CHAPTER

1

Looking at Data—Distributions

Students planning a referendum on college fees. See Example 1.1.

Introduction

Statistics is the science of learning from data. Data are numerical facts. Here is an example of a situation where students used the results of a referendum to convince their university Board of Trustees to make a decision.

EXAMPLE

1.1 Students vote for service learning scholarships. According to the National Service-Learning Clearinghouse: "Service-learning is a teaching and learning strategy that integrates meaningful community service with instruction and reflection to enrich the learning experience, teach civic responsibility, and strengthen communities."[1] University of Illinois at Urbana–Champaign students decided that they wanted to become involved in this national movement. They proposed a $15.00 per semester Legacy of Service and Learning Scholarship fee. Each year, $10.00 would be invested in an endowment and $5.00 would be used to fund current-use scholarships. In a referendum, students voted 3785 to 2977 in favor of the proposal. On April 11, 2006, the university Board of Trustees approved the proposal. Approximately $370,000 in current-use scholarship funds will be generated each year, and with the endowment, it is expected that in 20 years there will be more than a million dollars per year for these scholarships.

To learn from data, we need more than just the numbers. The numbers in a medical study, for example, mean little without some knowledge of the goals of the study and of what blood pressure, heart rate, and other measurements contribute to those goals. That is, *data are numbers with a context,* and we need to understand the context if we are to make sense of the numbers. On the other hand, measurements from the study's several hundred subjects are of little value to even the most knowledgeable medical expert until the tools of statistics organize, display, and summarize them. We begin our study of statistics by mastering the art of examining data.

Variables

Any set of data contains information about some group of *individuals*. The information is organized in *variables*.

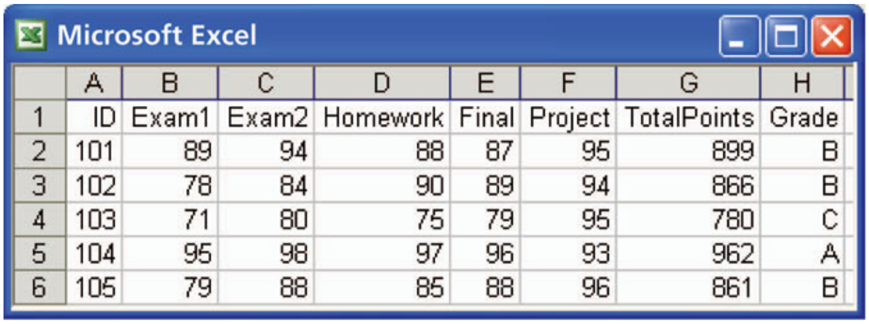
INDIVIDUALS AND VARIABLES

Individuals are the objects described in a set of data. Individuals are sometimes people. When the objects that we want to study are not people, we often call them **cases.**

A **variable** is any characteristic of an individual. A variable can take different values for different individuals.

EXAMPLE

1.2 Data for students in a statistics class. Figure 1.1 shows part of a data set for students enrolled in an introductory statistics class. Each row gives the data on one student. The values for the different variables are in the columns. This data set has eight variables. ID is an identifier for each student. Exam1, Exam2, Homework, Final, and Project give the points earned, out of a total of 100 possible, for each of these course requirements. Final grades are based on a possible 200 points for each exam and the final, 300 points for Homework, and 100 points for Project. TotalPoints is the variable that gives the composite score. It is computed by adding 2 times Exam1, Exam2, and Final, 3 times Homework plus 1 times Project. Grade is the grade earned in the course. This instructor used cut-offs of 900, 800, 700, etc. for the letter grades.



Microsoft Excel

	A	B	C	D	E	F	G	H
1	ID	Exam1	Exam2	Homework	Final	Project	TotalPoints	Grade
2	101	89	94	88	87	95	899	B
3	102	78	84	90	89	94	866	B
4	103	71	80	75	79	95	780	C
5	104	95	98	97	96	93	962	A
6	105	79	88	85	88	96	861	B

FIGURE 1.1 Spreadsheet for Example 1.2.

spreadsheet

The display in Figure 1.1 is from an Excel **spreadsheet.** Most statistical software packages use similar spreadsheets and many are able to import Excel spreadsheets.

USE YOUR KNOWLEDGE

1.1 **Read the spreadsheet.** Refer to Figure 1.1. Give the values of the variables Exam1, Exam2, and Final for the student with ID equal to 104.

1.2 **Calculate the grade.** A student whose data do not appear on the spreadsheet scored 88 on Exam1, 85 on Exam2, 77 for Homework, 90 on the Final, and 80 on the Project. Find TotalPoints for this student and give the grade earned.

Spreadsheets are very useful for doing the kind of simple computations that you did in Exercise 1.2. You can type in a formula and have the same computation performed for each row.

Note that the names we have chosen for the variables in our spreadsheet do not have spaces. For example, we could have used the name "Exam 1" for the first exam score rather than Exam1. In many statistical software packages, however, spaces are not allowed in variable names. For this reason, when creating spreadsheets for eventual use with statistical software, it is best to avoid spaces in variable names. Another convention is to use an underscore (_) where you would normally use a space. For our data set, we could use Exam_1, Exam_2, and Final_Exam.

In practice, any set of data is accompanied by background information that helps us understand the data. When you plan a statistical study or explore data from someone else's work, ask yourself the following questions:

1. **Why? What purpose** do the data have? Do we hope to answer some specific questions? Do we want to draw conclusions about individuals other than those for whom we actually have data?
2. **Who?** What **individuals** do the data describe? **How many** individuals appear in the data?
3. **What?** How many **variables** do the data contain? What are the **exact definitions** of these variables? Some variables have units. Weights, for example, might be recorded in pounds, in thousands of pounds, or in kilograms. For these kinds of variables, you need to know the **unit of measurement.**

EXAMPLE

1.3 Individuals and variables. The data set in Figure 1.1 was constructed to keep track of the grades for students in an introductory statistics course. The individuals are the students in the class. There are 8 variables in this data set. These include an identifier for each student and scores for the various course requirements. There are no units for ID and grade. The other variables all have "points" as the unit.

Some variables, like gender and college major, simply place individuals into categories. Others, like height and grade point average, take numerical values for which we can do arithmetic. It makes sense to give an average salary for a company's employees, but it does not make sense to give an "average" gender. We can, however, count the numbers of female and male employees and do arithmetic with these counts.

CATEGORICAL AND QUANTITATIVE VARIABLES

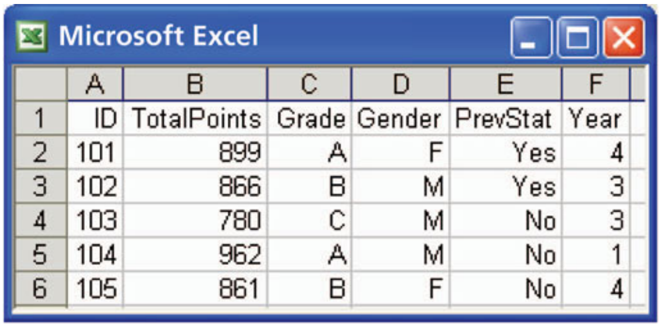
A **categorical variable** places an individual into one of two or more groups or categories.

A **quantitative variable** takes numerical values for which arithmetic operations such as adding and averaging make sense.

The **distribution** of a variable tells us what values it takes and how often it takes these values.

EXAMPLE

1.4 Variables for students in a statistics course. Suppose the data for the students in the introductory statistics class were also to be used to study relationships between student characteristics and success in the course. For this purpose, we might want to use a data set like the spreadsheet in Figure 1.2. Here, we have decided to focus on the TotalPoints and Grade as the outcomes of interest. Other variables of interest have been included: Gender, PrevStat (whether or not the student has taken a statistics course previously), and Year (student classification as first, second, third, or fourth year). ID is a categorical variable, total points is a quantitative variable, and the remaining variables are all categorical.



	A	B	C	D	E	F
1	ID	TotalPoints	Grade	Gender	PrevStat	Year
2	101	899	A	F	Yes	4
3	102	866	B	M	Yes	3
4	103	780	C	M	No	3
5	104	962	A	M	No	1
6	105	861	B	F	No	4

FIGURE 1.2 Spreadsheet for Example 1.4.

In our example, the possible values for the grade variable are A, B, C, D, and F. When computing grade point averages, many colleges and universities translate these letter grades into numbers using A = 4, B = 3, C = 2, D = 1, and F = 0. The transformed variable with numeric values is considered to be quantitative because we can average the numerical values across different courses to obtain a grade point average.

Sometimes, experts argue about numerical scales such as this. They ask whether or not the difference between an A and a B is the same as the difference between a D and an F. Similarly, many questionnaires ask people to respond on a 1 to 5 scale with 1 representing strongly agree, 2 representing agree, etc. Again we could ask about whether or not the five possible values for this scale are equally spaced in some sense. From a practical point of view, the averages that can be computed when we convert categorical scales such as these to numerical values frequently provide a very useful way to summarize data.

USE YOUR KNOWLEDGE

1.3 **Apartment rentals.** A data set lists apartments available for students to rent. Information provided includes the monthly rent, whether or not cable is included free of charge, whether or not pets are allowed, the number of bedrooms, and the distance to the campus. Describe the individuals or cases in the data set, give the number of variables, and specify whether each variable is categorical or quantitative.

Measurement: know your variables

instrument

The context of data includes an understanding of the variables that are recorded. Often the variables in a statistical study are easy to understand: height in centimeters, study time in minutes, and so on. But each area of work also has its own special variables. A psychologist uses the Minnesota Multiphasic Personality Inventory (MMPI), and a physical fitness expert measures "VO2 max," the volume of oxygen consumed per minute while exercising at your maximum capacity. Both of these variables are measured with special **instruments.** VO2 max is measured by exercising while breathing into a mouthpiece connected to an apparatus that measures oxygen consumed. Scores on the MMPI are based on a long questionnaire, which is also an instrument. Part of mastering your field of work is learning what variables are important and how they are best measured. Because details of particular measurements usually require knowledge of the particular field of study, we will say little about them.

CAUTION

rate

Be sure that each variable really does measure what you want it to. A poor choice of variables can lead to misleading conclusions. Often, for example, the **rate** at which something occurs is a more meaningful measure than a simple count of occurrences.

EXAMPLE

1.5 Accidents for passenger cars and motorcycles. The government's Fatal Accident Reporting System says that 27,102 passenger cars were involved in fatal accidents in 2002. Only 3339 motorcycles had fatal accidents that year.[2] Does this mean that motorcycles are safer than cars? Not at all—there are many more cars than motorcycles, so we expect cars to have a higher *count* of fatal accidents.

A better measure of the dangers of driving is a *rate,* the number of fatal accidents divided by the number of vehicles on the road. In 2002, passenger cars had about 21 fatal accidents for each 100,000 vehicles registered. There were about 67 fatal accidents for each 100,000 motorcycles registered. The rate for motorcycles is more than three times the rate for cars. Motorcycles are, as we might guess, much more dangerous than cars.

1.1 Displaying Distributions with Graphs

exploratory data analysis

Statistical tools and ideas help us examine data in order to describe their main features. This examination is called **exploratory data analysis.** Like an explorer crossing unknown lands, we want first to simply describe what we see. Here are two basic strategies that help us organize our exploration of a set of data:

- Begin by examining each variable by itself. Then move on to study the relationships among the variables.
- Begin with a graph or graphs. Then add numerical summaries of specific aspects of the data.

We will follow these principles in organizing our learning. This chapter presents methods for describing a single variable. We will study relationships among several variables in Chapter 2. Within each chapter, we will begin with graphical displays, then add numerical summaries for more complete description.

Graphs for categorical variables

The values of a categorical variable are labels for the categories, such as "female" and "male." The **distribution** of a categorical variable lists the categories and gives either the **count** or the **percent** of individuals who fall in each category. For example, how well educated are 30-something young adults? Here is the distribution of the highest level of education for people aged 25 to 34 years:[3]

Education	Count (millions)	Percent
Less than high school	4.6	12.1
High school graduate	11.6	30.5
Some college	7.4	19.5
Associate degree	3.3	8.7
Bachelor's degree	8.6	22.6
Advanced degree	2.5	6.6

Are you surprised that only 29.2% of young adults have at least a bachelor's degree?

bar graph

pie chart

The graphs in Figure 1.3 display these data. The **bar graph** in Figure 1.3(a) quickly compares the sizes of the six education groups. The heights of the bars show the percents in the six categories. The **pie chart** in Figure 1.3(b) helps us see what part of the whole each group forms. For example, the "Bachelor's" slice makes up 22.6% of the pie because 22.6% of young adults have a bachelor's degree but no higher degree. We have moved that slice out to call attention to it. Because pie charts lack a scale, we have added the percents to the labels for the slices. *Pie charts require that you include all the categories that make up a whole. Use them only when you want to emphasize each category's relation to the whole.* Bar graphs are easier to read and are also more flexible. For example, you can use a bar graph to compare the numbers of students at your college majoring in biology, business, and political science. A pie chart cannot make this comparison because not all students fall into one of these three majors.

