PART I

Organizing Data:
Looking for Patterns and Departures from Patterns

The Granger Collection, New York

FLORENCE NIGHTINGALE

Using Statistics to Save Lives

Florence Nightingale (1820–1910) won fame as a founder of the nursing profession and as a reformer of health care. As chief nurse for the British army during the Crimean War, from 1854 to 1856, she found that lack of sanitation and disease killed large numbers of soldiers hospitalized by wounds. Her reforms reduced the death rate at her military hospital from 42.7% to 2.2%, and she returned from the war famous. She at once began a fight to reform the entire military health care system, with considerable success.

One of the chief weapons Florence Nightingale used in her efforts was data. She had the facts, because she reformed record keeping as well as medical care. She was a pioneer in using graphs to present data in a vivid form that even generals and members of Parliament could understand. Her inventive graphs are a landmark in the growth of the new science of statistics. She considered statistics essential to understanding any social issue and tried to introduce the study of statistics into higher education.

In beginning our study of statistics, we will follow Florence Nightingale's lead. This chapter and the next will stress the analysis of data as a path to understanding. Like her, we will start with graphs to see what data can teach us. Along with the graphs we will present numerical summaries, just as Florence Nightingale calculated detailed death rates and other summaries. Data for Florence Nightingale were not dry or abstract, because they showed her, and helped her show others, how to save lives. That remains true today.

One of the chief weapons Florence Nightingale used in her efforts was data.

chapter 1

Exploring Data

ACTIVITY 1 How Fast Is Your Heart Beating?

Materials: Clock or watch with second hand

A person's pulse rate provides information about the health of his or her heart. Would you expect to find a difference between male and female pulse rates? In this activity, you and your classmates will collect some data to try to answer this question.

1. To determine your pulse rate, hold the *fingers* of one hand on the artery in your neck or on the inside of the wrist. (The thumb should not be used, because there is a pulse in the thumb.) Count the number of pulse beats in one minute. Do this three times, and calculate your *average* individual pulse rate (add your three pulse rates and divide by 3.) Why is doing this three times better than doing it once?

2. Record the pulse rates for the class in a table, with one column for males and a second column for females. Are there any unusual pulse rates?

3. For now, simply calculate the average pulse rate for the males and the average pulse rate for the females, and compare.

INTRODUCTION

Statistics is the science of data. We begin our study of statistics by mastering the art of examining data. 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 by a set of data. Individuals may be people, but they may also be animals or things.

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

A college's student data base, for example, includes data about every currently enrolled student. The students are the *individuals* described by the data set. For each individual, the data contain the values of *variables* such as age, gender (female or male), choice of major, and grade point average. In practice, any set of data is accompanied by background information that helps us understand the data.

When you meet a new set of data, ask yourself the following questions:

1. Who? What **individuals** do the data describe? **How many** individuals appear in the data?

2. What? How many **variables** are there? What are the **exact definitions** of these variables? In what **units** is each variable recorded? Weights, for example, might be recorded in pounds, in thousands of pounds, or in kilograms. Is there any reason to mistrust the values of any variable?

3. Why? What is the reason the data were gathered? Do we hope to answer some specific questions? Do we want to draw conclusions about individuals other than the ones we actually have data for?

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

CATEGORICAL AND QUANTITATIVE VARIABLES

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

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

EXAMPLE 1.1 EDUCATION IN THE UNITED STATES

Here is a small part of a data set that describes public education in the United States:

State	Region	Population (1000)	SAT Verbal	SAT Math	Percent taking	Percent no HS	Teachers' pay ($1000)
⋮							
CA	PAC	33,871	497	514	49	23.8	43.7
CO	MTN	4,301	536	540	32	15.6	37.1
CT	NE	3,406	510	509	80	20.8	50.7
⋮							

Let's answer the three "W" questions about these data.

1. Who? The *individuals* described are the states. There are 51 of them, the 50 states and the District of Columbia, but we give data for only 3. Each row in the table describes one individual. You will often see each row of data called a ***case.***

case

2. What? Each column contains the values of one variable for all the individuals. This is the usual arrangement in data tables. Seven variables are recorded for each state. The first column identifies the state by its two-letter post office code. We give data for California, Colorado, and Connecticut. The second column says which region of the country the state is in. The Census Bureau divides the nation into nine regions. These three are Pacific, Mountain, and New England. The third column contains state populations, in thousands of people. Be sure to notice that the *units* are thousands of people. California's 33,871 stands for 33,871,000 people. The population data come from the 2000 census. They are therefore quite accurate as of April 1, 2000, but don't show later changes in population.

The remaining five variables are the average scores of the states' high school seniors on the SAT verbal and mathematics exams, the percent of seniors who take the SAT, the percent of students who did not complete high school, and average teachers' salaries in thousands of dollars. Each of these variables needs more explanation before we can fully understand the data.

3. Why? Some people will use these data to evaluate the quality of individual states' educational programs. Others may compare states on one or more of the variables. Future teachers might want to know how much they can expect to earn.

A variable generally takes values that vary. One variable may take values that are very close together while another variable takes values that are quite spread out. We say that the *pattern of variation* of a variable is its *distribution*.

DISTRIBUTION

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

exploratory data analysis

Statistical tools and ideas can help you examine data in order to describe their main features. This examination is called ***exploratory data analysis.*** Like an explorer crossing unknown lands, we first simply describe what we see. Each example we meet will have some background information to help us, but our emphasis is on examining the data. 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 relationships among the variables.
- Begin with a graph or graphs. Then add numerical summaries of specific aspects of the data.

We will organize our learning the same way. Chapters 1 and 2 examine single-variable data, and Chapters 3 and 4 look at relationships among variables. In both settings, we begin with graphs and then move on to numerical summaries.

EXERCISES

1.1 FUEL-EFFICIENT CARS Here is a small part of a data set that describes the fuel economy (in miles per gallon) of 1998 model motor vehicles:

Make and Model	Vehicle type	Transmission type	Number of cylinders	City MPG	Highway MPG
⋮					
BMW 318I	Subcompact	Automatic	4	22	31
BMW 318I	Subcompact	Manual	4	23	32
Buick Century	Midsize	Automatic	6	20	29
Chevrolet Blazer	Four-wheel drive	Automatic	6	16	20
⋮					

(a) What are the individuals in this data set?

(b) For each individual, what variables are given? Which of these variables are categorical and which are quantitative?

1.2 MEDICAL STUDY VARIABLES Data from a medical study contain values of many variables for each of the people who were the subjects of the study. Which of the following variables are categorical and which are quantitative?

(a) Gender (female or male)

(b) Age (years)

(c) Race (Asian, black, white, or other)

(d) Smoker (yes or no)

(e) Systolic blood pressure (millimeters of mercury)

(f) Level of calcium in the blood (micrograms per milliliter)

1.3 You want to compare the "size" of several statistics textbooks. Describe at least three possible numerical variables that describe the "size" of a book. In what *units* would you measure each variable?

1.4 Popular magazines often rank cities in terms of how desirable it is to live and work in each city. Describe five variables that you would measure for each city if you were designing such a study. Give reasons for each of your choices.

