CHAPTER 3

Data Description

3.1 Introduction and Abstract of Research Study
3.2 Calculators, Computers, and Software Systems
3.3 Describing Data on a Single Variable: Graphical Methods
3.4 Describing Data on a Single Variable: Measures of Central Tendency
3.5 Describing Data on a Single Variable: Measures of Variability
3.6 The Boxplot
3.7 Summarizing Data from More Than One Variable: Graphs and Correlation
3.8 Research Study: Controlling for Student Background in the Assessment of Teaching
3.9 Summary and Key Formulas
3.10 Exercises

3.1 Introduction and Abstract of Research Study

In the previous chapter, we discussed how to gather data intelligently for an experiment or survey, Step 2 in Learning from Data. We turn now to Step 3, summarizing the data.

The field of statistics can be divided into two major branches: descriptive statistics and inferential statistics. In both branches, we work with a set of measurements. For situations in which data description is our major objective, the set of measurements available to us is frequently the entire population. For example, suppose that we wish to describe the distribution of annual incomes for all families registered in the 2000 census. Because all these data are recorded and are available on computer tapes, we do not need to obtain a random sample from the population; the complete set of measurements is at our disposal. Our major problem is in organizing, summarizing, and describing these data—that is, making sense of the data. Similarly, vast amounts of monthly, quarterly, and yearly data of medical costs are available for the managed health care industry, HMOs. These data are broken down by type of illness, age of patient, inpatient or outpatient care, prescription

costs, and out-of-region reimbursements, along with many other types of expenses. However, in order to present such data in formats useful to HMO managers, congressional staffs, doctors, and the consuming public, it is necessary to organize, summarize, and describe the data. Good descriptive statistics enable us to make sense of the data by reducing a large set of measurements to a few summary measures that provide a good, rough picture of the original measurements.

In situations in which we are unable to observe all units in the population, a sample is selected from the population and the appropriate measurements are made. We use the information in the sample to draw conclusions about the population from which the sample was drawn. However, in order for these inferences about the population to have a valid interpretation, the sample should be a random sample of one of the forms discussed in Chapter 2. During the process of making inferences, we also need to organize, summarize, and describe the data.

For example, the tragedy surrounding isolated incidents of product tampering has brought about federal legislation requiring tamper-resistant packaging for certain drug products sold over the counter. These same incidents also brought about increased industry awareness of the need for rigid standards of product and packaging quality that must be maintained while delivering these products to the store shelves. In particular, one company is interested in determining the proportion of packages out of total production that are improperly sealed or have been damaged in transit. Obviously, it would be impossible to inspect all packages at all stores where the product is sold, but a random sample of the production could be obtained, and the proportion defective in the sample could be used to estimate the actual proportion of improperly sealed or damaged packages.

Similarly, in order to monitor changes in the purchasing power of consumer's income, the federal government uses the Consumer Price Index (CPI) to measure the average change in prices over time in a market of goods and services purchased by urban wage earners. The current CPI is based on prices of food, clothing, shelter, fuels, transportation fares, charges for doctors' and dentists' services, drugs, and so on, purchased for day-to-day living. Prices are sampled from 85 areas across the country from over 57,000 housing units and 19,000 business establishments. Forecasts and inferences are then made using this information.

A third situation involves an experiment in which a drug company wants to study the effects of two factors on the level of blood sugar in diabetic patients. The factors are the type of drug (a new drug and two drugs currently being used) and the method of administering (two different delivery modes) the drug to the diabetic patient. The experiment involves randomly selecting a method of administering the drug and randomly selecting a type of drug then giving the drug to the patient. The fasting blood sugar of the patient is then recorded for at the time the patient receives the drug and at 6 hours intervals over a 2-day period of time. The six unique combinations of a type of drug and method of delivery are given to 10 different patients. In this experiment, the drug company wants to make inferences from the results of the experiment to determine if the new drug is commercially viable. In many experiments of this type, the use of the proper graphical displays provides valuable insights to the scientists with respect to unusual occurrences and in making comparisons of the responses to the different treatment combinations.

Whether we are describing an observed population or using sampled data to draw an inference from the sample to the population, an insightful description of the data is an important step in drawing conclusions from it. No matter what our objective, statistical inference or population description, we must first adequately describe the set of measurements at our disposal.

The two major methods for describing a set of measurements are graphical techniques and numerical descriptive techniques. Section 3.3 deals with graphical methods for describing data on a single variable. In Sections 3.4, 3.5, and 3.6, we discuss numerical techniques for describing data. The final topics on data description are presented in Section 3.7, in which we consider a few techniques for describing (summarizing) data on more than one variable. A research study involving the evaluation of primary school teachers will be used to illustrate many of the summary statistics and graphs introduced in this chapter.

Abstract of Research Study: Controlling for Student Background in the Assessment of Teachers

By way of background, there was a movement to introduce achievement standards and school/teacher accountability in the public schools of our nation long before the "No Child Left Behind" bill was passed by the Congress during the first term of President George W. Bush. However, even after an important federal study entitled "A Nation at Risk" (1983) spelled out the grave trend toward mediocrity in our schools and the risk this poses for the future, neither Presidents Reagan, H. W. Bush, nor Clinton ventured into this potentially sensitive area to champion meaningful change.

Many politicians, teachers, and educational organizations have criticized the No Child Left Behind (NCLB) legislation, which requires rigid testing standards in exchange for money to support low-income students. A recent survey conducted by the Educational Testing Service (ETS) with bipartisan sponsorship from the Congress showed the following:

- Those surveyed identified the value of our education as the most important source of America's success in the world. (Also included on the list of alternatives were our military strength, our geographical and natural resources, our democratic system of government, our entrepreneurial spirit, etc.)
- 45% of the parents surveyed viewed the NCLB reforms favorably; 34% viewed it unfavorably.
- Only 19% of the high school teachers surveyed viewed the NCLB reforms favorably, while 75% viewed it unfavorably.

Given the importance placed on education, the difference or gap between the responses of parents and those of educators is troubling. The tone of much of the criticism seems to run against the empirical results seen to date with the NCLB program. For example, in 2004 the Center on Education Policy, an independent research organization, reported that 36 of 49 (73.5%) schools surveyed showed improvement in student achievement.

One of the possible sources of criticism coming from the educators is that there is a risk of being placed on a "watch list" if the school does not meet the performance standards set. This would reflect badly on the teacher, the school, and the community. But another important source of the criticism by the teachers and of the gap between what parents and teachers favor relates to the performance standards themselves. In the previously mentioned ETS survey, those polled were asked whether the same standard should be used for all students of a given grade, regardless of their background, because of the view that it is wrong to have lower expectations for students from disadvantaged backgrounds. The opposing view is that it is not reasonable to expect teachers to be able to bring the level of achievement for disadvantaged students to the same level as students from more affluent areas. While more than 50% of the parents favored a single standards, only 25% of the teachers suggested this view.

Next we will examine some data that may offer some way to improve the NCLB program while maintaining the important concepts of performance standards and accountability.

In an article in the Spring 2004 issue of *Journal of Educational and Behavioral Statistics*, "An empirical comparison of statistical models for value-added assessment of school performance," data were presented from three elementary school grade cohorts (3rd–5th grades) in 1999 in a medium-sized Florida school district with 22 elementary schools. The data are given in Table 3.1. The minority status of a student was defined as black or non-black race. In this school district, almost all students are non-Hispanic blacks or whites. Most of the relatively small numbers of Hispanic students are white. Most students of other races are Asian but are relatively few in number. They were grouped in the minority category because of the similarity of their test score profiles. Poverty status was based on whether or not the student received free or reduced lunch subsidy. The math and reading scores are from the Iowa Test of Basic Skills. The number of students by class in each school is given by *N* in the table.

The superintendent of the schools presented the school board members with the data and they wanted an assessment of whether poverty and minority status had any effect on the math and reading scores. Just looking at the data in the table presented very little insight to answering this question. At the end of this chapter, we will present a discussion of what types of graphs and summary statistics would be beneficial to the school board in reaching a conclusion about the impact of these two variables on student performance.

TABLE 3.1
Assessment of elementary school performance

Third Grade					
School	**Math**	**Reading**	**% Minority**	**% Poverty**	***N***
1	166.4	165.0	79.2	91.7	48
2	159.6	157.2	73.8	90.2	61
3	159.1	164.4	75.4	86.0	57
4	155.5	162.4	87.4	83.9	87
5	164.3	162.5	37.3	80.4	51
6	169.8	164.9	76.5	76.5	68
7	155.7	162.0	68.0	76.0	75
8	165.2	165.0	53.7	75.8	95
9	175.4	173.7	31.3	75.6	45
10	178.1	171.0	13.9	75.0	36
11	167.1	169.4	36.7	74.7	79
12	177.1	172.9	26.5	63.2	68
13	174.2	172.7	28.3	52.9	191
14	175.6	174.9	23.7	48.5	97
15	170.8	174.9	14.5	39.1	110
16	175.1	170.1	25.6	38.4	86
17	182.8	181.4	22.9	34.3	70
18	180.3	180.6	15.8	30.3	165
19	178.8	178.0	14.6	30.3	89
20	181.4	175.9	28.6	29.6	98
21	182.8	181.6	21.4	26.5	98
22	186.1	183.8	12.3	13.8	130

(continued)

TABLE 3.1
Assessment of elementary school performance *(continued)*

Fourth Grade					
School	**Math**	**Reading**	**% Minority**	**% Poverty**	***N***
1	181.1	177.0	78.9	89.5	38
2	181.1	173.8	75.9	79.6	54
3	180.9	175.5	64.1	71.9	64
4	169.9	166.9	94.4	91.7	72
5	183.6	178.7	38.6	61.4	57
6	178.6	170.3	67.9	83.9	56
7	182.7	178.8	65.8	63.3	79
8	186.1	180.9	48.0	64.7	102
9	187.2	187.3	33.3	62.7	51
10	194.5	188.9	11.1	77.8	36
11	180.3	181.7	47.4	70.5	78
12	187.6	186.3	19.4	59.7	72
13	194.0	189.8	21.6	46.2	171
14	193.1	189.4	28.8	36.9	111
15	195.5	188.0	20.2	38.3	94
16	191.3	186.6	39.7	47.4	78
17	200.1	199.7	23.9	23.9	67
18	196.5	193.5	22.4	32.8	116
19	203.5	204.7	16.0	11.7	94
20	199.6	195.9	31.1	33.3	90
21	203.3	194.9	23.3	25.9	116
22	206.9	202.5	13.1	14.8	122

Fifth Grade					
School	**Math**	**Reading**	**% Minority**	**% Poverty**	***N***
1	197.1	186.6	81.0	92.9	42
2	194.9	200.1	83.3	88.1	42
3	192.9	194.5	56.0	80.0	50
4	193.3	189.9	92.6	75.9	54
5	197.7	199.6	21.7	67.4	46
6	193.2	193.6	70.4	76.1	71
7	198.0	200.9	64.1	67.9	78
8	205.2	203.5	45.5	61.0	77
9	210.2	223.3	34.7	73.5	49
10	204.8	199.0	29.4	55.9	34
11	205.7	202.8	42.3	71.2	52
12	201.2	207.8	15.8	51.3	76
13	205.2	203.3	19.8	41.2	131
14	212.7	211.4	26.7	41.6	101
15	—	—	—	—	—
16	209.6	206.5	22.4	37.3	67
17	223.5	217.7	14.3	30.2	63
18	222.8	218.0	16.8	24.8	137
19	—	—	—	—	—
20	228.1	222.4	20.6	23.5	102
21	221.0	221.0	10.5	13.2	114
22	—	—	—	—	—

3.2 Calculators, Computers, and Software Systems

Electronic calculators can be great aids in performing some of the calculations mentioned later in this chapter, especially for small data sets. For larger data sets, even hand-held calculators are of little use because of the time required to enter data. A computer can help in these situations. Specific programs or more general software systems can be used to perform statistical analyses almost instantaneously even for very large data sets after the data are entered into the computer. It is not necessary to know computer programming to make use of specific programs or software systems for planned analyses—most have user's manuals that give detailed directions for their use or provide pull-down menus that lead the user through the analysis of choice.

Many statistical software packages are available. A few of the more commonly used are SAS, SPSS, Minitab, R, JMP, and STATA. Because a software system is a group of programs that work together, it is possible to obtain plots, data descriptions, and complex statistical analyses in a single job. Most people find that they can use any particular system easily, although they may be frustrated by minor errors committed on the first few tries. The ability of such packages to perform complicated analyses on large amounts of data more than repays the initial investment of time and irritation.

In general, to use a system you need to learn about only the programs in which you are interested. Typical steps in a job involve describing your data to the software system, manipulating your data if they are not in the proper format or if you want a subset of your original data set, and then calling the appropriate set of programs or procedures using the key words particular to the software system you are using. The results obtained from calling a program are then displayed at your terminal or sent to your printer.

If you have access to a computer and are interested in using it, find out how to obtain an account, what programs and software systems are available for doing statistical analyses, and where to obtain instruction on data entry for these programs and software systems.

Because computer configurations, operating systems, and text editors vary from site to site, it is best to talk to someone knowledgeable about gaining access to a software system. Once you have mastered the commands to begin executing programs in a software system, you will find that running a job within a given software system is similar from site to site.

Because this isn't a text on computer use, we won't spend additional time and space on the mechanics, which are best learned by doing. Our main interest is in interpreting the output from these programs. The designers of these programs tend to include in the output everything that a user could conceivably want to know; as a result, in any particular situation, some of the output is irrelevant. When reading computer output look for the values you want; if you don't need or don't understand an output statistic, don't worry. Of course, as you learn more about statistics, more of the output will be meaningful. In the meantime, look for what you need and disregard the rest.

There are dangers in using such packages carelessly. A computer is a mindless beast, and will do anything asked of it, no matter how absurd the result might be. For instance, suppose that the data include age, gender (1 = female, 2 = male), religion (1 = Catholic, 2 = Jewish, 3 = Protestant, 4 = other or none), and monthly income of a group of people. If we asked the computer to calculate averages, we would get averages for the variables gender and religion, as well as for age and monthly income,

even though these averages are meaningless. Used intelligently, these packages are convenient, powerful, and useful—but be sure to examine the output from any computer run to make certain the results make sense. Did anything go wrong? Was something overlooked? In other words, be *skeptical.* One of the important acronyms of computer technology still holds; namely, GIGO: garbage in, garbage out.

Throughout the textbook, we will use computer software systems to do most of the more tedious calculations of statistics *after* we have explained how the calculations can be done. Used in this way, computers (and associated graphical and statistical analysis packages) will enable us to spend additional time on interpreting the results of the analyses rather than on doing the analyses.

3.3 Describing Data on a Single Variable: Graphical Methods

After the measurements of interest have been collected, ideally the data are organized, displayed, and examined by using various graphical techniques. As a general rule, the data should be arranged into categories so that *each measurement is classified into one, and only one, of the categories.* This procedure eliminates any ambiguity that might otherwise arise when categorizing measurements. For example, suppose a sex discrimination lawsuit is filed. The law firm representing the plaintiffs needs to summarize the salaries of all employees in a large corporation. To examine possible inequities in salaries, the law firm decides to summarize the 2005 yearly income rounded to the nearest dollar for all female employees into the categories listed in Table 3.2.

TABLE 3.2
Format for summarizing salary data

Income Level	Salary
1	less than $20,000
2	$20,000 to $39,999
3	$40,000 to $59,999
4	$60,000 to $79,999
5	$80,000 to $99,999
6	$100,000 or more

The yearly salary of each female employee falls into one, and only one, income category. However, if the income categories had been defined as shown in Table 3.3, then there would be confusion as to which category should be checked. For example, an employee earning $40,000 could be placed in either category 2 or 3. To reiterate: If the data are organized into categories, it is important to define the categories so that a measurement can be placed into only one category.

When data are organized according to this general rule, there are several ways to display the data graphically. The first and simplest graphical procedure for

TABLE 3.3
Format for summarizing salary data

Income Level	Salary
1	less than $20,000
2	$20,000 to $40,000
3	$40,000 to $60,000
4	$60,000 to $80,000
5	$80,000 to $100,000
6	$100,000 or more

pie chart

data organized in this manner is the **pie chart.** It is used to display the percentage of the total number of measurements falling into each of the categories of the variable by partitioning a circle (similar to slicing a pie).

The data of Table 3.4 represent a summary of a study to determine which types of employment may be the most dangerous to their employees. Using data from the National Safety Council, it was reported that in 1999, approximately 3,240,000 workers suffered disabling injuries (an injury that results in death, some degree of physical impairment, or renders the employee unable to perform regular activities for a full day beyond the day of the injury). Each of the 3,240,000 disabled workers was classified according to the industry group in which they were employed.

Although you can scan the data in Table 3.4, the results are more easily interpreted by using a pie chart. From Figure 3.1, we can make certain inferences about which industries have the highest number of injured employees and thus may require a closer scrutiny of their practices. For example, the services industry had nearly one-quarter, 24.3%, of all disabling injuries during 1999, whereas, government employees constituted only 14.9%. At this point, we must carefully consider what is being displayed in both Table 3.4 and Figure 3.1. These are the number of disabling injuries, and these figures do not take into account the number of workers employed in the various industry groups. To realistically reflect the risk of a disabling injury to the employees in each of the industry groups, we need to take into account the total number of employees in each of the industries. A rate of disabling injury could then be computed that would be a more informative index of the risk to a worked employed in each of the groups. For example, although the services group had the highest percentage of workers with a disabling injury, it had also the

TABLE 3.4
Disabling injuries by industry group

Industry Group	Number of Disabling Injuries (in 1,000s)	Percent of Total
Agriculture	130	3.4
Construction	470	12.1
Manufacturing	630	16.2
Transportation & Utilities	300	9.8
Trade	380	19.3
Services	750	24.3
Government	580	14.9

Source: Statistical Abstract of the United States—2002, 122nd Edition.

FIGURE 3.1
Pie chart for the data of Table 3.4

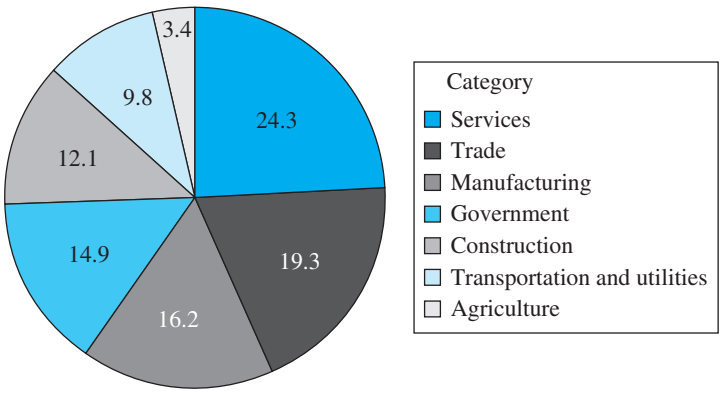
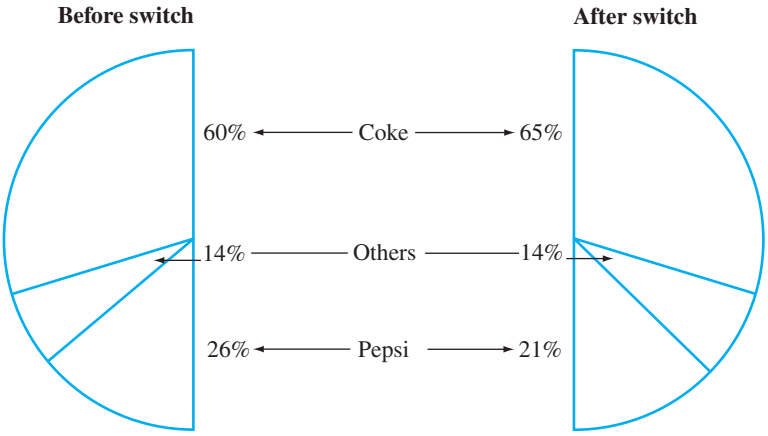


FIGURE 3.2
Estimated U.S. market share before and after switch in soft drink accounts



largest number of workers. Taking into account the number of workers employed in each of the industry groups, the services group had the lowest rate of disabling injuries in the seven groups. This illustrates the necessity of carefully examining tables of numbers and graphs prior to drawing conclusions.

Another variation of the pie chart is shown in Figure 3.2. It shows the loss of market share by PepsiCo as a result of the switch by a major fast-food chain from Pepsi to Coca-Cola for its fountain drink sales. In summary, the pie chart can be used to display percentages associated with each category of the variable. The following guidelines should help you to obtain clarity of presentation in pie charts.

Guidelines for Constructing Pie Charts

1. Choose a small number (five or six) of categories for the variable because too many make the pie chart difficult to interpret.
2. Whenever possible, construct the pie chart so that percentages are in either ascending or descending order.

bar chart

A second graphical technique is the **bar chart,** or bar graph. Figure 3.3 displays the number of workers in the Cincinnati, Ohio, area for the largest five foreign investors. There are many variations of the bar chart. Sometimes the bars are

FIGURE 3.3
Number of workers by major foreign investors

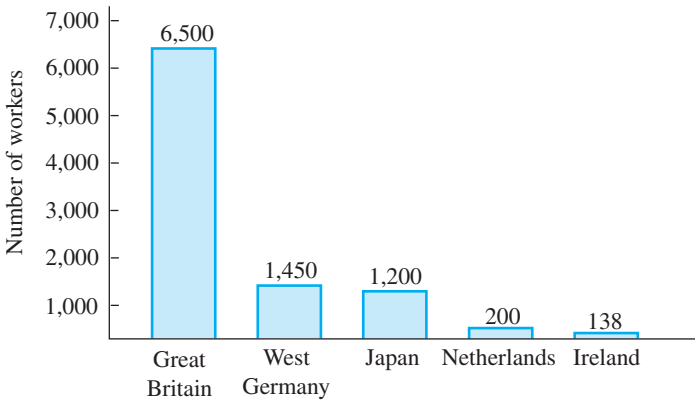
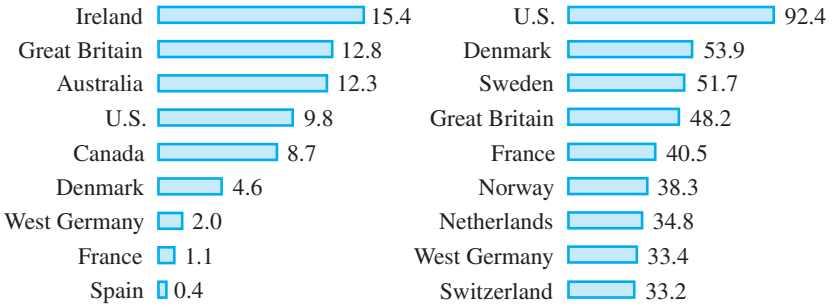
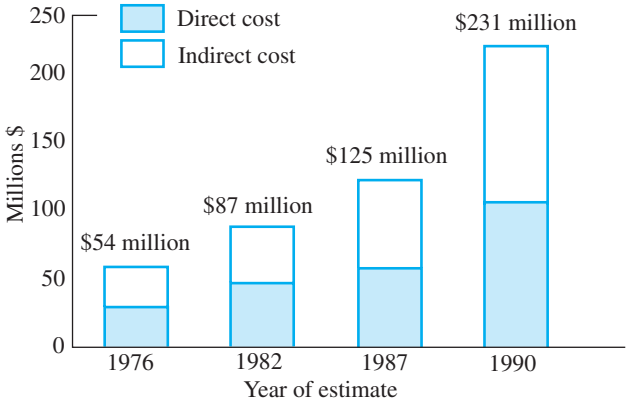


FIGURE 3.4
Greatest per capita consumption by country



(a) Breakfast cereals (in pounds) (b) Frozen foods (in pounds)

FIGURE 3.5
Estimated direct and indirect costs for developing a new drug by selected years



displayed horizontally, as in Figures 3.4(a) and (b). They can also be used to display data across time, as in Figure 3.5. Bar charts are relatively easy to construct if you use the following guidelines.

Guidelines for Constructing Bar Charts

1. Label frequencies on one axis and categories of the variable on the other axis.
2. Construct a rectangle at each category of the variable with a height equal to the frequency (number of observations) in the category.
3. Leave a space between each category to connote distinct, separate categories and to clarify the presentation.

frequency histogram, relative frequency histogram

The next two graphical techniques that we will discuss are the **frequency histogram** and the **relative frequency histogram.** Both of these graphical techniques are applicable only to quantitative (measured) data. As with the pie chart, we must organize the data before constructing a graph.

Gulf Coast ticks are significant pests of grazing cattle that require new strategies of population control. Some particular species of ticks are not only the source of considerable economic losses to the cattle industry due to weight loss in the cattle, but also are recognized vectors for a number of diseases in cattle. An entomologist carries out an experiment to investigate whether a new repellant for ticks is effective in preventing ticks from attaching to grazing cattle. The researcher determines that 100 cows will provide sufficient information to validate the results of the experiment and

TABLE 3.5
Number of attached ticks

17	18	19	20	20	20	21	21	21	22	22	22	22	23	23
23	24	24	24	24	24	25	25	25	25	25	25	25	26	26
27	27	27	27	27	27	28	28	28	28	28	28	28	28	28
28	28	29	29	29	29	29	29	29	29	29	29	30	30	30
30	30	30	30	30	31	31	31	31	31	31	32	32	32	32
32	32	32	32	33	33	33	34	34	34	34	35	35	35	36
36	36	36	37	37	38	39	40	41	42					

convince a commercial enterprise to manufacture and market the repellant. (In Chapter 5, we will present techniques for determining the appropriate sample size for a study to achieve specified goals.) The scientist will expose the cows to a specified number of ticks in a laboratory setting and then record the number of attached ticks after 1 hour of exposure. The average number of attached ticks on cows using a currently marketed repellant is 34 ticks. The scientist wants to demonstrate that using the new repellant will result in a reduction of the number of attached ticks. The numbers of attached ticks for the 100 cows are presented in Table 3.5.

An initial examination of the tick data reveals that the largest number of ticks is 42 and the smallest is 17. Although we might examine the table very closely to determine whether the number of ticks per cow is substantially less than 34, it is difficult to describe how the measurements are distributed along the interval 17 to 42. One way to obtain the answers to these questions is to organize the data in a **frequency table.**

frequency table

To construct a frequency table, we begin by dividing the range from 17 to 42 into an arbitrary number of subintervals called **class intervals.** The number of subintervals chosen depends on the number of measurements in the set, but we generally recommend using from 5 to 20 class intervals. The more data we have, the larger the number of classes we tend to use. The guidelines given here can be used for constructing the appropriate class intervals.

class intervals

Guidelines for Constructing Class Intervals

1. Divide the *range* of the measurements (the difference between the largest and the smallest measurements) by the approximate number of class intervals desired. Generally, we want to have from 5 to 20 class intervals.
2. After dividing the range by the desired number of subintervals, round the resulting number to a convenient (easy to work with) unit. This unit represents a common width for the class intervals.
3. Choose the first class interval so that it contains the smallest measurement. It is also advisable to choose a starting point for the first interval so that no measurement falls on a point of division between two subintervals, which eliminates any ambiguity in placing measurements into the class intervals. (One way to do this is to choose boundaries to one more decimal place than the data).

For the data in Table 3.5,

$$\text{range} = 42 - 17 = 25$$

Assume that we want to have approximately 10 subintervals. Dividing the range by 10 and rounding to a convenient unit, we have $25/10 = 2.5$. Thus, the class interval width is 2.5.

TABLE 3.6
Frequency table for number of attached ticks

Class	Class Interval	Frequency f_i	Relative Frequency f_i/n
1	16.25–18.75	2	.02
2	18.75–21.25	7	.07
3	21.25–23.75	7	.07
4	23.75–26.25	14	.14
5	26.25–28.75	17	.17
6	28.75–31.25	24	.24
7	31.25–33.75	11	.11
8	33.75–36.25	11	.11
9	36.25–38.75	3	.03
10	38.75–41.25	3	.03
11	41.25–43.75	1	.01
Totals		$n = 100$	1.00

It is convenient to choose the first interval to be 16.25–18.75, the second to be 18.75–21.25, and so on. Note that the smallest measurement, 17, falls in the first interval and that no measurement falls on the endpoint of a class interval. (See Tables 3.5 and 3.6.)

