 | 42 | 40 | 28 | 27 |
| *Length* | 26−<28 | 28−<30 | 30−<35 | 35−<40 | 40−<45 |
| *Frequency* | 26 | 14 | 27 | 11 | 2 |

Comparative boxplot for Exercise 77



**a.** Draw a histogram corresponding to these frequencies.
**b.** What proportion of these route lengths are less than 20? What proportion of these routes have lengths of at least 30?
**c.** Roughly what is the value of the 90th percentile of the route length distribution?
**d.** Roughly what is the median route length?

**81.** A study carried out to investigate the distribution of total braking time (reaction time plus accelerator-to-brake movement time, in ms) during real driving conditions at 60 km/hr gave the following summary information on the distribution of times ("A Field Study on Braking Responses During Driving," *Ergonomics,* 1995: 1903–1910):

mean = 535 median = 500 mode = 500
sd = 96 minimum = 220 maximum = 925
5th percentile = 400 10th percentile = 430
90th percentile = 640 95th percentile = 720

What can you conclude about the shape of a histogram of this data? Explain your reasoning.

**82.** The sample data $x_1, x_2, \ldots, x_n$ sometimes represents a **time series,** where $x_t$ = the observed value of a response variable $x$ at time $t$. Often the observed series shows a great deal of random variation, which makes it difficult to study longer-term behavior. In such situations, it is desirable to produce a smoothed version of the series. One technique for doing so involves **exponential smoothing.** The value of a smoothing constant $\alpha$ is chosen $(0 < \alpha < 1)$. Then with $\bar{x}_t$ = smoothed value at time $t$, we set $\bar{x}_1 = x_1$, and for $t = 2, 3, \ldots, n$, $\bar{x}_t = \alpha x_t + (1 - \alpha)\bar{x}_{t-1}$.

**a.** Consider the following time series in which $x_t$ = temperature (°F) of effluent at a sewage treatment plant on day $t$: 47, 54, 53, 50, 46, 46, 47, 50, 51, 50, 46, 52, 50, 50. Plot each $x_t$ against $t$ on a two-dimensional coordinate system (a time-series plot). Does there appear to be any pattern?
**b.** Calculate the $\bar{x}_t$'s using $\alpha = .1$. Repeat using $\alpha = .5$. Which value of $\alpha$ gives a smoother $\bar{x}_t$ series?
**c.** Substitute $\bar{x}_{t-1} = \alpha x_{t-1} + (1 - \alpha)\bar{x}_{t-2}$ on the right-hand side of the expression for $\bar{x}_t$, then substitute $\bar{x}_{t-2}$ in terms of $x_{t-2}$ and $\bar{x}_{t-3}$, and so on. On how many of the values $x_t, x_{t-1}, \ldots, x_1$ does $\bar{x}_t$ depend? What happens to the coefficient on $x_{t-k}$ as $k$ increases?
**d.** Refer to part (c). If $t$ is large, how sensitive is $\bar{x}_t$ to the initialization $\bar{x}_1 = x_1$? Explain.

[*Note:* A relevant reference is the article "Simple Statistics for Interpreting Environmental Data," *Water Pollution Control Fed. J.,* 1981: 167–175.]

**83.** Consider numerical observations $x_1, \ldots, x_n$. It is frequently of interest to know whether the $x_i$ s are (at least approximately) symmetrically distributed about some value. If $n$ is at least moderately large, the extent of symmetry can be assessed from a stem-and-leaf display or histogram. However, if $n$ is not very large, such pictures are not particularly informative. Consider the following alternative. Let $y_1$ denote the smallest $x_i$, $y_2$ the second smallest $x_i$, and so on. Then plot the following pairs as points on a two-dimensional coordinate system: $(y_n - \tilde{x}, \tilde{x} - y_1)$, $(y_{n-1} - \tilde{x}, \tilde{x} - y_2)$, $(y_{n-2} - \tilde{x}, \tilde{x} - y_3)$, . . . There are $n/2$ points when $n$ is even and $(n - 1)/2$ when $n$ is odd.