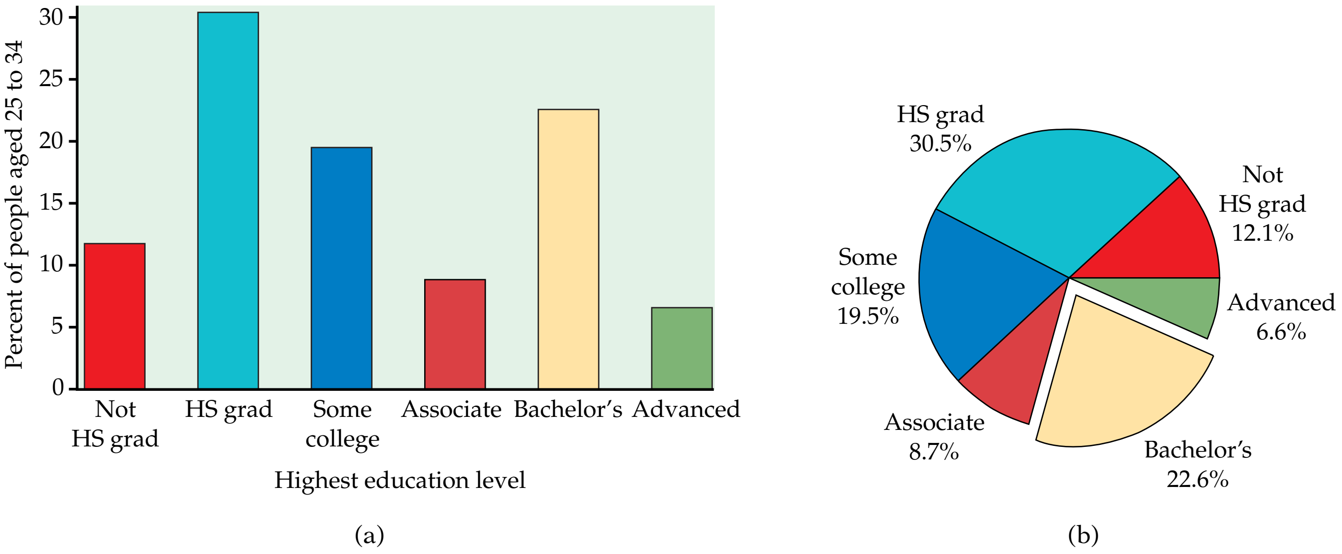


FIGURE 1.3 (a) Bar graph of the educational attainment of people aged 25 to 34 years. (b) Pie chart of the education data, with bachelor's degree holders emphasized.

USE YOUR KNOWLEDGE

1.4 Read the pie chart. Refer to Figure 1.3(b). What percent of young adults have either an associate degree or a bachelor's degree?

Bar graphs and pie charts help an audience grasp a distribution quickly. They are, however, of limited use for data analysis because it is easy to understand data on a single categorical variable, such as highest level of education, without a graph. We will move on to quantitative variables, where graphs are essential tools.

Data analysis in action: don't hang up on me

Many businesses operate call centers to serve customers who want to place an order or make an inquiry. Customers want their requests handled thoroughly. Businesses want to treat customers well, but they also want to avoid wasted time on the phone. They therefore monitor the length of calls and encourage their representatives to keep calls short. Here is an example of the difficulties this policy can cause.

EXAMPLE

1.6 Individuals and variables for the customer service center. We have data on the length of all 31,492 calls made to the customer service center of a small bank in a month. Table 1.1 displays the lengths of the first 80 calls. The file for the complete data set is *eg01-004,* which you can find on the text CD and Web site.[4]

Take a look at the data in Table 1.1. The numbers are meaningless without some background information. The *individuals* are calls made to the bank's call center. The *variable* recorded is the length of each call. The *units* are

TABLE 1.1

Service times (seconds) for calls to a customer service center

77	289	128	59	19	148	157	203
126	118	104	141	290	48	3	2
372	140	438	56	44	274	479	211
179	1	68	386	2631	90	30	57
89	116	225	700	40	73	75	51
148	9	115	19	76	138	178	76
67	102	35	80	143	951	106	55
4	54	137	367	277	201	52	9
700	182	73	199	325	75	103	64
121	11	9	88	1148	2	465	25

seconds. We see that the call lengths vary a great deal. The longest call lasted 2631 seconds, almost 44 minutes. More striking is that 8 of these 80 calls lasted less than 10 seconds. What's going on?

Figure 1.4 is a histogram of the lengths of all 31,492 calls. We did not plot the few lengths greater than 1200 seconds (20 minutes). As expected, the graph shows that most calls last between about a minute and 5 minutes, with some lasting much longer when customers have complicated problems. More striking is the fact that 7.6% of all calls are no more than 10 seconds long. It turned out that the bank penalized representatives whose average call length was too long—so some representatives just hung up on customers in order to bring their average length down. Neither the customers nor the bank were happy about this. The bank changed its policy, and later data showed that calls under 10 seconds had almost disappeared.

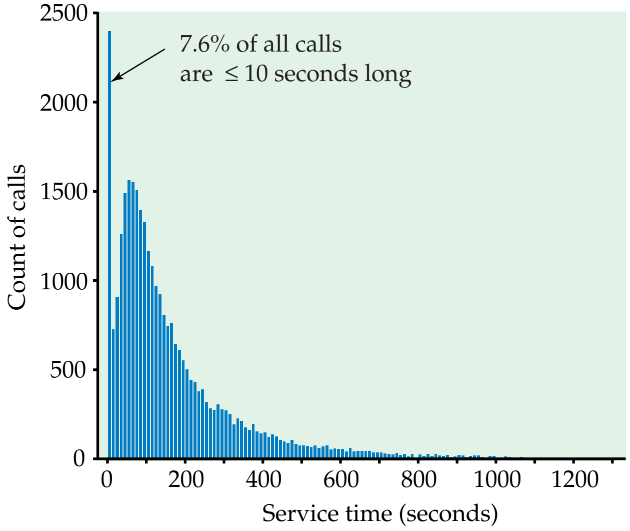


FIGURE 1.4 The distribution of call lengths for 31,492 calls to a bank's customer service center, for Example 1.6. The data show a surprising number of very short calls. These are mostly due to representatives deliberately hanging up in order to bring down their average call length.

tails

The extreme values of a distribution are in the **tails** of the distribution. The high values are in the upper, or right, tail and the low values are in the lower, or left, tail. The overall pattern in Figure 1.4 is made up of the many moderate call lengths and the long right tail of more lengthy calls. The striking departure from the overall pattern is the surprising number of very short calls in the left tail.

Our examination of the call center data illustrates some important principles:

- After you understand the background of your data (individuals, variables, units of measurement), the first thing to do is almost always **plot your data.**
- When you look at a plot, look for an **overall pattern** and also for any **striking departures** from the pattern.

We now turn to the kinds of graphs that are used to describe the distribution of a quantitative variable. We will explain how to make the graphs by hand, because knowing this helps you understand what the graphs show. However, making graphs by hand is so tedious that software is almost essential for effective data analysis unless you have just a few observations.

Stemplots

A *stemplot* (also called a stem-and-leaf plot) gives a quick picture of the shape of a distribution while including the actual numerical values in the graph. Stemplots work best for small numbers of observations that are all greater than 0.

STEMPLOT

To make a **stemplot:**

1. Separate each observation into a **stem** consisting of all but the final (rightmost) digit and a **leaf,** the final digit. Stems may have as many digits as needed, but each leaf contains only a single digit.

2. Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column.

3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

EXAMPLE

1.7 Literacy of men and women. The Islamic world is attracting increased attention in Europe and North America. Table 1.2 shows the percent of men and women at least 15 years old who were literate in 2002 in the major Islamic nations. We omitted countries with populations less than 3 million. Data for a few nations, such as Afghanistan and Iraq, are not available.[5]

To make a stemplot of the percents of females who are literate, use the first digits as stems and the second digits as leaves. Algeria's 60% literacy rate, for example, appears as the leaf 0 on the stem 6. Figure 1.5 shows the steps in making the plot.

TABLE 1.2

Literacy rates (percent) in Islamic nations

Country	Female percent	Male percent	Country	Female percent	Male percent
Algeria	60	78	Morocco	38	68
Bangladesh	31	50	Saudi Arabia	70	84
Egypt	46	68	Syria	63	89
Iran	71	85	Tajikistan	99	100
Jordan	86	96	Tunisia	63	83
Kazakhstan	99	100	Turkey	78	94
Lebanon	82	95	Uzbekistan	99	100
Libya	71	92	Yemen	29	70
Malaysia	85	92			

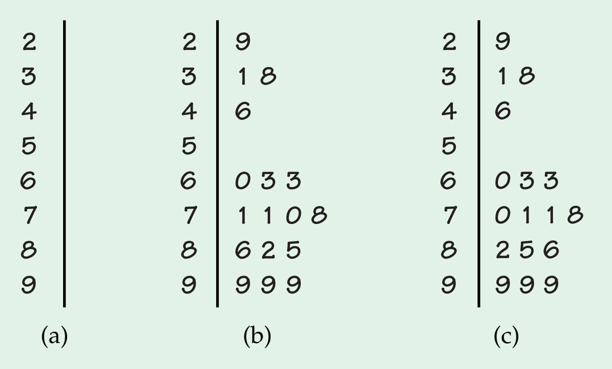


FIGURE 1.5 Making a stemplot of the data in Example 1.7. (a) Write the stems. (b) Go through the data and write each leaf on the proper stem. For example, the values on the 8 stem are 86, 82, and 85 in the order of the table. (c) Arrange the leaves on each stem in order out from the stem. The 8 stem now has leaves 2 5 6.

cluster

The overall pattern of the stemplot is irregular, as is often the case when there are only a few observations. There do appear to be two **clusters** of countries. The plot suggests that we might ask what explains the variation in literacy. For example, why do the three central Asian countries (Kazakhstan, Tajikistan, and Uzbekistan) have very high literacy rates?

USE YOUR KNOWLEDGE

1.5 **Make a stemplot.** Here are the scores on the first exam in an introductory statistics course for 30 students in one section of the course:

80	73	92	85	75	98	93	55	80	90	92	80	87	90	72
65	70	85	83	60	70	90	75	75	58	68	85	78	80	93

Use these data to make a stemplot. Then use the stemplot to describe the distribution of the first-exam scores for this course.

back-to-back stemplot

When you wish to compare two related distributions, a **back-to-back stemplot** with common stems is useful. The leaves on each side are ordered out from the common stem. Here is a back-to-back stemplot comparing the distributions of female and male literacy rates in the countries of Table 1.2.

Female		Male
9	2	
81	3	
6	4	
	5	0
330	6	88
8110	7	08
652	8	3459
999	9	22456
	10	000

The values on the left are the female percents, as in Figure 1.5, but ordered out from the stem from right to left. The values on the right are the male percents. It is clear that literacy is generally higher among males than among females in these countries.

CAUTION

splitting stems
trimming

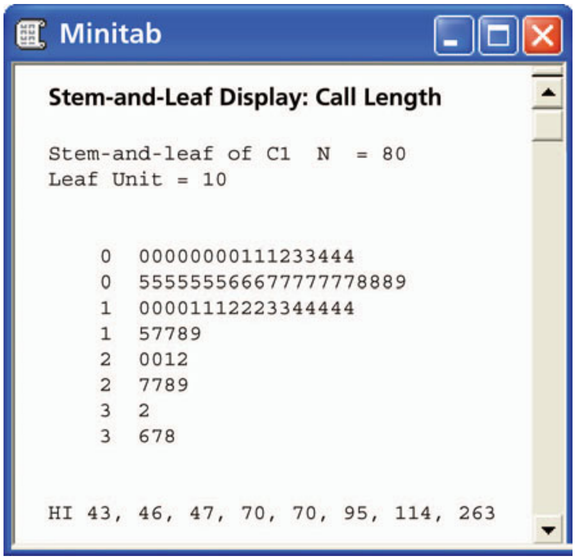
Stemplots do not work well for large data sets, where each stem must hold a large number of leaves. Fortunately, there are two modifications of the basic stemplot that are helpful when plotting the distribution of a moderate number of observations. You can double the number of stems in a plot by **splitting each stem** into two: one with leaves 0 to 4 and the other with leaves 5 through 9. When the observed values have many digits, it is often best to **trim** the numbers by removing the last digit or digits before making a stemplot. You must use your judgment in deciding whether to split stems and whether to trim, though statistical software will often make these choices for you. Remember that the purpose of a stemplot is to display the shape of a distribution. If a stemplot has fewer than about five stems, you should usually split the stems unless there are few observations. If there are many stems with no leaves or only one leaf, trimming will reduce the number of stems. Here is an example that makes use of both of these modifications.