Having determined the class interval, we construct a frequency table for the data. The first column labels the classes by number and the second column indicates the class intervals. We then examine the 100 measurements of Table 3.5, keeping a tally of the number of measurements falling in each interval. The number of measurements falling in a given class interval is called the **class frequency.** These data are recorded in the third column of the frequency table. (See Table 3.6.)

class frequency

relative frequency

The **relative frequency** of a class is defined to be the frequency of the class divided by the total number of measurements in the set (total frequency). Thus, if we let f_i denote the frequency for class i and let n denote the total number of measurements, the relative frequency for class i is f_i/n. The relative frequencies for all the classes are listed in the fourth column of Table 3.6.

The data of Table 3.5 have been organized into a frequency table, which can now be used to construct a *frequency histogram* or a *relative frequency histogram.* To construct a frequency histogram, draw two axes: a horizontal axis labeled with the class intervals and a vertical axis labeled with the frequencies. Then construct a rectangle over each class interval with a height equal to the number of measurements falling in a given subinterval. The frequency histogram for the data of Table 3.6 is shown in Figure 3.6(a).

The relative frequency histogram is constructed in much the same way as a frequency histogram. In the relative frequency histogram, however, the vertical axis is labeled as relative frequency, and a rectangle is constructed over each class interval with a height equal to the class relative frequency (the fourth column of Table 3.6). The relative frequency histogram for the data of Table 3.6 is shown in Figure 3.6(b). Clearly, the two histograms of Figures 3.6(a) and (b) are of the same shape and would be identical if the vertical axes were equivalent. We will frequently refer to either one as simply a **histogram.**

histogram

There are several comments that should be made concerning histograms. First, the distinction between bar charts and histograms is based on the distinction between *qualitative* and *quantitative* variables. Values of qualitative variables vary in kind but not degree and hence are not measurements. For example, the variable political party affiliation can be categorized as Republican, Democrat, or other,

FIGURE 3.6(a)
Frequency histogram for the tick data of Table 3.6

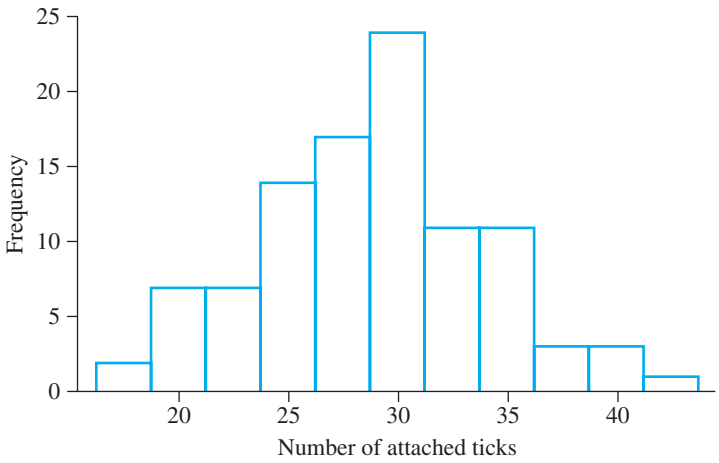
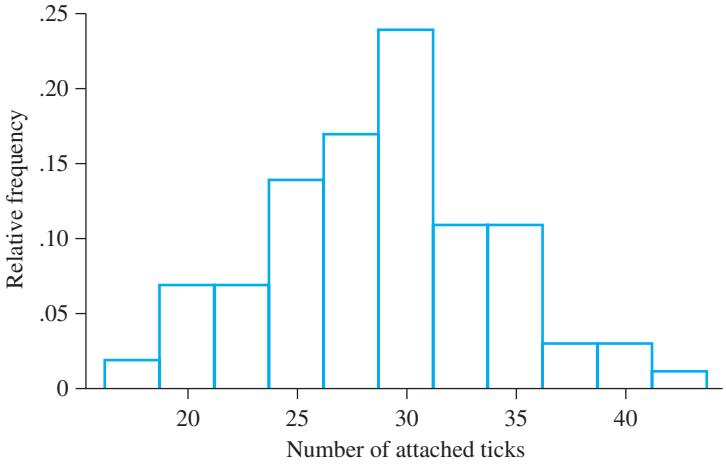


FIGURE 3.6(b)
Relative frequency histogram for the tick data of Table 3.6



and, although we could label the categories as one, two, or three, these values are only codes and have no quantitative interpretation. In contrast, quantitative variables have actual units of measure. For example, the variable yield (in bushels) per acre of corn can assume specific values. *Pie charts and bar charts are used to display frequency data from qualitative variables; histograms are appropriate for displaying frequency data for quantitative variables.*

Second, the histogram is the most important graphical technique we will present because of the role it plays in statistical inference, a subject we will discuss in later chapters. Third, if we had an extremely large set of measurements, and if we constructed a histogram using many class intervals, each with a very narrow width, the histogram for the set of measurements would be, for all practical purposes, a smooth curve. Fourth, the fraction of the total number of measurements in an interval is equal to the fraction of the total area under the histogram over the interval.

For example, suppose we consider those intervals having cows with fewer numbers of ticks than the average under the previously used repellant. That is, the intervals containing cows having a number of attached ticks less than 34. From Table 3.6, we observe that exactly 82 of the 100 cows had fewer than 34 attached ticks. Thus, the proportion of the total measurements falling in those intervals—82/100 = .82—is equal to the proportion of the total area under the histogram over those intervals.

probability

Fifth, if a single measurement is selected at random from the set of sample measurements, the chance, or **probability,** that the selected measurement lies in a particular interval is equal to the fraction of the total number of sample measurements falling in that interval. This same fraction is used to estimate the probability that a measurement selected from the population lies in the interval of interest. For example, from the sample data of Table 3.5, the chance or probability of selecting a cow with less than 34 attached ticks is .82. The value .82 is an approximation of the proportion of all cows treated with new repellant that would have fewer than 34 attached ticks after exposure to a similar tick population as was used in the study. In Chapters 5 and 6, we will introduce the process by which we can make a statement of our certainty that the new repellant is a significant improvement over the old repellant.

Because of the arbitrariness in the choice of number of intervals, starting value, and length of intervals, histograms can be made to take on different shapes for the same set of data, especially for small data sets. Histograms are most useful for describing data sets when the number of data points is fairly large, say 50 or more. In Figures 3.7(a)–(d), a set of histograms for the tick data constructed using 5, 9, 13, and 18 class intervals illustrates the problems that can be encountered in attempting to construct a histogram. These graphs were obtained using the Minitab software program.

FIGURE 3.7(a)
Relative frequency histogram for tick data (5 intervals)

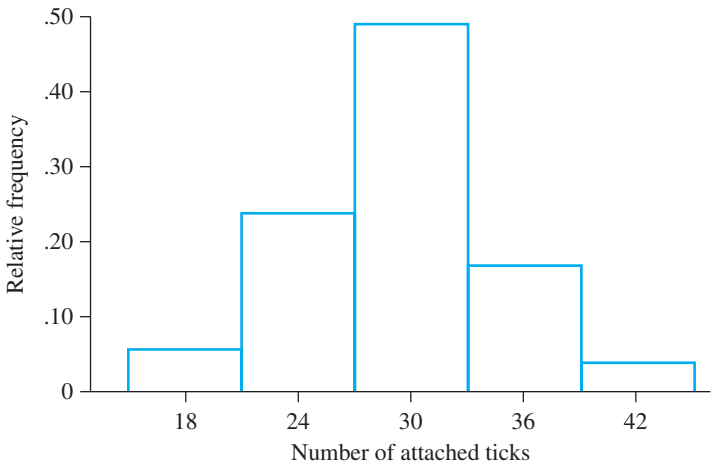


FIGURE 3.7(b)
Relative frequency histogram for tick data (9 intervals)

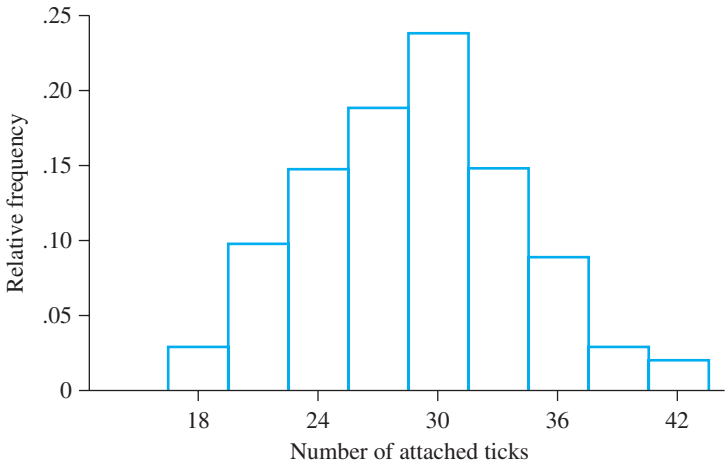


FIGURE 3.7(c)
Relative frequency histogram for tick data (13 intervals)

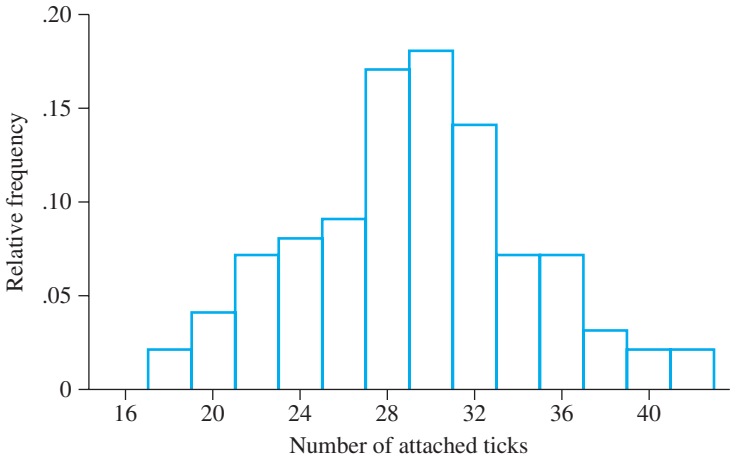
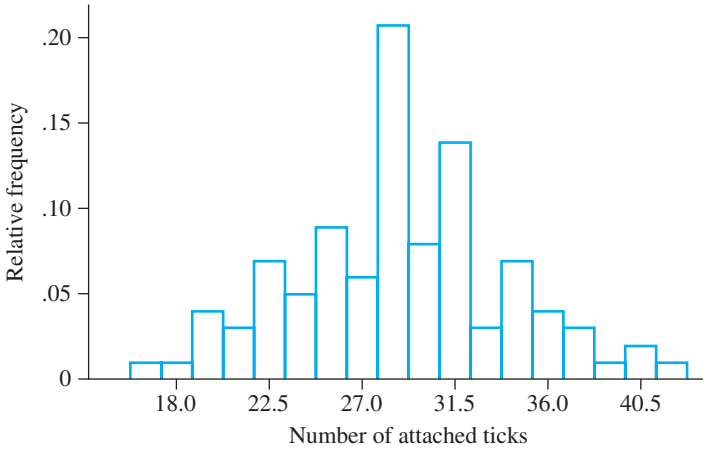


FIGURE 3.7(d)
Relative frequency histogram for tick data (18 intervals)



When the number of data points is relatively small and the number of intervals is large, the histogram fluctuates too much—that is, responds to a very few data values; see Figure 3.7(d). This results in a graph that is not a realistic depiction of the histogram for the whole population. When the number of class intervals is too small, most of the patterns or trends in the data are not displayed; see Figure 3.7(a). In the set of graphs in Figure 3.7, the histogram with 13 class intervals appears to be the most appropriate graph.

Finally, because we use proportions rather than frequencies in a relative frequency histogram, we can compare two different samples (or populations) by examining their relative frequency histograms even if the samples (populations) are of different sizes. When describing relative frequency histograms and comparing the plots from a number of samples, we examine the overall shape in the histogram. Figure 3.8 depicts many of the common shapes for relative frequency histograms.

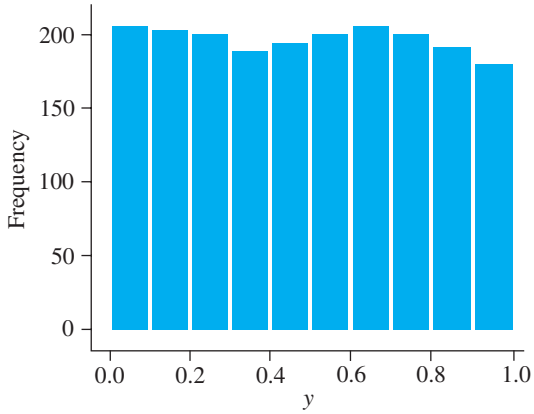
unimodal

A histogram with one major peak is called **unimodal,** see Figures 3.8(b), (c), and (d). When the histogram has two major peaks, such as in Figures 3.8(e) and (f), we state that the histogram is **bimodal.** In many instances, bimodal histograms are an indication that the sampled data are in fact from two distinct populations. Finally, when every interval has essentially the same number of observations, the histogram is called a **uniform** histogram; see Figure 3.8(a).

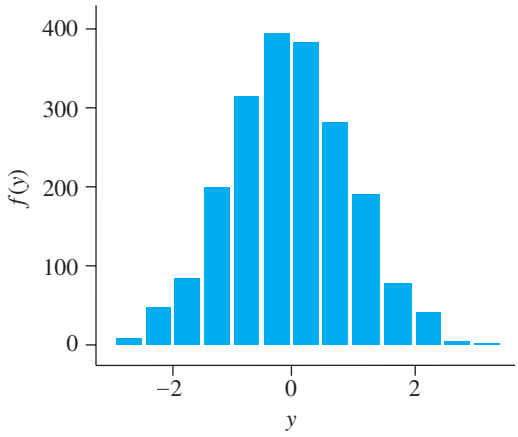
bimodal

uniform

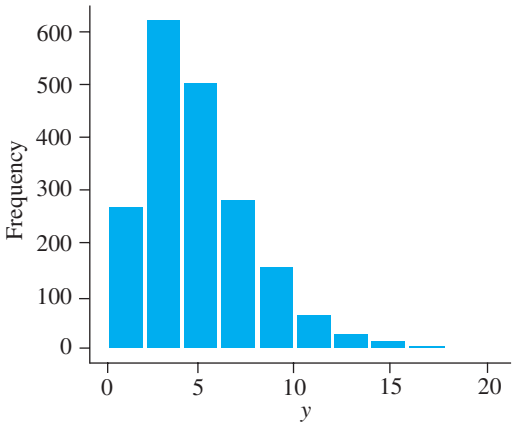
FIGURE 3.8 Some common shapes of distributions



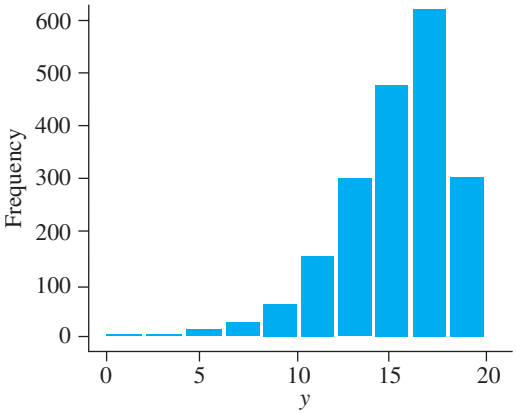
(a) Uniform distribution



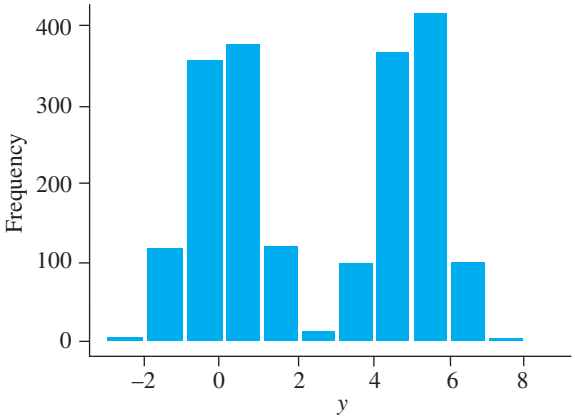
(b) Symmetric, unimodal (normal) distribution



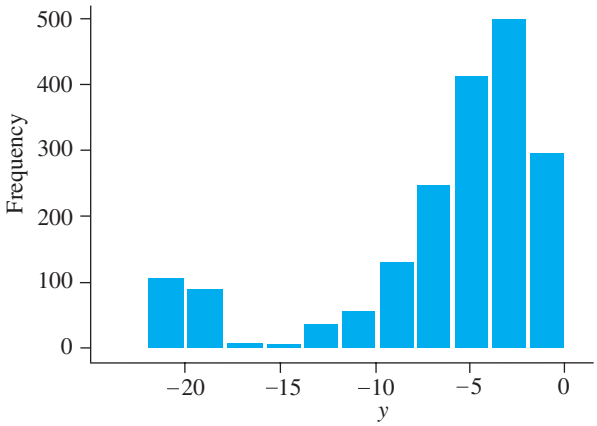
(c) Right-skewed distribution



(d) Left-skewed distribution



(e) Bimodal distribution



(f) Bimodal distribution skewed to left

symmetric

A histogram is **symmetric** in shape if the right and left sides have essentially the same shape. Thus, Figures 3.8(a), (b), and (e) have symmetric shapes. When the right side of the histogram, containing the larger half of the observations in the data, extends a greater distance than the left side, the histogram is referred to as **skewed to the right;** see Figure 3.8 (c). The histogram is **skewed to the left** when its left side extends a much larger distance than the right side; see Figure 3.8(d). We will see later in the text that knowing the shape of the distribution will help us choose the appropriate measures to summarize the data (Sections 3.4–3.7) and the methods for analyzing the data (Chapter 5 and beyond).

skewed to the right
skewed to the left

exploratory data analysis

The next graphical technique presented in this section is a display technique taken from an area of statistics called **exploratory data analysis (EDA).** Professor John Tukey (1977) has been the leading proponent of this practical philosophy of data analysis aimed at exploring and understanding data.

stem-and-leaf plot

The **stem-and-leaf plot** is a clever, simple device for constructing a histogramlike picture of a frequency distribution. It allows us to use the information contained in a frequency distribution to show the range of scores, where the scores are concentrated, the shape of the distribution, whether there are any specific values or scores not represented, and whether there are any stray or extreme scores. The stem-and-leaf plot does not follow the organization principles stated previously for histograms. We will use the data shown in Table 3.7 to illustrate how to construct a stem-and-leaf plot.

The data in Table 3.7 are the maximum ozone readings (in parts per billion (ppb)) taken on 80 summer days in a large city. The readings are either two- or three-digit numbers. We will use the first digit of the two-digit numbers and the first two digits of the three-digit numbers as the stem number (see Figure 3.9) and the remaining digits as the leaf number. For example, one of the readings was 85. Thus, 8 will be recorded as the stem number and 5 as the leaf number. A second maximum ozone reading was 111. Thus, 11 will be recorded as the stem number and 1 as the leaf number. If our data had been recorded in different units and resulted in, say, six-digit numbers such as 104,328, we might use the first two digits as stem numbers, the second digits as the leaf numbers, and ignore the last two digits. This would result in some loss of information but would produce a much more useful graph.

For the data on maximum ozone readings, the smallest reading was 60 and the largest was 169. Thus, the stem numbers will be 6, 7, 8, . . . , 15, 16. In the same way that a class interval determines where a measurement is placed in a frequency table, the leading digits (stem of a measurement) determine the row in which a measurement is placed in a stem-and-leaf graph. The trailing digits for a measurement are then written in the appropriate row. In this way, each measurement is recorded in the stem-and-leaf plot, as in Figure 3.9 for the ozone data. The stem-and-leaf plot in

TABLE 3.7
Maximum ozone readings (ppb)

60	61	61	64	64	64	64	66	66	68
68	68	69	71	71	71	71	71	71	72
72	73	75	75	80	80	80	80	80	80
82	82	83	85	86	86	87	87	87	89
91	92	94	94	98	99	99	100	101	103
103	103	108	111	113	113	114	118	119	119
122	122	124	124	124	125	125	131	133	134
136	141	142	143	146	150	152	155	169	169

FIGURE 3.9
Stem-and-leaf plot for maximum ozone readings (ppb) of Table 3.7

```
 6 0114444
 6 668889
 7 111111223
 7 55
 8 000000223
 8 5667779
 9 1244
 9 899
10 01333
10 8
11 1334
11 899
12 22444
12 55
13 134
13 6
14 123
14 6
15 02
15 5
16
16 99
```

Figure 3.9 was obtained using Minitab. Note that most of the stems are repeated twice, with leaf digits split into two groups: 0 to 4 and 5 to 9.

We can see that each stem defines a class interval and that the limits of each interval are the largest and smallest possible scores for the class. The values represented by each leaf must be between the lower and upper limits of the interval.

Note that a stem-and-leaf plot is a graph that looks much like a histogram turned sideways, as in Figure 3.9. The plot can be made a bit more useful by ordering the data (leaves) within a row (stem) from lowest to highest as we did in Figure 3.9. The advantage of such a graph over the histogram is that it reflects not only frequencies, concentration(s) of scores, and shapes of the distribution but also the actual scores. The disadvantage is that for large data sets, the stem-and-leaf plot can be more unwieldy than the histogram.

Guidelines for Constructing Stem-and-Leaf Plots

1. Split each score or value into two sets of digits. The first or leading set of digits is the stem and the second or trailing set of digits is the leaf.
2. List all possible stem digits from lowest to highest.
3. For each score in the mass of data, write the leaf values on the line labeled by the appropriate stem number.
4. If the display looks too cramped and narrow, stretch the display by using two lines per stem so that, for example, leaf digits 0, 1, 2, 3, and 4 are placed on the first line of the stem and leaf digits 5, 6, 7, 8, and 9 are placed on the second line.
5. If too many digits are present, such as in a six- or seven-digit score, drop the right-most trailing digit(s) to maximize the clarity of the display.
6. The rules for developing a stem-and-leaf plot are somewhat different from the rules governing the establishment of class intervals for the traditional frequency distribution and for a variety of other procedures that we will consider in later sections of the text. Class intervals for stem-and-leaf plots are, then, in a sense slightly atypical.

The following data display and stem and leaf plot (Figure 3.10) is obtained from Minitab. The data consist of the number of employees in the wholesale and retail trade industries in Wisconsin measured each month for a 5-year period.

```
Data Display

Trade
    322   317   319   323   327   328   325   326   330   334
    337   341   322   318   320   326   332   334   335   336
    335   338   342   348   330   325   329   337   345   350
    351   354   355   357   362   368   348   345   349   355
    362   367   366   370   371   375   380   385   361   354
    357   367   376   381   381   383   384   387   392   396
```

FIGURE 3.10
Character stem-and-leaf display for trade data

```
Stem-and-leaf of Trade     N = 60
Leaf Unit = 1.0

      31 789
      32 0223
      32 5666789
      33 00244
      33 556778
      34 12
      34 55889
      35 0144
      35 5577
      36 122
      36 6778
      37 01
      37 56
      38 01134
      38 57
      39 2
      39 6
```

Note that most of the stems are repeated twice, with the leaf digits split into two groups: 0 to 4 and 5 to 9.

The last graphical technique to be presented in this section deals with how certain variables change over time. For macroeconomic data such as disposable income and microeconomic data such as weekly sales data of one particular product at one particular store, plots of data over time are fundamental to business management. Similarly, social researchers are often interested in showing how variables change over time. They might be interested in changes with time in attitudes toward various racial and ethnic groups, changes in the rate of savings in the United States, or changes in crime rates for various cities. A pictorial method of presenting changes in a variable over time is called a **time series.** Figure 3.11 is a time series showing the number of homicides, forcible rapes, robberies, and aggravated assaults included in the Uniform Crime Reports of the FBI.

time series

Usually, time points are labeled chronologically across the horizontal axis (abscissa), and the numerical values (frequencies, percentages, rates, etc.) of the

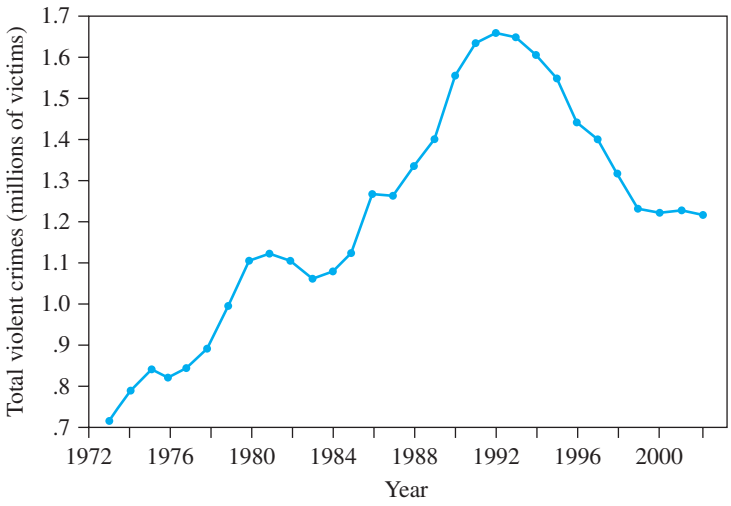


FIGURE 3.11
Total violent crimes in the United States, 1973–2002

variable of interest are labeled along the vertical axis (ordinate). Time can be measured in days, months, years, or whichever unit is most appropriate. As a rule of thumb, a time series should consist of no fewer than four or five time points; typically, these time points are equally spaced. Many more time points than this are desirable, though, in order to show a more complete picture of changes in a variable over time.

How we display the time axis in a time series frequently depends on the time intervals at which data are available. For example, the U.S. Census Bureau reports average family income in the United States only on a yearly basis. When information about a variable of interest is available in different units of time, we must decide which unit or units are most appropriate for the research. In an election year, a political scientist would most likely examine weekly or monthly changes in candidate preferences among registered voters. On the other hand, a manufacturer of machine-tool equipment might keep track of sales (in dollars and number of units) on a monthly, quarterly, and yearly basis. Figure 3.12 shows the quarterly

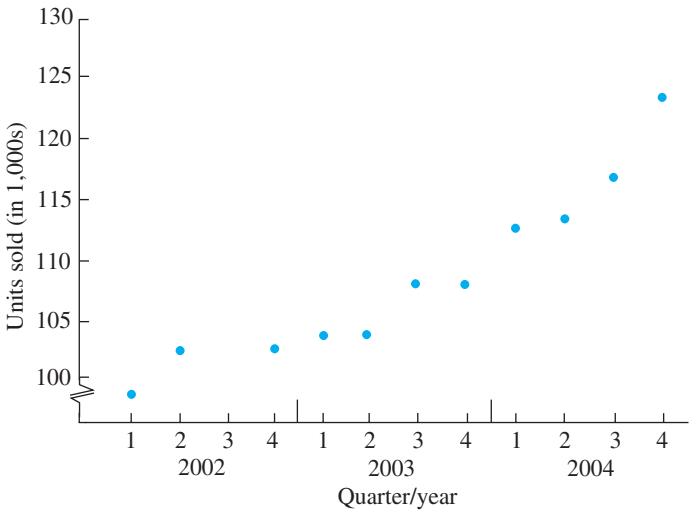


FIGURE 3.12
Quarterly sales (in thousands)

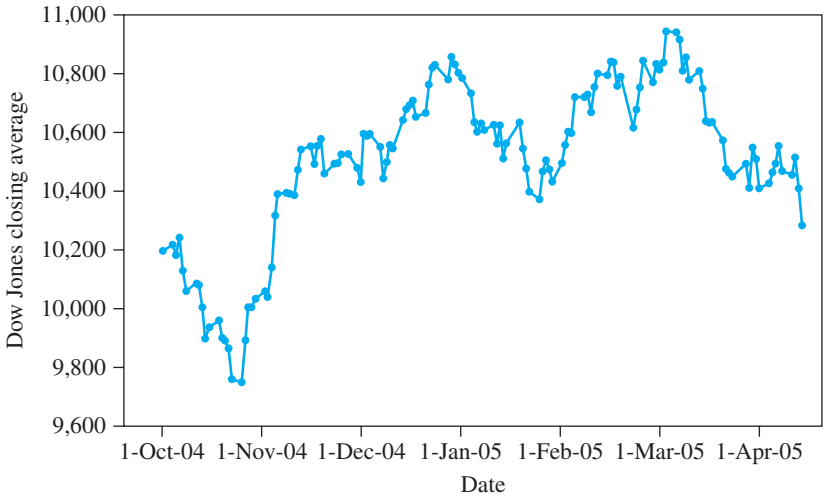


FIGURE 3.13
Time-series plot of the Dow Jones Average, October 2004 to April 2005

sales (in thousands of units) of a machine-tool product over the past 3 years. Note that from this time series it is clear that the company has experienced a gradual but steady growth in the number of units over the past 3 years.