1.1 DISPLAYING DISTRIBUTIONS WITH GRAPHS

Displaying categorical variables: bar graphs and pie charts

The values of a categorical variable are labels for the categories, such as "male" and "female." The distribution of a categorical variable lists the categories and gives either the **count** or the **percent** of individuals who fall in each category.

EXAMPLE 1.2 THE MOST POPULAR SOFT DRINK

The following table displays the sales figures and market share (percent of total sales) achieved by several major soft drink companies in 1999. That year, a total of 9930 million cases of soft drink were sold.[1]

Company	Cases sold (millions)	Market share (percent)
Coca-Cola Co.	4377.5	44.1
Pepsi-Cola Co.	3119.5	31.4
Dr. Pepper/7-Up (Cadbury)	1455.1	14.7
Cott Corp.	310.0	3.1
National Beverage	205.0	2.1
Royal Crown	115.4	1.2
Other	347.5	3.4

How to construct a bar graph:

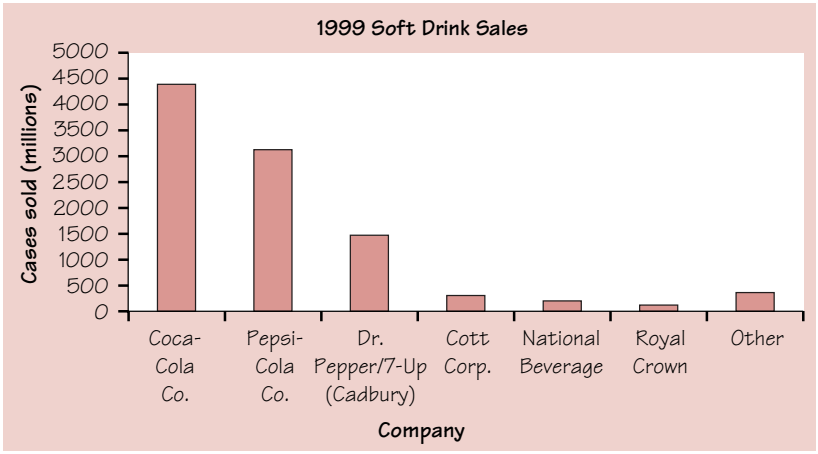
Step 1: Label your axes and title your graph. Draw a set of axes. Label the horizontal axis "Company" and the vertical axis "Cases sold." Title your graph.

Step 2: Scale your axes. Use the counts in each category to help you scale your vertical axis. Write the category names at equally spaced intervals beneath the horizontal axis.

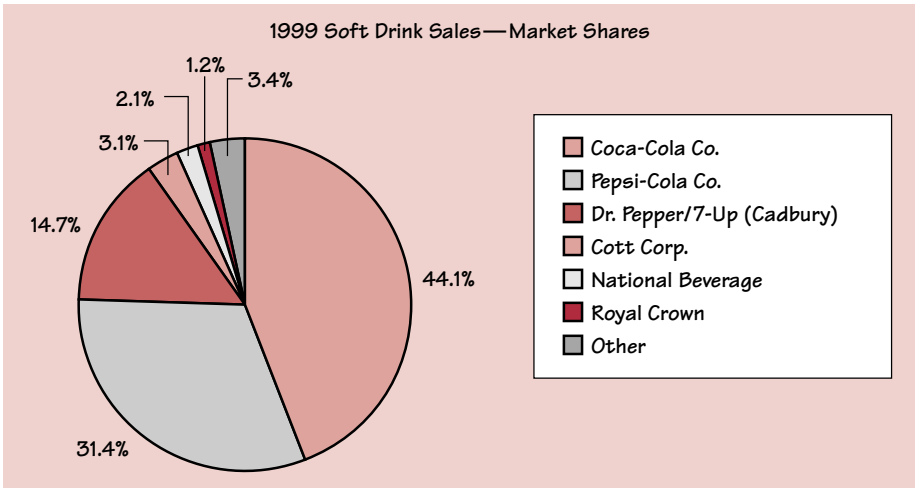
Step 3: Draw a vertical bar above each category name to a height that corresponds to the count in that category. For example, the height of the "Pepsi-Cola Co." bar should be at 3119.5 on the vertical scale. *Leave a space between the bars in a bar graph.*

Figure 1.1(a) displays the completed bar graph.

How to construct a pie chart: Use a computer! Any statistical software package and many spreadsheet programs will construct these plots for you. Figure 1.1(b) is a pie chart for the soft drink sales data.



(a)



(b)

FIGURE 1.1 A bar graph (a) and a pie chart (b) displaying soft drink sales by companies in 1999.

The **bar graph** in Figure 1.1(a) quickly compares the soft drink sales of the companies. The heights of the bars show the counts in the seven categories. The **pie chart** in Figure 1.1(b) helps us see what part of the whole each group forms. For example, the Coca-Cola "slice" makes up 44.1% of the pie because the Coca-Cola Company sold 44.1% of all soft drinks in 1999.

Bar graphs and pie charts help an audience grasp the distribution quickly. To make a pie chart, you must include all the categories that make up a whole. Bar graphs are more flexible.

EXAMPLE 1.3 DO YOU WEAR YOUR SEAT BELT?

In 1998, the National Highway and Traffic Safety Administration (NHTSA) conducted a study on seat belt use. The table below shows the percentage of automobile drivers who were observed to be wearing their seat belts in each region of the United States.[2]

Region	Percent wearing seat belts
Northeast	66.4
Midwest	63.6
South	78.9
West	80.8

Figure 1.2 shows a bar graph for these data. Notice that the vertical scale is measured in percents.

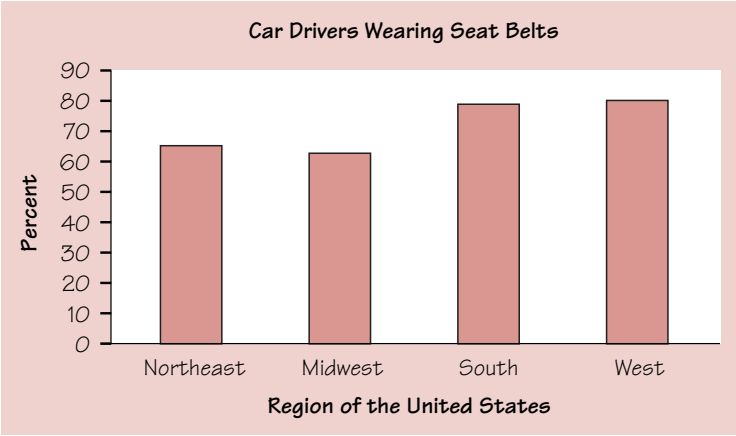


FIGURE 1.2 A bar graph showing the percentage of drivers who wear their seat belts in each of four U.S. regions.

Drivers in the South and West seem to be more concerned about wearing seat belts than those in the Northeast and Midwest. It is not possible to display these data in a single pie chart, because the four percentages cannot be combined to yield a whole (their sum is well over 100%).