EXAMPLE

1.8 Stemplot for length of service calls. Return to the 80 customer service call lengths in Table 1.1. To make a stemplot of this distribution, we first trim the call lengths to tens of seconds by dropping the last digit. For example, 56 seconds trims to 5 and 143 seconds trims to 14. (We might also round to the nearest 10 seconds, but trimming is faster than rounding if you must do it by hand.)

We can then use tens of seconds as our leaves, with the digits to the left forming stems. This gives us the single-digit leaves that a stemplot requires. For example, 56 trimmed to 5 becomes leaf 5 on the 0 stem; 143 trimmed to 14 becomes leaf 4 on the 1 stem.

Because we have 80 observations, we split the stems. Thus, 56 trimmed to 5 becomes leaf 5 on the second 0 stem, along with all leaves 5 to 9. Leaves



Stem-and-Leaf Display: Call Length

```
Stem-and-leaf of C1  N  = 80
Leaf Unit = 10


    0  00000000111233444
    0  555555566677777778889
    1  00001112223344444
    1  57789
    2  0012
    2  7789
    3  2
    3  678


HI 43, 46, 47, 70, 70, 95, 114, 263
```

FIGURE 1.6 Stemplot from Minitab of the 80 call lengths in Table 1.1, for Example 1.8. The software has trimmed the data by removing the last digit. It has also split stems and listed the highest observations apart from the plot.

0 to 4 go on the first 0 stem. Figure 1.6 is a stemplot of these data made by software. The software automatically did what we suggest: trimmed to tens of seconds and split stems. To save space, the software also listed the largest values as "HI" rather than create stems all the way up to 26. The stemplot shows the overall pattern of the distribution, with many short to moderate lengths and some very long calls.

Histograms

Stemplots display the actual values of the observations. This feature makes stemplots awkward for large data sets. Moreover, the picture presented by a stemplot divides the observations into groups (stems) determined by the number system rather than by judgment. Histograms do not have these limitations. A **histogram** breaks the range of values of a variable into classes and displays only the count or percent of the observations that fall into each class. You can choose any convenient number of classes, but you should always choose classes of equal width. Histograms are slower to construct by hand than stemplots and do not display the actual values observed. For these reasons we prefer stemplots for small data sets. The construction of a histogram is best shown by example. Any statistical software package will of course make a histogram for you.

histogram

EXAMPLE

1.9 Distribution of IQ scores. You have probably heard that the distribution of scores on IQ tests is supposed to be roughly "bell-shaped." Let's look at some actual IQ scores. Table 1.3 displays the IQ scores of 60 fifth-grade students chosen at random from one school.[6]

1. Divide the range of the data into classes of equal width. The scores in Table 1.3 range from 81 to 145, so we choose as our classes

TABLE 1.3

IQ test scores for 60 randomly chosen fifth-grade students

145	139	126	122	125	130	96	110	118	118
101	142	134	124	112	109	134	113	81	113
123	94	100	136	109	131	117	110	127	124
106	124	115	133	116	102	127	117	109	137
117	90	103	114	139	101	122	105	97	89
102	108	110	128	114	112	114	102	82	101

$$75 \leq \text{IQ score} < 85$$

$$85 \leq \text{IQ score} < 95$$

$$\vdots$$

$$145 \leq \text{IQ score} < 155$$

Be sure to specify the classes precisely so that each individual falls into exactly one class. A student with IQ 84 would fall into the first class, but IQ 85 falls into the second.

frequency
frequency table

2. Count the number of individuals in each class. These counts are called **frequencies,** and a table of frequencies for all classes is a **frequency table.**

Class	Count	Class	Count
75 to 84	2	115 to 124	13
85 to 94	3	125 to 134	10
95 to 104	10	135 to 144	5
105 to 114	16	145 to 154	1

3. Draw the histogram. First, on the horizontal axis mark the scale for the variable whose distribution you are displaying. That's IQ score. The scale runs from 75 to 155 because that is the span of the classes we chose. The vertical axis contains the scale of counts. Each bar represents a class. The base of the bar covers the class, and the bar height is the class count. There is no horizontal space between the bars unless a class is empty, so that its bar has height zero. Figure 1.7 is our histogram. It does look roughly "bell-shaped."

Large sets of data are often reported in the form of frequency tables when it is not practical to publish the individual observations. In addition to the frequency (count) for each class, we may be interested in the fraction or percent of the observations that fall in each class. A histogram of percents looks just like a frequency histogram such as Figure 1.7. Simply relabel the vertical scale to read in percents. Use histograms of percents for comparing several distributions that have different numbers of observations.

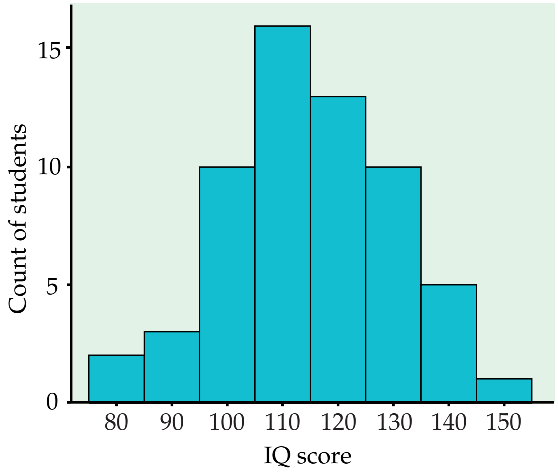


FIGURE 1.7 Histogram of the IQ scores of 60 fifth-grade students, for Example 1.9.

USE YOUR KNOWLEDGE

1.6 **Make a histogram.** Refer to the first-exam scores from Exercise 1.5. Use these data to make a histogram using classes 50–59, 60–69, etc. Compare the histogram with the stemplot as a way of describing this distribution. Which do you prefer for these data?

Our eyes respond to the *area* of the bars in a histogram. Because the classes are all the same width, area is determined by height and all classes are fairly represented. There is no one right choice of the classes in a histogram. Too few classes will give a "skyscraper" graph, with all values in a few classes with tall bars. Too many will produce a "pancake" graph, with most classes having one or no observations. Neither choice will give a good picture of the shape of the distribution. You must use your judgment in choosing classes to display the shape. Statistical software will choose the classes for you. The software's choice is often a good one, but you can change it if you want.

CAUTION

You should be aware that the appearance of a histogram can change when you change the classes. Figure 1.8 is a histogram of the customer service call lengths

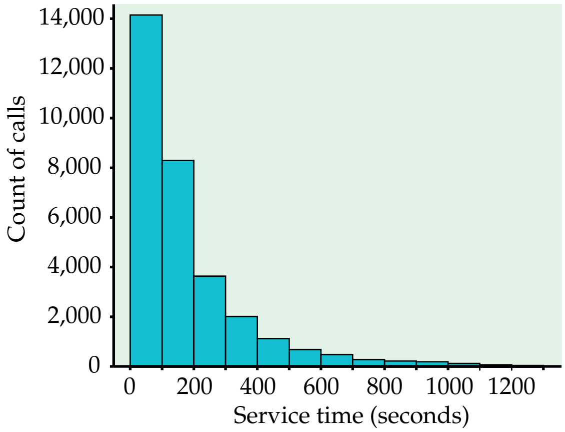


FIGURE 1.8 The "default" histogram produced by software for the call lengths in Example 1.6. This choice of classes hides the large number of very short calls that is revealed by the histogram of the same data in Figure 1.4.

that are also displayed in Figure 1.4. It was produced by software with no special instructions from the user. The software's "default" histogram shows the overall shape of the distribution, but it hides the spike of very short calls by lumping all calls of less than 100 seconds into the first class. We produced Figure 1.4 by asking for smaller classes after Table 1.1 suggested that very short calls might be a problem. Software automates making graphs, but it can't replace thinking about your data. The histogram function in the *One-Variable Statistical Calculator* applet on the text CD and Web site allows you to change the number of classes by dragging with the mouse, so that it is easy to see how the choice of classes affects the histogram.

USE YOUR KNOWLEDGE

1.7 Change the classes in the histogram. Refer to the first-exam scores from Exercise 1.5 and the histogram you produced in Exercise 1.6. Now make a histogram for these data using classes 40–59, 60–79, and 80–100. Compare this histogram with the one that you produced in Exercise 1.6.

1.8 Use smaller classes. Repeat the previous exercise using classes 55–59, 60–64, 65–69, etc.

Although histograms resemble bar graphs, their details and uses are distinct. A histogram shows the distribution of counts or percents among the values of a single variable. A bar graph compares the size of different items. The horizontal axis of a bar graph need not have any measurement scale but simply identifies the items being compared. Draw bar graphs with blank space between the bars to separate the items being compared. Draw histograms with no space, to indicate that all values of the variable are covered. *Some spreadsheet programs, which are not primarily intended for statistics, will draw histograms as if they were bar graphs, with space between the bars. Often, you can tell the software to eliminate the space to produce a proper histogram.*

Examining distributions

Making a statistical graph is not an end in itself. The purpose of the graph is to help us understand the data. After you make a graph, always ask, "What do I see?" Once you have displayed a distribution, you can see its important features as follows.

EXAMINING A DISTRIBUTION

In any graph of data, look for the **overall pattern** and for striking **deviations** from that pattern.

You can describe the overall pattern of a distribution by its **shape, center,** and **spread.**

An important kind of deviation is an **outlier,** an individual value that falls outside the overall pattern.

In Section 1.2, we will learn how to describe center and spread numerically. For now, we can describe the center of a distribution by its *midpoint*, the value with roughly half the observations taking smaller values and half taking larger values. We can describe the spread of a distribution by giving the *smallest and largest values*. Stemplots and histograms display the shape of a distribution in the same way. Just imagine a stemplot turned on its side so that the larger values lie to the right. Some things to look for in describing shape are:

modes
unimodal

- Does the distribution have one or several major peaks, called **modes**? A distribution with one major peak is called **unimodal.**

symmetric
skewed

- Is it approximately symmetric or is it skewed in one direction? A distribution is **symmetric** if the values smaller and larger than its midpoint are mirror images of each other. It is **skewed to the right** if the right tail (larger values) is much longer than the left tail (smaller values).

Some variables commonly have distributions with predictable shapes. Many biological measurements on specimens from the same species and sex—lengths of bird bills, heights of young women—have symmetric distributions. Money amounts, on the other hand, usually have right-skewed distributions. There are many moderately priced houses, for example, but the few very expensive mansions give the distribution of house prices a strong right-skew.

EXAMPLE

1.10 Examine the histogram. What does the histogram of IQ scores (Figure 1.7) tell us? **Shape:** The distribution is *roughly symmetric* with a *single peak* in the center. We don't expect real data to be perfectly symmetric, so we are satisfied if the two sides of the histogram are roughly similar in shape and extent. **Center:** You can see from the histogram that the midpoint is not far from 110. Looking at the actual data shows that the midpoint is 114. **Spread:** The spread is from 81 to 145. There are no outliers or other strong deviations from the symmetric, unimodal pattern.

The distribution of call lengths in Figure 1.8, on the other hand, is strongly *skewed to the right*. The midpoint, the length of a typical call, is about 115 seconds, or just under 2 minutes. The spread is very large, from 1 second to 28,739 seconds.

The longest few calls are *outliers*. They stand apart from the long right tail of the distribution, though we can't see this from Figure 1.8, which omits the largest observations. The longest call lasted almost 8 hours—that may well be due to equipment failure rather than an actual customer call.

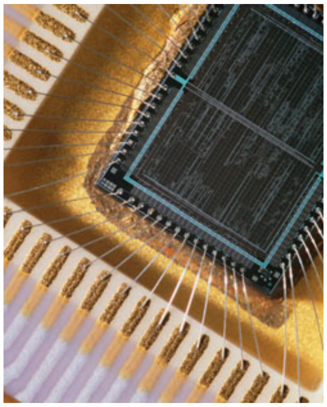
USE YOUR KNOWLEDGE

1.9 Describe the first-exam scores. Refer to the first-exam scores from Exercise 1.5. Use your favorite graphical display to describe the shape, the center, and the spread of these data. Are there any outliers?

Dealing with outliers

CAUTION

In data sets smaller than the service call data, you can spot outliers by looking for observations that stand apart (either high or low) from the overall pattern of a histogram or stemplot. *Identifying outliers is a matter for judgment. Look for points that are clearly apart from the body of the data, not just the most extreme observations in a distribution.* You should search for an explanation for any outlier. Sometimes outliers point to errors made in recording the data. In other cases, the outlying observation may be caused by equipment failure or other unusual circumstances.