Time-series plots are useful for examining general trends and seasonal or cyclic patterns. For example, the "Money and Investing" section of the *Wall Street Journal* gives the daily workday values for the Dow Jones Industrials Averages. The values given in the April 8, 2005, issue are displayed in Figure 3.13. Exercise 3.58 provides more details on how the Dow Industrial Average is computed. An examination of the plot reveals a somewhat increasing trend from July to November, followed by a sharp increase from November through January 8. In order to detect seasonal or cyclical patterns, it is necessary to have daily, weekly, or monthly data over a large number of years.

Sometimes it is important to compare trends over time in a variable for two or more groups. Figure 3.14 reports the values of two ratios from 1985 to 2000: the ratio of the median family income of African Americans to the median income of Anglo-Americans and the ratio of the median income of Hispanics to the median income of Anglo-Americans.

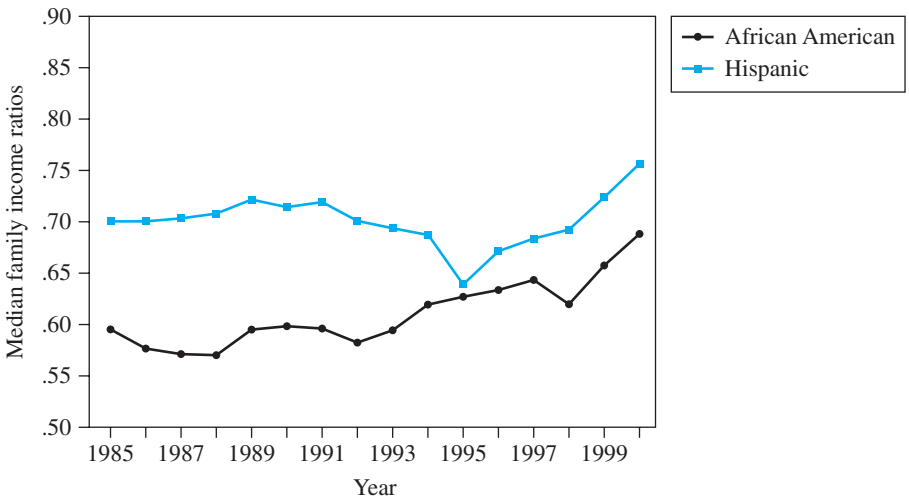


FIGURE 3.14
Ratio of African American and Hispanic median family income to Anglo-American median family income.
Source: U.S. Census Bureau

Median family income represents the income amount that divides family incomes into two groups—the top half and the bottom half. For example, in 1987, the median family income for African Americans was $18,098, meaning that 50% of all African American families had incomes above $18,098, and 50% had incomes below $18,098. The median, one of several measures of central tendency, is discussed more fully later in this chapter.

Figure 3.14 shows that the ratio of African American to Anglo-American family income and the ratio of Hispanic to Anglo-American family income remained fairly constant from 1985 to 1991. From 1995 to 2000, there was an increase in both ratios and a narrowing of the difference between the ratio for African American family income and the ratio of Hispanic family income. We can interpret this trend to mean that the income of African American and Hispanic families has generally increased relative to the income of Anglo-American families.

Sometimes information is not available in equal time intervals. For example, polling organizations such as Gallup or the National Opinion Research Center do not necessarily ask the American public the same questions about their attitudes or behavior on a yearly basis. Sometimes there is a time gap of more than 2 years before a question is asked again.

When information is not available in equal time intervals, it is important for the interval width between time points (the horizontal axis) to reflect this fact. If, for example, a social researcher is plotting values of a variable for 1995, 1996, 1997, and 2000, the interval width between 1997 and 2000 on the horizontal axis should be three times the width of that between the other years. If these interval widths were spaced evenly, the resulting trend line could be seriously misleading.

Before leaving graphical methods for describing data, there are several general guidelines that can be helpful in developing graphs with an impact. These guidelines pay attention to the design and presentation techniques and should help you make better, more informative graphs.

General Guidelines for Successful Graphics

1. Before constructing a graph, set your priorities. What messages should the viewer get?
2. Choose the type of graph (pie chart, bar graph, histogram, and so on).
3. Pay attention to the title. One of the most important aspects of a graph is its title. The title should immediately inform the viewer of the point of the graph and draw the eye toward the most important elements of the graph.
4. Fight the urge to use many type sizes, styles, and color changes. The indiscriminate and excessive use of different type sizes, styles, and colors will confuse the viewer. Generally, we recommend using only two typefaces; color changes and italics should be used in only one or two places.
5. Convey the tone of your graph by using colors and patterns. Intense, warm colors (yellows, oranges, reds) are more dramatic than the blues and purples and help to stimulate enthusiasm by the viewer. On the other hand, pastels (particularly grays) convey a conservative, businesslike tone. Similarly, simple patterns convey a conservative tone, whereas busier patterns stimulate more excitement.
6. Don't underestimate the effectiveness of a simple, straightforward graph.

3.4 Describing Data on a Single Variable: Measures of Central Tendency

Numerical descriptive measures are commonly used to convey a mental image of pictures, objects, and other phenomena. There are two main reasons for this. First, graphical descriptive measures are inappropriate for statistical inference, because it is difficult to describe the similarity of a sample frequency histogram and the corresponding population frequency histogram. The second reason for using numerical descriptive measures is one of expediency—we never seem to carry the appropriate graphs or histograms with us, and so must resort to our powers of verbal communication to convey the appropriate picture. We seek several numbers, called *numerical descriptive measures,* that will create a mental picture of the frequency distribution for a set of measurements.

central tendency
variability

The two most common numerical descriptive measures are measures of **central tendency** and measures of **variability;** that is, we seek to describe the center of the distribution of measurements and also how the measurements vary about the center of the distribution. We will draw a distinction between numerical descriptive measures for a population, called **parameters,** and numerical descriptive measures for a sample, called **statistics.** In problems requiring statistical inference, we will not be able to calculate values for various parameters, but we will be able to compute corresponding statistics from the sample and use these quantities to estimate the corresponding population parameters.

parameters
statistics

In this section, we will consider various measures of central tendency, followed in Section 3.5 by a discussion of measures of variability.

mode

The first measure of central tendency we consider is the **mode.**

DEFINITION 3.1

The **mode** of a set of measurements is defined to be the measurement that occurs most often (with the highest frequency).

We illustrate the use and determination of the mode in an example.

EXAMPLE 3.1

A consumer investigator is interested in the differences in the selling prices of a new popular compact automobile at various dealers in a 100 mile radius of Houston, Texas. She asks for a quote from 25 dealers for this car with exactly the same options. The selling prices (in $1,000) are given here.

26.6	25.3	23.8	24.0	27.5
21.1	25.9	22.6	23.8	25.1
22.6	27.5	26.8	23.4	27.5
20.8	20.4	22.4	27.5	23.7
22.2	23.8	23.2	28.7	27.5

Determine the modal selling price.

Solution For these data, the price 23.8 occurred three times in the sample but the price 27.5 occurred five times. Because no other value occurred more than once, we would state the data had a modal selling price of $27,500.

Identification of the mode for Example 3.1 was quite easy because we were able to count the number of times each measurement occurred. When dealing with grouped data—data presented in the form of a frequency table—we can define the modal interval to be the class interval with the highest frequency. However, because we would not know the actual measurements but only how many measurements fall into each interval, the mode is taken as the midpoint of the modal interval; it is an approximation to the mode of the actual sample measurements.

The mode is also commonly used as a measure of popularity that reflects central tendency or opinion. For example, we might talk about the most preferred stock, a most preferred model of washing machine, or the most popular candidate. In each case, we would be referring to the mode of the distribution. In Figure 3.8 of the previous section, frequency histograms (b), (c), and (d) had a single mode with the mode located at the center of the class having the highest frequency. Thus, the modes would be $-.25$ for histogram (b), 3 for histogram (c), and 17 for histogram (d). It should be noted that some distributions have more than one measurement that occurs with the highest frequency. Thus, we might encounter bimodal, trimodal, and so on, distributions. In Figure 3.8, histogram (e) is essentially bimodal, with nearly equal peaks at $y = 0.5$ and $y = 5.5$.

median

The second measure of central tendency we consider is the **median.**

DEFINITION 3.2

The **median** of a set of measurements is defined to be the middle value when the measurements are arranged from lowest to highest.

The median is most often used to measure the midpoint of a large set of measurements. For example, we may read about the median wage increase won by union members, the median age of persons receiving Social Security benefits, and the median weight of cattle prior to slaughter during a given month. Each of these situations involves a large set of measurements, and the median would reflect the central value of the data—that is, the value that divides the set of measurements into two groups, with an equal number of measurements in each group.

However, we may use the definition of median for small sets of measurements by using the following convention: The median for an even number of measurements is the average of the two middle values when the measurements are arranged from lowest to highest. When there are an odd number of measurements, the median is still the middle value. Thus, whether there are an even or odd number of measurements, there are an equal number of measurements above and below the median.

EXAMPLE 3.2

After the third-grade classes in a school district received low overall scores on a statewide reading test, a supplemental reading program was implemented in order to provide extra help to those students who were below expectations with respect to their reading proficiency. Six months after implementing the program, the 10 third-grade classes in the district were reexamined. For each of the 10 schools, the percentage of students reading above the statewide standard was determined. These data are shown here.

95 86 78 90 62 73 89 92 84 76

Determine the median percentage of the 10 schools.

Solution First we must arrange the percentage in order of magnitude.

62 73 76 78 84 86 89 90 92 95

Because there are an even number of measurements, the median is the average of the two midpoint scores.

$$\text{median} = \frac{84 + 86}{2} = 85$$

EXAMPLE 3.3

An experiment was conducted to measure the effectiveness of a new procedure for pruning grapes. Each of 13 workers was assigned the task of pruning an acre of grapes. The productivity, measured in worker-hours/acre, is recorded for each person.

4.4 4.9 4.2 4.4 4.8 4.9 4.8 4.5 4.3 4.8 4.7 4.4 4.2

Determine the mode and median productivity for the group.

Solution First arrange the measurements in order of magnitude:

4.2 4.2 4.3 4.4 4.4 4.4 4.5 4.7 4.8 4.8 4.8 4.9 4.9

For these data, we have two measurements appearing three times each. Hence, the data are bimodal, with modes of 4.4 and 4.8. The median for the odd number of measurements is the middle score, 4.5.

grouped data median

The **median for grouped data** is slightly more difficult to compute. Because the actual values of the measurements are unknown, we know that the median occurs in a particular class interval, but we do not know where to locate the median within the interval. If we assume that the measurements are spread evenly throughout the interval, we get the following result. Let

L = lower class limit of the interval that contains the median

n = total frequency

cf_b = the sum of frequencies (cumulative frequency) for all classes before the median class

f_m = frequency of the class interval containing the median

w = interval width

Then, for grouped data,

$$\text{median} = L + \frac{w}{f_m}(.5n - cf_b)$$

The next example illustrates how to find the median for grouped data.

EXAMPLE 3.4

Table 3.8 is a repeat of the frequency table (Table 3.6) with some additional columns for the tick data of Table 3.5. Compute the median number of ticks per cow for these data.

TABLE 3.8
Frequency table for number of attached ticks, Table 3.5

Class	Class Interval	f_i	Cumulative f_i	f_i/n	Cumulative f_i/n
1	16.25–18.75	2	2	.02	.02
2	18.75–21.25	7	9	.07	.09
3	21.25–23.75	7	16	.07	.16
4	23.75–26.25	14	30	.14	.30
5	26.25–28.75	17	47	.17	.47
6	28.75–31.25	24	71	.24	.71
7	31.25–33.75	11	82	.11	.82
8	33.75–36.25	11	93	.11	.93
9	36.25–38.75	3	96	.03	.96
10	38.75–41.25	3	99	.03	.99
11	41.25–43.75	1	100	.01	1.00

Solution Let the cumulative relative frequency for class j equal the sum of the relative frequencies for class 1 through class j. To determine the interval that contains the median, we must find the first interval for which the cumulative relative frequency exceeds .50. This interval is the one containing the median. For these data, the interval from 28.75 to 31.25 is the first interval for which the cumulative relative frequency exceeds .50, as shown in Table 3.8, Class 6. So this interval contains the median. Then

$$L = 28.75 \quad f_m = 24$$
$$n = 100 \quad w = 2.5$$
$$cf_b = 47$$

and

$$\text{median} = L + \frac{w}{f_m}(.5n - cf_b) = 28.75 + \frac{2.5}{24}(50 - 47) = 29.06$$

Note that the value of the median from the ungrouped data of Table 3.5 is 29. Thus, the approximated value and the value from the ungrouped data are nearly equal. The difference between the two values for the sample median decreases as the number of class intervals increases.

The third, and last, measure of central tendency we will discuss in this text is the arithmetic mean, known simply as the **mean.**

mean

DEFINITION 3.3

The **arithmetic mean,** or **mean,** of a set of measurements is defined to be the sum of the measurements divided by the total number of measurements.

When people talk about an "average," they quite often are referring to the mean. It is the balancing point of the data set. Because of the important role that the mean will play in statistical inference in later chapters, we give special symbols to the population mean and the sample mean. The *population mean* is denoted by the Greek letter μ (read "mu"), and the *sample mean* is denoted by the symbol $\bar{y}$ (read "y-bar"). As indicated in Chapter 1, a population of measurements is the complete set of

μ
$\bar{y}$

measurements of interest to us; a sample of measurements is a subset of measurements selected from the population of interest. If we let $y_1, y_2, \ldots, y_n$ denote the measurements observed in a sample of size n, then the sample mean $\bar{y}$ can be written as

$$\bar{y} = \frac{\Sigma_i y_i}{n}$$

where the symbol appearing in the numerator, $\Sigma_i y_i$, is the notation used to designate a sum of n measurements, y_i:

$$\sum_i y_i = y_1 + y_2 + \cdots + y_n$$

The corresponding population mean is μ.

In most situations, we will not know the population mean; the sample will be used to make inferences about the corresponding unknown population mean. For example, the accounting department of a large department store chain is conducting an examination of its overdue accounts. The store has thousands of such accounts, which would yield a population of overdue values having a mean value, μ. The value of μ could only be determined by conducting a large-scale audit that would take several days to complete. The accounting department monitors the overdue accounts on a daily basis by taking a random sample of n overdue accounts and computing the sample mean, $\bar{y}$. The sample mean, $\bar{y}$, is then used as an estimate of the mean value, μ, in *all* overdue accounts for that day. The accuracy of the estimate and approaches for determining the appropriate sample size will be discussed in Chapter 5.

EXAMPLE 3.5

A sample of $n = 15$ overdue accounts in a large department store yields the following amounts due:

$55.20	$ 4.88	$271.95
18.06	180.29	365.29
28.16	399.11	807.80
44.14	97.47	9.98
61.61	56.89	82.73

a. Determine the mean amount due for the 15 accounts sampled.
b. If there are a total of 150 overdue accounts, use the sample mean to predict the total amount overdue for all 150 accounts.

Solution

a. The sample mean is computed as follows:

$$\bar{y} = \frac{\Sigma_i y_i}{15} = \frac{55.20 + 18.06 + \cdots + 82.73}{15} = \frac{2{,}483.56}{15} = \$165.57$$

b. From part (a), we found that the 15 accounts sampled averaged $165.57 overdue. Using this information, we would predict, or estimate, the total amount overdue for the 150 accounts to be 150(165.57) = $24,835.50.

The sample mean formula for grouped data is only slightly more complicated than the formula just presented for ungrouped data. In certain situations, the original data will be presented in a frequency table or a histogram. Thus, we will not know the individual sample measurements, only the interval to which a measurement is assigned. In this type of situation, the formula for the mean from the grouped data will be an approximation to the actual sample mean. Hence, when

the sample measurements are known, the formula for ungrouped data should be used. If there are k class intervals and

y_i = midpoint of the ith class interval

f_i = frequency associated with the ith class interval

n = the total number of measurements

then

$$\bar{y} \cong \frac{\Sigma_i f_i y_i}{n},$$

where $\cong$ denotes "is approximately equal to."

EXAMPLE 3.6

The data of Example 3.4 are reproduced in Table 3.9, along with three additional columns: y_i, $f_i y_i$, $f_i(y_i - \bar{y})^2$. These values will be needed in order to compute approximations to the sample mean and the sample standard deviation. Using the information in Table 3.9, compute an approximation to the sample mean for this set of grouped data.

TABLE 3.9
Class information for number of attached ticks

Class	Class Interval	f_i	y_i	$f_i y_i$	$f_i(y_i - \bar{y})^2$
1	16.25–18.75	2	17.5	35.0	258.781
2	18.75–21.25	7	20.0	140.0	551.359
3	21.25–23.75	7	22.5	157.5	284.484
4	23.75–26.25	14	25.0	350.0	210.219
5	26.25–28.75	17	27.5	467.5	32.141
6	28.75–31.25	24	30.0	720.0	30.375
7	31.25–33.75	11	32.5	357.5	144.547
8	33.75–36.25	11	35.0	385.0	412.672
9	36.25–38.75	3	37.5	112.5	223.172
10	38.75–41.25	3	40.0	120.0	371.297
11	41.25–43.75	1	42.5	42.5	185.641
	Totals	100		2,887.5	2,704.688

Solution After adding the entries in the $f_i y_i$ column and substituting into the formula, we determine that an approximation to the sample mean is

$$\bar{y} \cong \frac{\Sigma_{i=1}^{11} f_i y_i}{100} = \frac{2{,}887.5}{100} = 28.875$$

Using the 100 values, y_i's, from Table 3.5, the actual value of the sample mean is

$$\bar{y} = \frac{\Sigma_{i=1}^{100} y_i}{100} = \frac{2{,}881}{100} = 28.81$$

which demonstrates that the approximation from the grouped data formula can be very close to the actual value. When the number of class intervals is relatively large, the approximation from the grouped data formula will be very close to the actual sample mean.

The mean is a useful measure of the central value of a set of measurements, but it is subject to distortion due to the presence of one or more extreme values in the set. In these situations, the extreme values (called **outliers**) pull the mean in the direction of the outliers to find the balancing point, thus distorting the mean as a measure of the central value. A variation of the mean, called a **trimmed mean,**

outliers

trimmed mean

drops the highest and lowest extreme values and averages the rest. For example, a 5% trimmed mean drops the highest 5% and the lowest 5% of the measurements and averages the rest. Similarly, a 10% trimmed mean drops the highest and the lowest 10% of the measurements and averages the rest. In Example 3.5, a 10% trimmed mean would drop the smallest and largest account, resulting in a mean of

$$\bar{y} = \frac{2{,}483.56 - 4.88 - 807.8}{13} = \$128.53$$

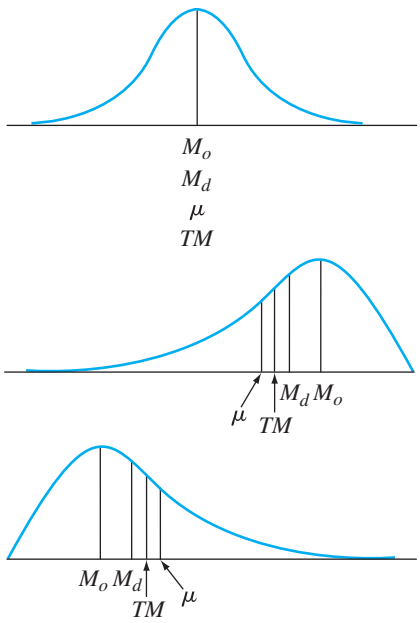
By trimming the data, we are able to reduce the impact of very large (or small) values on the mean, and thus get a more reliable measure of the central value of the set. This will be particularly important when the sample mean is used to predict the corresponding population central value.

Note that in a limiting sense the median is a 50% trimmed mean. Thus, the median is often used in place of the mean when there are extreme values in the data set. In Example 3.5, the value \$807.80 is considerably larger than the other values in the data set. This results in 10 of the 15 accounts having values less than the mean and only 5 having values larger than the mean. The median value for the 15 accounts is \$61.61. There are 7 accounts less than the median and 7 accounts greater than the median. Thus, in selecting a typical overdue account, the median is a more appropriate value than the mean. However, if we want to estimate the total amount overdue in all 150 accounts, we would want to use the mean and not the median. When estimating the sum of all measurements in a population, we would not want to exclude the extremes in the sample. Suppose a sample contains a few extremely large values. If the extremes are trimmed, then the population sum will be grossly underestimated using the sample trimmed mean or sample median in place of the sample mean.

In this section, we discussed the mode, median, mean, and trimmed mean. How are these measures of central tendency related for a given set of measurements? The answer depends on the **skewness** of the data. If the distribution is mound-shaped and symmetrical about a single peak, the mode (M_o), median (M_d), mean (μ), and trimmed mean (TM) will all be the same. This is shown using a smooth curve and population quantities in Figure 3.15(a). If the distribution is skewed, having a long tail in

skewness

FIGURE 3.15
Relation among the mean μ, the trimmed mean TM, the median M_d, and the mode M_o



(a) A mound-shaped distribution

(b) A distribution skewed to the left

(c) A distribution skewed to the right

one direction and a single peak, the mean is pulled in the direction of the tail; the median falls between the mode and the mean; and depending on the degree of trimming, the trimmed mean usually falls between the median and the mean. Figures 3.15(b) and (c) illustrate this for distributions skewed to the left and to the right.

The important thing to remember is that we are not restricted to using only one measure of central tendency. For some data sets, it will be necessary to use more than one of these measures to provide an accurate descriptive summary of central tendency for the data.

Major Characteristics of Each Measure of Central Tendency

Mode

1. It is the most frequent or probable measurement in the data set.
2. There can be more than one mode for a data set.
3. It is not influenced by extreme measurements.
4. Modes of subsets cannot be combined to determine the mode of the complete data set.
5. For grouped data its value can change depending on the categories used.
6. It is applicable for both qualitative and quantitative data.

Median

1. It is the central value; 50% of the measurements lie above it and 50% fall below it.
2. There is only one median for a data set.
3. It is not influenced by extreme measurements.
4. Medians of subsets cannot be combined to determine the median of the complete data set.
5. For grouped data, its value is rather stable even when the data are organized into different categories.
6. It is applicable to quantitative data only.

Mean

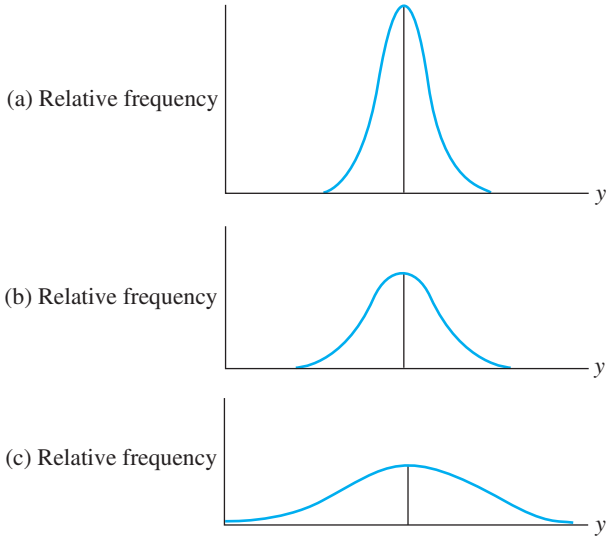
1. It is the arithmetic average of the measurements in a data set.
2. There is only one mean for a data set.
3. Its value is influenced by extreme measurements; trimming can help to reduce the degree of influence.
4. Means of subsets can be combined to determine the mean of the complete data set.
5. It is applicable to quantitative data only.

Measures of central tendency do not provide a complete mental picture of the frequency distribution for a set of measurements. In addition to determining the center of the distribution, we must have some measure of the spread of the data. In the next section, we discuss measures of variability, or dispersion.

3.5 Describing Data on a Single Variable: Measures of Variability

It is not sufficient to describe a data set using only measures of central tendency, such as the mean or the median. For example, suppose we are monitoring the production of plastic sheets that have a nominal thickness of 3 mm. If we randomly

FIGURE 3.16
Relative frequency histograms with different variabilities but the same mean



select 100 sheets from the daily output of the plant and find that the average thickness of the 100 sheets is 3 mm, does this indicate that all 100 sheets have the desired thickness of 3 mm? We may have a situation in which 50 sheets have a thickness of 1 mm and the remaining 50 sheets have a thickness of 5 mm. This would result in an average thickness of 3 mm, but none of the 100 sheets would have a thickness close to the specified 3 mm. Thus, we need to determine how dispersed are the sheet thicknesses about the mean of 3 mm.

Graphically, we can observe the need for some measure of variability by examining the relative frequency histograms of Figure 3.16. All the histograms have the same mean but each has a different spread, or **variability,** about the mean. For illustration, we have shown the histograms as smooth curves. Suppose the three histograms represent the amount of PCB (ppb) found in a large number of 1-liter samples taken from three lakes that are close to chemical plants. The average amount of PCB, μ, in a 1-liter sample is the same for all three lakes. However, the variability in the PCB quantity is considerably different. Thus, the lake with PCB quantity depicted in histogram (a) would have fewer samples containing very small or large quantities of PCB as compared to the lake with PCB values depicted in histogram (c). Knowing only the mean PCB quantity in the three lakes would mislead the investigator concerning the level of PCB present in all three lakes.

variability

The simplest but least useful measure of data variation is the **range,** which we alluded to in Section 3.2. We now present its definition.

range

DEFINITION 3.4

The **range** of a set of measurements is defined to be the difference between the largest and the smallest measurements of the set.

EXAMPLE 3.7

Determine the range of the 15 overdue accounts of Example 3.5.

Solution The smallest measurement is \$4.88 and the largest is \$807.80. Hence, the range is

$$807.80 - 4.88 = \$802.92$$

grouped data

For **grouped data,** because we do not know the individual measurements, the **range** is taken to be the difference between the upper limit of the last interval and the lower limit of the first interval.

Although the range is easy to compute, it is sensitive to outliers because it depends on the most extreme values. It does not give much information about the pattern of variability. Referring to the situation described in Example 3.5, if in the current budget period the 15 overdue accounts consisted of 10 accounts having a value of \$4.88, 3 accounts of \$807.80, and 1 account of \$11.36, then the mean value would be \$165.57 and the range would be \$802.92. The mean and range would be identical to the mean and range calculated for the data of Example 3.5. However, the data in the current budget period are more spread out about the mean than the data in the earlier budget period. What we seek is a measure of variability that discriminates between data sets having different degrees of concentration of the data about the mean.

percentiles

A second measure of variability involves the use of **percentiles.**

DEFINITION 3.5

The ***p*th percentile** of a set of n measurements arranged in order of magnitude is that value that has at most p% of the measurements below it and at most $(100 - p)$% above it.

For example, Figure 3.17 illustrates the 60th percentile of a set of measurements. Percentiles are frequently used to describe the results of achievement test scores and the ranking of a person in comparison to the rest of the people taking an examination. Specific percentiles of interest are the 25th, 50th, and 75th percentiles, often called the *lower quartile,* the *middle quartile* (median), and the *upper quartile,* respectively (see Figure 3.18).

The computation of percentiles is accomplished as follows: Each data value corresponds to a percentile for the percentage of the data values that are less than or equal to it. Let $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ denote the ordered observations for a data set; that is,

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$$

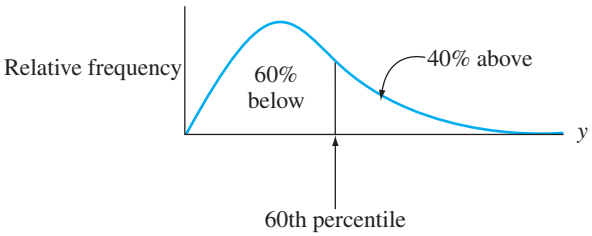


FIGURE 3.17
The 60th percentile of a set of measurements

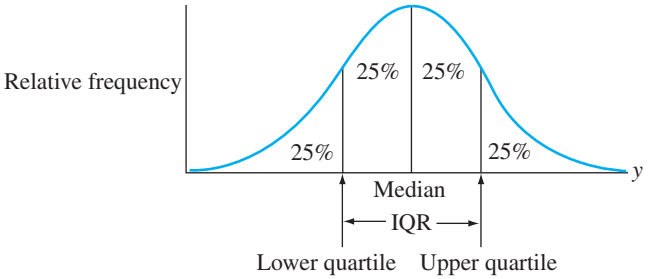


FIGURE 3.18
Quartiles of a distribution

The ith ordered observation, $y_{(i)}$, corresponds to the $100(i - .5)/n$ percentile. We use this formula in place of assigning the percentile $100i/n$ so that we avoid assigning the 100th percentile to $y_{(n)}$, which would imply that the largest possible data value in the population was observed in the data set, an unlikely happening. For example, a study of serum total cholesterol (mg/l) levels recorded the levels given in Table 3.10 for 20 adult patients. Thus, each ordered observation is a data percentile corresponding to a multiple of the fraction $100(i - .5)/n = 100(2i - 1)/2n = 100(2i - 1)/40$.