EXERCISES

1.5 FEMALE DOCTORATES Here are data on the percent of females among people earning doctorates in 1994 in several fields of study:[3]

Computer science	15.4%	Life sciences	40.7%
Education	60.8%	Physical sciences	21.7%
Engineering	11.1%	Psychology	62.2%

(a) Present these data in a well-labeled bar graph.

(b) Would it also be correct to use a pie chart to display these data? If so, construct the pie chart. If not, explain why not.

1.6 ACCIDENTAL DEATHS In 1997 there were 92,353 deaths from accidents in the United States. Among these were 42,340 deaths from motor vehicle accidents, 11,858 from falls, 10,163 from poisoning, 4051 from drowning, and 3601 from fires.[4]

(a) Find the percent of accidental deaths from each of these causes, rounded to the nearest percent. What percent of accidental deaths were due to other causes?

(b) Make a well-labeled bar graph of the distribution of causes of accidental deaths. Be sure to include an "other causes" bar.

(c) Would it also be correct to use a pie chart to display these data? If so, construct the pie chart. If not, explain why not.

Displaying quantitative variables: dotplots and stemplots

Several types of graphs can be used to display quantitative data. One of the simplest to construct is a **dotplot**.

EXAMPLE 1.4 GOOOOOOOOAAAAALLLLLLLL!!!

The number of goals scored by each team in the first round of the California Southern Section Division V high school soccer playoffs is shown in the following table.[5]

5	0	1	0	7	2	1	0	4	0	3	0	2	0
3	1	5	0	3	0	1	0	1	0	2	0	3	1

How to construct a dotplot:

Step 1: Label your axis and title your graph. Draw a horizontal line and label it with the variable (in this case, number of goals scored). Title your graph.

Step 2: Scale the axis based on the values of the variable.

Step 3: Mark a dot above the number on the horizontal axis corresponding to each data value. Figure 1.3 displays the completed dotplot.

FIGURE 1.3 Goals scored by teams in the California Southern Section Division V high school soccer playoffs.

Making a statistical graph is not an end in itself. After all, a computer or graphing calculator can make graphs faster than we can. The purpose of the graph is to help us understand the data. After you (or your calculator) make a graph, always ask, "What do I see?" Here is a general tactic for looking at graphs: *Look for an overall pattern and also for striking deviations from that pattern.*

OVERALL PATTERN OF A DISTRIBUTION

To describe the overall pattern of a distribution:

- Give the **center** and the **spread**.
- See if the distribution has a simple **shape** that you can describe in a few words.

Section 1.2 tells in detail how to measure center and spread. For now, describe the *center* by finding a value that divides the observations so that about half take larger values and about half have smaller values. In Figure 1.3, the center is 1. That is, a typical team scored about 1 goal in its playoff soccer game. You can describe the *spread* by giving the smallest and largest values. The spread in Figure 1.3 is from 0 goals to 7 goals scored.

The dotplot in Figure 1.3 shows that in most of the playoff games, Division V soccer teams scored very few goals. There were only four teams that scored 4 or more goals. We can say that the distribution has a "long tail" to the right, or that its *shape* is "skewed right." You will learn more about describing shape shortly.

Is the one team that scored 7 goals an *outlier*? This value certainly differs from the overall pattern. To some extent, deciding whether an observation is an outlier is a matter of judgment. We will introduce an objective criterion for determining outliers in Section 1.2.

OUTLIERS

An **outlier** in any graph of data is an individual observation that falls outside the overall pattern of the graph.

Once you have spotted outliers, look for an explanation. Many outliers are due to mistakes, such as typing 4.0 as 40. Other outliers point to the special nature of some observations. Explaining outliers usually requires some background information. Perhaps the soccer team that scored seven goals has some very talented offensive players. Or maybe their opponents played poor defense.

Sometimes the values of a variable are too spread out for us to make a reasonable dotplot. In these cases, we can consider another simple graphical display: a **stemplot.**

EXAMPLE 1.5 WATCH THAT CAFFEINE!

The U.S. Food and Drug Administration limits the amount of caffeine in a 12-ounce can of carbonated beverage to 72 milligrams (mg). Data on the caffeine content of popular soft drinks are provided in Table 1.1. How does the caffeine content of these drinks compare to the USFDA's limit?

TABLE 1.1 Caffeine content (in milligrams) for an 8-ounce serving of popular soft drinks

Brand	Caffeine (mg per 8-oz. serving)	Brand	Caffeine (mg per 8-oz. serving)
A&W Cream Soda	20	IBC Cherry Cola	16
Barq's root beer	15	Kick	38
Cherry Coca-Cola	23	KMX	36
Cherry RC Cola	29	Mello Yello	35
Coca-Cola Classic	23	Mountain Dew	37
Diet A&W Cream Soda	15	Mr. Pibb	27
Diet Cherry Coca-Cola	23	Nehi Wild Red Soda	33
Diet Coke	31	Pepsi One	37
Diet Dr. Pepper	28	Pepsi-Cola	25
Diet Mello Yello	35	RC Edge	47
Diet Mountain Dew	37	Red Flash	27
Diet Mr. Pibb	27	Royal Crown Cola	29
Diet Pepsi-Cola	24	Ruby Red Squirt	26
Diet Ruby Red Squirt	26	Sun Drop Cherry	43
Diet Sun Drop	47	Sun Drop Regular	43
Diet Sunkist Orange Soda	28	Sunkist Orange Soda	28
Diet Wild Cherry Pepsi	24	Surge	35
Dr. Nehi	28	TAB	31
Dr. Pepper	28	Wild Cherry Pepsi	25

Source: National Soft Drink Association, 1999.

The caffeine levels spread from 15 to 47 milligrams for these soft drinks. You could make a dotplot for these data, but a stemplot might be preferable due to the large spread.

How to construct a stemplot:

Step 1: Separate each observation into a *stem* consisting of all but the rightmost digit and a *leaf*, the final digit. A&W Cream Soda has 20 milligrams of caffeine per 8-ounce serving. The number 2 is the stem and 0 is the leaf.

Step 2: Write the stems vertically in increasing order from top to bottom, and draw a vertical line to the right of the stems. Go through the data, writing each leaf to the right of its stem and spacing the leaves equally.

```
1 | 5 5 6
2 | 0 3 9 3 3 8 7 4 6 8 4 8 8 7 5 7 9 6 8 5
3 | 1 5 7 8 6 5 7 3 7 5 1
4 | 7 7 3 3
```

Step 3: Write the stems again, and rearrange the leaves in increasing order out from the stem.

Step 4: Title your graph and add a key describing what the stems and leaves represent. Figure 1.4(a) shows the completed stemplot.

What *shape* does this distribution have? It is difficult to tell with so few stems. We can get a better picture of the caffeine content in soft drinks by "splitting stems." In Figure 1.4(a), the values from 10 to 19 milligrams are placed on the "1" stem. Figure 1.4(b) shows another stemplot of the same data. This time, values having leaves 0 through 4 are placed on one stem, while values ending in 5 through 9 are placed on another stem.