**a.** What does this plot look like when there is perfect symmetry in the data? What does it look like when observations stretch out more above the median than below it (a long upper tail)?
**b.** The accompanying data on rainfall (acre-feet) from 26 seeded clouds is taken from the article "A Bayesian Analysis of a Multiplicative Treatment Effect in Weather Modification" (*Technometrics,* 1975: 161–166). Construct the plot and comment on the extent of symmetry or nature of departure from symmetry.

| | | | | | | |
|---|---|---|---|---|---|---|
| 4.1 | 7.7 | 17.5 | 31.4 | 32.7 | 40.6 | 92.4 |
| 115.3 | 118.3 | 119.0 | 129.6 | 198.6 | 200.7 | 242.5 |
| 255.0 | 274.7 | 274.7 | 302.8 | 334.1 | 430.0 | 489.1 |
| 703.4 | 978.0 | 1656.0 | 1697.8 | 2745.6 | | |

## Bibliography


Chambers, John, William Cleveland, Beat Kleiner, and Paul Tukey, *Graphical Methods for Data Analysis,* Brooks/Cole, Pacific Grove, CA, 1983. A highly recommended presentation of various graphical and pictorial methodology in statistics.

Cleveland, William, *Visualizing Data,* Hobart Press, Summit, NJ, 1993. An entertaining tour of pictorial techniques.

Peck, Roxy, and Jay Devore, *Statistics: The Exploration and Analysis of Data* (6th ed.), Thomson Brooks/Cole, Belmont, CA, 2008. The first few chapters give a very nonmathematical survey of methods for describing and summarizing data.

Freedman, David, Robert Pisani, and Roger Purves, *Statistics* (4th ed.), Norton, New York, 2007. An excellent, very nonmathematical survey of basic statistical reasoning and methodology.

Hoaglin, David, Frederick Mosteller, and John Tukey, *Understanding Robust and Exploratory Data Analysis,* Wiley, New York, 1983. Discusses why, as well as how, exploratory methods should be employed; it is good on details of stem-and-leaf displays and boxplots.

Moore, David, and William Notz, *Statistics: Concepts and Controversies* (7th ed.), Freeman, San Francisco, 2009. An extremely readable and entertaining paperback that contains an intuitive discussion of problems connected with sampling and designed experiments.

Peck, Roxy, et al. (eds.), *Statistics: A Guide to the Unknown* (4th ed.), Thomson Brooks/Cole, Belmont, CA, 2006. Contains many short nontechnical articles describing various applications of statistics.

Verzani, John, *Using R for Introductory Statistics,* Chapman and Hall/CRC, Boca Raton, FL, 2005. A very nice introduction to the R software package.

# Answers to Selected Odd-Numbered Exercises

## Chapter 1

**1. a.** No. All students taking a large statistics course who participate in an SI program of this sort.
**b.** Randomization protects against various biases and helps ensure that those in the SI group are as similar as possible to the students in the control group.
**c.** There would be no firm basis for assessing the effectiveness of SI (nothing to which the SI scores could reasonably be compared).

**3. a.** How likely is it that more than half of the sampled computers will need or have needed warranty service? What is the expected number among the 100 that need warranty service? How likely is it that the number needing warranty service will exceed the expected number by more than 10?
**b.** Suppose that 15 of the 100 sampled needed warranty service. How confident can we be that the proportion of *all* such computers needing warranty service is between .08 and .22? Does the sample provide compelling evidence for concluding that more than 10% of all such computers need warranty service?

**5. a.** *Los Angeles Times, Oberlin Tribune, Gainesville Sun, Washington Post*
**b.** Duke Energy, Clorox, Seagate, Neiman Marcus
**c.** Vince Correa, Catherine Miller, Michael Cutler, Ken Lee
**d.** 2.97, 3.56, 2.20, 2.97

**7.** One could generate a simple random sample of all single-family homes in the city, or a stratified random sample by taking a simple random sample from each of the 10 district neighborhoods. From each of the selected homes, values of all desired variables would be determined. This would be an enumerative study because there exists a finite, identifiable population of objects from which to sample.