TABLE 3.10
Serum cholesterol levels

Observation (j)	Cholesterol (mg/l)	Percentile
1	133	2.5
2	137	7.5
3	148	12.5
4	149	17.5
5	152	22.5
6	167	27.5
7	174	32.5
8	179	37.5
9	189	42.5
10	192	47.5
11	201	52.5
12	209	57.5
13	210	62.5
14	211	67.5
15	218	72.5
16	238	77.5
17	245	82.5
18	248	87.5
19	253	92.5
20	257	97.5

The 22.5th percentile is 152 (mg/l). Thus, 22.5% of persons in the study have a serum cholesterol less than or equal to 152. Also, the median of the above data set, which is the 50th percentile, is halfway between 192 and 201; that is, median = $(192 + 201)/2 = 196.5$. Thus, approximately half of the persons in the study have a serum cholesterol level less than 196.5 and half have a level greater than 196.5.

When dealing with large data sets, the percentiles are generalized to quantiles, where a quantile, denoted $Q(u)$, is a number that divides a sample of n data values into two groups so that the specified fraction u of the data values is less than or equal to the value of the quantile, $Q(u)$. Plots of the quantiles $Q(u)$ versus the data fraction u provide a method of obtaining estimated quantiles for the population from which the data were selected. We can obtain a quantile plot using the following steps:

1. Place a scale on the horizontal axis of a graph covering the interval $(0, 1)$.
2. Place a scale on the vertical axis covering the range of the observed data, y_1 to y_n.
3. Plot $y_{(i)}$ versus $u_i = (i - .5)/n = (2i - 1)/2n$, for $i = 1, \ldots, n$.

Using the Minitab software, we obtain the plot shown in Figure 3.19 for the cholesterol data. Note that, with Minitab, the vertical axis is labeled $Q(u)$ rather than $y_{(i)}$. We plot $y_{(i)}$ versus u to obtain a quantile plot. Specific quantiles can be read from the plot.

We can obtain the quantile, $Q(u)$, for any value of u as follows. First, place a smooth curve through the plotted points in the quantile plot and then read the value off the graph corresponding to the desired value of u.

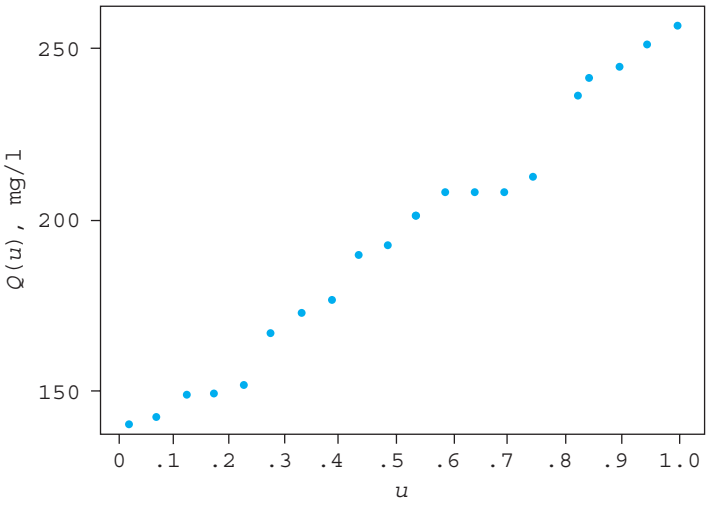


FIGURE 3.19
Quantile plot of cholesterol data

To illustrate the calculations, suppose we want to determine the 80th percentile for the cholesterol data—that is, the cholesterol level such that 80% of the persons in the population have a cholesterol level less than this value, $Q(.80)$.

Referring to Figure 3.19, locate the point $u = .8$ on the horizontal axis and draw a perpendicular line up to the quantile plot and then a horizontal line over to the vertical axis. The point where this line touches the vertical axis is our estimate of the 80th quantile. (See Figure 3.20.) Roughly 80% of the population have a cholesterol level less than 243.

When the data are grouped, the following formula can be used to approximate the percentiles for the original data. Let

P = percentile of interest
L = lower limit of the class interval that includes percentile of interest
n = total frequency
cf_b = cumulative frequency for all class intervals before the percentile class
f_p = frequency of the class interval that includes the percentile of interest
w = interval width

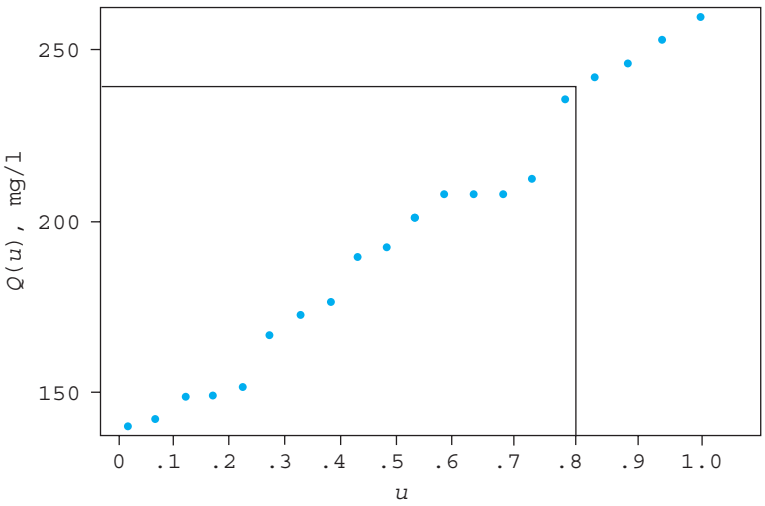


FIGURE 3.20
80th quantile of cholesterol data

Then, for example, the 65th percentile for a set of grouped data would be computed using the formula

$$P = L + \frac{w}{f_p}(.65n - cf_b)$$

To determine L, f_p, and cf_b, begin with the lowest interval and find the first interval for which the cumulative relative frequency exceeds .65. This interval would contain the 65th percentile.

EXAMPLE 3.8

Refer to the tick data of Table 3.8. Compute the 90th percentile.

Solution Because the eighth interval is the first interval for which the cumulative relative frequency exceeds .90, we have

$$L = 33.75$$
$$n = 100$$
$$cf_b = 82$$
$$f_{90} = 11$$
$$w = 2.5$$

Thus, the 90th percentile is

$$P_{90} = 33.75 + \frac{2.5}{11}[.9(100) - 82] = 35.57$$

This means that 90% of the cows have 35 or fewer attached ticks and 10% of the cows have 36 or more attached ticks.

interquartile range

The second measure of variability, the **interquartile range,** is now defined. A slightly different definition of the interquartile range is given along with the boxplot (Section 3.6).

DEFINITION 3.6

The **interquartile range (IQR)** of a set of measurements is defined to be the difference between the upper and lower quartiles; that is,

$$\text{IQR} = \text{75th percentile} - \text{25th percentile}$$

The interquartile range, although more sensitive to data pileup about the midpoint than the range, is still not sufficient for our purposes. In fact, the IQR can be very misleading when the data set is highly concentrated about the median. For example, suppose we have a sample consisting of 10 data values:

20, 50, 50, 50, 50, 50, 50, 50, 50, 80

The mean, median, lower quartile, and upper quartile would all equal 50. Thus, IQR equals $50 - 50 = 0$. This is very misleading because a measure of variability equal to 0 should indicate that the data consist of *n identical* values, which is not the case in our example. The IQR ignores the extremes in the data set completely. In fact, the IQR only measures the distance needed to cover the middle 50% of the data values and, hence, totally ignores the spread in the lower and upper 25% of the data. In summary, the IQR does not provide a lot of useful information about the variability of a single set of measurements, but it can be quite useful when

comparing the variabilities of two or more data sets. This is especially true when the data sets have some skewness. The IQR will be discussed further as part of the boxplot (Section 3.6).

In most data sets, we would typically need a minimum of five summary values to provide a minimal description of the data set: smallest value, $y_{(1)}$, lower quartile, $Q(.25)$, median, upper quartile, $Q(.75)$, and the largest value, $y_{(n)}$. When the data set has a unimodal, bell-shaped, and symmetric relative frequency histogram, just the sample mean and a measure of variability, the sample variance, can represent the data set. We will now develop the sample variance.

We seek now a sensitive measure of variability, not only for comparing the variabilities of two sets of measurements but also for interpreting the variability of a single set of measurements. To do this, we work with the **deviation** $y - \bar{y}$ of a measurement y from the mean $\bar{y}$ of the set of measurements.

deviation

To illustrate, suppose we have five sample measurements $y_1 = 68$, $y_2 = 67$, $y_3 = 66$, $y_4 = 63$, and $y_5 = 61$, which represent the percentages of registered voters in five cities who exercised their right to vote at least once during the past year. These measurements are shown in the dot diagram of Figure 3.21. Each measurement is located by a dot above the horizontal axis of the diagram. We use the sample mean

$$\bar{y} = \frac{\Sigma_i y_i}{n} = \frac{325}{5} = 65$$

to locate the center of the set and we construct horizontal lines in Figure 3.21 to represent the deviations of the sample measurements from their mean. The deviations of the measurements are computed by using the formula $y - \bar{y}$. The five measurements and their deviations are shown in Figure 3.21.

A data set with very little variability would have most of the measurements located near the center of the distribution. Deviations from the mean for a more variable set of measurements would be relatively large.

Many different measures of variability can be constructed by using the deviations $y - \bar{y}$. A first thought is to use the mean deviation, but this will always equal zero, as it does for our example. A second possibility is to ignore the minus signs and compute the average of the absolute values. However, a more easily interpreted function of the deviations involves the sum of the squared deviations of the measurements from their mean. This measure is called the **variance.**

variance

DEFINITION 3.7

The **variance** of a set of n measurements $y_1, y_2, \ldots, y_n$ with mean $\bar{y}$ is the sum of the squared deviations divided by $n - 1$:

$$\frac{\sum_i (y_i - \bar{y})^2}{n - 1}$$

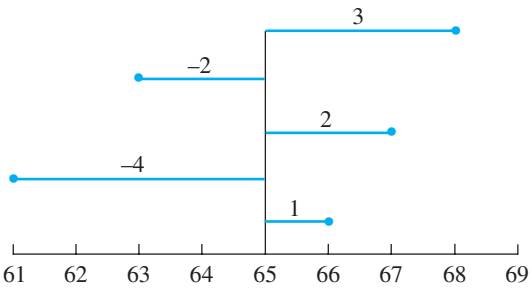


FIGURE 3.21
Dot diagram of the percentages of registered voters in five cities

s^2
σ^2

As with the sample and population means, we have special symbols to denote the sample and population variances. The symbol $\boldsymbol{s^2}$ represents the sample variance, and the corresponding population variance is denoted by the symbol $\boldsymbol{\sigma^2}$.

The definition for the variance of a set of measurements depends on whether the data are regarded as a sample or population of measurements. The definition we have given here assumes we are working with the sample, because the population measurements usually are not available. Many statisticians define the sample variance to be the average of the squared deviations, $\Sigma (y - \bar{y})^2/n$. However, the use of $(n - 1)$ as the denominator of s^2 is not arbitrary. This definition of the sample variance makes it an *unbiased estimator* of the population variance σ^2. This means roughly that if we were to draw a very large number of samples, each of size n, from the population of interest and if we computed s^2 for each sample, the average sample variance would equal the population variance σ^2. Had we divided by n in the definition of the sample variance s^2, the average sample variance computed from a large number of samples would be less than the population variance; hence, s^2 would tend to underestimate σ^2.

standard deviation

Another useful measure of variability, the **standard deviation,** involves the square root of the variance. One reason for defining the standard deviation is that it yields a measure of variability having the same units of measurement as the original data, whereas the units for variance are the square of the measurement units.

DEFINITION 3.8

The **standard deviation** of a set of measurements is defined to be the positive square root of the variance.

s
σ

We then have $\boldsymbol{s}$ denoting the sample standard deviation and $\boldsymbol{\sigma}$ denoting the corresponding population standard deviation.

EXAMPLE 3.9

The time between an electric light stimulus and a bar press to avoid a shock was noted for each of five conditioned rats. Use the given data to compute the sample variance and standard deviation.

Shock avoidance times (seconds): 5, 4, 3, 1, 3

Solution The deviations and the squared deviations are shown in Table 3.11. The sample mean $\bar{y}$ is 3.2.

TABLE 3.11
Shock avoidance data

	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
	5	1.8	3.24
	4	.8	.64
	3	−.2	.04
	1	−2.2	4.84
	3	−.2	.04
Totals	16	0	8.80

Using the total of the squared deviations column, we find the sample variance to be

$$s^2 = \frac{\Sigma_i (y_i - \bar{y})^2}{n - 1} = \frac{8.80}{4} = 2.2$$

We can make a simple modification of our formula for the sample variance to approximate the sample variance if only grouped data are available. Recall that in approximating the sample mean for grouped data, we let y_i and f_i denote the midpoint and frequency, respectively, for the ith class interval. With this notation, the sample variance for grouped data is $s^2 = \Sigma_i f_i(y_i - \bar{y})^2/(n-1)$. The sample standard deviation is $\sqrt{s^2}$.

EXAMPLE 3.10

Refer to the tick data from Table 3.9 of Example 3.6. Calculate the sample variance and standard deviation for these data.

Solution From Table 3.9, the sum of the $f_i(y_i - \bar{y})^2$ calculations is 2,704.688. Using this value, we can approximate s^2 and s.

$$s^2 \cong \frac{1}{n-1}\sum_i f_i(y_i - \bar{y})^2 = \frac{1}{99}(2{,}704.688) = 27.32008$$

$$s \cong \sqrt{27.32008} = 5.227$$

If we compute s from the original 100 data values, the value of s (using Minitab) is computed to be 5.212. The values of s computed from the original data and from the grouped data are very close. However, when the frequency table has a small number of classes, the approximation of s from the frequency table values will not generally be as close as in this example.

We have now discussed several measures of variability, each of which can be used to compare the variabilities of two or more sets of measurements. The standard deviation is particularly appealing for two reasons: (1) we can compare the variabilities of *two or more* sets of data using the standard deviation, and (2) we can also use the results of the rule that follows to interpret the standard deviation of a single set of measurements. This rule applies to data sets with roughly a "mound-shaped" histogram—that is, a histogram that has a single peak, is symmetrical, and tapers off gradually in the tails. Because so many data sets can be classified as mound-shaped, the rule has wide applicability. For this reason, it is called the *Empirical Rule.*

EMPIRICAL RULE

Give a set of n measurements possessing a mound-shaped histogram, then

the interval $\bar{y} \pm s$ contains approximately 68% of the measurements
the interval $\bar{y} \pm 2s$ contains approximately 95% of the measurements
the interval $\bar{y} \pm 3s$ contains approximately 99.7% of the measurements.

EXAMPLE 3.11

The yearly report from a particular stockyard gives the average daily wholesale price per pound for steers as \$.61, with a standard deviation of \$.07. What conclusions can we reach about the daily steer prices for the stockyard? Because the original daily price data are not available, we are not able to provide much further information about the daily steer prices. However, from past experience it is known that the daily price measurements have a mound-shaped relative frequency histogram. Applying the Empirical Rule, what conclusions can we reach about the distribution of daily steer prices?

Solution Applying the Empirical Rule, the interval

$.61 \pm .07$ or \$.54 to \$.68

contains approximately 68% of the measurements. The interval

$.61 \pm .14$ or \$.47 to \$.75

contains approximately 95% of the measurements. The interval

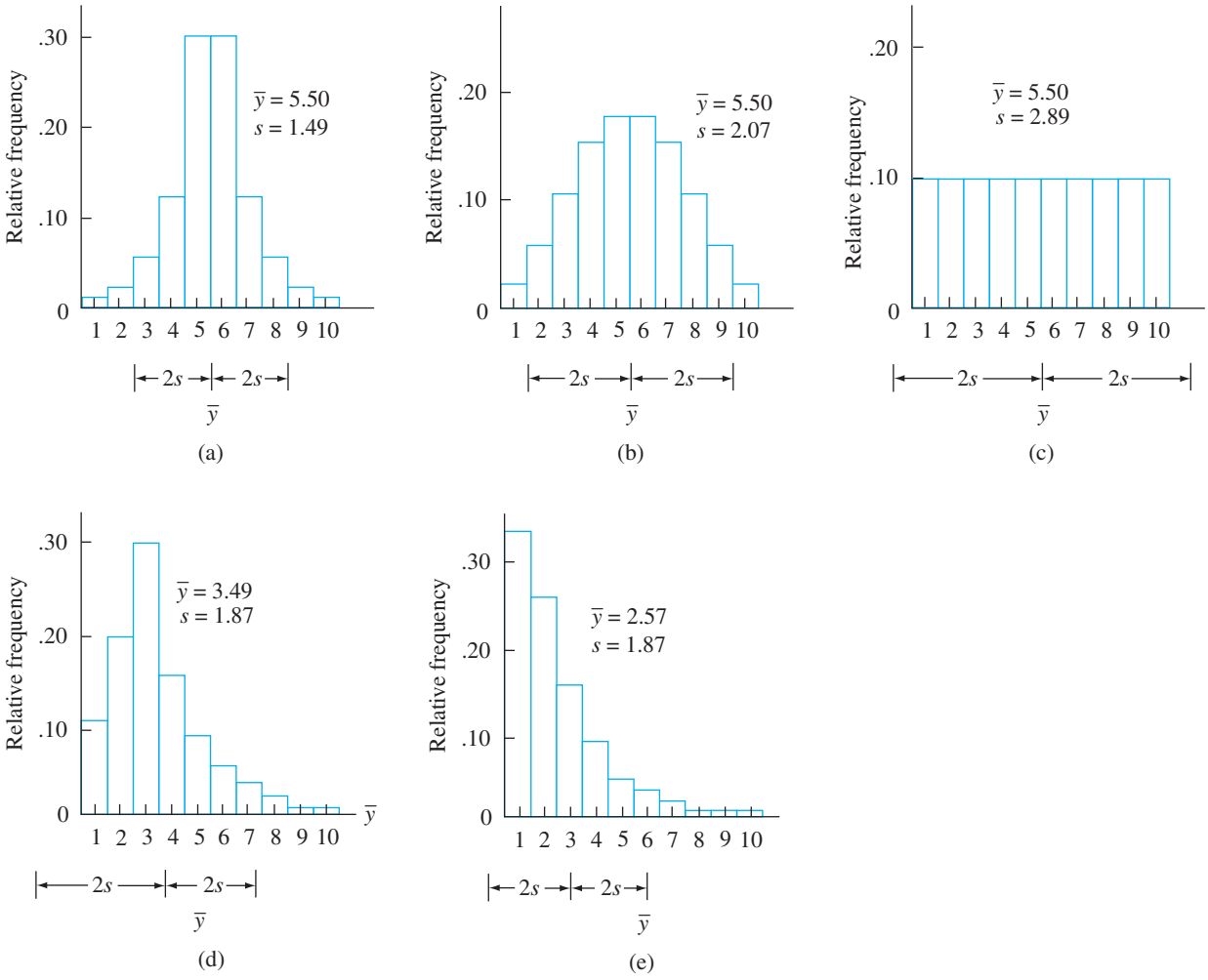
$.61 \pm .21$ or \$.40 to \$.82

contains approximately 99.7% of the measurements.

In English, approximately two-thirds of the steers sold for between \$.54 and \$.68 per pound; and 95% sold for between \$.47 and \$.75 per pound, with minimum and maximum prices being approximately \$.40 and \$.82.

To increase our confidence in the Empirical Rule, let us see how well it describes the five frequency distributions of Figure 3.22. We calculated the mean and standard deviation for each of the five data sets (not given), and these are shown next to each frequency distribution. Figure 3.22(a) shows the frequency distribution

FIGURE 3.22 A demonstration of the utility of the Empirical Rule



for measurements made on a variable that can take values $y = 0, 1, 2, \ldots, 10$. The mean and standard deviation $\bar{y} = 5.50$ and $s = 1.49$ for this symmetric mound-shaped distribution were used to calculate the interval $\bar{y} \pm 2s$, which is marked below the horizontal axis of the graph. We found 94% of the measurements falling in this interval—that is, lying within two standard deviations of the mean. Note that this percentage is very close to the 95% specified in the Empirical Rule. We also calculated the percentage of measurements lying within one standard deviation of the mean. We found this percentage to be 60%, a figure that is not too far from the 68% specified by the Empirical Rule. Consequently, we think the Empirical Rule provides an adequate description for Figure 3.22(a).

Figure 3.22(b) shows another mound-shaped frequency distribution, but one that is less peaked than the distribution of Figure 3.22(a). The mean and standard deviation for this distribution, shown to the right of the figure, are 5.50 and 2.07, respectively. The percentages of measurements lying within one and two standard deviations of the mean are 64% and 96%, respectively. Once again, these percentages agree very well with the Empirical Rule.

Now let us look at three other distributions. The distribution in Figure 3.22(c) is perfectly flat, whereas the distributions of Figures 3.22(d) and (e) are nonsymmetric and skewed to the right. The percentages of measurements that lie within two standard deviations of the mean are 100%, 96%, and 95%, respectively, for these three distributions. All these percentages are reasonably close to the 95% specified by the Empirical Rule. The percentages that lie within one standard deviation of the mean (60%, 75%, and 87%, respectively) show some disagreement with the 68% of the Empirical Rule.

To summarize, you can see that the Empirical Rule accurately forecasts the percentage of measurements falling within two standard deviations of the mean for all five distributions of Figure 3.22, even for the distributions that are flat, as in Figure 3.22(c), or highly skewed to the right, as in Figure 3.22(e). The Empirical Rule is less accurate in forecasting the percentages within one standard deviation of the mean, but the forecast, 68%, compares reasonably well for the three distributions that might be called mound-shaped, Figures 3.22(a), (b), and (d).

The results of the Empirical Rule enable us to obtain a quick approximation to the sample standard deviation s. The Empirical Rule states that approximately 95% of the measurements lie in the interval $\bar{y} \pm 2s$. The length of this interval is, therefore, $4s$. Because the range of the measurements is approximately $4s$, we obtain an **approximate value for s** by dividing the range by 4:

approximating s

$$\text{approximate value of } s = \frac{\text{range}}{4}$$

Some people might wonder why we did not equate the range to $6s$, because the interval $\bar{y} \pm 3s$ should contain almost all the measurements. This procedure would yield an approximate value for s that is smaller than the one obtained by the preceding procedure. If we are going to make an error (as we are bound to do with any approximation), it is better to overestimate the sample standard deviation so that we are not led to believe there is less variability than may be the case.

EXAMPLE 3.12

The Texas legislature planned on expanding the items on which the state sales tax was imposed. In particular, groceries were previously exempt from sales tax. A consumer advocate argued that low-income families would be impacted because they spend a much larger percentage of their income on groceries than do middle- and

upper-income families. The U.S. Bureau of Labor Statistics publication *Consumer Expenditures in 2000* reported that an average family in Texas spent approximately 14% of their family income on groceries. The consumer advocate randomly selected 30 families with income below the poverty level and obtained the following percentages of family incomes allocated to groceries.

26	28	30	37	33	30
29	39	49	31	38	36
33	24	34	40	29	41
40	29	35	44	32	45
35	26	42	36	37	35

For these data, $\Sigma y_i = 1{,}043$ and $\Sigma(y_i - \bar{y})^2 = 1{,}069.3667$. Compute the mean, variance, and standard deviation of the percentage of income spent on food. Check your calculation of s.

Solution The sample mean is

$$\bar{y} = \frac{\Sigma_i y_i}{30} = \frac{1{,}043}{30} = 34.77$$

The corresponding sample variance and standard deviation are

$$s^2 = \frac{1}{n-1}\Sigma(y_i - \bar{y})^2$$
$$= \frac{1}{29}(1{,}069.3667) = 36.8747$$
$$s = \sqrt{36.8747} = 6.07$$

We can check our calculation of s by using the range approximation. The largest measurement is 49 and the smallest is 24. Hence, an approximate value of s is

$$s \approx \frac{\text{range}}{4} = \frac{49 - 24}{4} = 6.25$$

Note how close the approximation is to our computed value.

Although there will not always be the close agreement found in Example 3.12, the range approximation provides a useful and quick check on the calculation of s.

The standard deviation can be deceptive when comparing the amount of variability of different types of populations. A unit of variation in one population might be considered quite small, whereas that same amount of variability in a different population would be considered excessive. For example, suppose we want to compare two production processes that fill containers with products. Process A is filling fertilizer bags, which have a nominal weight of 80 pounds. The process produces bags having a mean weight of 80.6 pounds with a standard deviation of 1.2 pounds. Process B is filling 24-ounce cornflakes boxes, which have a nominal weight of 24 ounces. Process B produces boxes having a mean weight of 24.3 ounces with a standard deviation of 0.4 ounces. Is process A much more variable than process B because 1.2 is three times larger than 0.4? To compare the variability in two considerably different processes or populations, we need to define another measure of variability. The

coefficient of variation

coefficient of variation measures the variability in the values in a population relative to the magnitude of the population mean. In a process or population with mean μ and standard deviation σ, the coefficient of variation is defined as

$$\text{CV} = \frac{\sigma}{|\mu|}$$

provided $\mu \neq 0$. Thus, the coefficient of variation is the standard deviation of the population or process expressed in units of μ. The two filling processes would have equivalent degrees of variability if the two processes had the same CV. For the fertilizer process, the CV = 1.2/80 = .015. The cornflakes process has CV = 0.4/24 = .017. Hence, the two processes have very similar variability relative to the size of their means. The CV is a unit-free number because the standard deviation and mean are measured using the same units. Hence, the CV is often used as an index of process or population variability. In many applications, the CV is expressed as a percentage: CV = $100(\sigma/|\mu|)\%$. Thus, if a process has a CV of 15%, the standard deviation of the output of the process is 15% of the process mean. Using sampled data from the population, we estimate CV with $100(s/|\bar{y}|)\%$.

3.6 The Boxplot

boxplot

As mentioned earlier in this chapter, a stem-and-leaf plot provides a graphical representation of a set of scores that can be used to examine the shape of the distribution, the range of scores, and where the scores are concentrated. The **boxplot,** which builds on the information displayed in a stem-and-leaf plot, is more concerned with the symmetry of the distribution and incorporates numerical measures of central tendency and location to study the variability of the scores and the concentration of scores in the tails of the distribution.

Before we show how to construct and interpret a boxplot, we need to introduce several new terms that are peculiar to the language of exploratory data analysis (EDA).We are familiar with the definitions for the first, second (median), and third quartiles of a distribution presented earlier in this chapter. The boxplot uses the median and **quartiles** of a distribution.

quartiles

We can now illustrate a *skeletal boxplot* using an example.

EXAMPLE 3.13

A criminologist is studying whether there are wide variations in violent crime rates across the United States. Using Department of Justice data from 2000, the crime rates in 90 cities selected from across the United States were obtained. Use the data given in Table 3.12 to construct a skeletal boxplot to demonstrate the degree of variability in crime rates.