Now the bimodal (two-peaked) *shape* of the distribution is clear. Most soft drinks seem to have between 25 and 29 milligrams or between 35 and 38 milligrams of caffeine per 8-ounce serving. The center of the distribution is 28 milligrams per 8-ounce serving. At first glance, it looks like none of these soft drinks even comes close to the USFDA's caffeine limit of 72 milligrams per 12-ounce serving. Be careful! The values in the stemplot are given in milligrams per 8-ounce serving. Two soft drinks have caffeine levels of 47 milligrams per 8-ounce serving. A 12-ounce serving of these beverages would have 1.5(47) = 70.5 milligrams of caffeine. Always check the units of measurement!

CAFFEINE CONTENT (MG) PER 8-OUNCE SERVING OF VARIOUS SOFT DRINKS

```
1 | 5 5 6
2 | 0 3 3 3 4 4 5 5 6 6 7 7 7 8 8 8 8 8 9 9
3 | 1 1 3 5 5 5 5 6 7 7 7 8
4 | 3 3 7 7
```

(a)

Key:
3|5 means the soft drink contains 35 mg of caffeine per 8-ounce serving.

```
1 | 5 5 6
2 | 0 3 3 3 4 4
2 | 5 5 6 6 7 7 7 8 8 8 8 8 9 9
3 | 1 1 3
3 | 5 5 5 6 7 7 7 8
4 | 3 3
4 | 7 7
```

(b)

Key:
2|8 means the soft drink contains 28 mg of caffeine per 8-ounce serving.

FIGURE 1.4 Two stemplots showing the caffeine content (mg) of various soft drinks. Figure 1.4(b) improves on the stemplot of Figure 1.4(a) by splitting stems.

Here are a few tips for you to consider when you want to construct a stemplot:

- Whenever you split stems, be sure that each stem is assigned an equal number of possible leaf digits.
- There is no magic number of stems to use. Too few stems will result in a skyscraper-shaped plot, while too many stems will yield a very flat "pancake" graph.

- Five stems is a good minimum.
- You can get more flexibility by *rounding* the data so that the final digit after rounding is suitable as a leaf. Do this when the data have too many digits.

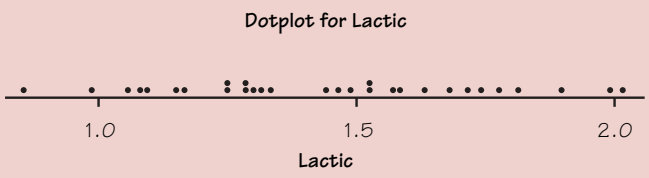
The chief advantages of dotplots and stemplots are that they are easy to construct and they display the actual data values (unless we round). Neither will work well with large data sets. Most statistical software packages will make dotplots and stemplots for you. That will allow you to spend more time making sense of the data.

TECHNOLOGY TOOLBOX *Interpreting computer output*

As cheddar cheese matures, a variety of chemical processes take place. The taste of mature cheese is related to the concentration of several chemicals in the final product. In a study of cheddar cheese from the Latrobe Valley of Victoria, Australia, samples of cheese were analyzed for their chemical composition. The final concentrations of lactic acid in the 30 samples, as a multiple of their initial concentrations, are given below.[6]

A dotplot and a stemplot from the Minitab statistical software package are shown in Figure 1.5. The dots in the dotplot are so spread out that the distribution seems to have no distinct shape. The stemplot does a better job of summarizing the data.

0.86	1.53	1.57	1.81	0.99	1.09	1.29	1.78	1.29	1.58
1.68	1.90	1.06	1.30	1.52	1.74	1.16	1.49	1.63	1.99
1.15	1.33	1.44	2.01	1.31	1.46	1.72	1.25	1.08	1.25




```
Stem-and-leaf of Lactic       N = 30
Leaf Unit = 0.010

 1    8    6
 2    9    9
 5   10    689
 7   11    56
11   12    5599
14   13    013
(3)  14    469
13   15    2378
 9   16    38
 7   17    248
 4   18    1
 3   19    09
 1   20    1
```

FIGURE 1.5 Minitab dotplot and stemplot for cheese data.

TECHNOLOGY TOOLBOX *Interpreting computer output (continued)*

Notice how the data are recorded in the stemplot. The "leaf unit" is 0.01, which tells us that the stems are given in tenths and the leaves are given in hundredths. We can see that the *spread* of the lactic acid concentrations is from 0.86 to 2.01. Where is the *center* of the distribution? Minitab counts the number of observations from the bottom up and from the top down and lists those counts to the left of the stemplot. Since there are 30 observations, the "middle value" would fall between the 15th and 16th data values from either end—at 1.45. The (3) to the far left of this stem is Minitab's way of marking the location of the "middle value." So a typical sample of mature cheese has 1.45 times as much lactic acid as it did initially. The distribution is roughly symmetrical in *shape*. There appear to be no *outliers*.

EXERCISES

1.7 OLYMPIC GOLD Athletes like Cathy Freeman, Rulon Gardner, Ian Thorpe, Marion Jones, and Jenny Thompson captured public attention by winning gold medals in the 2000 Summer Olympic Games in Sydney, Australia. Table 1.2 displays the total number of gold medals won by several countries in the 2000 Summer Olympics.

TABLE 1.2 Gold medals won by selected countries in the 2000 Summer Olympics

Country	Gold medals	Country	Gold medals
Sri Lanka	0	Netherlands	12
Qatar	0	India	0
Vietnam	0	Georgia	0
Great Britain	28	Kyrgyzstan	0
Norway	10	Costa Rica	0
Romania	26	Brazil	0
Switzerland	9	Uzbekistan	1
Armenia	0	Thailand	1
Kuwait	0	Denmark	2
Bahamas	1	Latvia	1
Kenya	2	Czech Republic	2
Trinidad and Tobago	0	Hungary	8
Greece	13	Sweden	4
Mozambique	1	Uruguay	0
Kazakhstan	3	United States	39

Source: BBC Olympics Web site.

Make a dotplot to display these data. Describe the distribution of number of gold medals won.

1.8 ARE YOU DRIVING A GAS GUZZLER? Table 1.3 displays the highway gas mileage for 32 model year 2000 midsize cars.

TABLE 1.3 Highway gas mileage for model year 2000 midsize cars

Model	MPG	Model	MPG
Acura 3.5RL	24	Lexus GS300	24
Audi A6 Quattro	24	Lexus LS400	25
BMW 740I Sport M	21	Lincoln-Mercury LS	25
Buick Regal	29	Lincoln-Mercury Sable	28
Cadillac Catera	24	Mazda 626	28
Cadillac Eldorado	28	Mercedes-Benz E320	30
Chevrolet Lumina	30	Mercedes-Benz E430	24
Chrysler Cirrus	28	Mitsubishi Diamante	25
Dodge Stratus	28	Mitsubishi Galant	28
Honda Accord	29	Nissan Maxima	28
Hyundai Sonata	28	Oldsmobile Intrigue	28
Infiniti I30	28	Saab 9-3	26
Infiniti Q45	23	Saturn LS	32
Jaguar Vanden Plas	24	Toyota Camry	30
Jaguar S/C	21	Volkswagon Passat	29
Jaguar X200	26	Volvo S70	27

(a) Make a dotplot of these data.