EXAMPLE

1.11 Semiconductor wires. Manufacturing an electronic component requires attaching very fine wires to a semiconductor wafer. If the strength of the bond is weak, the component may fail. Here are measurements on the breaking strength (in pounds) of 23 connections:[7]

0	0	550	750	950	950	1150	1150
1150	1150	1150	1250	1250	1350	1450	1450
1450	1550	1550	1550	1850	2050	3150	

Figure 1.9 is a histogram of these data. We expect the breaking strengths of supposedly identical connections to have a roughly symmetric overall pattern, showing chance variation among the connections. Figure 1.9 does show a symmetric pattern centered at about 1250 pounds—but it also shows three *outliers* that stand apart from this pattern, two low and one high.

The engineers were able to explain all three outliers. The two low outliers had strength 0 because the bonds between the wire and the wafer were not made. The high outlier at 3150 pounds was a measurement error. Further study of the data can simply omit the three outliers. One immediate finding is that the variation in breaking strength is too large—550 pounds to 2050 pounds when we ignore the outliers. The process of bonding wire to wafer must be improved to give more consistent results.

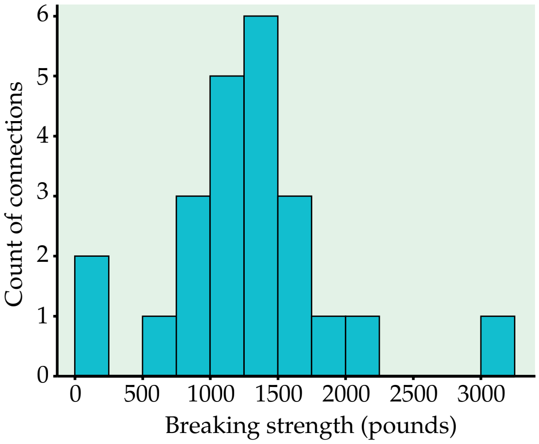


FIGURE 1.9 Histogram of a distribution with both low and high outliers, for Example 1.11.

Time plots

Whenever data are collected over time, it is a good idea to plot the observations in time order. *Displays of the distribution of a variable that ignore time order, such as stemplots and histograms, can be misleading when there is systematic change over time.*

TIME PLOT

A **time plot** of a variable plots each observation against the time at which it was measured. Always put time on the horizontal scale of your plot and the variable you are measuring on the vertical scale. Connecting the data points by lines helps emphasize any change over time.

EXAMPLE

1.12 Water from the Mississippi River. Table 1.4 lists the volume of water discharged by the Mississippi River into the Gulf of Mexico for each year from 1954 to 2001.[8] The units are cubic kilometers of water—the Mississippi is a big river. Both graphs in Figure 1.10 describe these data. The histogram in Figure 1.10(a) shows the distribution of the volume discharged. The histogram is symmetric and unimodal, with center near 550 cubic kilometers. We might think that the data show just chance year-to-year fluctuation in river level about its long-term average.

Figure 1.10(b) is a time plot of the same data. For example, the first point lies above 1954 on the "Year" scale at height 290, the volume of water discharged by the Mississippi in 1954. The time plot tells a more interesting story

TABLE 1.4

Yearly discharge of the Mississippi River (in cubic kilometers of water)

Year	Discharge	Year	Discharge	Year	Discharge	Year	Discharge
1954	290	1966	410	1978	560	1990	680
1955	420	1967	460	1979	800	1991	700
1956	390	1968	510	1980	500	1992	510
1957	610	1969	560	1981	420	1993	900
1958	550	1970	540	1982	640	1994	640
1959	440	1971	480	1983	770	1995	590
1960	470	1972	600	1984	710	1996	670
1961	600	1973	880	1985	680	1997	680
1962	550	1974	710	1986	600	1998	690
1963	360	1975	670	1987	450	1999	580
1964	390	1976	420	1988	420	2000	390
1965	500	1977	430	1989	630	2001	580

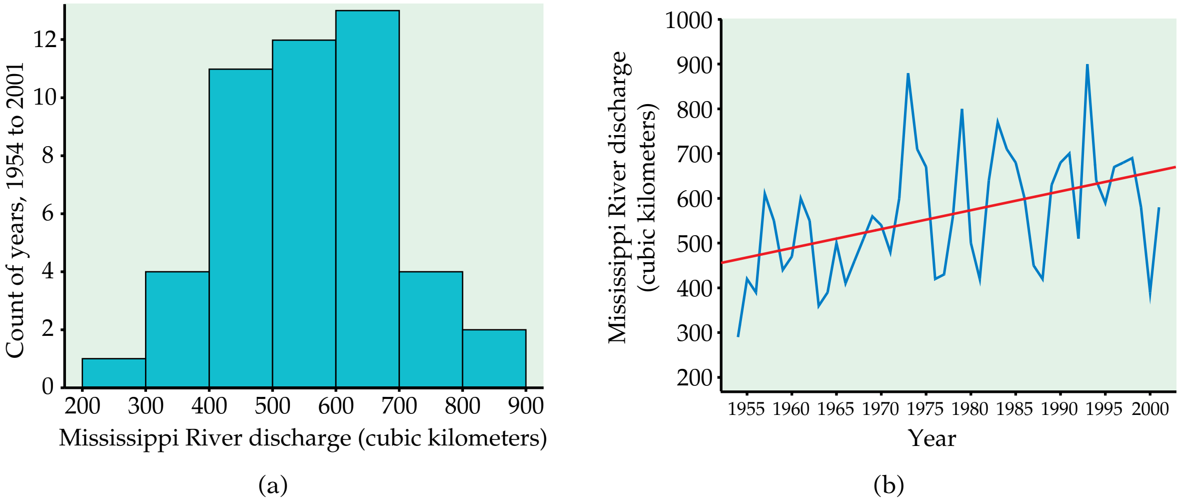


FIGURE 1.10 (a) Histogram of the volume of water discharged by the Mississippi River over the 48 years from 1954 to 2001, for Example 1.12. Data are from Table 1.4. (b) Time plot of the volume of water discharged by the Mississippi River for the years 1954 to 2001. The line shows the trend toward increasing river flow, a trend that cannot be seen in the histogram in Figure 1.10(a).

trend

than the histogram. There is a great deal of year-to-year variation, but there is also a clear increasing **trend** over time. That is, there is a long-term rise in the volume of water discharged. The line on the graph is a "trend line" calculated from the data to describe this trend. The trend reflects climate change: rainfall and river flows have been increasing over most of North America.

time series

Many interesting data sets are **time series,** measurements of a variable taken at regular intervals over time. Government, economic, and social data are often published as time series. Some examples are the monthly unemployment rate and the quarterly gross domestic product. Weather records, the demand for electricity, and measurements on the items produced by a manufacturing process are other examples of time series. Time plots can reveal the main features of a time series.

BEYOND THE BASICS

Decomposing Time Series*

When you examine a time plot, again look first for overall patterns and then for striking deviations from those patterns. Here are two important types of overall patterns to look for in a time series.

*"Beyond the Basics" sections briefly discuss supplementary topics. Your software may make some of these topics available to you. For example, the results plotted in Figures 1.11 to 1.13 come from the Minitab statistical software.

TREND AND SEASONAL VARIATION

A **trend** in a time series is a persistent, long-term rise or fall.

A pattern in a time series that repeats itself at known regular intervals of time is called **seasonal variation.**

seasonally adjusted

Because many economic time series show strong seasonal variation, government agencies often adjust for this variation before releasing economic data. The data are then said to be **seasonally adjusted.** Seasonal adjustment helps avoid misinterpretation. A rise in the unemployment rate from December to January, for example, does not mean that the economy is slipping. Unemployment almost always rises in January as temporary holiday help is laid off and outdoor employment in the North drops because of bad weather. The seasonally adjusted unemployment rate reports an increase only if unemployment rises more than normal from December to January.

EXAMPLE

1.13 Gasoline prices. Figure 1.11 is a time plot of the average retail price of regular gasoline each month for the years 1990 to 2003.[9] The prices are *not* seasonally adjusted. You can see the upward spike in prices due to the 1990 Iraqi invasion of Kuwait, the drop in 1998 when an economic crisis in Asia reduced demand for fuel, and rapid price increases in 2000 and 2003 due to instability in the Middle East and OPEC production limits. These deviations are so large that overall patterns are hard to see.

There is nonetheless a clear *trend* of increasing price. Much of this trend just reflects inflation, the rise in the overall price level during these years. In addition, a close look at the plot shows *seasonal variation,* a regular rise and fall that recurs each year. Americans drive more in the summer vacation season, so the price of gasoline rises each spring, then drops in the fall as demand goes down.

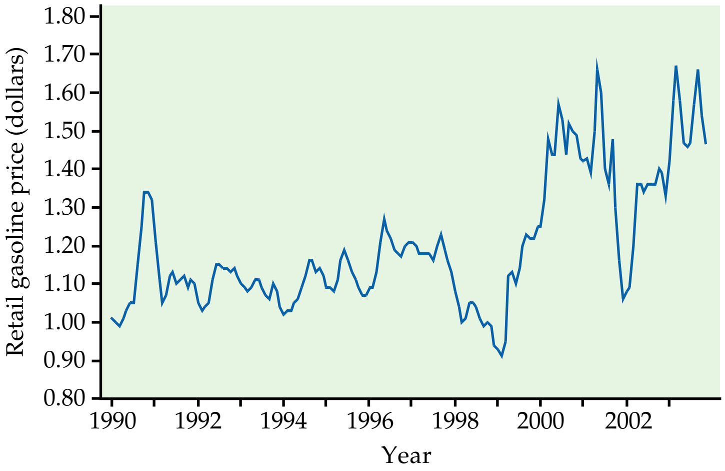


FIGURE 1.11 Time plot of the average monthly price of regular gasoline from 1990 to 2003, for Example 1.13.

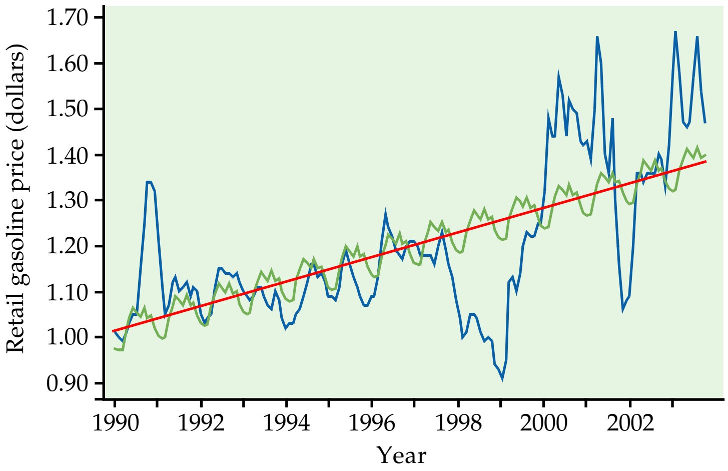


FIGURE 1.12 Time plot of gasoline prices with a trend line and seasonal variation added. These are overall patterns extracted from the data by software.

Statistical software can help us examine a time series by "decomposing" the data into systematic patterns, such as trends and seasonal variation, and the *residuals* that remain after we remove these patterns. Figure 1.12 superimposes the trend and seasonal variation on the time plot of gasoline prices. The red line shows the increasing trend. The seasonal variation appears as the colored line that regularly rises and falls each year. This is an average of the seasonal pattern for all the years in the original data, automatically extracted by software.

The trend and seasonal variation in Figure 1.12 are overall patterns in the data. Figure 1.13 is a plot of what remains when we subtract both the trend and the seasonal variation from the original data. That is, Figure 1.13 emphasizes the deviations from the pattern. In the case of gasoline prices, the deviations are large (as much as 30 cents both up and down). It is clear that we can't use trend and seasonal variation to predict gasoline prices at all accurately.

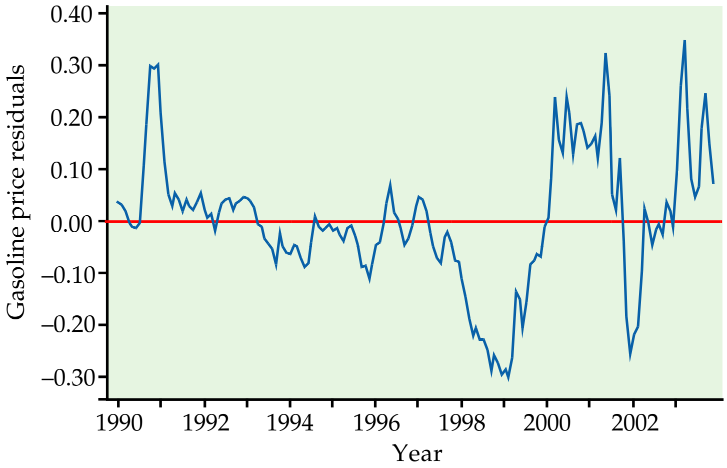


FIGURE 1.13 The residuals that remain when we subtract both trend and seasonal variation from monthly gasoline prices.