TABLE 3.12
Violent crime rates for 90 standard metropolitan statistical areas selected from around the United States

South	Rate	North	Rate	West	Rate
Albany, GA	876	Allentown, PA	189	Abilene, TX	570
Anderson, SC	578	Battle Creek, MI	661	Albuquerque, NM	928
Anniston, AL	718	Benton Harbor, MI	877	Anchorage, AK	516
Athens, GA	388	Bridgeport, CT	563	Bakersfield, CA	885
Augusta, GA	562	Buffalo, NY	647	Brownsville, TX	751
Baton Rouge, LA	971	Canton, OH	447	Denver, CO	561
Charleston, SC	698	Cincinnati, OH	336	Fresno, CA	1,020
Charlottesville, VA	298	Cleveland, OH	526	Galveston, TX	592
Chattanooga, TN	673	Columbus, OH	624	Houston, TX	814
Columbus, GA	537	Dayton, OH	605	Kansas City, MO	843

(*continued*)

TABLE 3.12
Violent crime rates for 90 standard metropolitan statistical areas selected from around the United States (*continued*)

South	Rate	North	Rate	West	Rate
Dothan, AL	642	Des Moines, IA	496	Lawton, OK	466
Florence, SC	856	Dubuque, IA	296	Lubbock, TX	498
Fort Smith, AR	376	Gary, IN	628	Merced, CA	562
Gadsden, AL	508	Grand Rapids, MI	481	Modesto, CA	739
Greensboro, NC	529	Janesville, WI	224	Oklahoma City, OK	562
Hickery, NC	393	Kalamazoo, MI	868	Reno, NV	817
Knoxville, TN	354	Lima, OH	804	Sacramento, CA	690
Lake Charles, LA	735	Madison, WI	210	St. Louis, MO	720
Little Rock, AR	811	Milwaukee, WI	421	Salinas, CA	758
Macon, GA	504	Minneapolis, MN	435	San Diego, CA	731
Monroe, LA	807	Nassau, NY	291	Santa Ana, CA	480
Nashville, TN	719	New Britain, CT	393	Seattle, WA	559
Norfolk, VA	464	Philadelphia, PA	605	Sioux City, IA	505
Raleigh, NC	410	Pittsburgh, PA	341	Stockton, CA	703
Richmond, VA	491	Portland, ME	352	Tacoma, WA	809
Savannah, GA	557	Racine, WI	374	Tucson, AZ	706
Shreveport, LA	771	Reading, PA	267	Victoria, TX	631
Washington, DC	685	Saginaw, MI	684	Waco, TX	626
Wilmington, DE	448	Syracuse, NY	685	Wichita Falls, TX	639
Wilmington, NC	571	Worcester, MA	460	Yakima, WA	585

Note: Rates represent the number of violent crimes (murder, forcible rape, robbery, and aggravated assault) per 100,000 inhabitants, rounded to the nearest whole number.
Source: Department of Justice, Crime Reports and the United States, 2000.

Solution The data were summarized using a stem-and-leaf plot as depicted in Figure 3.23. Use this plot to construct a skeletal boxplot.

FIGURE 3.23
Stem-and-leaf plot of crime data

```
 1  40  45
 1  63
 2  11  20  21
 2  60  61  64  85
 3  12  26  30  39  43  44
 3  50  54  57  58  67  68  97
 4  03  05  13  15  18  26  27  28  47
 4  51  53  63  65  65  70  77  77  86  90  94  96  97  98
 5  21  22  23  28  28  49
 5  55  64  68  71  74  76  79  84
 6  10  25  37  45  46  49
 6  57  60  76  89  92
 7  17
 7  54  54  61  72  73  92  98
 8
 8  71  72  87  97
 9  25  44  46
 9  73  98
10
10  64  94
```

When the scores are ordered from lowest to highest, the median is computed by averaging the 45th and 46th scores. For these data, the 45th score (counting

from the lowest to the highest in Figure 3.23) is 497 and the 46th is 498, hence, the median is

$$M = \frac{497 + 498}{2} = 497.5$$

To find the lower and upper quartiles for this distribution of scores, we need to determine the 25th and 75th percentiles. We can use the method given on page 87 to compute $Q(.25)$ and $Q(.75)$. A quick method that yields essentially the same values for the two quartiles consists of the following steps:

1. Order the data from smallest to largest value.
2. Divide the ordered data set into two data sets using the median as the dividing value.
3. Let the lower quartile be the median of the set of values consisting of the smaller values.
4. Let the upper quartile be the median of the set of values consisting of the larger values.

In the example, the data set has 90 values. Thus, we create two data sets, one containing the $90/2 = 45$ smallest values and the other containing the 45 largest values. The lower quartile is the $(45 + 1)/2 = 23$rd smallest value and the upper quartile is the 23rd value counting from the largest value in the data set. The 23rd-lowest score and 23rd-highest scores are 397 and 660.

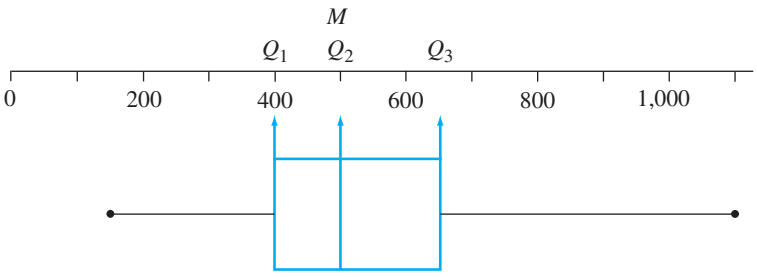
lower quartile, $Q_1 = 397$
upper quartile, $Q_3 = 660$

skeletal boxplot

These three descriptive measures and the smallest and largest values in a data set are used to construct a skeletal boxplot (see Figure 3.24). The **skeletal boxplot** is constructed by drawing a box between the lower and upper quartiles with a solid line drawn across the box to locate the median. A straight line is then drawn connecting the box to the largest value; a second line is drawn from the box to the smallest value. These straight lines are sometimes called whiskers, and the entire graph is called a **box-and-whiskers plot.**

box-and-whiskers plot

FIGURE 3.24
Skeletal boxplot for the data of Figure 3.23



With a quick glance at a skeletal boxplot, it is easy to obtain an impression about the following aspects of the data:

1. The lower and upper quartiles, Q_1 and Q_3
2. The interquartile range (IQR), the distance between the lower and upper quartiles
3. The most extreme (lowest and highest) values
4. The symmetry or asymmetry of the distribution of scores

If we were presented with Figure 3.24 without having seen the original data, we would have observed that

$$Q_1 \approx 400$$
$$Q_3 \approx 675$$
$$\text{IQR} \approx 675 - 400 = 275$$
$$M \approx 500$$

most extreme values: 140 and 1,075

Also, because the median is closer to the lower quartile than the upper quartile and because the upper whisker is a little longer than the lower whisker, the distribution is slightly nonsymmetrical. To see that this conclusion is true, construct a frequency histogram for these data.

The skeletal boxplot can be expanded to include more information about extreme values in the tails of the distribution. To do so, we need the following additional quantities:

lower inner fence: $Q_1 - 1.5(\text{IQR})$
upper inner fence: $Q_3 + 1.5(\text{IQR})$
lower outer fence: $Q_1 - 3(\text{IQR})$
upper outer fence: $Q_3 + 3(\text{IQR})$

Any data value beyond an inner fence on either side is called a *mild outlier,* and a data value beyond an outer fence on either side is called an *extreme outlier.* The smallest and largest data values that are *not* outliers are called the *lower adjacent value* and *upper adjacent value,* respectively.

EXAMPLE 3.14

Compute the inner and outer fences for the data of Example 3.13. Identify any mild and extreme outliers.

Solution For these data, we found the lower and upper quartiles to be 397 and 660, respectively; IQR = 660 − 397 = 263. Then

lower inner fence = 397 − 1.5(263) = 2.5
upper inner fence = 660 + 1.5(263) = 1,054.5
lower outer fence = 397 − 3(263) = −392
upper outer fence = 660 + 3(263) = 1,449

Also, from the stem-and-leaf plot we can determine that the lower and upper adjacent values are 140 and 998. There are two mild outliers, 1,064 and 1,094, because both values fall between the upper inner fence, 1054.5, and upper outer fence, 1449.

We now have all the quantities necessary for constructing a boxplot.

Steps in Constructing a Boxplot

1. As with a skeletal boxplot, mark off a box from the lower quartile to the upper quartile.
2. Draw a solid line across the box to locate the median.
3. Mark the location of the upper and lower adjacent values with an *x*.
4. Draw a line between each quartile and its adjacent value.
5. Mark each outlier with the symbol o.

EXAMPLE 3.15

Construct a boxplot for the data of Example 3.13.

Solution The boxplot is shown in Figure 3.25.

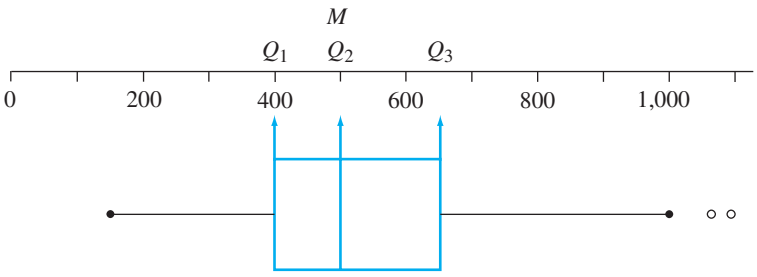


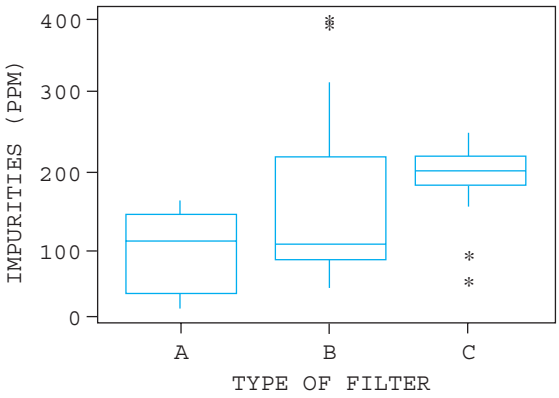
FIGURE 3.25
Skeletal boxplot for the data of Example 3.13

What information can be drawn from a boxplot? First, the center of the distribution of scores is indicated by the median line (Q_2) in the boxplot. Second, a measure of the variability of the scores is given by the interquartile range, the length of the box. Recall that the box is constructed between the lower and upper quartiles so it contains the middle 50% of the scores in the distribution, with 25% on either side of the median line inside the box. Third, by examining the relative position of the median line, we can gauge the symmetry of the middle 50% of the scores. For example, if the median line is closer to the lower quartile than the upper, there is a greater concentration of scores on the lower side of the median within the box than on the upper side; a symmetric distribution of scores would have the median line located in the center of the box. Fourth, additional information about skewness is obtained from the lengths of the whiskers; the longer one whisker is relative to the other one, the more skewness there is in the tail with the longer whisker. Fifth, a general assessment can be made about the presence of outliers by examining the number of scores classified as mild outliers and the number classified as extreme outliers.

Boxplots provide a powerful graphical technique for comparing samples from several different treatments or populations. We will illustrate these concepts using the following example. Several new filtration systems have been proposed for use in small city water systems. The three systems under consideration have very similar initial and operating costs, and will be compared on the basis of the amount of impurities that remain in the water after passing through the system. After careful assessment, it is determined that monitoring 20 days of operation will provide sufficient information to determine any significant difference among the three systems. Water samples are collected on a hourly basis. The amount of impurities, in ppm, remaining in the water after the water passes through the filter is recorded. The average daily values for the three systems are plotted using a side-by-side boxplot, as presented in Figure 3.26.

An examination of the boxplots in Figure 3.26 reveals the shapes of the relative frequency histograms for the three types of filters based on their boxplots. Filter A has a symmetric distribution, filter B is skewed to the right, and filter C is skewed to the left. Filters A and B have nearly equal medians. However, filter B is much more variable than both filters A and C. Filter C has a larger median than both filters A and B but smaller variability than A with the exception of the two very small values obtained using filter C. The extreme values obtained by filters C and B, identified by *, would be examined to make sure that they are valid measurements. These measurements could be either recording errors or operational

FIGURE 3.26
Removing impurities using three filter types



errors. They must be carefully checked because they have such a large influence on the summary statistics. Filter A would produce a more consistent filtration than filter B. Filter A generally filters the water more thoroughly than filter C. We will introduce statistical techniques in Chapter 8 that will provide us with ways to differentiate among the three filter types.

3.7 Summarizing Data from More Than One Variable: Graphs and Correlation

In the previous sections, we've discussed graphical methods and numerical descriptive methods for summarizing data from a single variable. Frequently, more than one variable is being studied at the same time, and we might be interested in summarizing the data on each variable separately, and also in studying relations among the variables. For example, we might be interested in the prime interest rate and in the consumer price index, as well as in the relation between the two. In this section, we'll discuss a few techniques for summarizing data from two (or more) variables. Material in this section will provide a brief preview and introduction to contingency tables (Chapter 10), analysis of variance (Chapters 8 and 14–18), and regression (Chapters 11, 12, and 13).

contingency table

Consider first the problem of summarizing data from two qualitative variables. Cross-tabulations can be constructed to form a **contingency table.** The rows of the table identify the categories of one variable, and the columns identify the categories of the other variable. The entries in the table are the number of times each value of one variable occurs with each possible value of the other. For example, a study of episodic or "binge" drinking—the consumption of large quantities of alcohol at a single session resulting in intoxication—among eighteen-to-twenty-four-year-olds can have a wide range of adverse effects—medical, personal, and social. A survey was conducted on 917 eighteen-to-twenty-four-year-olds by the Institute of Alcohol Studies. Each individual surveyed was asked questions about their alcohol consumption in the prior 6 months. The criminal background of the individuals was also obtained from a police data base. The results of the survey are displayed in Table 3.13. From this table, it is observed that 114 of binge drinkers were involved in violent crimes, whereas, 27 occasional drinkers and 7 nondrinkers were involved in violent crimes.

One method for examining the relationships between variables in a contingency table is a percentage comparison based on row totals, column totals, or the overall total. If we calculate percentages within each column, we can compare the

TABLE 3.13
Data from a survey of drinking behavior of eighteen-to-twenty-four-year-old youths

	Level of Drinking			
Criminal Offenses	**Binge/Regular Drinker**	**Occasional Drinker**	**Never Drinks**	**Total**
Violent Crime	114	27	7	148
Theft/Property Damage	53	27	7	87
Other Criminal Offenses	138	53	15	206
No Criminal Offenses	50	274	152	476
Total	355	381	181	917

TABLE 3.14
Comparing the distribution of criminal activity for each level of alcohol consumption

	Level of Drinking		
Criminal Offenses	**Binge/Regular Drinker**	**Occasional Drinker**	**Never Drinks**
Violent Crime	32.1%	7.1%	3.9%
Theft/Property Damage	14.9%	7.1%	3.9%
Other Criminal Offenses	38.9%	13.9%	8.2%
No Criminal Offenses	14.1%	71.9%	84.0%
Total	100% (n = 355)	100% (n = 381)	100% (n = 181)

distribution of criminal activity within each level of drinking. A percentage comparison based on column totals is shown in Table 3.14.

For all three types of criminal activities, the binge/regular drinkers had more than double the level of activity than did the occassional or nondrinkers. For binge/regular drinkers, 32.1% had committed a violent crime, whereas, only 7.1% of occasional drinkers and 3.9% of nondrinkers had committed a violent crime. This pattern is repeated across the other two levels of criminal activity. In fact, 85.9% of binge/regular drinkers had committed some form of criminal violation. The level of criminal activity among occasional drinkers was 28.1%, and only 16% for nondrinkers. In Chapter 10, we will use statistical methods to explore further relations between two (or more) qualitative variables.

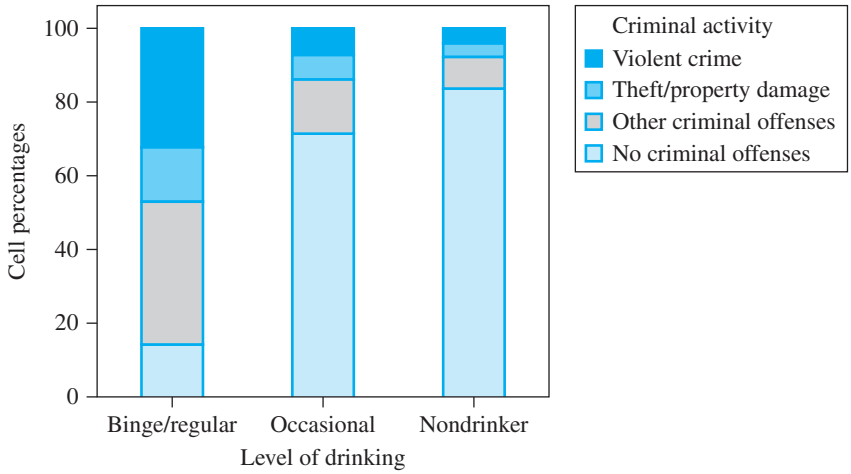
Stacked bar graph

An extension of the bar graph provides a convenient method for displaying data from a pair of qualitative variables. Figure 3.27 is a **stacked bar graph,** which displays the data in Table 3.14.

The graph represents the distribution of criminal activity for three levels of alcohol consumption by young adults. This type of information is useful in making youths aware of the dangers involved in the consumption of large amounts of alcohol. While the heaviest drinkers are at the greatest risk of committing a criminal offense, the risk of increased criminal behavior is also present for the occasional drinker when compared to those youths who are nondrinkers. This type of data may lead to programs that advocate prevention policies and assistance from the beer/alcohol manufacturers to include messages about appropriate consumption in their advertising.

A second extension of the bar graph provides a convenient method for displaying the relationship between a single quantitative and a qualitative variable. A food scientist is studying the effects of combining different types of fats with different

FIGURE 3.27
Chart of cell percentages versus level of drinking, criminal activity



surfactants on the specific volume of baked bread loaves. The experiment is designed with three levels of surfactant and three levels of fat, a 3×3 factorial experiment with varying number of loaves baked from each of the nine treatments. She bakes bread from dough mixed from the nine different combinations of the types of fat and types of surfactants and then measures the specific volume of the bread. The data and summary statistics are displayed in Table 3.15.

cluster bar graph

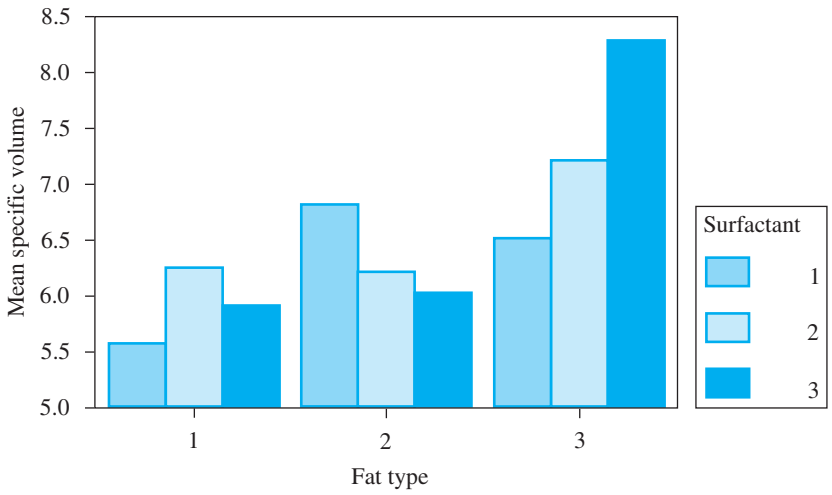
In this experiment, the scientist wants to make inferences from the results of the experiment to the commercial production process. Figure 3.28 is a **cluster bar graph** from the baking experiment. This type of graph allows the experimenter to examine the simultaneous effects of two factors, type of fat and type of surfactant, on the specific volume of the bread. Thus, the researcher can examine the differences in the specific volumes of the nine different ways in which the bread was formulated. A quantitative assessment of the effects of fat type and type of surfactant on the mean specific volume will be addressed in Chapter 15.

We can also construct data plots to summarize the relation between two quantitative variables. Consider the following example. A manager of a small

TABLE 3.15
Descriptive statistics with the dependent variable, specific volume

Fat	Surfactant	Mean	Standard Deviation	*N*
1	1	5.567	1.206	3
	2	6.200	.794	3
	3	5.900	.458	3
	Total	5.889	.805	9
2	1	6.800	.794	3
	2	6.200	.849	2
	3	6.000	.606	4
	Total	6.311	.725	9
3	1	6.500	.849	2
	2	7.200	.668	4
	3	8.300	1.131	2
	Total	7.300	.975	8
Total	1	6.263	1.023	8
	2	6.644	.832	9
	3	6.478	1.191	9
	Total	6.469	.997	26

FIGURE 3.28
Specific volumes from baking experiment



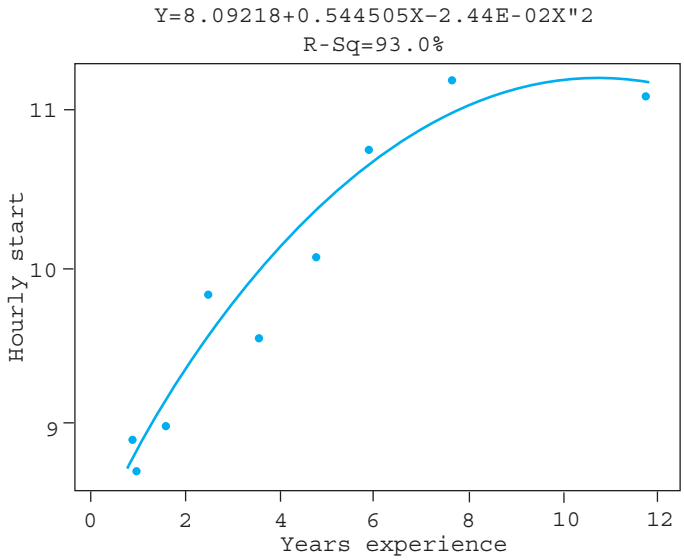
machine shop examined the starting hourly wage y offered to machinists with x years of experience. The data are shown here:

y (dollars)	8.90	8.70	9.10	9.00	9.79	9.45	10.00	10.65	11.10	11.05
x (years)	1.25	1.50	2.00	2.00	2.75	4.00	5.00	6.00	8.00	12.00

scatterplot

Is there a relationship between hourly wage offered and years of experience? One way to summarize these data is to use a **scatterplot,** as shown in Figure 3.29. Each point on the plot represents a machinist with a particular starting wage and years of experience. The smooth curve fitted to the data points, called the *least squares line,* represents a summarization of the relationship between y and x. This line allows the prediction of hourly starting wages for a machinist having years of experience not represented in the data set. How this curve is obtained will be discussed in Chapters 11 and 12. In general, the fitted curve indicates that, as the years of experience x increases, the hourly starting wage increases to a point and then levels off. The basic idea of relating several quantitative variables is discussed in the chapters on regression (11–13).

FIGURE 3.29
Scatterplot of starting hourly wage and years of experience



Using a scatterplot, the general shape and direction of the relationship between two quantitative variables can be displayed. In many instances the relationship can be summarized by fitting a straight line through the plotted points. Thus, the strength of the relationship can be described in the following manner. There is a strong relationship if the plotted points are positioned close to the line, and a weak relationship if the points are widely scattered about the line. It is fairly difficult to "eyeball" the strength using a scatterplot. In particular, if we wanted to compare two different scatterplots, a numerical measure of the strength of the relationship would be advantagous. The following example will illustrate the difficulty of using scatterplots to compare the strength of relationship between two quantitative variables.

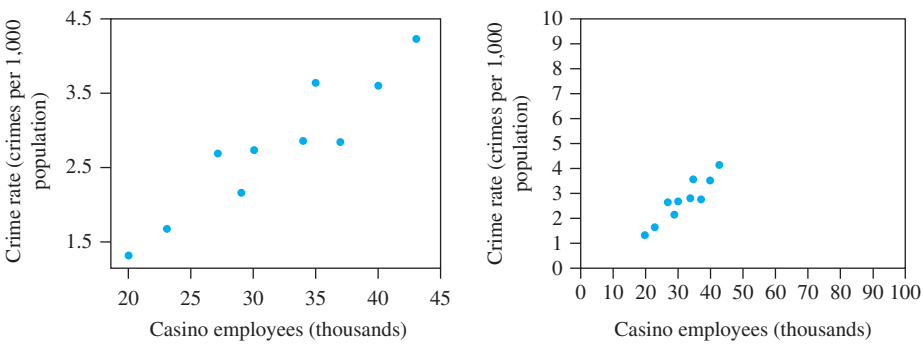
Several major cities in the United States are now considering allowing gambling casinos to operate under their jurisdiction. A major argument in opposition to casino gambling is the perception that there will be a subsequent increase in the crime rate. Data were collected over a 10-year period in a major city where casino gambling had been legalized. The results are listed in Table 3.16 and plotted in Figure 3.30. The two scatterplots are depicting exactly the same data, but the scales of the plots differ considerably. This results in one scatterplot appearing to show a stronger relationship than the other scatterplot.

Because of the difficulty of determining the strength of relationship between two quantitative variables by visually examining a scatterplot, a numerical measure of the strength of relationship will be defined as a supplement to a graphical display. The *correlation coefficient* was first introduced by Francis Galton in 1888. He applied the correlation coefficient to study the relationship between forearm length and the heights of particular groups of people.

TABLE 3.16
Crime rate as a function of number of casino employees

Year	Number of Casino Employees x (thousands)	Crime Rate y (Number of crimes) per 1,000 population)
1994	20	1.32
1995	23	1.67
1996	29	2.17
1997	27	2.70
1998	30	2.75
1999	34	2.87
2000	35	3.65
2001	37	2.86
2002	40	3.61
2003	43	4.25

FIGURE 3.30
Crime rate as a function of number of casino employees



DEFINITION 3.9

The **correlation coefficient** measures the strength of the linear relationship between two quantitative variables. The correlation coefficient is usually denoted as r.

Suppose we have data on two variables x and y collected from n individuals or objects with means and standard deviations of the variables given as $\bar{x}$ and s_x for the x-variable and $\bar{y}$ and s_y for the y-variable. The correlation r between x and y is computed as

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

In computing the correlation coefficient, the two variables x and y are standardized to be unit-free variables. The standardized x-variable for the ith individual, $\left(\frac{x_i - \bar{x}}{s_x}\right)$, measures how many standard deviations x_i is above or below the x-mean. Thus, the correlation coefficient, r, is a unit-free measure of the strength of linear relationship between the quantitative variables, x and y.

EXAMPLE 3.16

For the data in Table 3.16, compute the value of the correlation coefficient.

Solution The computation of r can be obtained from any of the statistical software packages or from Excel. The required calculations in obtaining the value of r for the data in Table 3.16 are given in Table 3.17, with $\bar{x} = 31.80$ and $\bar{y} = 2.785$. The first row is computed as

$$x - \bar{x} = 20 - 31.8 = -11.8, \qquad y - \bar{y} = 1.32 - 2.785 = -1.465,$$
$$(x - \bar{x})(y - \bar{y}) = (-11.8)(-1.465) = 17.287,$$
$$(x - \bar{x})^2 = (-11.8)^2 = 139.24, \qquad (y - \bar{y})^2 = (-1.465)^2 = 2.14623$$

TABLE 3.17
Data and calculations for computing r

	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$
	20	1.32	−11.8	−1.465	17.287	139.24	2.14623
	23	1.67	−8.8	−1.115	9.812	77.44	1.24323
	29	2.17	−2.8	−0.615	1.722	7.84	0.37823
	27	2.70	−4.8	−0.085	0.408	23.04	0.00722
	30	2.75	−1.8	−0.035	0.063	3.24	0.00123
	34	2.87	2.2	0.085	0.187	4.84	0.00722
	35	3.65	3.2	0.865	2.768	10.24	0.74822
	37	2.86	5.2	0.075	0.390	27.04	0.00562
	40	3.61	8.2	0.825	6.765	67.24	0.68062
	43	4.25	11.2	1.465	16.408	125.44	2.14622
Total	318	27.85	0	0	55.810	485.60	7.3641
Mean	31.80	2.785					

A form of r that is somewhat more direct in its calculation is given by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{55.810}{\sqrt{(485.6)(7.3641)}} = .933$$

The above calculations depict a positive correlation between the number of casino employees and the crime rate. However, this result does not prove that an increase

in the number of casino workers *causes* an increase in the crime rate. There may be many other associated factors involved in the increase of the crime rate.

Generally, the correlation coefficient, r, is a positive number if y tends to increase as x increases; r is negative if y tends to decrease as x increases; and r is nearly zero if there is either no relation between changes in x and changes in y or there is a nonlinear relation between x and y such that the patterns of increase and decrease in y (as x increases) cancel each other.

Some properties of r that assist us in the interpretation of relationship between two variables include the following:

1. A positive value for r indicates a positive association between the two variables, and a negative value for r indicates a negative association between the two variables.
2. The value of r is a number between -1 and $+1$. When the value of r is very close to ± 1, the points in the scatterplot will lie close to a straight line.
3. Because the two variables are standardized in the calculation of r, the value of r does not change if we alter the units of x or y. The same value of r will be obtained no matter what units are used for x and y. Correlation is a unit-free measure of association.
4. Correlation measures the degree of straight line relationship between two variables. The correlation coefficient does *not* describe the closeness of the points (x, y) to a curved relationship, no matter how strong the relationship.