(b) Describe the shape, center, and spread of the distribution of gas mileages. Are there any potential outliers?

1.9 MICHIGAN COLLEGE TUITIONS There are 81 colleges and universities in Michigan. Their tuition and fees for the 1999 to 2000 school year run from $1260 at Kalamazoo Valley Community College to $19,258 at Kalamazoo College. Figure 1.6 (next page) shows a stemplot of the tuition charges.

(a) What do the stems and leaves represent in the stemplot? Have the data been rounded?

(b) Describe the shape, center, and spread of the tuition distribution. Are there any outliers?

1.10 DRP TEST SCORES There are many ways to measure the reading ability of children. One frequently used test is the Degree of Reading Power (DRP). In a research study on third-grade students, the DRP was administered to 44 students.[7] Their scores were:

40	26	39	14	42	18	25	43	46	27	19
47	19	26	35	34	15	44	40	38	31	46
52	25	35	35	33	29	34	41	49	28	52
47	35	48	22	33	41	51	27	14	54	45

Display these data graphically. Write a paragraph describing the distribution of DRP scores.

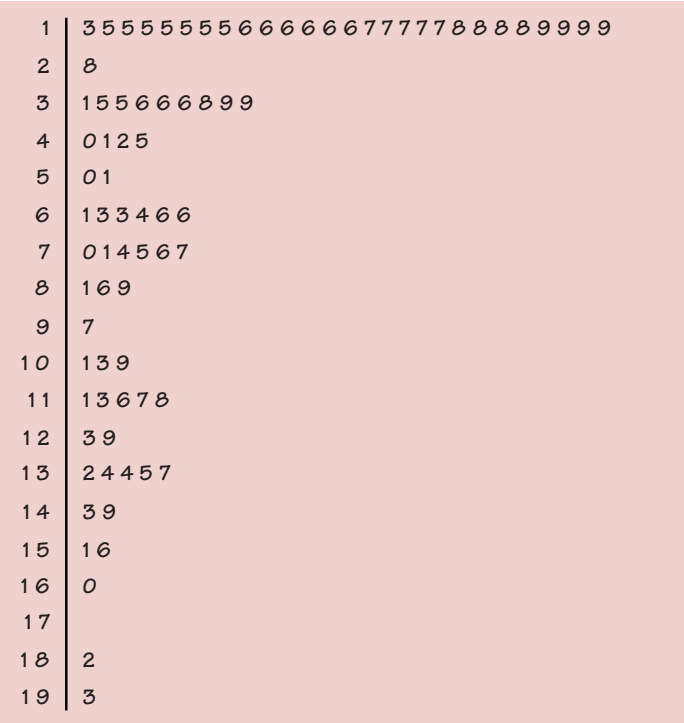


FIGURE 1.6 Stemplot of the Michigan tuition and fee data, for Exercise 1.9.

1.11 SHOPPING SPREE! A marketing consultant observed 50 consecutive shoppers at a supermarket. One variable of interest was how much each shopper spent in the store. Here are the data (in dollars), arranged in increasing order:

3.11	8.88	9.26	10.81	12.69	13.78	15.23	15.62	17.00	17.39
18.36	18.43	19.27	19.50	19.54	20.16	20.59	22.22	23.04	24.47
24.58	25.13	26.24	26.26	27.65	28.06	28.08	28.38	32.03	34.98
36.37	38.64	39.16	41.02	42.97	44.08	44.67	45.40	46.69	48.65
50.39	52.75	54.80	59.07	61.22	70.32	82.70	85.76	86.37	93.34

(a) Round each amount to the nearest dollar. Then make a stemplot using tens of dollars as the stem and dollars as the leaves.

(b) Make another stemplot of the data by splitting stems. Which of the plots shows the shape of the distribution better?

(c) Describe the shape, center, and spread of the distribution. Write a few sentences describing the amount of money spent by shoppers at this supermarket.

Displaying quantitative variables: histograms

Quantitative variables often take many values. A graph of the distribution is clearer if nearby values are grouped together. The most common graph of the distribution of one quantitative variable is a **histogram.**

EXAMPLE 1.6 PRESIDENTIAL AGES AT INAUGURATION

How old are presidents at their inaugurations? Was Bill Clinton, at age 46, unusually young? Table 1.4 gives the data, the ages of all U.S presidents when they took office.

TABLE 1.4 Ages of the Presidents at inauguration

President	Age	President	Age	President	Age
Washington	57	Lincoln	52	Hoover	54
J. Adams	61	A. Johnson	56	F. D. Roosevelt	51
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	61
Monroe	58	Garfield	49	Kennedy	43
J. Q. Adams	57	Arthur	51	L. B. Johnson	55
Jackson	61	Cleveland	47	Nixon	56
Van Buren	54	B. Harrison	55	Ford	61
W. H. Harrison	68	Cleveland	55	Carter	52
Tyler	51	McKinley	54	Reagan	69
Polk	49	T. Roosevelt	42	G. Bush	64
Taylor	64	Taft	51	Clinton	46
Fillmore	50	Wilson	56	G. W. Bush	54
Pierce	48	Harding	55		
Buchanan	65	Coolidge	51		

How to make a histogram:

Step 1: Divide the range of the data into classes of equal width. Count the number of observations in each class. The data in Table 1.4 range from 42 to 69, so we choose as our classes

$$40 \le \text{president's age at inauguration} < 45$$
$$45 \le \text{president's age at inauguration} < 50$$
$$\vdots$$
$$65 \le \text{president's age at inauguration} < 70$$

Be sure to specify the classes precisely so that each observation falls into exactly one class. Martin Van Buren, who was age 54 at the time of his inauguration, would fall into the third class interval. Grover Cleveland, who was age 55, would be placed in the fourth class interval.

Here are the counts:

Class	Count
40–44	2
45–49	6
50–54	13
55–59	12
60–64	7
65–69	3

Step 2: Label and scale your axes and title your graph. Label the horizontal axis "Age at inauguration" and the vertical axis "Number of presidents." For the classes we chose, we should scale the horizontal axis from 40 to 70, with tick marks 5 units apart. The vertical axis contains the scale of counts and should range from 0 to at least 13.

Step 3: Draw a bar that represents the count in each class. The base of a bar should cover its class, and the bar height is the class count. Leave no horizontal space between the bars (unless a class is empty, so that its bar has height 0). Figure 1.7 shows the completed histogram.

Graphing note: It is common to add a "break-in-scale" symbol (//) on an axis that does not start at 0, like the horizontal axis in this example.

Interpretation:

Center: It appears that the typical age of a new president is about 55 years, because 55 is near the center of the histogram.

Spread: As the histogram in Figure 1.7 shows, there is a good deal of variation in the ages at which presidents take office. Teddy Roosevelt was the youngest, at age 42, and Ronald Reagan, at age 69, was the oldest.

Shape: The distribution is roughly symmetric and has a single peak (unimodal).

Outliers: There appear to be no outliers.