**9. a.** Possibly measurement error, recording error, differences in environmental conditions at the time of measurement, etc.
**b.** No. There is no sampling frame.

**11.**

| | |
|---|---|
| 6L | 430 |
| 6H | 769689 |
| 7L | 42014202 |
| 7H | |
| 8L | 011211410342 |
| 8H | 9595578 |
| 9L | 30 |
| 9H | 58 |

The gap in the data—no scores in the high 70s.

**13. a.**

| $y$ | Freq. | Rel. freq. |
|---|---|---|
| 0 | 17 | .362 |
| 1 | 22 | .468 |
| 2 | 6 | .128 |
| 3 | 1 | .021 |
| 4 | 0 | .000 |
| 5 | 1 | .021 |
| | 47 | 1.000 |

.362, .638

**b.**

| $z$ | Freq. | Rel. freq. |
|---|---|---|
| 0 | 13 | .277 |
| 1 | 11 | .234 |
| 2 | 3 | .064 |
| 3 | 7 | .149 |
| 4 | 5 | .106 |
| 5 | 3 | .064 |
| 6 | 3 | .064 |
| 7 | 0 | .000 |
| 8 | 2 | .043 |
| | 47 | 1.001 |

.894, .830

**15.**

| Am | | Fr |
|---|---|---|
| | 8 | 1 |
| 157020153504 | 9 | 00645632 |
| 9324 | 10 | 2563 |
| 6306 | 11 | 6913 |
| Stem: Hundreds and tens 058 | 12 | 325528 |
| Leaf: Ones 8 | 13 | 7 |
| | 14 | |
| | 15 | 8 |
| 2 | 16 | |

Representative values: low 100s for Am and low 110's for Fr. Somewhat more variability in Fr times than in Am times. More extreme positive skew for Am than for Fr. 162 is an Am outlier, and 158 is perhaps a Fr outlier.

**17. a.** 12 2 leaf: ones digit
12 445
12 6667777
12 889999
13 00011111111
13 22222222233333333333333
13 44444444444444444555555555555555555555
13 6666666666677777777777
13 888888888888999999
14 0000001111
14 2333333
14 444
14 77
symmetry

**b.** Close to bell-shaped, center ≈ 135, not insignificant dispersion, no gaps or outliers.

**19. a.** .99 (99%), .71 (71%) **b.** .64 (64%), .44 (44%)

**c.** Strictly speaking, the histogram is not unimodal, but is close to being so with a moderate positive skew. A much larger sample size would likely give a smoother picture.

**21. a.**

| # Nonconforming | Frequency | Rel. freq. |
|---|---|---|
| 0 | 7 | .117 |
| 1 | 12 | .200 |
| 2 | 13 | .217 |
| 3 | 14 | .233 |
| 4 | 6 | .100 |
| 5 | 3 | .050 |
| 6 | 3 | .050 |
| 7 | 1 | .017 |
| 8 | 1 | .017 |
| | 60 | 1.001 |

**b.** .917, .867, 1 − .867 = .133

**c.** The histogram has a substantial positive skew. It is centered somewhere between 2 and 3 and spreads out quite a bit about its center.

**23.**

| Complaint | Freq. | Rel. freq. |
|---|---|---|
| J | 10 | .1667 |
| F | 9 | .1500 |
| B | 7 | .1167 |
| M | 4 | .0667 |
| C | 3 | .0500 |
| N | 6 | .1000 |
| O | 21 | .3500 |
| | 60 | 1.0001 |

**25.**

| Class | Freq. | Class | Freq. |
|---|---|---|---|
| 10–<20 | 8 | 1.1–<1.2 | 2 |
| 20–<30 | 14 | 1.2–<1.3 | 6 |
| 30–<40 | 8 | 1.3–<1.4 | 7 |
| 40–<50 | 4 | 1.4–<1.5 | 9 |
| 50–<60 | 3 | 1.5–<1.6 | 6 |
| 60–<70 | 2 | 1.6–<1.7 | 4 |
| 70–<80 | 1 | 1.7–<1.8 | 5 |
| | 40 | 1.8–<1.9 | 1 |
| | | | 40 |

Original: positively skewed;
Transformed much more symmetric, not far from bell-shaped.