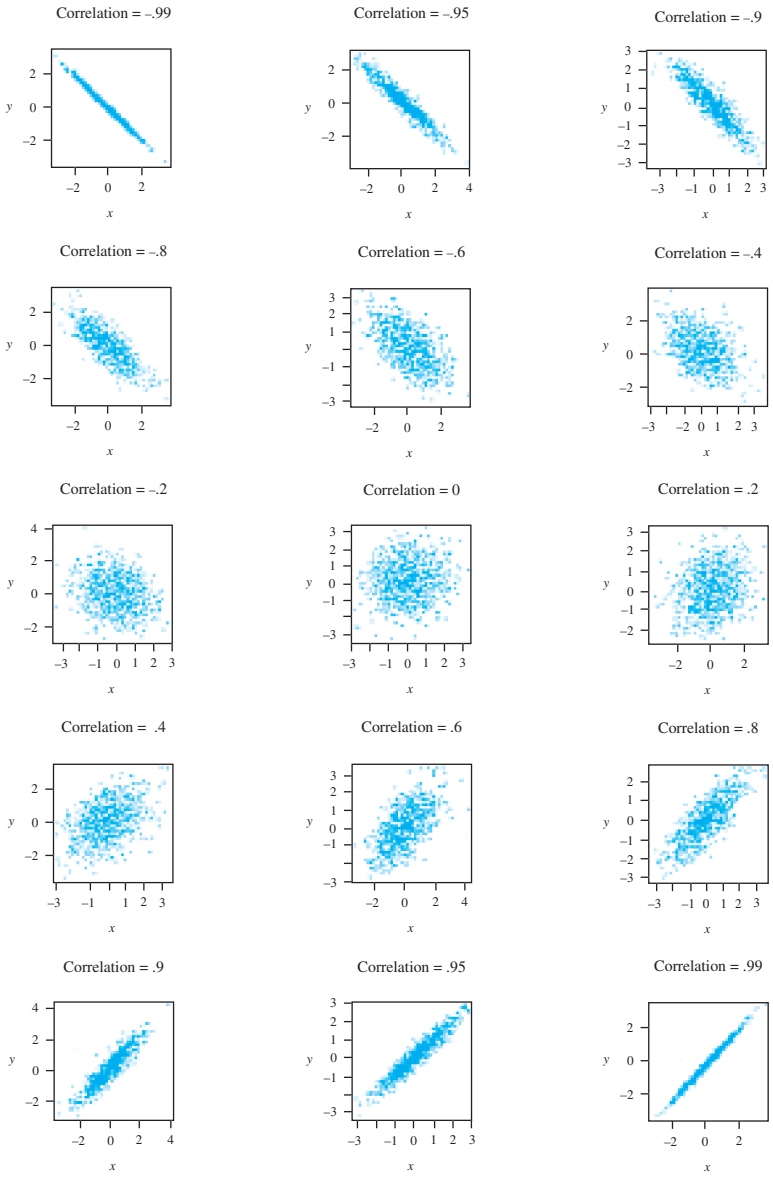
What values of r indicate a "strong" relationship between y and x? Figure 3.31 displays 15 scatterplots obtained by randomly selecting 1,000 pairs (x_i, y_i) from 15 populations having bivariate normal distributions with correlations ranging from $-.99$ to .99. We can observe that unless $|r|$ is greater than .6, there is very little trend in the scatterplot.

Finally, we can construct data plots for summarizing the relation between several quantitative variables. Consider the following example. Thall and Vail (1990) described a study to evaluate the effectiveness of the anti-epileptic drug progabide as an adjuvant to standard chemotherapy. A group of 59 epileptics was selected to be used in the clinical trial. The patients suffering from simple or complex partial seizures were randomly assigned to receive either the anti-epileptic drug progabide or a placebo. At each of four successive postrandomization clinic visits, the number of seizures occurring over the previous 2 weeks was reported. The measured variables were y_i ($i = 1, 2, 3, 4$—the seizure counts recorded at the four clinic visits); Trt (x_1)—0 is the placebo, 1 is progabide; Base (x_2), the baseline seizure rate; Age (x_3), the patient's age in years. The data and summary statistics are given in Tables 3.18 and 3.19.

side-by-side boxplots

The first plots are **side-by-side boxplots** that compare the base number of seizures and ages of the treatment patients to the patients assigned to the placebo. These plots provide a visual assessment of whether the treatment patients and placebo patients had similar distributions of age and base seizure counts prior to the start of the clinical trials. An examination of Figure 3.32(a) reveals that the number of seizures prior to the beginning of the clinical trials has similar patterns for the two groups of patients. There is a single patient with a base seizure count greater than 100 in both groups. The base seizure count for the placebo group is somewhat more variable than for the treatment group—its box is wider than the box for the treatment group. The descriptive statistics table contradicts this

FIGURE 3.31
Scatterplots showing various values for r



observation. The sample standard deviation is 26.10 for the placebo group and 27.37 for the treatment group. This seemingly inconsistent result occurs due to the large base count for a single patient in the treatment group. The median number of base seizures is higher for the treatment group than for the placebo group. The means are nearly identical for the two groups. The means are in greater agreement than are the medians due to the skewed-to-the-right distribution of the middle 50% of the data for the placebo group, whereas the treatment group is nearly symmetric for the middle 50% of its data. Figure 3.32(b) displays the nearly identical distribution of age for the two treatment groups; the only difference is that the treatment group has a slightly smaller median age and is slightly more variable than the placebo group. Thus, the two groups appear to have similar age and base-seizure distributions prior to the start of the clinical trials.

TABLE 3.18
Data for epilepsy study: successive 2-week seizure counts for 59 epileptics. Covariates are adjuvant treatment (0 = placebo, 1 = Progabide), 8-week baseline seizure counts, and age (in years)

ID	y_1	y_2	y_3	y_4	Trt	Base	Age
104	5	3	3	3	0	11	31
106	3	5	3	3	0	11	30
107	2	4	0	5	0	6	25
114	4	4	1	4	0	8	36
116	7	18	9	21	0	66	22
118	5	2	8	7	0	27	29
123	6	4	0	2	0	12	31
126	40	20	23	12	0	52	42
130	5	6	6	5	0	23	37
135	14	13	6	0	0	10	28
141	26	12	6	22	0	52	36
145	12	6	8	4	0	33	24
201	4	4	6	2	0	18	23
202	7	9	12	14	0	42	36
205	16	24	10	9	0	87	26
206	11	0	0	5	0	50	26
210	0	0	3	3	0	18	28
213	37	29	28	29	0	111	31
215	3	5	2	5	0	18	32
217	3	0	6	7	0	20	21
219	3	4	3	4	0	12	29
220	3	4	3	4	0	9	21
222	2	3	3	5	0	17	32
226	8	12	2	8	0	28	25
227	18	24	76	25	0	55	30
230	2	1	2	1	0	9	40
234	3	1	4	2	0	10	19
238	13	15	13	12	0	47	22
101	11	14	9	8	1	76	18
102	8	7	9	4	1	38	32
103	0	4	3	0	1	19	20
108	3	6	1	3	1	10	30
110	2	6	7	4	1	19	18
111	4	3	1	3	1	24	24
112	22	17	19	16	1	31	30
113	5	4	7	4	1	14	35
117	2	4	0	4	1	11	27
121	3	7	7	7	1	67	20
122	4	18	2	5	1	41	22
124	2	1	1	0	1	7	28
128	0	2	4	0	1	22	23
129	5	4	0	3	1	13	40
137	11	14	25	15	1	46	33
139	10	5	3	8	1	36	21
143	19	7	6	7	1	38	35
147	1	1	2	3	1	7	25
203	6	10	8	8	1	36	26
204	2	1	0	0	1	11	25
207	102	65	72	63	1	151	22
208	4	3	2	4	1	22	32
209	8	6	5	7	1	41	25
211	1	3	1	5	1	32	35
214	18	11	28	13	1	56	21
218	6	3	4	0	1	24	41
221	3	5	4	3	1	16	32
225	1	23	19	8	1	22	26
228	2	3	0	1	1	25	21
232	0	0	0	0	1	13	36
236	1	4	3	2	1	12	37

TABLE 3.19
Descriptive statistics: Minitab output for epilepsy example (worksheet size: 100,000 cells)

```
0=PLACEBO
1=TREATED

Variable    TREATMENT       N     Mean   Median  Tr Mean   StDev  SE Mean
Y1          0              28     9.36     5.00     8.54   10.14     1.92
            1              31     8.58     4.00     5.26   18.24     3.28

Y2          0              28     8.29     4.50     7.81    8.16     1.54
            1              31     8.42     5.00     6.37   11.86     2.13

Y3          0              28     8.79     5.00     6.54   14.67     2.77
            1              31     8.13     4.00     5.63   13.89     2.50

Y4          0              28     7.96     5.00     7.46    7.63     1.44
            1              31     6.71     4.00     4.78   11.26     2.02

BASE        0              28    30.79    19.00    28.65   26.10     4.93
            1              31    31.61    24.00    27.37   27.98     5.03

AGE         0              28    29.00    29.00    28.88    6.00     1.13
            1              31    27.74    26.00    27.52    6.60     1.19

Variable    TREATMENT     Min      Max       Q1       Q3
Y1          0            0.00    40.00     3.00    12.75
            1            0.00   102.00     2.00     8.00

Y2          0            0.00    29.00     3.00    12.75
            1            0.00    65.00     3.00    10.00

Y3          0            0.00    76.00     2.25     8.75
            1            0.00    72.00     1.00     8.00

Y4          0            0.00    29.00     3.00    11.25
            1            0.00    63.00     2.00     8.00

BASE        0            6.00   111.00    11.00    49.25
            1            7.00   151.00    13.00    38.00

AGE         0           19.00    42.00    24.25    32.00
            1           18.00    41.00    22.00    33.00
```

FIGURE 3.32(a)
Boxplot of base by treatment

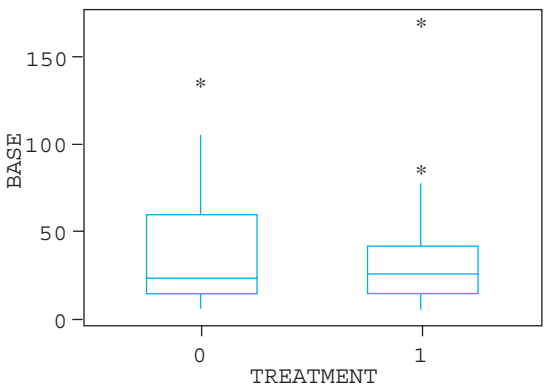
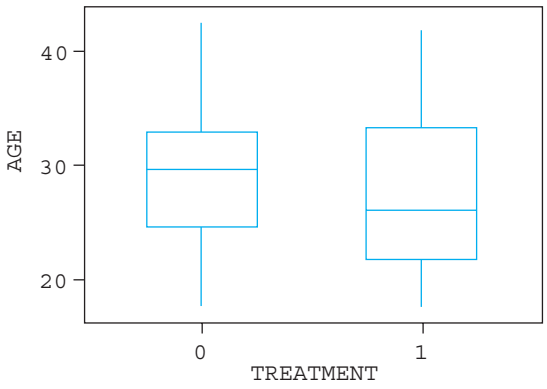


FIGURE 3.32(b)
Boxplot of age by treatment



3.8 Research Study: Controlling for Student Background in the Assessment of Teaching

At the beginning of this chapter, we described a situation faced by many school administrators having a large minority population in their school and/or a large proportion of their students classified as from a low-income family. The implications of such demographics on teacher evaluations through the performance of their students on standardized reading and math tests generates much controversy in the educational community. The task of achieving goals set by the national *Leave no student behind* mandate are much more difficult for students from disadvantaged backgrounds. Requiring teachers and administrators from school districts with a high proportion of disadvantaged students to meet the same standards as those from schools with a more advantaged student body is inherently unfair. This type of policy may prove to be counterproductive. It may lead to the alienation of teachers and administrators and the flight of the most qualified and most productive educators from disadvantaged school districts, resulting in a staff with only those educators with an overwhelming commitment to students with a disadvantaged background and/or educators who lack the qualifications to move to the higher-rated schools. A policy that mandates that educators should be held accountable for the success of their students without taking into account the backgrounds of those students is destined for failure.

The data from a medium-sized Florida school district with 22 elementary schools were presented at the beginning of this chapter. The minority status of a student was defined as black or non-black race. In this school district, almost all students are non-Hispanic blacks or whites. Most of the relatively small numbers of Hispanic students are white. Most students of other races are Asian but they are relatively few in number. They were grouped in the minority category because of the similarity of their test score profiles. Poverty status was based on whether or not the student received free or reduced lunch subsidy. The math and reading scores are from the Iowa Test of Basic Skills. The number of students by class in each school is given by N in Table 3.20.

The superintendent of schools presented the school board with the data, and they wanted an assessment of whether poverty and minority status had any effect on the math and reading scores. Just looking at the data presented very little insight in reaching an answer to this question. Using a number of the graphs and summary statistics introduced in this chapter, we will attempt to assist the superintendent in

TABLE 3.20
Summary statistics for reading scores and math scores by grade level

Variable	Grade	N	Mean	St. Dev	Minimum	Q_1	Median	Q_3	Maximum
Math	3	22	171.87	9.16	155.50	164.98	174.65	179.18	186.10
	4	22	189.88	9.64	169.90	181.10	189.45	197.28	206.90
	5	19	206.16	11.14	192.90	197.10	205.20	212.70	228.10
Reading	3	22	171.10	7.46	157.20	164.78	171.85	176.43	183.80
	4	22	185.96	10.20	166.90	178.28	186.95	193.85	204.70
	5	19	205.36	11.04	186.60	199.00	203.30	217.70	223.30
%Minority	3	22	39.43	25.32	12.30	20.00	28.45	69.45	87.40
	4	22	40.22	24.19	11.10	21.25	32.20	64.53	94.40
	5	19	40.42	26.37	10.50	19.80	29.40	64.10	92.60
%Poverty	3	22	58.76	24.60	13.80	33.30	68.95	77.48	91.70
	4	22	54.00	24.20	11.70	33.18	60.55	73.38	91.70
	5	19	56.47	23.48	13.20	37.30	61.00	75.90	92.90

providing insight to the school board concerning the impact of poverty and minority status on student performance.

In order to access the degree of variability in the mean math and reading scores between the 22 schools, a boxplot of the math and reading scores for each of the three grade levels is given in Figure 3.33. There are 22 third- and fourth-grade classes, and only 19 fifth-grade classes.

From these plots, we observe that for each of the three grade levels there is a wide variation in mean math and reading scores. However, the level of variability within a grade appears to be about the same for math and reading scores but with a wide level of variability for fourth and fifth grades in comparison to third graders. Furthermore, there is an increase in the median scores from the third to the fifth grades. A detailed summary of the data is given in Table 3.20.

For the third-grade classes, the scores for math and reading had similar ranges: 155 to 185. The range for the 22 schools increased to 170 to 205 for the

FIGURE 3.33
Boxplot of math and reading scores for each grade

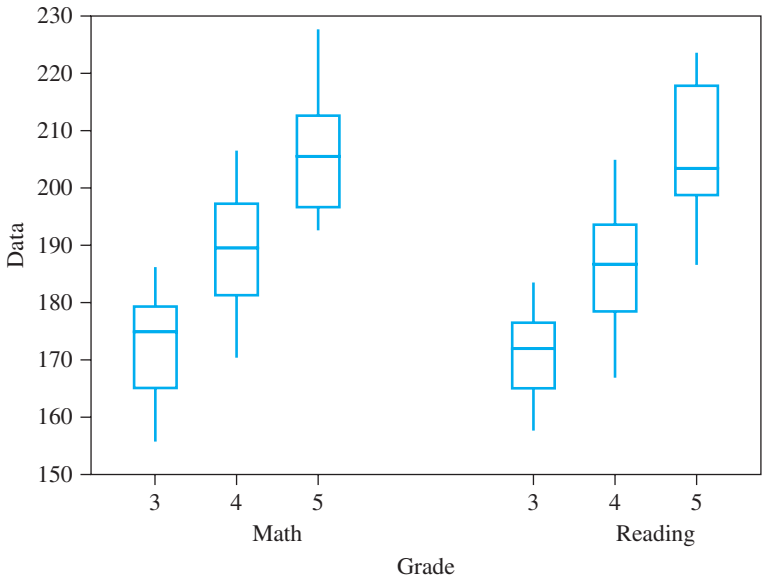
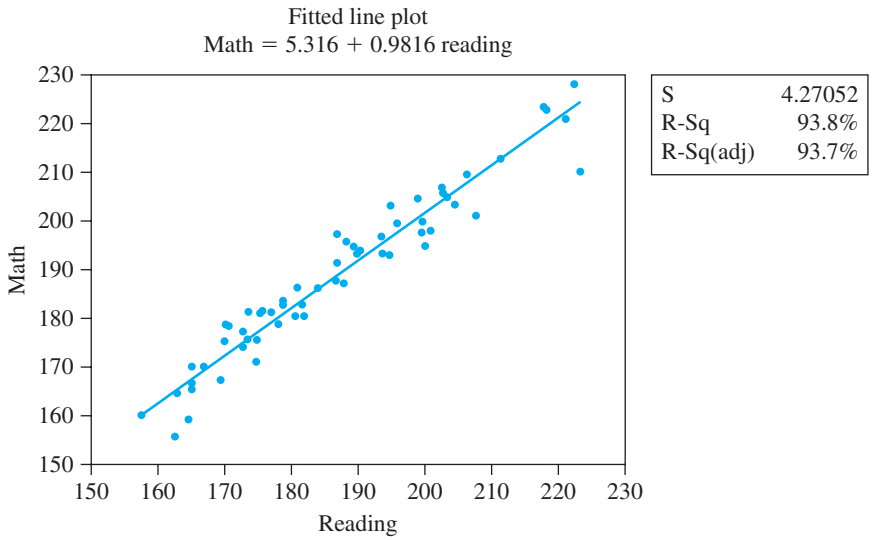


FIGURE 3.34
Scatterplot of reading scores versus math scores



fourth-grade students in both math and reading. This size of the range for the fifth-grade students was similar to the fourth graders: 190 to 225 for both math and reading. Thus, the level of variability in reading and math scores is increasing from third grade to fourth grade to fifth grade. This is confirmed by examining the standard deviations for the three grades. Also, the median scores for both math and reading are increasing across the three grades. The school board then asked the superintendent to identify possible sources of differences in the 22 schools that may help explain the differences in the mean math and reading scores.

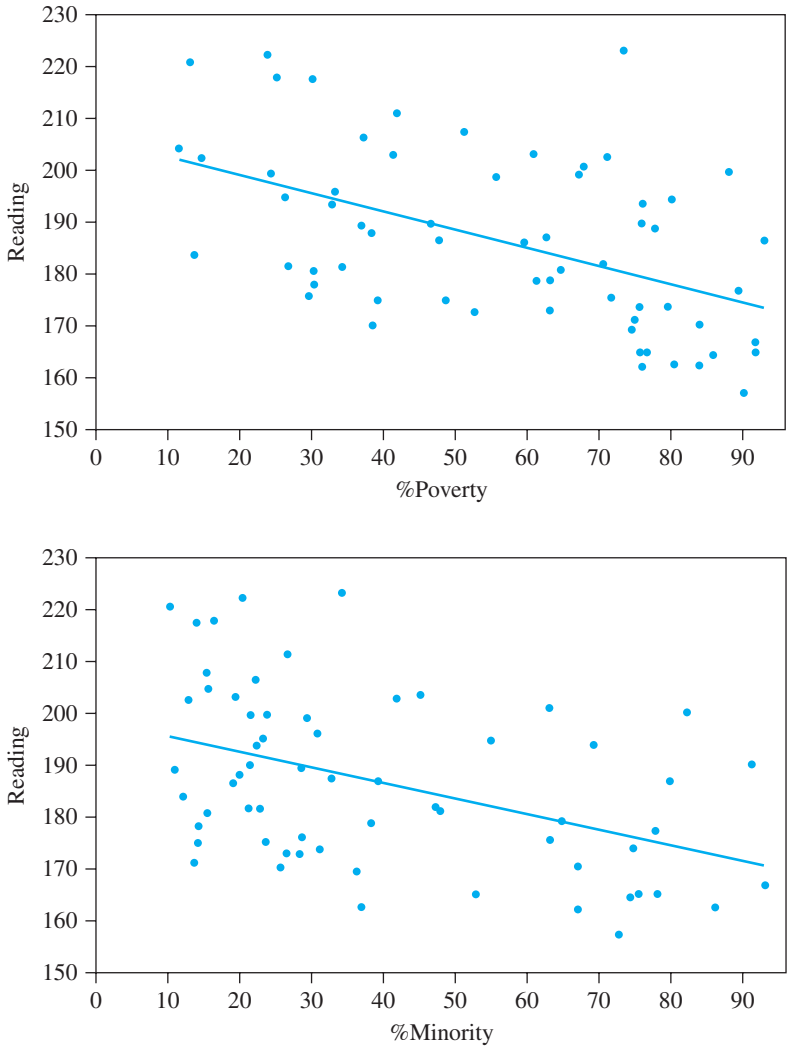
In order to simplify the analysis somewhat, it was proposed to analyze just the reading scores because it would appear that the math and reading scores had a similar variation between the 22 schools. To help justify this choice in analysis, a scatterplot of the 63 pairs of math and reading scores (recall there were only 19 fifth-grade classes) was generated (see Figure 3.34). From this plot we can observe a strong correlation between the reading and math scores for the 64 schools. In fact, the correlation coefficient between math and reading scores is computed to be .97. Thus, there is a very strong relationship between reading and math scores at the 22 schools. The remainder of the analysis will be with respect to the reading scores.

The next step in the process of examining if minority or poverty status are associated with the reading scores. Figure 3.35 is a scatterplot of reading versus %poverty and reading versus %minority.

Although there appears to be a general downward trend in reading scores as the level of %poverty and %minority in the schools increases, there is a wide scattering of individual scores about the fitted line. The correlation between reading and %poverty is $-.45$ and between reading and %minority is $-.53$. However, recall that there is a general upward shift in reading scores from the third grade to the fifth grade. Therefore, a more appropriate plot of the data would be to fit a separate line for each of the three grades. This plot is given in Figure 3.36.

From these plots, we can observe a much stronger association between reading scores and both %poverty and %minority. In fact, if we compute the correlation between the variables separately for each grade level, we will note a dramatic increase in the value of the correlation coefficient. The values are given in Table 3.21.

FIGURE 3.35
Scatterplot of reading scores versus %minority and %poverty



From Figure 3.36 and the values of the correlation coefficients, we can observe that as the proportion of minority students in the schools increases there is a steady decline in reading scores. The same pattern is observed with respect to the proportion of students who are classified as being from a low-income family.

What can we conclude from the information presented above? First, it would appear that scores on reading exams tend to decrease as the values of %poverty and %minority increase. Thus, we may be inclined to conclude that increasing values of %poverty and %minority *cause* a decline in reading scores and hence that the teachers in schools with high levels of %poverty and %minority should have special considerations when teaching evaluations are conducted. This type of thinking often leads to very misleading conclusions. There may be many other variables involved other than %poverty and %minority that may be impacting the reading scores. To conclude that the high levels %poverty and %minority in a school will often result in low reading scores cannot be supported by this data. Much more information is needed to reach any conclusion having this type of certainty.

FIGURE 3.36
Scatterplot of reading scores versus %minority and %poverty with separate lines for each grade

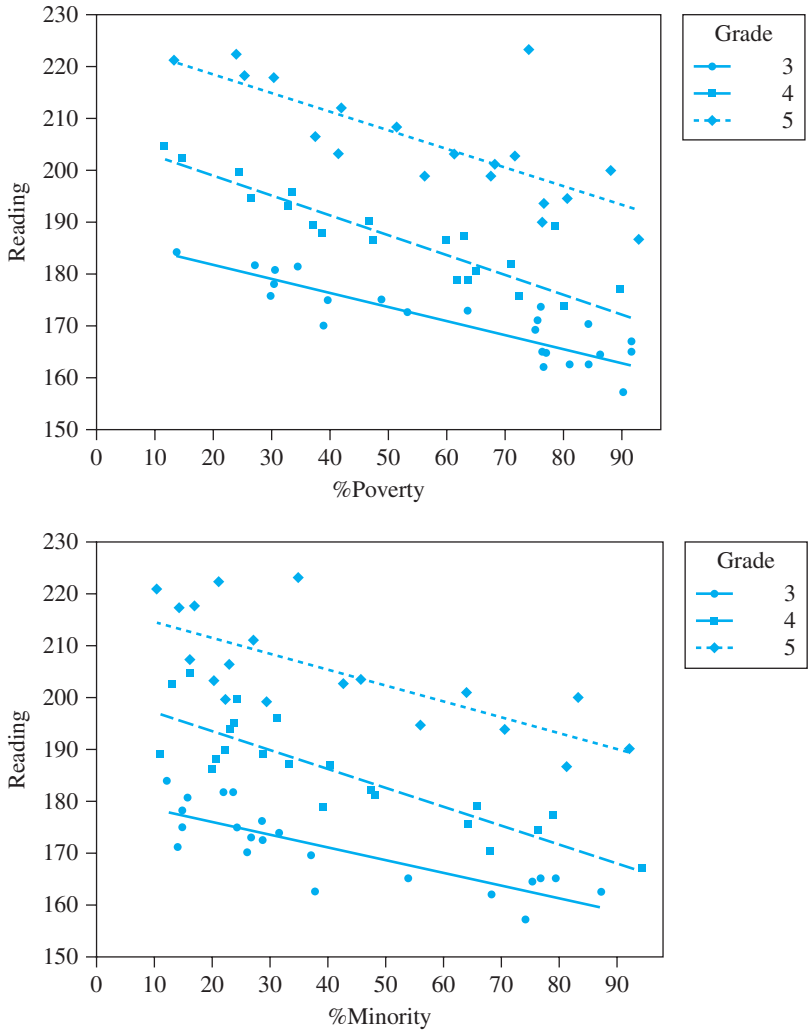


TABLE 3.21
Correlation between reading scores and %poverty and %minority

Correlation between	3rd Grade	4th Grade	5th Grade
Reading scores and			
%minority	−.83	−.87	−.75
%poverty	−.89	−.92	−.76

3.9 Summary and Key Formulas

This chapter was concerned with graphical and numerical description of data. The pie chart and bar graph are particularly appropriate for graphically displaying data obtained from a qualitative variable. The frequency and relative frequency histograms and stem-and-leaf plots are graphical techniques applicable only to quantitative data.

Numerical descriptive measures of data are used to convey a mental image of the distribution of measurements. Measures of central tendency include the mode, the median, and the arithmetic mean. Measures of variability include the range, the interquartile range, the variance, and the standard deviation of a set of measurements.

We extended the concept of data description to summarize the relations between two qualitative variables. Here cross-tabulations were used to develop percentage comparisons. We examined plots for summarizing the relations between quantitative and qualitative variables and between two quantitative variables. Material presented here (namely, summarizing relations among variables) will be discussed and expanded in later chapters on chi-square methods, on the analysis of variance, and on regression.

Key Formulas

1. Median, grouped data

$$\text{Median} = L + \frac{w}{f_m}(.5n - cf_b)$$

2. Sample mean

$$\bar{y} = \frac{\Sigma_i y_i}{n}$$

3. Sample mean, grouped data

$$\bar{y} \cong \frac{\Sigma_i f_i y_i}{n}$$

4. Sample variance

$$s^2 = \frac{1}{n-1}\sum_i (y_i - \bar{y})^2$$

5. Sample variance, grouped data

$$s^2 \cong \frac{1}{n-1}\sum_i f_i(y_i - \bar{y})^2$$

6. Sample standard deviation

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

7. Sample coefficient of variation

$$\text{CV} = \frac{s}{|\bar{y}|}$$

8. Correlation coefficient

$$r = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

3.10 Exercises

3.3 Describing Data on a Single Variable: Graphical Methods

Gov. **3.1** The U.S. government spent more than $2.5 trillion in the 2006 fiscal year. How is this incredible sum of money spent? The following table provides broad categories which demonstrate the expenditures of the Federal government for domestic and defense programs.

Federal Program	2006 Expenditures (Billions of Dollars)
National Defense	$525
Social Security	$500
Medicare & Medicaid	$500
National Debt Interest	$300
Major Social-Aid Programs	$200
Other	$475

a. Construct a pie chart for these data.
b. Construct a bar chart for these data.
c. Construct a pie chart and bar chart using percentages in place of dollars.
d. Which of the four charts is more informative to the tax-paying public?

Bus. **3.2** A major change appears to be taking place with respect to the type of vehicle the U.S. public is purchasing. The U.S. Bureau of Economic Analysis in their publication *Survey of Current*

Business (February 2002) provide the data given in the following table. The numbers reported are in thousands of units—that is, 9,436 represents 9,436,000 vehicles sold in 1990.