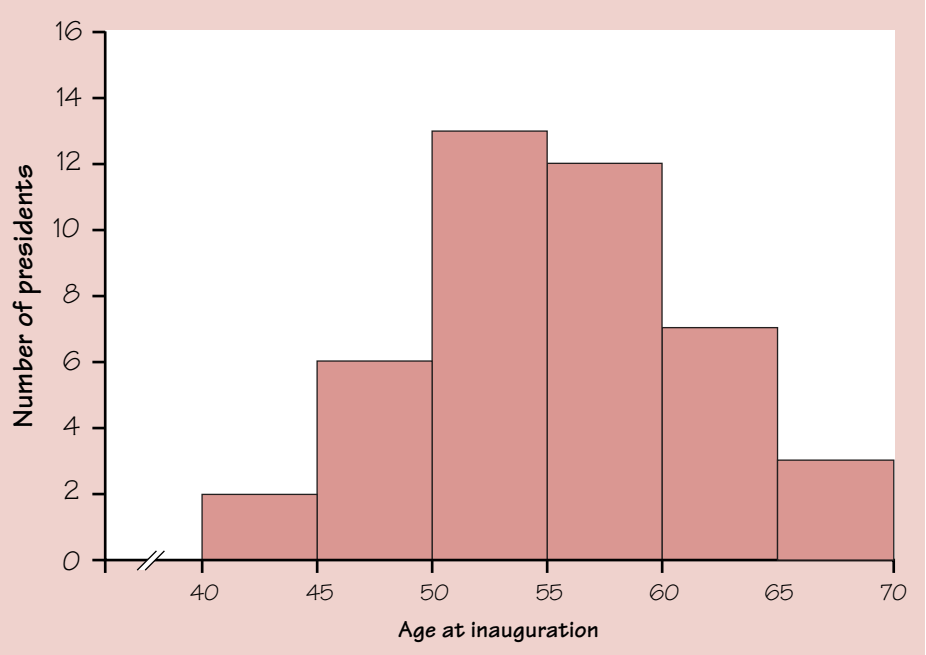


FIGURE 1.7 The distribution of the ages of presidents at their inaugurations, from Table 1.4.

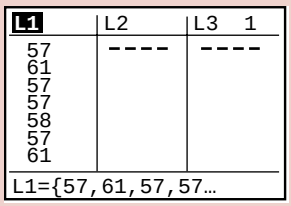
You can also use computer software or a calculator to construct histograms.

TECHNOLOGY TOOLBOX *Making calculator histograms*

1. Enter the presidential age data from Example 1.6 in your statistics list editor.

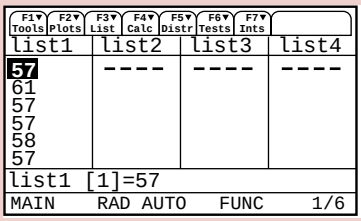
TI-83

- Press STAT and choose 1:Edit....
- Type the values into list L_1.



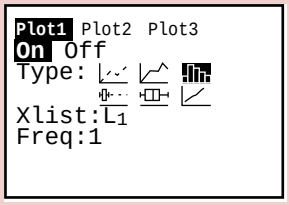
TI-89

- Press APPS, choose 1:FlashApps, then select Stats/List Editor and press ENTER.
- Type the values into list1.

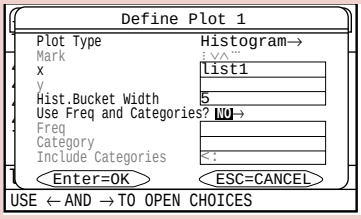


2. Set up a histogram in the statistics plots menu.

- Press 2nd Y= (STAT PLOT).
- Press ENTER to go into Plot1.
- Adjust your settings as shown.

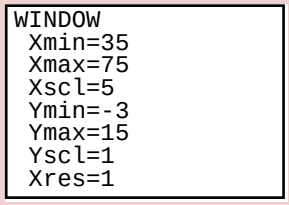


- Press F2 and choose 1:Plot Setup....
- With Plot 1 highlighted, press F1 to define.
- Change Hist. Bucket Width to 5, as shown.

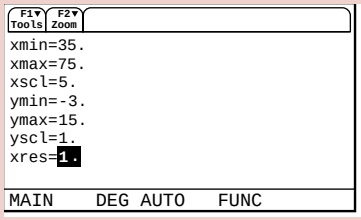


3. Set the window to match the class intervals chosen in Example 1.6.

- Press WINDOW.
- Enter the values shown.



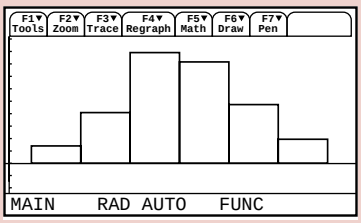
- Press ♦ F2 (WINDOW).
- Enter the values shown.



4. Graph the histogram. Compare with Figure 1.7.

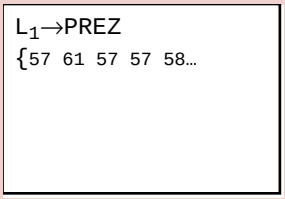
- Press GRAPH.
- Press ♦ F3 (GRAPH).

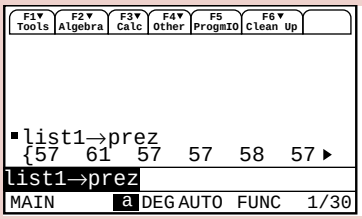
TECHNOLOGY TOOLBOX *Making calculator histograms (continued)*



5. Save the data in a named list for later use.

- From the home screen, type the command `L1→PREZ` (`list1→prez` on the TI-89) and press ENTER. The data are now stored in a list called PREZ.





Histogram tips:

- 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.
- Five classes is a good minimum.
- Our eyes respond to the *area* of the bars in a histogram, so be sure to choose classes that are all the same width. Then area is determined by height and all classes are fairly represented.
- If you use a computer or graphing calculator, beware of letting the device choose the classes.

EXERCISES

1.12 WHERE DO OLDER FOLKS LIVE? Table 1.5 gives the percentage of residents aged 65 or older in each of the 50 states.

TABLE 1.5 Percent of the population in each state aged 65 or older

State	Percent	State	Percent	State	Percent
Alabama	13.1	Louisiana	11.5	Ohio	13.4
Alaska	5.5	Maine	14.1	Oklahoma	13.4
Arizona	13.2	Maryland	11.5	Oregon	13.2
Arkansas	14.3	Massachusetts	14.0	Pennsylvania	15.9
California	11.1	Michigan	12.5	Rhode Island	15.6
Colorado	10.1	Minnesota	12.3	South Carolina	12.2
Connecticut	14.3	Mississippi	12.2	South Dakota	14.3
Delaware	13.0	Missouri	13.7	Tennessee	12.5
Florida	18.3	Montana	13.3	Texas	10.1
Georgia	9.9	Nebraska	13.8	Utah	8.8
Hawaii	13.3	Nevada	11.5	Vermont	12.3
Idaho	11.3	New Hampshire	12.0	Virginia	11.3
Illinois	12.4	New Jersey	13.6	Washington	11.5
Indiana	12.5	New Mexico	11.4	West Virginia	15.2
Iowa	15.1	New York	13.3	Wisconsin	13.2
Kansas	13.5	North Carolina	12.5	Wyoming	11.5
Kentucky	12.5	North Dakota	14.4		

Source: U.S. Census Bureau, 1998.