SECTION 1.1 Summary

A data set contains information on a collection of **individuals.** Individuals may be people, animals, or things. The data for one individual make up a **case.** For each individual, the data give values for one or more **variables.** A variable describes some characteristic of an individual, such as a person's height, gender, or salary.

Some variables are **categorical** and others are **quantitative.** A categorical variable places each individual into a category, such as male or female. A quantitative variable has numerical values that measure some characteristic of each individual, such as height in centimeters or annual salary in dollars.

Exploratory data analysis uses graphs and numerical summaries to describe the variables in a data set and the relations among them.

The **distribution** of a variable tells us what values it takes and how often it takes these values.

Bar graphs and **pie charts** display the distributions of categorical variables. These graphs use the counts or percents of the categories.

Stemplots and **histograms** display the distributions of quantitative variables. Stemplots separate each observation into a **stem** and a one-digit **leaf.** Histograms plot the **frequencies** (counts) or the percents of equal-width classes of values.

When examining a distribution, look for **shape, center,** and **spread** and for clear **deviations** from the overall shape.

Some distributions have simple shapes, such as **symmetric** or **skewed.** The number of **modes** (major peaks) is another aspect of overall shape. Not all distributions have a simple overall shape, especially when there are few observations.

Outliers are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.

When observations on a variable are taken over time, make a **time plot** that graphs time horizontally and the values of the variable vertically. A time plot can reveal **trends** or other changes over time.

SECTION 1.1 Exercises

For Exercises 1.1 to 1.2, see page 3; for Exercise 1.3, see page 5; for Exercise 1.4, see page 7; for Exercise 1.5, see page 10; for Exercise 1.6, see page 14; for Exercises 1.7 and 1.8, see page 15; and for Exercise 1.9, see page 16.

1.10 Survey of students. A survey of students in an introductory statistics class asked the following questions: (a) age; (b) do you like to dance? (yes, no); (c) can you play a musical instrument (not at all, a little, pretty well); (d) how much did you spend on food last week? (e) height; (f) do you like broccoli? (yes, no). Classify each of these variables as categorical or quantitative and give reasons for your answers.

1.11 What questions would you ask? Refer to the previous exercise. Make up your own survey questions with at least six questions. Include at least two categorical variables and at least two quantitative variables. Tell which variables are categorical and which are quantitative. Give reasons for your answers.

1.12 Study habits of students. You are planning a survey to collect information about the study habits of college students. Describe two categorical variables and two quantitative variables that you might measure for each student. Give the units of measurement for the quantitative variables.

1.13 Physical fitness of students. You want to measure the "physical fitness" of college students. Describe several variables you might use to measure fitness. What instrument or instruments does each measurement require?

1.14 Choosing a college or university. Popular magazines rank colleges and universities on their "academic quality" in serving undergraduate students. Describe five variables that you would like to see measured for each college if you were choosing where to study. Give reasons for each of your choices.

1.15 Favorite colors. What is your favorite color? One survey produced the following summary of responses to that question: blue, 42%; green, 14%; purple, 14%; red, 8%; black, 7%; orange, 5%; yellow, 3%; brown, 3%; gray, 2%; and white, 2%.[10] Make a bar graph of the percents and write a short summary of the major features of your graph.

1.16 Least-favorite colors. Refer to the previous exercise. The same study also asked people about their least-favorite color. Here are the results: orange, 30%; brown, 23%; purple, 13%; yellow, 13%; gray, 12%; green, 4%; white, 4%; red, 1%; black, 0%; and blue, 0%. Make a bar graph of these percents and write a summary of the results.

1.17 Ages of survey respondents. The survey about color preferences reported the age distribution of the people who responded. Here are the results:

Age group (years)	1–18	19–24	25–35	36–50	51–69	70 and over
Count	10	97	70	36	14	5

(a) Add the counts and compute the percents for each age group.

(b) Make a bar graph of the percents.

(c) Describe the distribution.

(d) Explain why your bar graph is not a histogram.

1.18 Garbage. The formal name for garbage is "municipal solid waste." The table at the top of the next column gives a breakdown of the materials that made up American municipal solid waste.[11]

Material	Weight (million tons)	Percent of total
Food scraps	25.9	11.2
Glass	12.8	5.5
Metals	18.0	7.8
Paper, paperboard	86.7	37.4
Plastics	24.7	10.7
Rubber, leather, textiles	15.8	6.8
Wood	12.7	5.5
Yard trimmings	27.7	11.9
Other	7.5	3.2
Total	231.9	100.0

(a) Add the weights for the nine materials given, including "Other." Each entry, including the total, is separately rounded to the nearest tenth. So the sum and the total may differ slightly because of **roundoff error.**

(b) Make a bar graph of the percents. The graph gives a clearer picture of the main contributors to garbage if you order the bars from tallest to shortest.

(c) If you use software, also make a pie chart of the percents. Comparing the two graphs, notice that it is easier to see the small differences among "Food scraps," "Plastics," and "Yard trimmings" in the bar graph.

1.19 Spam. Email spam is the curse of the Internet. Here is a compilation of the most common types of spam:[12]

Type of spam	Percent
Adult	14.5
Financial	16.2
Health	7.3
Leisure	7.8
Products	21.0
Scams	14.2

Make two bar graphs of these percents, one with bars ordered as in the table (alphabetical) and the other with bars in order from tallest to shortest. Comparisons are easier if you order the bars by height. A bar graph ordered from tallest to shortest bar is sometimes called a **Pareto chart,** after the Italian economist who recommended this procedure.

1.20 Women seeking graduate and professional degrees. The table on the next page gives the percents of women among students seeking various graduate and professional degrees:[13]

(a) Explain clearly why we cannot use a pie chart to display these data.

(b) Make a bar graph of the data. (Comparisons are easier if you order the bars by height.)

Degree	Percent female
Master's in business administration	39.8
Master's in education	76.2
Other master of arts	59.6
Other master of science	53.0
Doctorate in education	70.8
Other PhD degree	54.2
Medicine (MD)	44.0
Law	50.2
Theology	20.2

1.21 **An aging population.** The population of the United States is aging, though less rapidly than in other developed countries. Here is a stemplot of the percents of residents aged 65 and over in the 50 states, according to the 2000 census. The stems are whole percents and the leaves are tenths of a percent.

```
 5 | 7
 6 |
 7 |
 8 | 5
 9 | 679
10 | 6
11 | 02233677
12 | 0011113445789
13 | 00012233345568
14 | 034579
15 | 36
16 |
17 | 6
```

(a) There are two outliers: Alaska has the lowest percent of older residents, and Florida has the highest. What are the percents for these two states?

(b) Ignoring Alaska and Florida, describe the shape, center, and spread of this distribution.

1.22 **Split the stems.** Make another stemplot of the percent of residents aged 65 and over in the states other than Alaska and Florida by splitting stems 8 to 15 in the plot from the previous exercise. Which plot do you prefer? Why?

1.23 **Diabetes and glucose.** People with diabetes must monitor and control their blood glucose level. The goal is to maintain "fasting plasma glucose" between about 90 and 130 milligrams per deciliter (mg/dl). Here are the fasting plasma glucose levels for 18 diabetics enrolled in a diabetes control class, five months after the end of the class:[14]

141	158	112	153	134	95	96	78	148
172	200	271	103	172	359	145	147	255

Make a stemplot of these data and describe the main features of the distribution. (You will want to trim and also split stems.) Are there outliers? How well is the group as a whole achieving the goal for controlling glucose levels?

1.24 **Compare glucose of instruction and control groups.** The study described in the previous exercise also measured the fasting plasma glucose of 16 diabetics who were given individual instruction on diabetes control. Here are the data:

128	195	188	158	227	198	163	164
159	128	283	226	223	221	220	160

Make a back-to-back stemplot to compare the class and individual instruction groups. How do the distribution shapes and success in achieving the glucose control goal compare?

1.25 **Vocabulary scores of seventh-grade students.** Figure 1.14 displays the scores of all 947 seventh-grade students in the public schools of Gary, Indiana, on the vocabulary part of the Iowa Test of Basic Skills.[15] Give a brief description of the overall pattern (shape, center, spread) of this distribution.

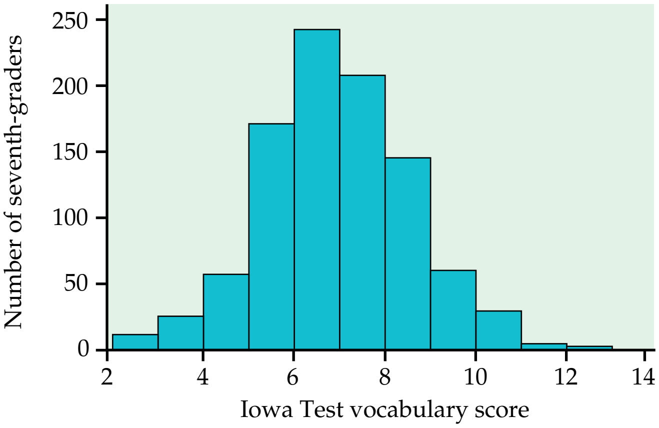


FIGURE 1.14 Histogram of the Iowa Test of Basic Skills vocabulary scores of seventh-grade students in Gary, Indiana, for Exercise 1.25.

1.26 **Shakespeare's plays.** Figure 1.15 is a histogram of the lengths of words used in Shakespeare's plays. Because there are so many words in the plays, we use a histogram of percents. What is the overall shape of this distribution? What does this shape say about word lengths in Shakespeare? Do you expect other authors to have word length distributions of the same general shape? Why?

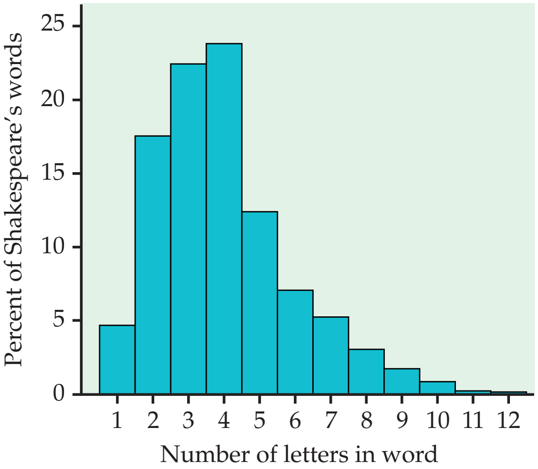


FIGURE 1.15 Histogram of lengths of words used in Shakespeare's plays, for Exercise 1.26.

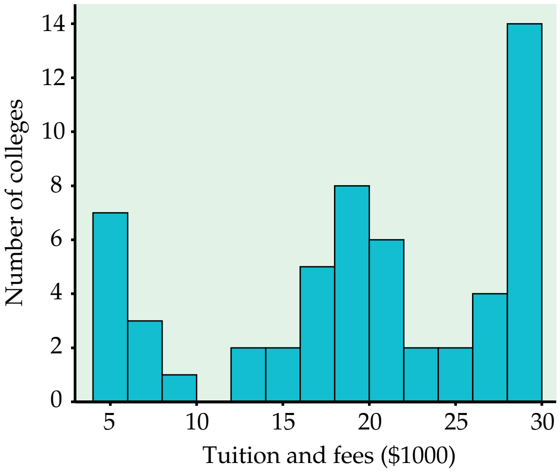


FIGURE 1.16 Histogram of the tuition and fees charged by four-year colleges in Massachusetts, for Exercise 1.27.

1.27 College tuition and fees. Jeanna plans to attend college in her home state of Massachusetts. She looks up the tuition and fees for all 56 four-year colleges in Massachusetts (omitting art schools and other special colleges). Figure 1.16 is a histogram of the data. For state schools, Jeanna used the in-state tuition. What is the most important aspect of the overall pattern of this distribution? Why do you think this pattern appears?

1.28 Tornado damage. The states differ greatly in the kinds of severe weather that afflict them. Table 1.5 shows the average property damage caused by tornadoes per year over the period from 1950 to 1999 in each of the 50 states and Puerto Rico.[16] (To adjust for the changing buying power of the dollar over time, all damages were restated in 1999 dollars.)

(a) What are the top five states for tornado damage? The bottom five?