	Year							
Type of Vehicle	**1990**	**1995**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**
Passenger Car	9,436	8,687	8,273	8,142	8,697	8,852	8,422	8,082
SUV/Light Truck	4,733	6,517	7,226	7,821	8,717	8,965	9,050	9,036

a. Would pie charts be appropriate graphical summaries for displaying these data? Why or why not?
b. Construct a bar chart that would display the changes across the 12 years in the public's choice in vehicle.
c. Do you observe a trend in the type of vehicles purchased? Do you feel this trend will continue if there was a substantial rise in gasoline prices?

Med. **3.3** It has been reported that there has been a change in the type of practice physicians are selecting for their career. In particular, there is concern that there will be a shortage of family practice physicians in future years. The following table contains data on the total number of office-based physicians and the number of those physicians declaring themselves to be family practice physicians. The numbers in the table are given in thousands of physicians. (Source: Statistical Abstract of the United States: 2003)

	Year						
	1980	**1990**	**1995**	**1998**	**1999**	**2000**	**2001**
Family Practice	47.8	57.6	59.9	64.6	66.2	67.5	70.0
Total Office-Based Physicians	271.3	359.9	427.3	468.8	473.2	490.4	514.0

a. Use a bar chart to display the increase in the number of family practice physicians from 1990 to 2002.
b. Calculate the percent of office-based physicians who are family practice physicians and then display this data in a bar chart.
c. Is there a major difference in the trend displayed by the two bar charts?

Env. **3.4** The regulations of the board of health in a particular state specify that the fluoride level must not exceed 1.5 parts per million (ppm). The 25 measurements given here represent the fluoride levels for a sample of 25 days. Although fluoride levels are measured more than once per day, these data represent the early morning readings for the 25 days sampled.

.75	.86	.84	.85	.97
.94	.89	.84	.83	.89
.88	.78	.77	.76	.82
.72	.92	1.05	.94	.83
.81	.85	.97	.93	.79

a. Determine the range of the measurements.
b. Dividing the range by 7, the number of subintervals selected, and rounding, we have a class interval width of .05. Using .705 as the lower limit of the first interval, construct a frequency histogram.
c. Compute relative frequencies for each class interval and construct a relative frequency histogram. Note that the frequency and relative frequency histograms for these data have the same shape.
d. If one of these 25 days were selected at random, what would be the chance (probability) that the fluoride reading would be greater than .90 ppm? Guess (predict) what proportion of days in the coming year will have a fluoride reading greater than .90 ppm.

Gov. **3.5** The National Highway Traffic Safety Administration has studied the use of rear-seat automobile lap and shoulder seat belts. The number of lives potentially saved with the use of lap and shoulder seat belts is shown for various percentages of use.

Percentage of Use	Lives Saved Wearing	
	Lap Belt Only	Lap and Shoulder Belt
100	529	678
80	423	543
60	318	407
40	212	271
20	106	136
10	85	108

Suggest several different ways to graph these data. Which one seems more appropriate and why?

Soc. **3.6** As the mobility of the population in the United States has increased and with the increase in home-based employment, there is an inclination to assume that the personal income in the United States would become fairly uniform across the country. The following table provides the per capita personal income for each of the 50 states and the District of Columbia.

Income (thousands of dollars)	Number of States
22.0–24.9	5
25.0–27.9	13
28.0–30.9	16
31.0–33.9	9
34.0–36.9	4
37.0–39.9	2
40.0–42.9	2
Total	51

a. Construct a relative frequency histogram for the income data.
b. Describe the shape of the histogram using the standard terminology of histograms.
c. Would you describe per capita income as being fairly homogenous across the United States?

Med. **3.7** The survival times (in months) for two treatments for patients with severe chronic left-ventricular heart failure are given in the following tables.

Standard Therapy						
4	15	24	10	1	27	31
14	2	16	32	7	13	36
29	6	12	18	14	15	18
6	13	21	20	8	3	24

New Therapy						
5	20	29	15	7	32	36
17	15	19	35	10	16	39
27	14	10	16	12	13	16
9	18	33	30	29	31	27

a. Construct separate relative frequency histograms for the survival times of both the therapies.
b. Compare the two histograms. Does the new therapy appear to generate a longer survival time? Explain your answer.

3.8 Combine the data from the separate therapies in Exercise 3.7 into a single data set and construct a relative frequency histogram for this combined data set. Does the plot indicate that the data are from two separate populations? Explain your answer.

Gov. **3.9** Liberal members of Congress have asserted that the U.S. federal government has been expending an increasing portion of the nation's resources on the military and intelligence agencies. The following table contains the outlays (in billions of dollars) for the Defense Department and associated intelligence agencies since 1980. The data are also given as a percentage of gross national product (% GNP).

Year	Expenditure	%GNP	Year	Expenditure	%GNP
1980	134	4.9	1993	291	4.4
1981	158	5.2	1994	282	4.1
1982	185	5.8	1995	272	3.7
1983	210	6.1	1996	266	3.5
1984	227	6.0	1997	271	3.3
1985	253	6.1	1998	269	3.1
1986	273	6.2	1999	275	3.0
1987	282	6.1	2000	295	3.0
1988	290	5.9	2001	306	3.0
1989	304	5.7	2002	349	3.4
1990	299	5.2	2003	376	3.5
1991	273	4.6	2004	391	3.5
1992	298	4.8			

Source: Statistical Abstract of the United States, 2003

a. Plot the defense expenditures time-series data and describe any trends across the time 1980 to 2004.
b. Plot the %GNP time-series data and describe any trends across the time 1980 to 2004.
c. Do the two time series have similar trends? Do either of the plots support the members of Congress assertions?

Gov. **3.10** There has been an increasing emphasis in recent years to make sure that young women are given the same opportunities to develop their mathematical skills as males in U.S. educational systems. The following table provides the average SAT scores for male and female students over the past 35 years. Plot the four separate time series.

	Year												
Gender/Type	**1967**	**1970**	**1975**	**1980**	**1985**	**1990**	**1993**	**1994**	**1995**	**1996**	**2000**	**2001**	**2002**
Male/Verbal	540	536	515	506	514	505	504	501	505	507	507	509	507
Female/Verbal	545	538	509	498	503	496	497	497	502	503	504	502	502
Male/Math	535	531	518	515	522	521	524	523	525	527	533	533	534
Female/Math	495	493	479	473	480	483	484	487	490	492	498	498	500

Source: Statistical Abstract of the United States, 2003

a. Plot the four separate time series and describe any trends in the separate time series.
b. Do the trends appear to imply a narrowing in the differences between male and female math scores?
c. Do the trends appear to imply a narrowing in the differences between male and female verbal scores?

Soc. **3.11** The following table presents the homeownership rates, in percentages, by state for the years 1985, 1996 and 2002. These values represent the proportion of homes owned by the occupant to the total number of occupied homes.

State	1985	1996	2002	State	1985	1996	2002
Alabama	70.4	71.0	73.5	Montana	66.5	68.6	69.3
Alaska	61.2	62.9	67.3	Nebraska	68.5	66.8	68.4
Arizona	64.7	62.0	65.9	Nevada	57.0	61.1	65.5
Arkansas	66.6	66.6	70.2	New Hampshire	65.5	65.0	69.5
California	54.2	55.0	58.0	New Jersey	62.3	64.6	67.2
Colorado	63.6	64.5	69.1	New Mexico	68.2	67.1	70.3
Connecticut	69.0	69.0	71.6	New York	50.3	52.7	55.0
Delaware	70.3	71.5	75.6	North Carolina	68.0	70.4	70.0
Dist. of Columbia	37.4	40.4	44.1	North Dakota	69.9	68.2	69.5
Florida	67.2	67.1	68.7	Ohio	67.9	69.2	72.0
Georgia	62.7	69.3	71.7	Oklahoma	70.5	68.4	69.4
Hawaii	51.0	50.6	57.4	Oregon	61.5	63.1	66.2
Idaho	71.0	71.4	73.0	Pennsylvania	71.6	71.7	74.0
Illinois	60.6	68.2	70.2	Rhode Island	61.4	56.6	59.6
Indiana	67.6	74.2	75.0	South Carolina	72.0	72.9	77.3
Iowa	69.9	72.8	73.9	South Dakota	67.6	67.8	71.5
Kansas	68.3	67.5	70.2	Tennessee	67.6	68.8	70.1
Kentucky	68.5	73.2	73.5	Texas	60.5	61.8	63.8
Louisiana	70.2	64.9	67.1	Utah	71.5	72.7	72.7
Maine	73.7	76.5	73.9	Vermont	69.5	70.3	70.2
Maryland	65.6	66.9	72.0	Virginia	68.5	68.5	74.3
Massachusetts	60.5	61.7	62.7	Washington	66.8	63.1	67.0
Michigan	70.7	73.3	76.0	West Virginia	75.9	74.3	77.0
Minnesota	70.0	75.4	77.3	Wisconsin	63.8	68.2	72.0
Mississippi	69.6	73.0	74.8	Wyoming	73.2	68.0	72.8
Missouri	69.2	70.2	74.6				

Source: U.S. Bureau of the Census, Internet site: http://www.census.gov/ftp/pub/hhes/www/hvs.html

a. Construct a relative frequency histogram plot for the homeownership data given in the table for the years 1985, 1996, and 2002.
b. What major differences exist between the plots for the three years?
c. Why do you think the plots have changed over these 17 years?
d. How could Congress use the information in these plots for writing tax laws that allow major tax deductions for homeownership?

3.12 Construct a stem-and-leaf plot for the data of Exercise 3.11.

3.13 Describe the shape of the stem-and-leaf plot and histogram for the homeownership data in Exercises 3.11 and 3.12, using the terms *modality, skewness,* and *symmetry* in your description.

Bus. **3.14** A supplier of high-quality audio equipment for automobiles accumulates monthly sales data on speakers and receiver–amplifier units for 5 years. The data (in thousands of units per month) are shown in the following table. Plot the sales data. Do you see any overall trend in the data? Do there seem to be any cyclic or seasonal effects?

Year	J	F	M	A	M	J	J	A	S	O	N	D
1	101.9	93.0	93.5	93.9	104.9	94.6	105.9	116.7	128.4	118.2	107.3	108.6
2	109.0	98.4	99.1	110.7	100.2	112.1	123.8	135.8	124.8	114.1	114.9	112.9
3	115.5	104.5	105.1	105.4	117.5	106.4	118.6	130.9	143.7	132.2	120.8	121.3
4	122.0	110.4	110.8	111.2	124.4	112.4	124.9	138.0	151.5	139.5	127.7	128.0
5	128.1	115.8	116.0	117.2	130.7	117.5	131.8	145.5	159.3	146.5	134.0	134.2

3.4 Describing Data on a Single Variable: Measures of Central Tendency

Basic **3.15** Compute the mean, median, and mode for the following data:

55	85	90	50	110	115	75	85	8	23
70	65	50	60	90	90	55	70	5	31

Basic **3.16** Refer to the data in Exercise 3.15 with the measurements 110 and 115 replaced by 345 and 467. Recompute the mean, median, and mode. Discuss the impact of these extreme measurements on the three measures of central tendency.

Basic **3.17** Refer to the data in Exercise 3.15 and 3.16. Compute a 10% trimmed mean for both data sets—that is, the original and the one with the two extreme values. Do the extreme values affect the 10% trimmed mean? Would a 5% trimmed mean be affected by the two extreme values?

Basic **3.18** Determine the mean, median, and mode for the data presented in the following frequency table.

Class Interval	Frequency
2.0–4.9	5
5.0–7.9	13
8.0–10.9	16
11.0–13.9	9
14.0–16.9	4
17.0–19.9	2
20.0–22.9	2

Engin. **3.19** A study of the reliability of buses ["Large sample simultaneous confidence intervals for the multinominal probabilities on transformations of the cell frequencies," *Technometrics* (1980) 22:588] examined the reliability of 191 buses. The distance traveled (in 1,000s of miles) prior to the first major motor failure was classified into intervals. A modified form of the table follows.

Distance Traveled (1,000 miles)	Frequency
0–20.0	6
20.1–40.0	11
40.1–60.0	16
60.1–100.0	59
100.1–120.0	46
120.1–140.0	33
140.1–160.0	16
160.1–200.0	4

a. Sketch the relative frequency histogram for the distance data and describe its shape.
b. Estimate the mode, median, and mean for the distance traveled by the 191 buses.
c. What does the relationship among the three measures of center indicate about the shape of the histogram for these data?
d. Which of the three measures would you recommend as the most appropriate representative of the distance traveled by one of the 191 buses? Explain your answer.

Med. **3.20** In a study of 1,329 American men reported in *American Statistician* [(1974) 28:115–122] the men were classified by serum cholesterol and blood pressure. The group of 408 men who had

blood pressure readings less than 127 mm Hg were then classified according to their serum cholesterol level.

Serum Cholesterol (mg/100cc)	Frequency
0.0–199.9	119
200.0–219.9	88
220.0–259.9	127
greater than 259	74

a. Estimate the mode, median, and mean for the serum cholesterol readings (if possible).
b. Which of the three summary statistics is more informative concerning a typical serum cholesterol level for the group of men? Explain your answer.

Env. **3.21** The ratio of DDE (related to DDT) to PCB concentrations in bird eggs has been shown to have had a number of biological implications. The ratio is used as an indication of the movement of contamination through the food chain. The paper "The ratio of DDE to PCB concentrations in Great Lakes herring gull eggs and its use in interpreting contaminants data" [*Journal of Great Lakes Research* (1998) 24(1):12–31] reports the following ratios for eggs collected at 13 study sites from the five Great Lakes. The eggs were collected from both terrestrial- and aquatic-feeding birds.

	DDE to PCB Ratio										
Terrestrial Feeders	76.50	6.03	3.51	9.96	4.24	7.74	9.54	41.70	1.84	2.50	1.54
Aquatic Feeders	0.27	0.61	0.54	0.14	0.63	0.23	0.56	0.48	0.16	0.18	

a. Compute the mean and median for the 21 ratios, ignoring the type of feeder.
b. Compute the mean and median separately for each type of feeder.
c. Using your results from parts (a) and (b), comment on the relative sensitivity of the mean and median to extreme values in a data set.
d. Which measure, mean or median, would you recommend as the most appropriate measure of the DDE to PCB level for both types of feeders? Explain your answer.

Med. **3.22** A study of the survival times, in days, of skin grafts on burn patients was examined in Woolson and Lachenbruch [*Biometrika* (1980) 67:597–606]. Two of the patients left the study prior to the failure of their grafts. The survival time for these individuals is some number greater than the reported value.

Survival time (days): 37, 19, 57*, 93, 16, 22, 20, 18, 63, 29, 60*

(The "*" indicates that the patient left the study prior to failure of the graft; values given are for the day the patient left the study.)

a. Calculate the measures of center (if possible) for the 11 patients.
b. If the survival times of the two patients who left the study were obtained, how would these new values change the values of the summary statistics calculated in (a)?

Engin. **3.23** A study of the reliability of diesel engines was conducted on 14 engines. The engines were run in a test laboratory. The time (in days) until the engine failed is given here. The study was terminated after 300 days. For those engines that did not fail during the study period, an asterisk is placed by the number 300. Thus, for these engines, the time to failure is some value greater than 300.

Failure time (days): 130, 67, 300*, 234, 90, 256, 87, 120, 201, 178, 300*, 106, 289, 74

a. Calculate the measures of center for the 14 engines.
b. What are the implications of computing the measures of center when some of the exact failure times are not known?

Gov. **3.24** Effective tax rates (per $100) on residential property for three groups of large cities, ranked by residential property tax rate, are shown in the following table.

Group 1	Rate	Group 2	Rate	Group 3	Rate
Detroit, MI	4.10	Burlington, VT	1.76	Little Rock, AR	1.02
Milwaukee, WI	3.69	Manchester, NH	1.71	Albuquerque, NM	1.01
Newark, NJ	3.20	Fargo, ND	1.62	Denver, CO	.94
Portland, OR	3.10	Portland ME	1.57	Las Vegas, NV	.88
Des Moines, IA	2.97	Indianapolis, IN	1.57	Oklahoma City, OK	.81
Baltimore, MD	2.64	Wilmington, DE	1.56	Casper, WY	.70
Sioux Falls, IA	2.47	Bridgeport, CT	1.55	Birmingham, AL	.70
Providence, RI	2.39	Chicago, IL	1.55	Phoenix, AZ	.68
Philadelphia, PA	2.38	Houston, TX	1.53	Los Angeles, CA	.64
Omaha, NE	2.29	Atlanta, GA	1.50	Honolulu, HI	.59

Source: Government of the District of Columbia, Department of Finance and Revenue, *Tax Rates and Tax Burdens in the District of Columbia: A Nationwide Comparison,* annual.

a. Compute the mean, median, and mode separately for the three groups.
b. Compute the mean, median, and mode for the complete set of 30 measurements.
c. What measure or measures best summarize the center of these distributions? Explain.

3.25 Refer to Exercise 3.24. Average the three group means, the three group medians, and the three group modes, and compare your results to those of part (b). Comment on your findings.

3.5 Describing Data on a Single Variable: Measures of Variability

Engin. **3.26** Pushing economy and wheelchair-propulsion technique were examined for eight wheelchair racers on a motorized treadmill in a paper by Goosey and Campbell [*Adapted Physical Activity Quarterly* (1998) 15:36–50]. The eight racers had the following years of racing experience:

Racing experience (years): 6, 3, 10, 4, 4, 2, 4, 7

a. Verify that the mean years' experience is 5 years. Does this value appear to adequately represent the center of the data set?
b. Verify that $\Sigma_i(y - \bar{y})^2 = \Sigma_i(y - 5)^2 = 46$.
c. Calculate the sample variance and standard deviation for the experience data. How would you interpret the value of the standard deviation relative to the sample mean?

3.27 In the study described in Exercise 3.26, the researchers also recorded the ages of the eight racers.

Age (years): 39, 38, 31, 26, 18, 36, 20, 31

a. Calculate the sample standard deviation of the eight racers' ages.
b. Why would you expect the standard deviation of the racers' ages to be larger than the standard deviation of their years of experience?

Engin. **3.28** For the data in Exercise 3.26,
a. Calculate the coefficient of variation (CV) for both the racer's age and their years of experience. Are the two CVs relatively the same? Compare their relative sizes to the relative sizes of their standard deviations.
b. Estimate the standard deviations for both the racer's age and their years of experience by dividing the ranges by 4. How close are these estimates to the standard deviations calculated in Exercise 3.27?

Med. **3.29** The treatment times (in minutes) for patients at a health clinic are as follows:

21	20	31	24	15	21	24	18	33	8
26	17	27	29	24	14	29	41	15	11
13	28	22	16	12	15	11	16	18	17
29	16	24	21	19	7	16	12	45	24
21	12	10	13	20	35	32	22	12	10

Construct the quantile plot for the treatment times for the patients at the health clinic.

a. Find the 25th percentile for the treatment times and interpret this value.

b. The health clinic advertises that 90% of all its patients have a treatment time of 40 minutes or less. Do the data support this claim?

Env. **3.30** To assist in estimating the amount of lumber in a tract of timber, an owner decided to count the number of trees with diameters exceeding 12 inches in randomly selected 50 × 50-foot squares. Seventy 50 × 50 squares were randomly selected from the tract and the number of trees (with diameters in excess of 12 inches) were counted for each. The data are as follows:

7	8	6	4	9	11	9	9	9	10
9	8	11	5	8	5	8	8	7	8
3	5	8	7	10	7	8	9	8	11
10	8	9	8	9	9	7	8	13	8
9	6	7	9	9	7	9	5	6	5
6	9	8	8	4	4	7	7	8	9
10	2	7	10	8	10	6	7	7	8

a. Construct a relative frequency histogram to describe these data.

b. Calculate the sample mean $\bar{y}$ as an estimate of μ, the mean number of timber trees with diameter exceeding 12 inches for all 50 × 50 squares in the tract.

c. Calculate s for the data. Construct the intervals $(\bar{y} \pm s)$, $(\bar{y} \pm 2s)$, and $(\bar{y} \pm 3s)$. Count the percentages of squares falling in each of the three intervals, and compare these percentages with the corresponding percentages given by the Empirical Rule.

Bus. **3.31** *Consumer Reports* in its June 1998 issue reports on the typical daily room rate at six luxury and nine budget hotels. The room rates are given in the following table.

Luxury Hotel	$175	$180	$120	$150	$120	$125			
Budget Hotel	$50	$50	$49	$45	$36	$45	$50	$50	$40

a. Compute the mean and standard deviation of the room rates for both luxury and budget hotels.

b. Verify that luxury hotels have a more variable room rate than budget hotels.

c. Give a practical reason why the luxury hotels are more variable than the budget hotels.

d. Might another measure of variability be better to compare luxury and budget hotel rates? Explain.

Env. **3.32** Many marine phanerogam species are highly sensitive to changes in environmental conditions. In the article "*Posidonia oceanica:* A biological indicator of past and present mercury contamination in the Mediterranean Sea" [*Marine Environmental Research,* 45:101–111], the researchers report the mercury concentrations over a period of about 20 years at several locations in the Mediterranean Sea. Samples of *Posidonia oceanica* were collected by scuba diving at a depth of 10 meters. For each site, 45 orthotropic shoots were sampled and the mercury concentration was determined. The average mercury concentration is recorded in the following table for each of the sampled years.

Mercury Concentration (ng/g dry weight)		
Year	Site 1 Calvi	Site 2 Marseilles-Coriou
1992	14.8	70.2
1991	12.9	160.5
1990	18.0	102.8
1989	8.7	100.3
1988	18.3	103.1
1987	10.3	129.0
1986	19.3	156.2
1985	12.7	117.6
1984	15.2	170.6
1983	24.6	139.6
1982	21.5	147.8
1981	18.2	197.7
1980	25.8	262.1
1979	11.0	123.3
1978	16.5	363.9
1977	28.1	329.4
1976	50.5	542.6
1975	60.1	369.9
1974	96.7	705.1
1973	100.4	462.0
1972	*	556.1
1971	*	461.4
1970	*	628.8
1969	*	489.2

a. Generate a time-series plot of the mercury concentrations and place lines for both sites on the same graph. Comment on any trends in the lines across the years of data. Are the trends similar for both sites?
b. Select the most appropriate measure of center for the mercury concentrations. Compare the center for the two sites.
c. Compare the variability in mercury concentrations at the two sites. Use the CV in your comparison and explain why it is more appropriate than using the standard deviations.
d. When comparing the center and variability of the two sites, should the years 1969–1972 be used for site 2?

3.6 The Boxplot

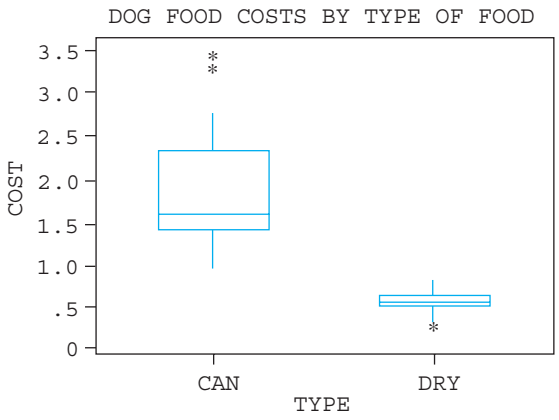
Basic **3.33** Find the median and the lower and upper quartiles for the following measurements: 13, 21, 9, 15, 13, 17, 21, 9, 19, 23, 11, 9, 21.

Med. **3.34** The number of persons who volunteered to give a pint of blood at a central donor center was recorded for each of 20 successive Fridays. The data are shown here:

320	370	386	334	325	315	334	301	270	310
274	308	315	368	332	260	295	356	333	250

a. Construct a stem-and-leaf plot.
b. Construct a boxplot and describe the shape of the distribution of the number of persons donating blood.

Bus. **3.35** *Consumer Reports* in its May 1998 issue provides cost per daily feeding for 28 brands of dry dog food and 23 brands of canned dog food. Using the Minitab computer program, the side-by-side boxplot for these data follow.



a. From these graphs, determine the median, lower quartile, and upper quartile for the daily costs of both dry and canned dog food.
b. Comment on the similarities and differences in the distributions of daily costs for the two types of dog food.

3.7 Summarizing Data from More Than One Variable: Graphs and Correlation

Soc. **3.36** For the homeownership rates given in Exercise 3.11, construct separate boxplots for the years 1985, 1996, and 2002.
a. Describe the distributions of homeownership rates for each of the 3 years.
b. Compare the descriptions given in part (a) to the descriptions given in Exercise 3.11.

Soc. **3.37** Compute the mean, median, and standard deviation for the homeownership rates given in Exercise 3.11.
a. Compare the mean and median for the 3 years of data. Which value, mean or median, is most appropriate for these data sets? Explain your answers.
b. Compare the degree of variability in homeownership rates over the 3 years.

Soc. **3.38** For the boxplots constructed for the homeownership rates given in Exercise 3.36, place the three boxplots on the same set of axes.
a. Use this side-by-side boxplot to discuss changes in the median homeownership rate over the 3 years.
b. Use this side-by-side boxplot to discuss changes in the variation in these rates over the 3 years.
c. Are there any states that have extremely low homeownership rates?
d. Are there any states that have extremely high homeownership rates?

Soc. **3.39** In the paper "Demographic implications of socioeconomic transition among the tribal populations of Manipur, India" [*Human Biology* (1998) 70(3): 597–619], the authors describe the tremendous changes that have taken place in all the tribal populations of Manipur, India, since the beginning of the twentieth century. The tribal populations of Manipur are in the process of socioeconomic transition from a traditional subsistence economy to a market-oriented economy. The following table displays the relation between literacy level and subsistence group for a sample of 614 married men and women in Manipur, India.

	Literacy Level		
Subsistence Group	Illiterate	Primary Schooling	At Least Middle School
Shifting cultivators	114	10	45
Settled agriculturists	76	2	53
Town dwellers	93	13	208

a. Graphically depict the data in the table using a stacked bar graph.
b. Do a percentage comparison based on the row and column totals. What conclusions do you reach with respect to the relation between literacy and subsistence group?

Engin. **3.40** In the manufacture of soft contact lenses, the power (the strength) of the lens needs to be very close to the target value. In the paper "An ANOM-type test for variances from normal populations" [*Technometrics* (1997) 39:274–283], a comparison of several suppliers is made relative to the consistency of the power of the lens. The following table contains the deviations from the target power value of lenses produced using materials from three different suppliers:

Supplier	Deviations from Target Power Value								
1	189.9	191.9	190.9	183.8	185.5	190.9	192.8	188.4	189.0
2	156.6	158.4	157.7	154.1	152.3	161.5	158.1	150.9	156.9
3	218.6	208.4	187.1	199.5	202.0	211.1	197.6	204.4	206.8

a. Compute the mean and standard deviation for the deviations of each supplier.
b. Plot the sample deviation data.
c. Describe the deviation from specified power for the three suppliers.
d. Which supplier appears to provide material that produces lenses having power closest to the target value?

Bus. **3.41** The federal government keeps a close watch on money growth versus targets that have been set for that growth. We list two measures of the money supply in the United States, M2 (private checking deposits, cash, and some savings) and M3 (M2 plus some investments), which are given here for 20 consecutive months.

	Money Supply (in trillions of dollars)			Money Supply (in trillions of dollars)	
Month	M2	M3	Month	M2	M3
1	2.25	2.81	11	2.43	3.05
2	2.27	2.84	12	2.42	3.05
3	2.28	2.86	13	2.44	3.08
4	2.29	2.88	14	2.47	3.10
5	2.31	2.90	15	2.49	3.10
6	2.32	2.92	16	2.51	3.13
7	2.35	2.96	17	2.53	3.17
8	2.37	2.99	18	2.53	3.18
9	2.40	3.02	19	2.54	3.19
10	2.42	3.04	20	2.55	3.20

a. Would a scatterplot describe the relation between M2 and M3?
b. Construct a scatterplot. Is there an obvious relation?

3.42 Refer to Exercise 3.41. What other data plot might be used to describe and summarize these data? Make the plot and interpret your results.

Supplementary Exercises

Env. **3.43** To control the risk of severe core damage during a commercial nuclear power station blackout accident, the reliability of the emergency diesel generators to start on demand must be maintained at a high level. The paper "Empirical Bayes estimation of the reliability of nuclear-power emergency diesel generators" [*Technometrics* (1996) 38:11–23] contains data on the failure history of seven nuclear power plants. The following data are the number of successful demands between failures for the diesel generators at one of these plants from 1982 to 1988.