(a) Construct a histogram to display these data. Record your class intervals and counts.

(b) Describe the distribution of people aged 65 and over in the states.

(c) Enter the data into your calculator's statistics list editor. Make a histogram using a window that matches your histogram from part (a). Copy the calculator histogram and mark the scales on your paper.

(d) Use the calculator's zoom feature to generate a histogram. Copy this histogram onto your paper and mark the scales.

(e) Store the data into the named list ELDER for later use.

1.13 DRP SCORES REVISITED Refer to Exercise 1.10 (page 17). Make a histogram of the DRP test scores for the sample of 44 children. Be sure to show your frequency table. Which do you prefer: the stemplot from Exercise 1.10 or the histogram that you just constructed? Why?

1.14 CEO SALARIES In 1993, *Forbes* magazine reported the age and salary of the chief executive officer (CEO) of each of the top 59 small businesses.[8] Here are the salary data, rounded to the nearest thousand dollars:

145	621	262	208	362	424	339	736	291	58	498	643	390	332
750	368	659	234	396	300	343	536	543	217	298	1103	406	254
862	204	206	250	21	298	350	800	726	370	536	291	808	543
149	350	242	198	213	296	317	482	155	802	200	282	573	388
250	396	572											

Construct a histogram for these data. Describe the shape, center, and spread of the distribution of CEO salaries. Are there any apparent outliers?

1.15 CHEST OUT, SOLDIER! In 1846, a published paper provided chest measurements (in inches) of 5738 Scottish militiamen. Table 1.6 displays the data in summary form.

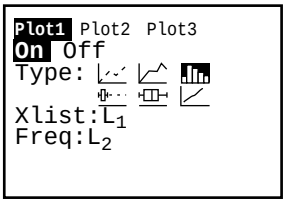
TABLE 1.6 Chest measurements (inches) of 5738 Scottish militiamen

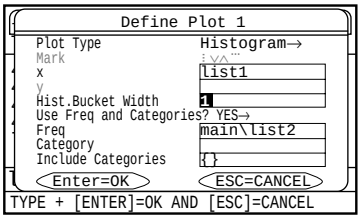
Chest size	Count	Chest size	Count
33	3	41	934
34	18	42	658
35	81	43	370
36	185	44	92
37	420	45	50
38	749	46	21
39	1073	47	4
40	1079	48	1

Source: Data and Story Library (DASL), http://lib.stat.cmu.edu/DASL/.

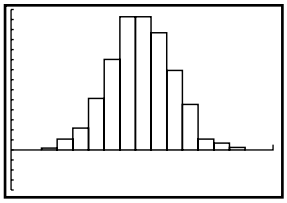
(a) You can use your graphing calculator to make a histogram of data presented in summary form like the chest measurements of Scottish militiamen.

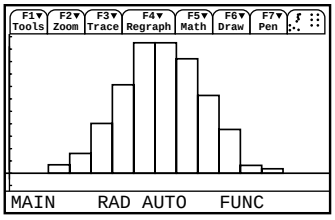
- Type the chest measurements into L_1/list1 and the corresponding counts into L_2/list2.
- Set up a statistics plot to make a histogram with *x*-values from L_1/list1 and *y*-values (bar heights) from L_2/list2.





- Adjust your viewing window settings as follows: xmin = 32, xmax = 49, xscl = 1, ymin = –300, ymax = 1100, yscl = 100. From now on, we will abbreviate in this form: $X[32,49]_1$ by $Y[-300,1100]_{100}$. Try using the calculator's built-in ZoomStat/ZoomData command. What happens?
- Graph.





(b) Describe the shape, center, and spread of the chest measurements distribution. Why might this information be useful?

More about shape

When you describe a distribution, concentrate on the main features. Look for major peaks, not for minor ups and downs in the bars of the histogram. Look for clear outliers, not just for the smallest and largest observations. Look for rough *symmetry* or clear *skewness*.

SYMMETRIC AND SKEWED DISTRIBUTIONS

A distribution is **symmetric** if the right and left sides of the histogram are approximately mirror images of each other.

A distribution is **skewed to the right** if the right side of the histogram (containing the half of the observations with larger values) extends much farther out than the left side. It is **skewed to the left** if the left side of the histogram extends much farther out than the right side.

In mathematics, symmetry means that the two sides of a figure like a histogram are exact mirror images of each other. Data are almost never exactly symmetric, so we are willing to call histograms like that in Exercise 1.15 approximately symmetric as an overall description. Here are more examples.

EXAMPLE 1.7 LIGHTNING FLASHES AND SHAKESPEARE

Figure 1.8 comes from a study of lightning storms in Colorado. It shows the distribution of the hour of the day during which the first lightning flash for that day occurred. The distribution has a single peak at noon and falls off on either side of this peak. The two sides of the histogram are roughly the same shape, so we call the distribution symmetric.

Figure 1.9 shows the distribution of lengths of words used in Shakespeare's plays.[9] This distribution also has a single peak but is skewed to the right. That is, there are many short words (3 and 4 letters) and few very long words (10, 11, or 12 letters), so that the right tail of the histogram extends out much farther than the left tail.

Notice that the vertical scale in Figure 1.9 is not the *count* of words but the *percent* of all of Shakespeare's words that have each length. A histogram of percents rather than counts is convenient when the counts are very large or when we want to compare several distributions. Different kinds of writing have different distributions of word lengths, but all are right-skewed because short words are common and very long words are rare.

The overall shape of a distribution is important information about a variable. Some types of data regularly produce distributions that are symmetric or skewed. For example, the sizes of living things of the same species (like lengths of cockroaches) tend to be symmetric. Data on incomes (whether of individuals, companies, or nations) are usually strongly skewed to the right. There are many moderate incomes, some large incomes, and a few very large incomes. Do remember that

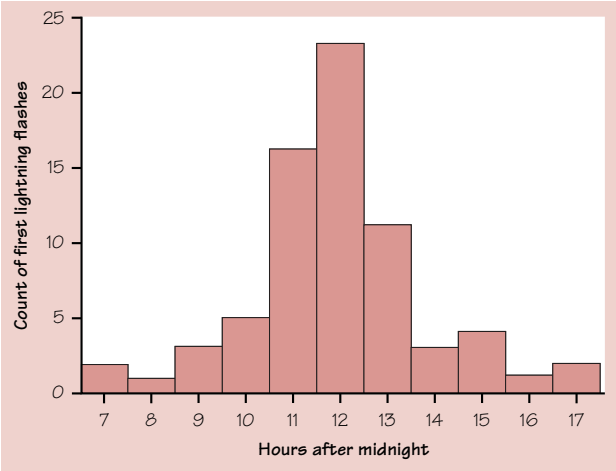


FIGURE 1.8 The distribution of the time of the first lightning flash each day at a site in Colorado, for Example 1.7.