(b) Make a histogram of the data, by hand or using software, with classes "0 ≤ damage < 10," "10 ≤ damage < 20," and so on. Describe the shape, center, and spread of the distribution. Which states

TABLE 1.5

Average property damage per year due to tornadoes

State	Damage ($millions)	State	Damage ($millions)	State	Damage ($millions)
Alabama	51.88	Louisiana	27.75	Ohio	44.36
Alaska	0.00	Maine	0.53	Oklahoma	81.94
Arizona	3.47	Maryland	2.33	Oregon	5.52
Arkansas	40.96	Massachusetts	4.42	Pennsylvania	17.11
California	3.68	Michigan	29.88	Puerto Rico	0.05
Colorado	4.62	Minnesota	84.84	Rhode Island	0.09
Connecticut	2.26	Mississippi	43.62	South Carolina	17.19
Delaware	0.27	Missouri	68.93	South Dakota	10.64
Florida	37.32	Montana	2.27	Tennessee	23.47
Georgia	51.68	Nebraska	30.26	Texas	88.60
Hawaii	0.34	Nevada	0.10	Utah	3.57
Idaho	0.26	New Hampshire	0.66	Vermont	0.24
Illinois	62.94	New Jersey	2.94	Virginia	7.42
Indiana	53.13	New Mexico	1.49	Washington	2.37
Iowa	49.51	New York	15.73	West Virginia	2.14
Kansas	49.28	North Carolina	14.90	Wisconsin	31.33
Kentucky	24.84	North Dakota	14.69	Wyoming	1.78

may be outliers? (To understand the outliers, note that most tornadoes in largely rural states such as Kansas cause little property damage. Damage to crops is not counted as property damage.)

(c) If you are using software, also display the "default" histogram that your software makes when you give it no instructions. How does this compare with your graph in (b)?

1.29 APPLET **Use an applet for the tornado damage data.** The *One-Variable Statistical Calculator* applet on the text CD and Web site will make stemplots and histograms. It is intended mainly as a learning tool rather than as a replacement for statistical software. The histogram function is particularly useful because you can change the number of classes by dragging with the mouse. The tornado damage data from Table 1.5 are available in the applet. Choose this data set and go to the "Histogram" tab.

(a) Sketch the default histogram that the applet first presents. If the default graph does not have nine classes, drag it to make a histogram with nine classes and sketch the result. This should agree with your histogram in part (b) of the previous exercise.

(b) Make a histogram with one class and also a histogram with the greatest number of classes that the applet allows. Sketch the results.

(c) Drag the graph until you find the histogram that you think best pictures the data. How many classes did you choose? Sketch your final histogram.

1.30 **Carbon dioxide from burning fuels.** Burning fuels in power plants or motor vehicles emits carbon dioxide (CO_2), which contributes to global warming. Table 1.6 displays CO_2 emissions per person from countries with population at least 20 million.[17]

(a) Why do you think we choose to measure emissions per person rather than total CO_2 emissions for each country?

(b) Display the data of Table 1.6 in a graph. Describe the shape, center, and spread of the distribution. Which countries are outliers?

1.31 **California temperatures.** Table 1.7 contains data on the mean annual temperatures (degrees Fahrenheit) for the years 1951 to 2000 at two locations in California: Pasadena and Redding.[18] Make time plots of both time series and compare their main features. You can see why discussions of climate change often bring disagreement.

1.32 **What do you miss in the histogram?** Make a histogram of the mean annual temperatures

TABLE 1.6

Carbon dioxide emissions (metric tons per person)

Country	CO_2	Country	CO_2
Algeria	2.3	Mexico	3.7
Argentina	3.9	Morocco	1.0
Australia	17.0	Myanmar	0.2
Bangladesh	0.2	Nepal	0.1
Brazil	1.8	Nigeria	0.3
Canada	16.0	Pakistan	0.7
China	2.5	Peru	0.8
Columbia	1.4	Tanzania	0.1
Congo	0.0	Philippines	0.9
Egypt	1.7	Poland	8.0
Ethiopia	0.0	Romania	3.9
France	6.1	Russia	10.2
Germany	10.0	Saudi Arabia	11.0
Ghana	0.2	South Africa	8.1
India	0.9	Spain	6.8
Indonesia	1.2	Sudan	0.2
Iran	3.8	Thailand	2.5
Iraq	3.6	Turkey	2.8
Italy	7.3	Ukraine	7.6
Japan	9.1	United Kingdom	9.0
Kenya	0.3	United States	19.9
Korea, North	9.7	Uzbekistan	4.8
Korea, South	8.8	Venezuela	5.1
Malaysia	4.6	Vietnam	0.5

TABLE 1.7

Mean annual temperatures (°F) in two California cities

	Mean Temperature			Mean Temperature	
Year	Pasadena	Redding	Year	Pasadena	Redding
1951	62.27	62.02	1976	64.23	63.51
1952	61.59	62.27	1977	64.47	63.89
1953	62.64	62.06	1978	64.21	64.05
1954	62.88	61.65	1979	63.76	60.38
1955	61.75	62.48	1980	65.02	60.04
1956	62.93	63.17	1981	65.80	61.95
1957	63.72	62.42	1982	63.50	59.14
1958	65.02	64.42	1983	64.19	60.66
1959	65.69	65.04	1984	66.06	61.72
1960	64.48	63.07	1985	64.44	60.50
1961	64.12	63.50	1986	65.31	61.76
1962	62.82	63.97	1987	64.58	62.94
1963	63.71	62.42	1988	65.22	63.70
1964	62.76	63.29	1989	64.53	61.50
1965	63.03	63.32	1990	64.96	62.22
1966	64.25	64.51	1991	65.60	62.73
1967	64.36	64.21	1992	66.07	63.59
1968	64.15	63.40	1993	65.16	61.55
1969	63.51	63.77	1994	64.63	61.63
1970	64.08	64.30	1995	65.43	62.62
1971	63.59	62.23	1996	65.76	62.93
1972	64.53	63.06	1997	66.72	62.48
1973	63.46	63.75	1998	64.12	60.23
1974	63.93	63.80	1999	64.85	61.88
1975	62.36	62.66	2000	66.25	61.58

at Pasadena for the years 1951 to 2000. (Data appear in Table 1.7.) Describe the distribution of temperatures. Then explain why this histogram misses very important facts about temperatures in Pasadena.

1.33 **Change the scale of the axis.** The impression that a time plot gives depends on the scales you use on the two axes. If you stretch the vertical axis and compress the time axis, change appears to be more rapid. Compressing the vertical axis and stretching the time axis make change appear slower. Make two more time plots of the data for Pasadena in Table 1.7, one that makes mean temperature appear to increase very rapidly and one that shows only a slow increase. The moral of this exercise is: *pay close attention to the scales when you look at a time plot.*

1.34 **Fish in the Bering Sea.** "Recruitment," the addition of new members to a fish population, is an important measure of the health of ocean ecosystems. Here are data on the recruitment of rock sole in the Bering Sea between 1973 and 2000:[19]

Year	Recruitment (millions)	Year	Recruitment (millions)
1973	173	1987	4700
1974	234	1988	1702
1975	616	1989	1119
1976	344	1990	2407
1977	515	1991	1049
1978	576	1992	505
1979	727	1993	998
1980	1411	1994	505
1981	1431	1995	304
1982	1250	1996	425
1983	2246	1997	214
1984	1793	1998	385
1985	1793	1999	445
1986	2809	2000	676

(a) Make a graph to display the distribution of rock sole recruitment, then describe the pattern and any striking deviations that you see.

(b) Make a time plot of recruitment and describe its pattern. As is often the case with time series data, a time plot is needed to understand what is happening.

1.35 **Thinness in Asia.** Asian culture does not emphasize thinness, but young Asians are often influenced by Western culture. In a study of concerns about weight among young Korean women, researchers administered the Drive for Thinness scale (a questionnaire) to 264 female college students in Seoul, South Korea.[20] Drive for Thinness measures excessive concern with weight and dieting and fear of weight gain. Roughly speaking, a score of 15 is typical of Western women with eating disorders but is unusually high (90th percentile) for other Western women. Graph the data and describe the shape, center, and spread of the distribution of Drive for Thinness scores for these Korean students. Are there any outliers?

1.36 **Acidity of rainwater.** Changing the choice of classes can change the appearance of a histogram. Here is an example in which a small shift in the classes, with no change in the number of classes, has an important effect on the histogram. The data are the acidity levels (measured by pH) in 105 samples of rainwater. Distilled water has pH 7.00. As the water becomes more acidic, the pH goes down. The pH of rainwater is important to environmentalists because of the problem of acid rain.[21]

4.33	4.38	4.48	4.48	4.50	4.55	4.59	4.59
4.61	4.61	4.75	4.76	4.78	4.82	4.82	4.83
4.86	4.93	4.94	4.94	4.94	4.96	4.97	5.00
5.01	5.02	5.05	5.06	5.08	5.09	5.10	5.12
5.13	5.15	5.15	5.15	5.16	5.16	5.16	5.18
5.19	5.23	5.24	5.29	5.32	5.33	5.35	5.37
5.37	5.39	5.41	5.43	5.44	5.46	5.46	5.47
5.50	5.51	5.53	5.55	5.55	5.56	5.61	5.62
5.64	5.65	5.65	5.66	5.67	5.67	5.68	5.69
5.70	5.75	5.75	5.75	5.76	5.76	5.79	5.80
5.81	5.81	5.81	5.81	5.85	5.85	5.90	5.90
6.00	6.03	6.03	6.04	6.04	6.05	6.06	6.07
6.09	6.13	6.21	6.34	6.43	6.61	6.62	6.65
6.81							

(a) Make a histogram of pH with 14 classes, using class boundaries 4.2, 4.4, . . . , 7.0. How many modes does your histogram show? More than one mode suggests that the data contain groups that have different distributions.

(b) Make a second histogram, also with 14 classes, using class boundaries 4.14, 4.34, . . . , 6.94. The classes are those from (a) moved 0.06 to the left. How many modes does the new histogram show?

(c) Use your software's histogram function to make a histogram without specifying the number of classes or their boundaries. How does the software's default histogram compare with those in (a) and (b)?

1.37 **Identify the histograms.** A survey of a large college class asked the following questions:

1. Are you female or male? (In the data, male = 0, female = 1.)

2. Are you right-handed or left-handed? (In the data, right = 0, left = 1.)

3. What is your height in inches?

4. How many minutes do you study on a typical weeknight?

Figure 1.17 shows histograms of the student responses, in scrambled order and without scale markings. Which histogram goes with each variable? Explain your reasoning.

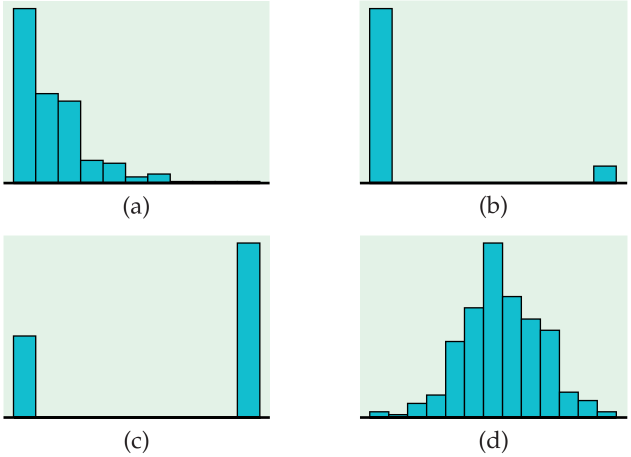


FIGURE 1.17 Match each histogram with its variable, for Exercise 1.37.

1.38 Sketch a skewed distribution. Sketch a histogram for a distribution that is skewed to the left. Suppose that you and your friends emptied your pockets of coins and recorded the year marked on each coin. The distribution of dates would be skewed to the left. Explain why.

1.39 Oil wells. How much oil the wells in a given field will ultimately produce is key information in deciding whether to drill more wells. Here are the estimated total amounts of oil recovered from 64 wells in the Devonian Richmond Dolomite area of the Michigan basin, in thousands of barrels:[22]

21.7	53.2	46.4	42.7	50.4	97.7	103.1	51.9
43.4	69.5	156.5	34.6	37.9	12.9	2.5	31.4
79.5	26.9	18.5	14.7	32.9	196.0	24.9	118.2
82.2	35.1	47.6	54.2	63.1	69.8	57.4	65.6
56.4	49.4	44.9	34.6	92.2	37.0	58.8	21.3
36.6	64.9	14.8	17.6	29.1	61.4	38.6	32.5
12.0	28.3	204.9	44.5	10.3	37.7	33.7	81.1
12.1	20.1	30.5	7.1	10.1	18.0	3.0	2.0

Graph the distribution and describe its main features.

1.40 The density of the earth. In 1798 the English scientist Henry Cavendish measured the density of the earth by careful work with a torsion balance. The variable recorded was the density of the earth as a multiple of the density of water. Here are Cavendish's 29 measurements:[23]

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29
5.58	5.65	5.57	5.53	5.62	5.29	5.44	5.34
5.79	5.10	5.27	5.39	5.42	5.47	5.63	5.34
5.46	5.30	5.75	5.68	5.85			

Present these measurements graphically by either a stemplot or a histogram and explain the reason for your choice. Then briefly discuss the main features of the distribution. In particular, what is your estimate of the density of the earth based on these measurements?