28	50	193	55	4	7	147	76	10	0	10	84	0	9	1	0	62
26	15	226	54	46	128	4	105	40	4	273	164	7	55	41	26	6

(*Note:* The failure of the diesel generator does not necessarily result in damage to the nuclear core because all nuclear power plants have several emergency diesel generators.)

a. Calculate the mean and median of the successful demands between failures.
b. Which measure appears to best represent the center of the data?
c. Calculate the range and standard deviation, s.
d. Use the range approximation to estimate s. How close is the approximation to the true value?
e. Construct the intervals

$$\bar{y} \pm s \qquad \bar{y} \pm 2s \qquad \bar{y} \pm 3s$$

Count the number of demands between failures falling in each of the three intervals. Convert these numbers to percentages and compare your results to the Empirical Rule.
f. Why do you think the Empirical Rule and your percentages do not match well?

Edu. **3.44** The College of Dentistry at the University of Florida has made a commitment to develop its entire curriculum around the use of self-paced instructional materials such as videotapes, slide tapes, and syllabi. It is hoped that each student will proceed at a pace commensurate with his or her ability and that the instructional staff will have more free time for personal consultation in student–faculty interaction. One such instructional module was developed and tested on the first 50 students proceeding through the curriculum. The following measurements represent the number of hours it took these students to complete the required modular material.

16	8	33	21	34	17	12	14	27	6
33	25	16	7	15	18	25	29	19	27
5	12	29	22	14	25	21	17	9	4
12	15	13	11	6	9	26	5	16	5
9	11	5	4	5	23	21	10	17	15

a. Calculate the mode, the median, and the mean for these recorded completion times.
b. Guess the value of s.
c. Compute s by using the shortcut formula and compare your answers to that of part (b).
d. Would you expect the Empirical Rule to describe adequately the variability of these data? Explain.

Bus. **3.45** The February 1998 issue of *Consumer Reports* provides data on the price of 24 brands of paper towels. The prices are given in both cost per roll and cost per sheet because the brands had varying numbers of sheets per roll.

Brand	Price per Roll	Number of Sheets per Roll	Cost per Sheet
1	1.59	50	.0318
2	0.89	55	.0162
3	0.97	64	.0152
4	1.49	96	.0155
5	1.56	90	.0173
6	0.84	60	.0140

(*continued*)

Brand	Price per Roll	Number of Sheets per Roll	Cost per Sheet
7	0.79	52	.0152
8	0.75	72	.0104
9	0.72	80	.0090
10	0.53	52	.0102
11	0.59	85	.0069
12	0.89	80	.0111
13	0.67	85	.0079
14	0.66	80	.0083
15	0.59	80	.0074
16	0.76	80	.0095
17	0.85	85	.0100
18	0.59	85	.0069
19	0.57	78	.0073
20	1.78	180	.0099
21	1.98	180	.0011
22	0.67	100	.0067
23	0.79	100	.0079
24	0.55	90	.0061

a. Compute the standard deviation for both the price per roll and the price per sheet.
b. Which is more variable, price per roll or price per sheet?
c. In your comparison in part (b), should you use s or CV? Justify your answer.

3.46 Refer to Exercise 3.45. Use a scatterplot to plot the price per roll and number of sheets per roll.
a. Do the 24 points appear to fall on a straight line?
b. If not, is there any other relation between the two prices?
c. What factors may explain why the ratio of price per roll to number of sheets is not a constant?

3.47 Construct boxplots for both price per roll and number of sheets per roll. Are there any "unusual" brands in the data?

Env.

3.48 The paper "Conditional simulation of waste-site performance" [*Technometrics* (1994) 36: 129–161] discusses the evaluation of a pilot facility for demonstrating the safe management, storage, and disposal of defense-generated, radioactive, transuranic waste. Researchers have determined that one potential pathway for release of radionuclides is through contaminant transport in groundwater. Recent focus has been on the analysis of transmissivity, a function of the properties and the thickness of an aquifer that reflects the rate at which water is transmitted through the aquifer. The following table contains 41 measurements of transmissivity, T, made at the pilot facility.

9.354	6.302	24.609	10.093	0.939	354.81	15399.27	88.17	1253.43	0.75	312.10
1.94	3.28	1.32	7.68	2.31	16.69	2772.68	0.92	10.75	0.000753	
1.08	741.99	3.23	6.45	2.69	3.98	2876.07	12201.13	4273.66	207.06	
2.50	2.80	5.05	3.01	462.38	5515.69	118.28	10752.27	956.97	20.43	

a. Draw a relative frequency histogram for the 41 values of T.
b. Describe the shape of the histogram.
c. When the relative frequency histogram is highly skewed to the right, the Empirical Rule may not yield very accurate results. Verify this statement for the data given.
d. Data analysts often find it easier to work with mound-shaped relative frequency histograms. A transformation of the data will sometimes achieve this shape. Replace the given 41 T values with the logarithm base 10 of the values and reconstruct the relative frequency histogram. Is the shape more mound-shaped than the original data? Apply

the Empirical Rule to the transformed data and verify that it yields more accurate results than it did with the original data.

Soc. **3.49** A random sample of 90 standard metropolitan statistical areas (SMSAs) was studied to obtain information on murder rates. The murder rate (number of murders per 100,000 people) was recorded, and these data are summarized in the following frequency table.

Class Interval	f_i	Class Interval	f_i
−.5–1.5	2	13.5–15.5	9
1.5–3.5	18	15.5–17.5	4
3.5–5.5	15	17.5–19.5	2
5.5–7.5	13	19.5–21.5	1
7.5–9.5	9	21.5–23.5	1
9.5–11.5	8	23.5–25.5	1
11.5–13.5	7		

Construct a relative frequency histogram for these data.

3.50 Refer to the data of Exercise 3.49.
- **a.** Compute the sample median and the mode.
- **b.** Compute the sample mean.
- **c.** Which measure of central tendency would you use to describe the center of the distribution of murder rates?

3.51 Refer to the data of Exercise 3.49.
- **a.** Compute the interquartile range.
- **b.** Compute the sample standard deviation.

3.52 Using the homeownership data in Exercise 3.11, construct a quantile plot for both years.
- **a.** Find the 20th percentile for the homeownership percentage and interpret this value for the 1996 data.
- **b.** Congress wants to designate those states that have the highest homeownership percentage in 1996. Which states fall into the upper 10th percentile of homeownership rates?
- **c.** Similarly identify those states that fall into the upper 10th percentile of homeownership rates during 1985. Are these states different from the states in this group during 1996?

Gov. **3.53** Per capita expenditure (dollars) for health and hospital services by state are shown here.

Dollars	f
45–59	1
60–74	4
75–89	9
90–104	9
105–119	12
120–134	6
135–149	4
150–164	1
165–179	3
180–194	0
195–209	1
Total	50

- **a.** Construct a relative frequency histogram.
- **b.** Compute approximate values for $\bar{y}$ and s from the grouped expenditure data.

Engin. **3.54** The Insurance Institute for Highway Safety published data on the total damage suffered by compact automobiles in a series of controlled, low-speed collisions. The data, in dollars, with brand names removed are as follows:

361	393	430	543	566	610	763	851
886	887	976	1,039	1,124	1,267	1,328	1,415
1,425	1,444	1,476	1,542	1,544	2,048	2,197	

a. Draw a histogram of the data using six or seven categories.
b. On the basis of the histogram, what would you guess the mean to be?
c. Calculate the median and mean.
d. What does the relation between the mean and median indicate about the shape of the data?

Soc. **3.55** Data are collected on the weekly expenditures of a sample of urban households on food (including restaurant expenditures). The data, obtained from diaries kept by each household, are grouped by number of members of the household. The expenditures are as follows:

1 member:	67	62	168	128	131	118	80	53	99	68	
	76	55	84	77	70	140	84	65	67	183	
2 members:	129	116	122	70	141	102	120	75	114	81	106	95
	94	98	85	81	67	69	119	105	94	94	92	
3 members:	79	99	171	145	86	100	116	125				
	82	142	82	94	85	191	100	116				
4 members:	139	251	93	155	158	114	108					
	111	106	99	132	62	129	91					
5+ members:	121	128	129	140	206	111	104	109	135	136		

a. Calculate the mean expenditure separately for each number of members.
b. Calculate the median expenditure separately for each number of members.

3.56 Answer the following for the data in Exercise 3.55:
a. Calculate the mean of the combined data, using the raw data.
b. Can the combined mean be calculated from the means for each number of members?
c. Calculate the median of the combined data using the raw data.
d. Can the combined median be calculated from the medians for each number of members?

Gov. **3.57** Federal authorities have destroyed considerable amounts of wild and cultivated marijuana plants. The following table shows the number of plants destroyed and the number of arrests for a 12-month period for 15 states.

State	Plants	Arrests
1	110,010	280
2	256,000	460
3	665	6
4	367,000	66
5	4,700,000	15
6	4,500	8
7	247,000	36
8	300,200	300
9	3,100	9
10	1,250	4
11	3,900,200	14
12	68,100	185
13	450	5
14	2,600	4
15	205,844	33

a. Discuss the appropriateness of using the sample mean to describe these two variables.

b. Compute the sample mean, 10% trimmed mean, and 20% trimmed mean. Which trimmed mean seems more appropriate for each variable? Why?

c. Does there appear to be a relation between the number of plants destroyed and the number of arrests? How might you examine this question? What other variable(s) might be related to the number of plants destroyed?

Bus. **3.58** The most widely reported index of the performance of the New York Stock Exchange (NYSE) is the Dow Jones Industrial Average (DJIA). This index is computed from the stock prices of 30 companies. When the DJIA was invented in 1896, the index was the average price of 12 stocks. The index was modified over the years as new companies were added and dropped from the index and was also altered to reflect when a company splits its stock. The closing New York Stock Exchange (NYSE) prices for the 30 components (as of May 2004) of the DJIA are given in the following table.

a. Compute the average price of the 30 stock prices in the DJIA.

b. Compute the range of the 30 stock prices in the DJIA.

c. The DJIA is no longer an average; the name includes the word "average" only for historical reasons. The index is computed by summing the stock prices and dividing by a constant, which is changed as stocks are added or removed from the index and when stocks split.

$$\text{DJIA} = \frac{\Sigma_{i=1}^{30} y_i}{C}$$

where y_i is the closing price for stock i, and $C = .1409017$. Using the stock prices given, compute the DJIA for May 27, 2004.

d. The DJIA is a summary of data. Does the DJIA provide information about a population using sampled data? If so, to what population? Is the sample a random sample?

Components of DJIA

Company	Percent of DJIA	NYSE Stock Price (5/27/04)
3M Co.	5.9078	84.95
Alcoa Inc.	2.1642	31.12
Altria Group Inc.	3.3673	48.42
American Express Co.	3.5482	51.02
American International Group Inc.	5.0628	72.8
Boeing Co.	3.213	46.2
Caterpillar Inc.	5.2277	75.17
Citigroup Inc.	3.2352	46.52
Coca-Cola Co.	3.569	51.32
E.I. DuPont de Numours & Co.	3.0057	43.22
Exxon Mobil Corp.	3.0161	43.37
General Electric Co.	2.174	31.26
General Motors Corp.	3.1601	45.44
Hewlett-Packard Co.	1.4702	21.14
Home Depot Inc.	2.4925	35.84
Honeywell International Inc.	2.3499	33.79
Intel Corp.	1.9785	28.45
International Business Machines Corp.	6.1609	88.59
J.P. Morgan Chase & Co.	2.5697	36.95
Johnson & Johnson	3.8799	55.79
McDonald's Corp.	1.8269	26.27
Merck & Co. Inc.	3.2985	47.43

(*continued*)

Components of DJIA

Company	Percent of DJIA	NYSE Stock Price (5/27/04)
Microsoft Corp.	1.8214	26.19
Pfizer Inc.	2.4619	35.4
Procter & Gamble Co.	7.5511	108.58
SBC Communications Inc.	1.6586	23.85
United Technologies Corp.	5.848	84.09
Verizon Communications Inc.	2.4396	35.08
Wal-Mart Stores Inc.	3.8924	55.97
Walt Disney Co.	1.6489	23.71

H.R. **3.59** As one part of a review of middle-manager selection procedures, a study was made of the relation between hiring source (promoted from within, hired from related business, hired from unrelated business) and the 3-year job history (additional promotion, same position, resigned, dismissed). The data for 120 middle managers follow.

	Source			
Job History	Within Firm	Related Business	Unrelated Business	Total
Promoted	13	4	10	27
Same position	32	8	18	58
Resigned	9	6	10	25
Dismissed	3	3	4	10
Total	57	21	42	120

a. Calculate job-history percentages within each source.
b. Would you say that there is a strong dependence between source and job history?

Env. **3.60** A survey was taken of 150 residents of major coal-producing states, 200 residents of major oil- and natural-gas–producing states, and 450 residents of other states. Each resident chose a most preferred national energy policy. The results are shown in the following SPSS printout.

		STATE		
	COUNT ROW PCT COL PCT TOT PCT	COAL	OIL AND GAS	OTHER	ROW TOTAL
OPINION		62	25	102	189
COAL ENCOURAGED		32.8	13.2	54.0	23.6
		41.3	12.5	22.7	
		7.8	3.1	12.8	
		3	12	26	41
FUSION DEVELOP		7.3	29.3	63.4	5.1
		2.0	6.0	5.8	
		0.4	1.5	3.3	
		8	6	22	36
NUCLEAR DEVELOP		22.2	16.7	61.1	4.5
		5.3	3.0	4.9	
		1.0	0.8	2.8	
		19	79	53	151
OIL DEREGULATION		12.6	52.3	35.1	18.9
		12.7	39.5	11.8	
		2.4	9.9	6.6	

```
                                    58          78         247         383
                SOLAR DEVELOP     15.1        20.4        64.5        47.9
                                  38.7        39.0        54.9
                                   7.3         9.8        30.9

                          COLUMN   150         200         450         800
                          TOTAL   18.8        25.0        56.3       100.0

     CHI SQUARE = 106.19406 WITH 8 DEGREES OF FREEDOM SIGNIFICANCE = 0.0000
     CRAMER'S V = 0.25763
     CONTINGENCY COEFFICIENT = 0.34233
     LAMBDA = 0.01199 WITH OPINION DEPENDENT, = 0.07429 WITH STATE DEPENDENT.
```

a. Interpret the values 62, 32.8, 41.3, and 7.8 in the upper left cell of the cross tabulation. Note the labels COUNT, ROW PCT, COL PCT, and TOT PCT at the upper left corner.
b. Which of the percentage calculations seems most meaningful to you?
c. According to the percentage calculations you prefer, does there appear to be a strong dependence between state and opinion?

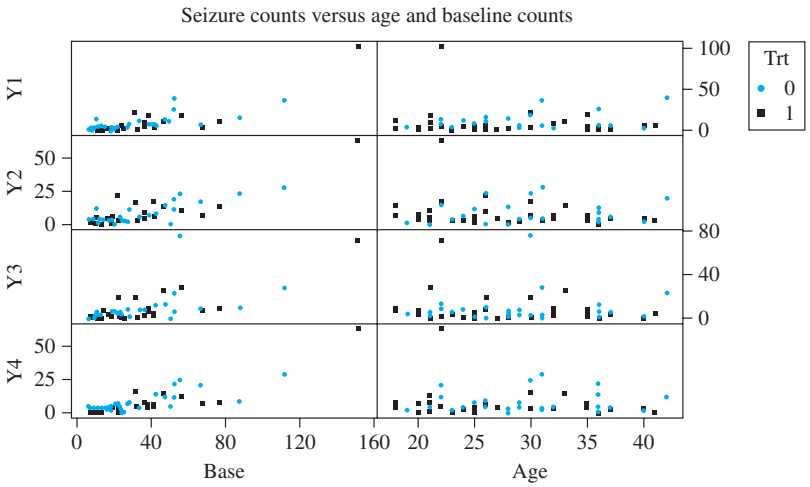
Bus. **3.61** A municipal workers' union that represents sanitation workers in many small midwestern cities studied the contracts that were signed in the previous years. The contracts were subdivided into those settled by negotiation without a strike, those settled by arbitration without a strike, and all those settled after a strike. For each contract, the first-year percentage wage increase was determined. Summary figures follow.

Contract Type	Negotation	Arbitration	Poststrike
Mean percentage wage increase	8.20	9.42	8.40
Variance	0.87	1.04	1.47
Standard deviation	0.93	1.02	1.21
Sample size	38	16	6

Does there appear to be a relationship between contract type and mean percent wage increase? If you were management rather than union affiliated, which posture would you take in future contract negotiations?

Med. **3.62** Refer to the epilepsy study data in Table 3.18. Examine the scatterplots of Y_1, Y_2, Y_3, and Y_4 versus baseline counts and age given here.

a. Does there appear to be a difference in the relationships between the seizure counts ($Y_1 - Y_4$) and either the baseline counts or age when considering the two groups (treatment and placebo)?
b. Describe the type of apparent differences, if any, that you found in (a).



Med. **3.63** The correlations computed for the six variables in the epilepsy study are given here. Do the sizes of the correlation coefficients reflect the relationships displayed in the graphs given in Exercise 3.62? Explain your answer.

	Placebo Group				
	Y_1	Y_2	Y_3	Y_4	Base
Y_2	.782				
Y_3	.507	.661			
Y_4	.675	.780	.676		
Base	.744	.831	.493	.818	
Age	.326	.108	.113	.117	.033

	Treatment Group				
	Y_1	Y_2	Y_3	Y_4	Base
Y_2	.907				
Y_3	.912	.925			
Y_4	.971	.947	.952		
Base	.854	.845	.834	.876	
Age	−.141	−.243	−.194	−.197	−.343

Med. **3.64** An examination of the scatterplots reveals one patient with a very large value for baseline count and all subsequent counts. The patient has ID 207.

a. Predict the effect of removing the patient with ID 207 from the data set on the size of the correlations in the treatment group.

b. Using a computer program, compute the correlations with patient ID 207 removed from the data. Do the values confirm your predictions?

Med. **3.65** Refer to the research study concerning the effect of social factors on reading and math scores. We justified studying just the reading scores because there was a strong correlation between reading and math scores. Construct the same plots for the math scores as were constructed for the reading scores.

a. Is there support for the same conclusions for the math scores as obtained for the reading scores?

b. If the conclusions are different, why do you suppose this has happened?

Med. **3.66** In the research study concerning the effect of social factors on reading and math scores, we found a strong negative correlation between %minority and %poverty and reading scores.

a. Why is it not possible to conclude that large relative values for %minority and %poverty in a school results in lower reading scores for children in these social classes?

b. List several variables related to the teachers and students in the schools which may be important in explaining why low reading scores were strongly associated with schools having large values of %minority and %poverty.

Soc. **3.67** In the January 2004 issue of *Consumer Reports* an article titled "Cut the fat" described some of the possible problems in the diets of the U.S. public. The following table gives data on the increase in daily calories in the food supply per person. Construct a time-series plot to display the increase in calorie intake.

Year	1970	1975	1980	1985	1990	1995	2000
Calories	3,300	3,200	3,300	3,500	3,600	3,700	3,900

a. Describe the trend in calorie intake over the 30 years.

b. What would you predict the calorie intake was in 2005? Justify your answer by explaining any assumptions you are making about calorie intake.

Soc. **3.68** In the January 2004 issue of *Consumer Reports* an article titled "Cut the fat" described some of the possible problems in the diets of the U.S. public. The following table gives data on the increase in pounds of added sugar produced per person. Construct a time-series plot to display the increase in sugar production.

Year	1970	1975	1980	1985	1990	1995	2000
Pounds of Sugar	119	114	120	128	132	144	149

a. Describe the trend in sugar production over the 30 years.
b. Compute the correlation coefficient between calorie intake (using the data in Exercise 3.67) and sugar production. Is there strong evidence that the increase in sugar production is causing the increased calorie intake by the U.S. public?

Med. **3.69** Certain types of diseases tend to occur in clusters. In particular, persons affected with AIDS, syphilis, and tuberculosis may have some common characteristics and associations which increase their chances of contracting these diseases. The following table lists the number of reported cases by state in 2001.

State	AIDS	Syphilis	Tuber.	State	AIDS	Syphilis	Tuber.
AL	438	720	265	MT	15	0	20
AK	18	9	54	NE	74	16	40
AZ	540	1147	289	NV	252	62	96
AR	199	239	162	NH	40	20	20
CA	4315	3050	3332	NJ	1756	1040	530
CO	288	149	138	NM	143	73	54
CT	584	165	121	NY	7476	3604	1676
DE	248	79	33	NC	942	1422	398
DC	870	459	74	ND	3	2	6
FL	5138	2914	1145	OH	581	297	306
GA	1745	1985	575	OK	243	288	194
HI	124	41	151	OR	259	48	123
ID	19	11	9	PA	1840	726	350
IL	1323	1541	707	RI	103	39	60
IN	378	529	115	SC	729	913	263
IA	90	44	43	SD	25	1	13
KS	98	88	63	TN	602	1478	313
KY	333	191	152	TX	2892	3660	1643
LA	861	793	294	UT	124	25	35
ME	48	16	20	VT	25	8	7
MD	1860	937	262	VA	951	524	306
MA	765	446	270	WA	532	174	261
MI	548	1147	330	WV	100	7	32
MN	157	132	239	WI	193	131	86
MS	418	653	154	WY	5	4	3
MO	445	174	157	All States	41,868	32,221	15,989

a. Construct a scatterplot of the number of AIDS cases versus the number of syphilis cases.
b. Compute the correlation between the number of AIDS cases and the number of syphilis cases.
c. Does the value of the correlation coefficient reflect the degree of association shown in the scatterplot?
d. Why do you think there may be a correlation between these two diseases?

Med. **3.70** Refer to the data in Exercise 3.69.
- **a.** Construct a scatterplot of the number of AIDS cases versus the number of tuberculosis cases.
- **b.** Compute the correlation between the number of AIDS cases and the number of tuberculosis cases.
- **c.** Why do you think there may be a correlation between these two diseases?

Med. **3.71** Refer to the data in Exercise 3.69.
- **a.** Construct a scatterplot of the number of syphilis cases versus the number of tuberculosis cases.
- **b.** Compute the correlation between the number of syphilis cases and the number of tuberculosis cases.
- **c.** Why do you think there may be a correlation between these two diseases?

Med. **3.72** Refer to the data in Exercise 3.69.
- **a.** Construct a quantile plot of the number of syphilis cases.
- **b.** From the quantile plot, determine the 90th percentile for the number of syphilis cases.
- **c.** Identify the states having number of syphilis cases that are above the 90th percentile.

Med. **3.73** Refer to the data in Exercise 3.69.
- **a.** Construct a quantile plot of the number of tuberculosis cases.
- **b.** From the quantile plot, determine the 90th percentile for the number of tuberculosis cases.
- **c.** Identify the states having number of tuberculosis cases that are above the 90th percentile.

Med. **3.74** Refer to the data in Exercise 3.69.
- **a.** Construct a quantile plot of the number of AIDS cases.
- **b.** From the quantile plot, determine the 90th percentile for the number of AIDS cases.
- **c.** Identify the states having number of AIDS cases that are above the 90th percentile.

Med. **3.75** Refer to the results from Exercises 3.72–3.74.
- **a.** How many states had number of AIDS, tuberculosis, and syphilis cases all above the 90th percentiles?
- **b.** Identify these states and comment on any common elements between the states.
- **c.** How could the U.S. government apply the results from Exercises 3.69–3.75 in making public health policy?

Med. **3.76** In the article "Viral load and heterosexual transmission of human immunodeficiency virus type 1" [*New England Journal of Medicine* (2000) 342:921–929], studied the question of whether people with high levels of HIV-1 are significantly more likely to transmit HIV to their uninfected partners. Measurements follow of the amount of HIV-1 RNA levels in the group whose partners who were initially uninfected became HIV positive during the course of the study: values are given in units of RNA copies/mL.

79725, 12862, 18022, 76712, 256440, 14013, 46083, 6808, 85781, 1251, 6081, 50397, 11020, 13633 1064, 496433, 25308, 6616, 11210, 13900

- **a.** Determine the mean, median, and standard deviation.
- **b.** Find the 25th, 50th, and 75th percentiles.
- **c.** Plot the data in a boxplot and histogram.
- **d.** Describe the shape of the distribution.

Med. **3.77** In many statistical procedures, it is often advantageous to have a symmetric distribution. When the data have a histogram that is highly right-skewed, it is often possible to obtain a symmetric distribution by taking a transformation of the data. For the data in Exercise 3.76, take the natural logarithm of the data and answer the following questions.
- **a.** Determine the mean, median, and standard deviation.
- **b.** Find the 25th, 50th, and 75th percentiles.
- **c.** Plot the data in a boxplot and histogram.
- **d.** Did the logarithm transformation result in a somewhat symmetric distribution?

Env. **3.78** PCBs are a class of chemicals often found near the disposal of electrical devices. PCBs tend to concentrate in human fat and have been associated with numerous health problems. In the article "Some other persistent organochlorines in Japanese human adipose tissue" [*Environmental Health Perspective,* Vol. 108, pp. 599–603], researchers examined the concentrations of PCB (ng/g) in the fat of a group of adults. They detected the following concentrations:

1800, 1800, 2600, 1300, 520, 3200, 1700, 2500, 560, 930, 2300, 2300, 1700, 720

a. Determine the mean, median, and standard deviation.
b. Find the 25th, 50th, and 75th percentiles.
c. Plot the data in a boxplot.
d. Would it be appropriate to apply the Empirical Rule to these data? Why or why not?

Agr. **3.79** The focal point of an agricultural research study was the relationship between when a crop is planted and the amount of crop harvested. If a crop is planted too early or too late, farmers may fail to obtain optimal yield and hence not make a profit. An ideal date for planting is set by the researchers, and the farmers then record the number of days either before or after the designated date. In the following data set, D is the number of days from the ideal planting date and Y is the yield (in bushels per acre) of a wheat crop:

D	−19	−18	−15	−12	−9	−6	−4	−3	−1	0
Y	30.7	29.7	44.8	41.4	48.1	42.8	49.9	46.9	46.4	53.5
D	1	3	6	8	12	15	17	19	21	24
Y	55.0	46.9	44.1	50.2	41.0	42.8	36.5	35.8	32.2	23.3

a. Plot the data in a scatterplot.
b. Describe the relationship between the number of days from the optimal planting date and the wheat yield.
c. Calculate the correlation coefficient between days from optimal planting and yield.
d. Explain why the correlation coefficient is relatively small for this data set.

Con. **3.80** Although an exhaust fan is present in nearly every bathroom, they often are not used due to the high noise level. This is an unfortunate practice because regular use of the fan results in a reduction of indoor moisture. Excessive indoor moisture often results in the development of mold which may lead to adverse health consequences. *Consumer Reports* in its January 2004 issue reports on a wide variety of bathroom fans. The following table displays the price (P) in dollars of the fans and the quality of the fan measured in airflow (AF), cubic feet per minute (cfm).

P	95	115	110	15	20	20	75	150	60	60
AF	60	60	60	55	55	55	85	80	80	75
P	160	125	125	110	130	125	30	60	110	85
AF	90	90	100	110	90	90	90	110	110	60

a. Plot the data in a scatterplot and comment on the relationship between price and airflow.
b. Compute the correlation coefficient for this data set. Is there a strong or weak relationship between price and airflow of the fans?
c. Is your conclusion in part (b) consistent with your answer in part (a)?
d. Based on your answers in parts (a) and (b), would it be reasonable to conclude that higher priced fans generate greater airflow?

CHAPTER 4

Probability and Probability Distributions

4.1 Introduction and Abstract of Research Study
4.2 Finding the Probability of an Event
4.3 Basic Event Relations and Probability Laws
4.4 Conditional Probability and Independence
4.5 Bayes' Formula
4.6 Variables: Discrete and Continuous
4.7 Probability Distributions for Discrete Random Variables
4.8 Two Discrete Random Variables: The Binomial and the Poisson
4.9 Probability Distributions for Continuous Random Variables
4.10 A Continuous Probability Distribution: The Normal Distribution
4.11 Random Sampling
4.12 Sampling Distributions
4.13 Normal Approximation to the Binomial
4.14 Evaluating Whether or Not a Population Distribution Is Normal
4.15 Research Study: Inferences about Performance-Enhancing Drugs among Athletes
4.16 Minitab Instructions
4.17 Summary and Key Formulas
4.18 Exercises

4.1 Introduction and Abstract of Research Study

We stated in Chapter 1 that a scientist uses inferential statistics to make statements about a population based on information contained in a sample of units selected from that population. Graphical and numerical descriptive techniques were presented in Chapter 3 as a means to summarize and describe a sample.