**27. a.** The observation 50 falls on a class boundary.

**b.**

| Class | Freq. | Rel. freq. |
|---|---|---|
| 0–<50 | 9 | .18 |
| 50–<100 | 19 | .38 |
| 100–<150 | 11 | .22 |
| 150–<200 | 4 | .08 |
| 200–<300 | 4 | .08 |
| 300–<400 | 2 | .04 |
| 400–<500 | 0 | .00 |
| 500–<600 | 1 | .02 |
| | 50 | 1.00 |

A representative (central) value is either a bit below or a bit above 100, depending on how one measures center. There is a great deal of variability in lifetimes, especially in values at the upper end of the data. There are several candidates for outliers.

**c.**

| Class | Freq. | Rel. freq. |
|---|---|---|
| 2.25–<2.75 | 2 | .04 |
| 2.75–<3.25 | 2 | .04 |
| 3.25–<3.75 | 3 | .06 |
| 3.75–<4.25 | 8 | .16 |
| 4.25–<4.75 | 18 | .36 |
| 4.75–<5.25 | 10 | .20 |
| 5.25–<5.75 | 4 | .08 |
| 5.75–<6.25 | 3 | .06 |
| | 50 | 1.00 |

There is much more symmetry in the distribution of the $\ln(x)$ values than in the $x$ values themselves, and less variability. There are no longer gaps or obvious outliers.

**d.** .38, .14

**29.** The class widths are not equal, so the density scale must be used. The densities for the six classes are .2030, .1373, .0303, .0086, .0021, and .0009, respectively. The resulting histogram is unimodal with a very substantial positive skew.

**31.**

| Class | Freq. | Cum. freq. | Cum. rel. freq. |
|---|---|---|---|
| 0–<4 | 2 | 2 | .050 |
| 4–<8 | 14 | 16 | .400 |
| 8–<12 | 11 | 27 | .675 |
| 12–<16 | 8 | 35 | .875 |
| 16–<20 | 4 | 39 | .975 |
| 20–<24 | 0 | 39 | .975 |
| 24–<28 | 1 | 40 | 1.000 |

**33. a.** 640.5, 582.5
**b.** 610.5, 582.5
**c.** 591.2
**d.** 593.71

**35.** $\tilde{x} = 68.0, \bar{x}_{tr(20)} = 66.2, \bar{x}_{tr(30)} = 67.5$

**37.** $\bar{x}_{tr(10)} = 11.46$

**39. a.** $\bar{x} = 12.55, \tilde{x} = 12.50, \bar{x}_{tr(12.5)} = 12.40$. Deletion of the largest observation (18.0) causes $\tilde{x}$ and $\bar{x}_{tr}$ to be a bit smaller than $\bar{x}$.
**b.** By at most 4.0 **c.** No; multiply the values of $\bar{x}$ and $\tilde{x}$ by the conversion factor 1/2.2.

**41. a.** .7 **b.** Also .7 **c.** 13

**43. a.** $\bar{x} = 1.0297, \tilde{x} = 1.009$ **b.** .383

**45. a.** 1264.766, 35.564 **b.** .351, .593

**47.** $\bar{x} = 116.2, s = 25.75$. The magnitude of $s$ indicates a substantial amount of variation about the center (a "representative" deviation of roughly 25).

**49. a.** 56.80, 197.8040 **b.** .5016, .708

**51. a.** Yes. 125.8 is an extreme outlier and 250.2 is a mild outlier.
**b.** In addition to the presence of outliers, there is positive skewness both in the middle 50% of the data and, excepting the outliers, overall. Except for the two outliers, there appears to be a relatively small amount of variability in the data.