1.41 Time spent studying. Do women study more than men? We asked the students in a large first-year college class how many minutes they studied on a typical weeknight. Here are the responses of random samples of 30 women and 30 men from the class:

Women					Men				
180	120	180	360	240	90	120	30	90	200
120	180	120	240	170	90	45	30	120	75
150	120	180	180	150	150	120	60	240	300
200	150	180	150	180	240	60	120	60	30
120	60	120	180	180	30	230	120	95	150
90	240	180	115	120	0	200	120	120	180

(a) Examine the data. Why are you not surprised that most responses are multiples of 10 minutes? We eliminated one student who claimed to study 30,000 minutes per night. Are there any other responses you consider suspicious?

(b) Make a back-to-back stemplot of these data. Report the approximate midpoints of both groups. Does it appear that women study more than men (or at least claim that they do)?

1.42 Guinea pigs. Table 1.8 gives the survival times in days of 72 guinea pigs after they were injected with tubercle bacilli in a medical experiment.[24] Make a suitable graph and describe the shape, center, and spread of the distribution of survival times. Are there any outliers?

1.43 Grades and self-concept. Table 1.9 presents data on 78 seventh-grade students in a rural midwestern school.[25] The researcher was interested in the relationship between the students' "self-concept" and their academic performance. The data we give here include each student's grade point average (GPA), score on a standard IQ test, and gender, taken from school records. Gender is coded as F for female and M for male. The students are identified only by an observation number (OBS). The missing OBS numbers show that some students dropped out of the study. The final variable is each student's score on the Piers-Harris Children's Self-Concept Scale, a psychological test administered by the researcher.

(a) How many variables does this data set contain? Which are categorical variables and which are quantitative variables?

(b) Make a stemplot of the distribution of GPA, after rounding to the nearest tenth of a point.

TABLE 1.8

Survival times (days) of guinea pigs in a medical experiment

43	45	53	56	56	57	58	66	67	73
74	79	80	80	81	81	81	82	83	83
84	88	89	91	91	92	92	97	99	99
100	100	101	102	102	102	103	104	107	108
109	113	114	118	121	123	126	128	137	138
139	144	145	147	156	162	174	178	179	184
191	198	211	214	243	249	329	380	403	511
522	598								

TABLE 1.9

Educational data for 78 seventh-grade students

OBS	GPA	IQ	Gender	Self-concept	OBS	GPA	IQ	Gender	Self-concept
001	7.940	111	M	67	043	10.760	123	M	64
002	8.292	107	M	43	044	9.763	124	M	58
003	4.643	100	M	52	045	9.410	126	M	70
004	7.470	107	M	66	046	9.167	116	M	72
005	8.882	114	F	58	047	9.348	127	M	70
006	7.585	115	M	51	048	8.167	119	M	47
007	7.650	111	M	71	050	3.647	97	M	52
008	2.412	97	M	51	051	3.408	86	F	46
009	6.000	100	F	49	052	3.936	102	M	66
010	8.833	112	M	51	053	7.167	110	M	67
011	7.470	104	F	35	054	7.647	120	M	63
012	5.528	89	F	54	055	0.530	103	M	53
013	7.167	104	M	54	056	6.173	115	M	67
014	7.571	102	F	64	057	7.295	93	M	61
015	4.700	91	F	56	058	7.295	72	F	54
016	8.167	114	F	69	059	8.938	111	F	60
017	7.822	114	F	55	060	7.882	103	F	60
018	7.598	103	F	65	061	8.353	123	M	63
019	4.000	106	M	40	062	5.062	79	M	30
020	6.231	105	F	66	063	8.175	119	M	54
021	7.643	113	M	55	064	8.235	110	M	66
022	1.760	109	M	20	065	7.588	110	M	44
024	6.419	108	F	56	068	7.647	107	M	49
026	9.648	113	M	68	069	5.237	74	F	44
027	10.700	130	F	69	071	7.825	105	M	67
028	10.580	128	M	70	072	7.333	112	F	64
029	9.429	128	M	80	074	9.167	105	M	73
030	8.000	118	M	53	076	7.996	110	M	59
031	9.585	113	M	65	077	8.714	107	F	37
032	9.571	120	F	67	078	7.833	103	F	63
033	8.998	132	F	62	079	4.885	77	M	36
034	8.333	111	F	39	080	7.998	98	F	64
035	8.175	124	M	71	083	3.820	90	M	42
036	8.000	127	M	59	084	5.936	96	F	28
037	9.333	128	F	60	085	9.000	112	F	60
038	9.500	136	M	64	086	9.500	112	F	70
039	9.167	106	M	71	087	6.057	114	M	51
040	10.140	118	F	72	088	6.057	93	F	21
041	9.999	119	F	54	089	6.938	106	M	56

(c) Describe the shape, center, and spread of the GPA distribution. Identify any suspected outliers from the overall pattern.

(d) Make a back-to-back stemplot of the rounded GPAs for female and male students. Write a brief comparison of the two distributions.

1.44 Describe the IQ scores. Make a graph of the distribution of IQ scores for the seventh-grade students in Table 1.9. Describe the shape, center, and spread of the distribution, as well as any outliers. IQ scores are usually said to be centered at 100. Is the midpoint for these students close to 100, clearly above, or clearly below?

1.45 Describe the self-concept scores. Based on a suitable graph, briefly describe the distribution of self-concept scores for the students in Table 1.9. Be sure to identify any suspected outliers.

1.46 The Boston Marathon. Women were allowed to enter the Boston Marathon in 1972. The following table gives the times (in minutes, rounded to the nearest minute) for the winning women from 1972 to 2006.

Year	Time	Year	Time	Year	Time	Year	Time
1972	190	1981	147	1990	145	1999	143
1973	186	1982	150	1991	144	2000	146
1974	167	1983	143	1992	144	2001	144
1975	162	1984	149	1993	145	2002	141
1976	167	1985	154	1994	142	2003	145
1977	168	1986	145	1995	145	2004	144
1978	165	1987	146	1996	147	2005	145
1979	155	1988	145	1997	146	2006	143
1980	154	1989	144	1998	143		

Make a graph that shows change over time. What overall pattern do you see? Have times stopped improving in recent years? If so, when did improvement end?

1.2 Describing Distributions with Numbers

Interested in a sporty car? Worried that it may use too much gas? The Environmental Protection Agency lists most such vehicles in its "two-seater" or "minicompact" categories. Table 1.10 gives the city and highway gas mileage for cars in these groups.[26] (The mileages are for the basic engine and transmission combination for each car.) We want to compare two-seaters with minicompacts and city mileage with highway mileage. We can begin with graphs, but numerical summaries make the comparisons more specific.

A brief description of a distribution should include its *shape* and numbers describing its *center* and *spread*. We describe the shape of a distribution based on inspection of a histogram or a stemplot. Now we will learn specific ways to use numbers to measure the center and spread of a distribution. We can calculate these numerical measures for any quantitative variable. But to interpret measures of center and spread, and to choose among the several measures we will learn, you must think about the shape of the distribution and the meaning of the data. The numbers, like graphs, are aids to understanding, not "the answer" in themselves.

Measuring center: the mean

Numerical description of a distribution begins with a measure of its center or average. The two common measures of center are the *mean* and the *median*. The mean is the "average value" and the median is the "middle value." These are two different ideas for "center," and the two measures behave differently. We need precise recipes for the mean and the median.

TABLE 1.10

Fuel economy (miles per gallon) for 2004 model vehicles

Two-Seater Cars			Minicompact Cars		
Model	City	Highway	Model	City	Highway
Acura NSX	17	24	Aston Martin Vanquish	12	19
Audi TT Roadster	20	28	Audi TT Coupe	21	29
BMW Z4 Roadster	20	28	BMW 325CI	19	27
Cadillac XLR	17	25	BMW 330CI	19	28
Chevrolet Corvette	18	25	BMW M3	16	23
Dodge Viper	12	20	Jaguar XK8	18	26
Ferrari 360 Modena	11	16	Jaguar XKR	16	23
Ferrari Maranello	10	16	Lexus SC 430	18	23
Ford Thunderbird	17	23	Mini Cooper	25	32
Honda Insight	60	66	Mitsubishi Eclipse	23	31
Lamborghini Gallardo	9	15	Mitsubishi Spyder	20	29
Lamborghini Murcielago	9	13	Porsche Cabriolet	18	26
Lotus Esprit	15	22	Porsche Turbo 911	14	22
Maserati Spyder	12	17			
Mazda Miata	22	28			
Mercedes-Benz SL500	16	23			
Mercedes-Benz SL600	13	19			
Nissan 350Z	20	26			
Porsche Boxster	20	29			
Porsche Carrera 911	15	23			
Toyota MR2	26	32			

THE MEAN $\bar{x}$

To find the **mean** $\bar{x}$ of a set of observations, add their values and divide by the number of observations. If the n observations are $x_1, x_2, \ldots, x_n$, their mean is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

or, in more compact notation,

$$\bar{x} = \frac{1}{n}\sum x_i$$

The $\sum$ (capital Greek sigma) in the formula for the mean is short for "add them all up." The bar over the x indicates the mean of all the x-values. Pronounce the mean $\bar{x}$ as "x-bar." This notation is so common that writers who are discussing data use $\bar{x}$, $\bar{y}$, etc. without additional explanation. The subscripts on the observations x_i are just a way of keeping the n observations separate. They do not necessarily indicate order or any other special facts about the data.

EXAMPLE

1.14 Highway mileage for two-seaters. The mean highway mileage for the 21 two-seaters in Table 1.10 is

$$\begin{aligned}\bar{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{24 + 28 + 28 + \cdots + 32}{21} \\ &= \frac{518}{21} = 24.7 \text{ miles per gallon}\end{aligned}$$

In practice, you can key the data into your calculator and hit the $\bar{x}$ key.

USE YOUR KNOWLEDGE

1.47 Find the mean. Here are the scores on the first exam in an introductory statistics course for 10 students:

80 73 92 85 75 98 93 55 80 90

Find the mean first-exam score for these students.

resistant measure

The data for Example 1.14 contain an outlier: the Honda Insight is a hybrid gas-electric car that doesn't belong in the same category as the 20 gasoline-powered two-seater cars. If we exclude the Insight, the mean highway mileage drops to 22.6 mpg. The single outlier adds more than 2 mpg to the mean highway mileage. This illustrates an important weakness of the mean as a measure of center: *the mean is sensitive to the influence of a few extreme observations.* These may be outliers, but a skewed distribution that has no outliers will also pull the mean toward its long tail. Because the mean cannot resist the influence of extreme observations, we say that it is not a **resistant measure** of center. A measure that is resistant does more than limit the influence of outliers. Its value does not respond strongly to changes in a few observations, no matter how large those changes may be. The mean fails this requirement because we can make the mean as large as we wish by making a large enough increase in just one observation.

Measuring center: the median

We used the midpoint of a distribution as an informal measure of center in the previous section. The *median* is the formal version of the midpoint, with a specific rule for calculation.

THE MEDIAN *M*

The **median *M*** is the midpoint of a distribution. Half the observations are smaller than the median and the other half are larger than the median. Here is a rule for finding the median:

1. Arrange all observations in order of size, from smallest to largest.

2. If the number of observations n is odd, the median M is the center observation in the ordered list. Find the location of the median by counting $(n+1)/2$ observations up from the bottom of the list.

3. If the number of observations n is even, the median M is the mean of the two center observations in the ordered list. The location of the median is again $(n+1)/2$ from the bottom of the list.

Note that the formula $(n+1)/2$ does *not* give the median, just the location of the median in the ordered list. Medians require little arithmetic, so they are easy to find by hand for small sets of data. Arranging even a moderate number of observations in order is tedious, however, so that finding the median by hand for larger sets of data is unpleasant. Even simple calculators have an $\bar{x}$ button, but you will need computer software or a graphing calculator to automate finding the median.

EXAMPLE

1.15 Find the median. To find the median highway mileage for 2004 model two-seater cars, arrange the data in increasing order:

13 15 16 16 17 19 20 22 23 23 **23** 24 25 25 26 28 28 28 29 32 66

Be sure to list *all* observations, even if they repeat the same value. The median is the bold 23, the 11th observation in the ordered list. You can find the median by eye—there are 10 observations to the left and 10 to the right. Or you can use the recipe $(n+1)/2 = 22/2 = 11$ to locate the median in the list.

What happens if we drop the Honda Insight? The remaining 20 cars have highway mileages

13 15 16 16 17 19 20 22 23 **23 23** 24 25 25 26 28 28 28 29 32

Because the number of observations $n = 20$ is even, there is no center observation. There is a center pair—the bold pair of 23s have 9 observations to their left and 9 to their right. The median M is the mean of the center pair, which is 23. The recipe $(n+1)/2 = 21/2 = 10.5$ for the position of the median in the list says that the median is at location "ten and one-half," that is, halfway between the 10th and 11th observations.

APPLET

You see that the median is more resistant than the mean. Removing the Honda Insight did not change the median at all. Even if we mistakenly enter the Insight's mileage as 660 rather than 66, the median remains 23. The very high value is simply one observation to the right of center. The *Mean and Median* applet on the text CD and Web site is an excellent way to compare the resistance of M and $\bar{x}$. See Exercises 1.75 to 1.77 for use of this applet.

USE YOUR KNOWLEDGE

1.48 Find the median. Here are the scores on the first exam in an introductory statistics course for 10 students:

80 73 92 85 75 98 93 55 80 90

Find the median first-exam score for these students.

Mean versus median

The median and mean are the most common measures of the center of a distribution. The mean and median of a symmetric distribution are close together. If the distribution is exactly symmetric, the mean and median are exactly the same. In a skewed distribution, the mean is farther out in the long tail than is the median. The endowment for a college or university is money set aside and invested. The income from the endowment is usually used to support various programs. The distribution of the sizes of the endowments of colleges and universities is strongly skewed to the right. Most institutions have modest endowments, but a few are very wealthy. The median endowment of colleges and universities in a recent year was $70 million—but the mean endowment was over $320 million. The few wealthy institutions pulled the mean up but did not affect the median. *Don't confuse the "average" value of a variable (the mean) with its "typical" value, which we might describe by the median.*

CAUTION

We can now give a better answer to the question of how to deal with outliers in data. First, look at the data to identify outliers and investigate their causes. You can then correct outliers if they are wrongly recorded, delete them for good reason, or otherwise give them individual attention. The three outliers in Figure 1.9 (page 17) can all be dropped from the data once we discover why they appear. If you have no clear reason to drop outliers, you may want to use resistant methods, so that outliers have little influence over your conclusions. The choice is often a matter for judgment. The government's fuel economy guide lists the Honda Insight with the other two-seaters in Table 1.10. We might choose to report median rather than mean gas mileage for all two-seaters to avoid giving too much influence to one car model. In fact, we think that the Insight doesn't belong, so we will omit it from further analysis of these data.

Measuring spread: the quartiles

A measure of center alone can be misleading. Two nations with the same median family income are very different if one has extremes of wealth and poverty and the other has little variation among families. A drug with the correct mean concentration of active ingredient is dangerous if some batches are much too high and others much too low. We are interested in the *spread* or *variability* of incomes and drug potencies as well as their centers. **The simplest useful numerical description of a distribution consists of both a measure of center and a measure of spread.**

quartile

We can describe the spread or variability of a distribution by giving several percentiles. The median divides the data in two; half of the observations are above the median and half are below the median. We could call the median the 50th percentile. The upper **quartile** is the median of the upper half of the data. Similarly, the lower quartile is the median of the lower half of the data. With the median, the quartiles divide the data into four equal parts; 25% of the data are in each part.

percentile

We can do a similar calculation for any percent. The ***p*th percentile** of a distribution is the value that has p percent of the observations fall at or below it. To calculate a percentile, arrange the observations in increasing order and count up the required percent from the bottom of the list. Our definition of percentiles is a bit inexact because there is not always a value with exactly p

percent of the data at or below it. We will be content to take the nearest observation for most percentiles, but the quartiles are important enough to require an exact rule.

THE QUARTILES Q_1 AND Q_3

To calculate the quartiles:

1. Arrange the observations in increasing order and locate the median *M* in the ordered list of observations.

2. The **first quartile Q_1** is the median of the observations whose position in the ordered list is to the left of the location of the overall median.

3. The **third quartile Q_3** is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

EXAMPLE

1.16 Find the median and the quartiles. The highway mileages of the 20 gasoline-powered two-seater cars, arranged in increasing order, are

13 15 16 16 17 19 20 22 23 23 **|** 23 24 25 25 26 28 28 28 29 32

The median is midway between the center pair of observations. We have marked its position in the list by **|**. The first quartile is the median of the 10 observations to the left of the position of the median. Check that its value is $Q_1 = 18$. Similarly, the third quartile is the median of the 10 observations to the right of the **|**. Check that $Q_3 = 27$.

When there is an odd number of observations, the median is the unique center observation, and the rule for finding the quartiles excludes this center value. The highway mileages of the 13 minicompact cars in Table 1.10 are (in order)

19 22 23 23 23 26 **26** 27 28 29 29 31 32

The median is the bold 26. The first quartile is the median of the 6 observations falling to the left of this point in the list, $Q_1 = 23$. Similarly, $Q_3 = 29$.

We find other percentiles more informally if we are working without software. For example, we take the 90th percentile of the 13 minicompact mileages to be the 12th in the ordered list, because $0.90 \times 13 = 11.7$, which we round to 12. The 90th percentile is therefore 31 mpg.

USE YOUR KNOWLEDGE

1.49 Find the quartiles. Here are the scores on the first-exam in an introductory statistics course for 10 students:

80 73 92 85 75 98 93 55 80 90

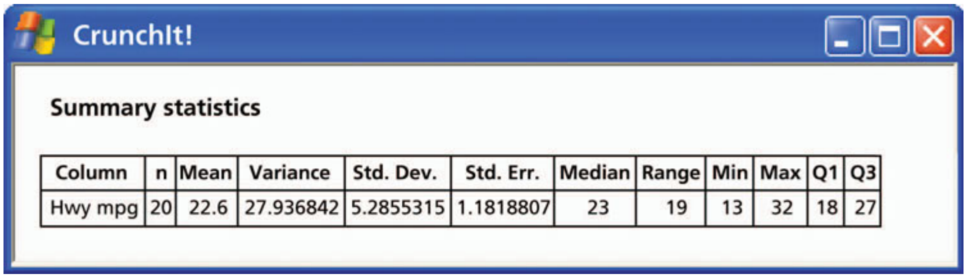
Find the quartiles for these first-exam scores.

EXAMPLE

1.17 Results from software. Statistical software often provides several numerical measures in response to a single command. Figure 1.18 displays such output from the CrunchIt! and Minitab software for the highway mileages of two-seater cars (without the Honda Insight).

Both tell us that there are 20 observations and give the mean, median, quartiles, and smallest and largest data values. Both also give other measures, some of which we will meet soon. CrunchIt! is basic online software that offers no choice of output. Minitab allows you to choose the descriptive measures you want from a long list.

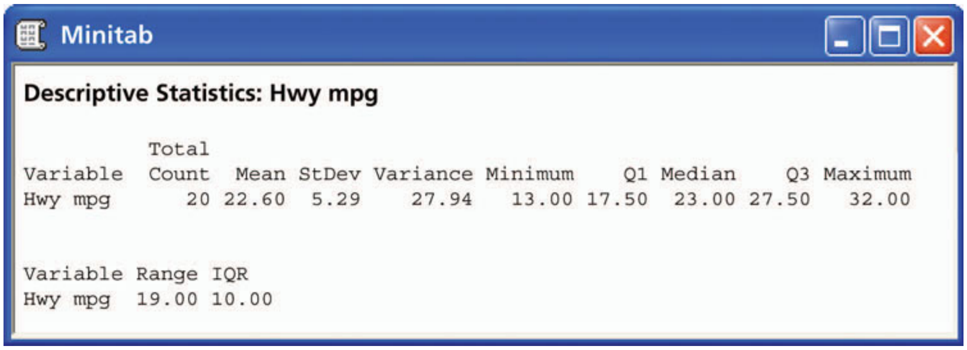
The quartiles from CrunchIt! agree with our values from Example 1.16. But Minitab's quartiles are a bit different. For example, our rule for hand calculation gives first quartile $Q_1 = 18$. Minitab's value is $Q_1 = 17.5$. *There are several rules for calculating quartiles, which often give slightly different values. The differences are always small. For describing data, just report the values that your software gives.*



Summary statistics

Column	n	Mean	Variance	Std. Dev.	Std. Err.	Median	Range	Min	Max	Q1	Q3
Hwy mpg	20	22.6	27.936842	5.2855315	1.1818807	23	19	13	32	18	27

(a)



Descriptive Statistics: Hwy mpg

```
          Total
Variable  Count   Mean  StDev  Variance  Minimum     Q1  Median     Q3  Maximum
Hwy mpg      20  22.60   5.29     27.94    13.00  17.50   23.00  27.50    32.00


Variable  Range    IQR
Hwy mpg   19.00  10.00
```

(b)

FIGURE 1.18 Numerical descriptions of the highway gas mileage of two-seater cars from software, for Example 1.17. (a) CrunchIt! (b) Minitab.

The five-number summary and boxplots

In Section 1.1, we used the smallest and largest observations to indicate the spread of a distribution. These single observations tell us little about the distribution as a whole, but they give information about the tails of the distribution that is missing if we know only Q_1, M, and Q_3. To get a quick summary of both center and spread, combine all five numbers.

THE FIVE-NUMBER SUMMARY

The **five-number summary** of a set of observations consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is

$$\text{Minimum} \quad Q_1 \quad M \quad Q_3 \quad \text{Maximum}$$

These five numbers offer a reasonably complete description of center and spread. The five-number summaries for highway gas mileages are

13 18 23 27 32

for two-seaters and

19 23 26 29 32

for minicompacts. The median describes the center of the distribution; the quartiles show the spread of the center half of the data; the minimum and maximum show the full spread of the data.

USE YOUR KNOWLEDGE

1.50 Find the five-number summary. Here are the scores on the first exam in an introductory statistics course for 10 students:

80 73 92 85 75 98 93 55 80 90

Find the five-number summary for these first-exam scores.

The five-number summary leads to another visual representation of a distribution, the *boxplot*. Figure 1.19 shows boxplots for both city and highway gas mileages for our two groups of cars.

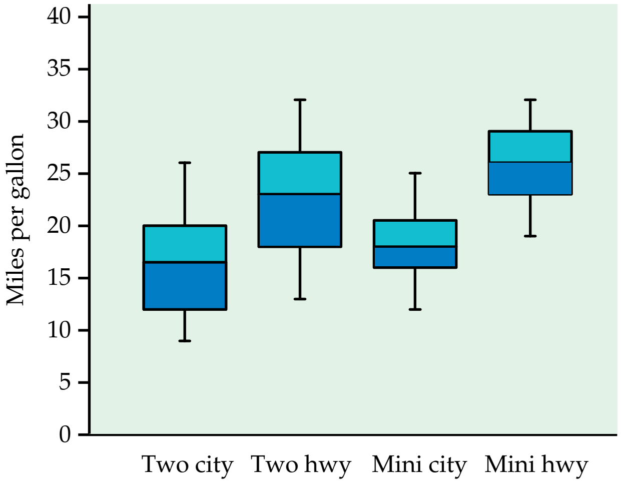


FIGURE 1.19 Boxplots of the highway and city gas mileages for cars classified as two-seaters and as minicompacts by the Environmental Protection Agency.

BOXPLOT

A **boxplot** is a graph of the five-number summary.

- A central box spans the quartiles Q_1 and Q_3.
- A line in the box marks the median M.
- Lines extend from the box out to the smallest and largest observations.

When you look at a boxplot, first locate the median, which marks the center of the distribution. Then look at the spread. The quartiles show the spread of the middle half of the data, and the extremes (the smallest and largest observations) show the spread of the entire data set.

USE YOUR KNOWLEDGE

1.51 Make a boxplot. Here are the scores on the first exam in an introductory statistics course for 10 students:

80 73 92 85 75 98 93 55 80 90

Make a boxplot for these first-exam scores.

Boxplots are particularly effective for comparing distributions as we did in Figure 1.19. We see at once that city mileages are lower than highway mileages. The minicompact cars have slightly higher median gas mileages than the two-seaters, and their mileages are markedly less variable. In particular, the low gas mileages of the Ferraris and Lamborghinis in the two-seater group pull the group minimum down.

The 1.5 × *IQR* rule for suspected outliers

Look again at the 80 service center call lengths in Table 1.1 (page 8). Figure 1.6 (page 12) is a stemplot of their distribution. You can check that the five-number summary is

1 54.5 103.5 200 2631

There is a clear outlier, a call lasting 2631 seconds, more than twice the length of any other call. How shall we describe the spread of this distribution? The smallest and largest observations are extremes that do not describe the spread of the majority of the data. The distance between the quartiles (the range of the center half of the data) is a more resistant measure of spread. This distance is called the *interquartile range*.

THE INTERQUARTILE RANGE *IQR*

The **interquartile range *IQR*** is the distance between the first and third quartiles,

$$IQR = Q_3 - Q_1$$