**53. a.** Bal: 1.121, 1.050, .536
Gr: 1.244, 1.100, .448
**b.** Typical ratios are quite similar for the two types. There is somewhat more variability in the Bal sample, due primarily to the two outliers (one mild, one extreme). For Bal, there is substantial symmetry in the middle 50% but positive skewness overall. For Gr, there is substantial positive skew in the middle 50% and mild positive skewness overall.

**55. a.** 33 **b.** No
**c.** Slight positive skewness in the middle half, but rather symmetric overall. The extent of variability appears substantial.
**d.** At most 32

**57. a.** $\bar{x} = 115.58$; the deviations are .82, .32, −.98, −.38, .22
**b.** .482, .694 **c.** .482 **d.** .482

**59. a.** ED: .4, .10, 2.75, 2.65;
Non-Ed: 1.60, .30, 7.90, 7.60
**b.** ED: 8.9 and 9.2 are mild outliers, and 11.7 and 21.0 are extreme outliers.
There are not outliers in the non-ED sample.
**c.** Four outliers for ED, none for non-ED. Substantial positive skewness in both samples; less variability in ED (smaller $f_s$), and non-ED observations tend to be somewhat larger than ED observations.

**61.** Outliers, both mild and extreme, only at 6 A.M. Distributions at other times are quite symmetric. Variability increases somewhat until 2 P.M. and then decreases slightly, and the same is true of "typical" gasoline-vapor coefficient values.

**63.** $\bar{x} = 64.89, \tilde{x} = 64.70, s = 7.803$, lower $4^{th} = 57.8$, upper $4^{th} = 70.4, f_s = 12.6$. A histogram consisting of 8 classes starting at 52, each of width 4, is bimodal but close to unimodal with a positive skew. A boxplot shows no outliers, there is a very mild negative skew in the middle 50%, and the upper whisker is much longer than the lower whisker.
**b.** .9231, .9053
**c.** .48

**65. a.** $\bar{x}_{n+1} = (n\bar{x}_n + x_{n+1})/(n + 1)$
**c.** 12.53, .532

**67. a.** M: $\bar{x} = 3.64, \tilde{x} = 3.70, s = .269, f_s = .40$
F: $\bar{x} = 3.28, \tilde{x} = 3.15, s = .478, f_s = .50$
Female values are typically somewhat smaller than male values, and show somewhat more variability. An M boxplot shows negative skew whereas an F boxplot shows positive skew.
**b.** F: $\bar{x}_{tr(10)} = 3.24$ M: $\bar{x}_{tr(10)} = 3.652 \approx 3.65$

**69. a.** $\bar{y} = a\bar{x} + b, s_y^2 = a^2 s_x^2$ **b.** 189.14, 1.87

**73.** $\bar{x} = .9255, s = .0809, \tilde{x} = .93$, small amount of variability, slight bit of skewness

**75. a.** The "five-number summaries" ($\tilde{x}$, the two fourths, and the smallest and largest observations) are identical and there are no outliers, so the three individual boxplots are identical.
**b.** Differences in variability, nature of gaps, and existence of clusters for the three samples.
**c.** No. Detail is lost.

**77. c.** Representative depths are quite similar for the four types of soils—between 1.5 and 2. Data from the C and CL soils shows much more variability than for the other two types. The boxplots for the first three types show substantial positive skewness both in the middle 50% and overall. The boxplot for the SYCL soil shows negative skewness in the middle 50% and mild positive skewness overall. Finally, there are multiple outliers for the first three types of soils, including extreme outliers.

**79. a.** The mean, median, and trimmed mean are virtually identical, suggesting a substantial amount of symmetry in the data; the fact that the quartiles are roughly the same distance from the median and that the smallest and largest observations are roughly equidistant from the center provides additional support for symmetry. The standard deviation is quite small relative to the mean and median.

**b.** See the comments of (a). In addition, using 1.5(Q3 − Q1) as a yardstick, the two largest and three smallest observations are mild outliers.

**81.** A substantial positive skew (assuming unimodality)

**83. a.** All points fall on a 45° line. Points fall below a 45° line.
**b.** Points fall well below a 45° line, indicating a substantial positive skew.