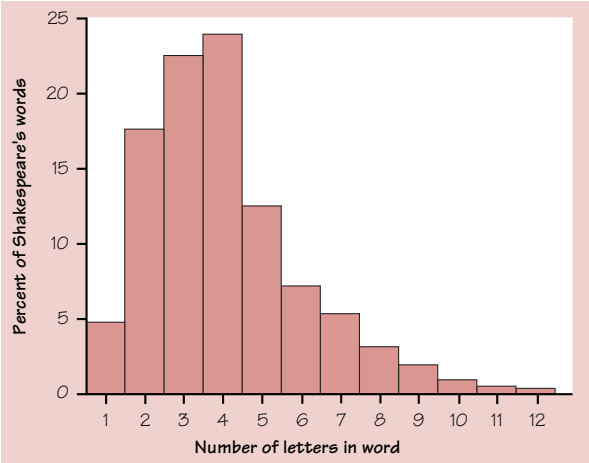


FIGURE 1.9 The distribution of lengths of words used in Shakespeare's plays, for Example 1.7.

many distributions have shapes that are neither symmetric nor skewed. Some data show other patterns. Scores on an exam, for example, may have a cluster near the top of the scale if many students did well. Or they may show two distinct peaks if a tough problem divided the class into those who did and didn't solve it. Use your eyes and describe what you see.

EXERCISES

1.16 STOCK RETURNS The total return on a stock is the change in its market price plus any dividend payments made. Total return is usually expressed as a percent of the beginning price. Figure 1.10 is a histogram of the distribution of total returns for all 1528 stocks listed on the New York Stock Exchange in one year.[10] Like

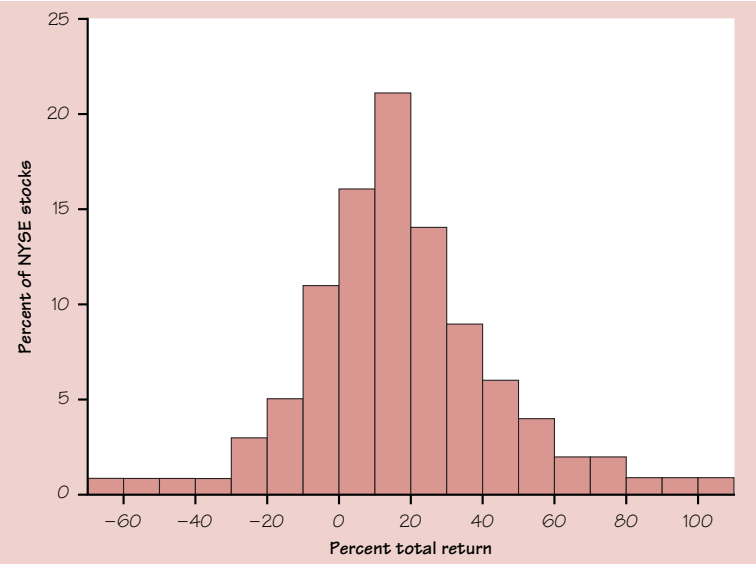


FIGURE 1.10 The distribution of percent total return for all New York Stock Exchange common stocks in one year.

Figure 1.9, it is a histogram of the percents in each class rather than a histogram of counts.

(a) Describe the overall shape of the distribution of total returns.

(b) What is the approximate center of this distribution? (For now, take the center to be the value with roughly half the stocks having lower returns and half having higher returns.)

(c) Approximately what were the smallest and largest total returns? (This describes the spread of the distribution.)

(d) A return less than zero means that an owner of the stock lost money. About what percent of all stocks lost money?

1.17 FREEZING IN GREENWICH, ENGLAND Figure 1.11 is a histogram of the number of days in the month of April on which the temperature fell below freezing at Greenwich, England.[11] The data cover a period of 65 years.

(a) Describe the shape, center, and spread of this distribution. Are there any outliers?

(b) In what percent of these 65 years did the temperature never fall below freezing in April?

1.18 How would you describe the center and spread of the distribution of first lightning flash times in Figure 1.8? Of the distribution of Shakespeare's word lengths in Figure 1.9?

Relative frequency, cumulative frequency, percentiles, and ogives

Sometimes we are interested in describing the relative position of an individual within a distribution. You may have received a standardized test score report that said you were in the 80th percentile. What does this mean? Put simply,

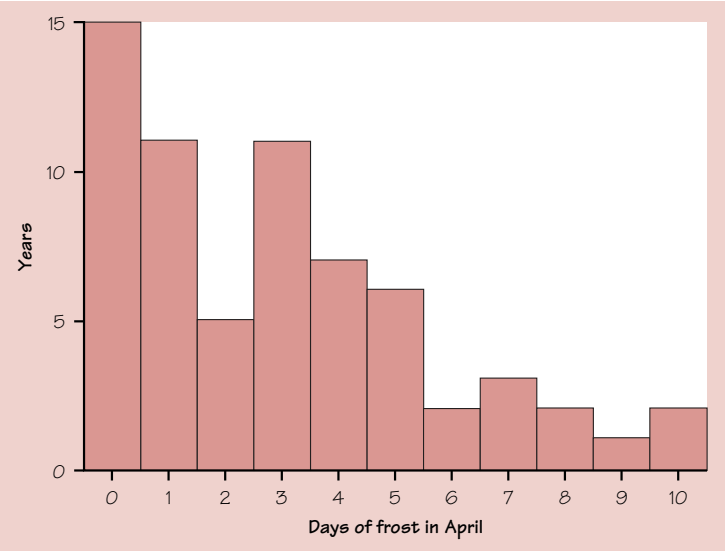


FIGURE 1.11 The distribution of the number of frost days during April at Greenwich, England, over a 65-year period, for Exercise 1.17.

80% of the people who took the test earned scores that were less than or equal to your score. The other 20% of students taking the test earned higher scores than you did.

PERCENTILE

The ***p*th percentile** of a distribution is the value such that p percent of the observations fall at or below it.

A histogram does a good job of displaying the distribution of values of a variable. But it tells us little about the relative standing of an individual observation. If we want this type of information, we should construct a **relative cumulative frequency graph,** often called an **ogive** (pronounced O-JIVE).

EXAMPLE 1.8 WAS BILL CLINTON A YOUNG PRESIDENT?

In Example 1.6, we made a histogram of the ages of U.S. presidents when they were inaugurated. Now we will examine where some specific presidents fall within the age distribution.

How to construct an ogive (relative cumulative frequency graph):

Step 1: Decide on class intervals and make a frequency table, just as in making a histogram. Add three columns to your frequency table: relative frequency, cumulative frequency, and relative cumulative frequency.

• To get the values in the *relative frequency* column, divide the count in each class interval by 43, the total number of presidents. Multiply by 100 to convert to a percentage.

• To fill in the *cumulative frequency* column, add the counts in the frequency column that fall in or below the current class interval.

• For the *relative cumulative frequency* column, divide the entries in the cumulative frequency column by 43, the total number of individuals.

Here is the frequency table from Example 1.6 with the relative frequency, cumulative frequency, and relative cumulative frequency columns added.

Class	Frequency	Relative frequency	Cumulative frequency	Relative cumulative frequency
40–44	2	$\frac{2}{43} = 0.047$, or 4.7%	2	$\frac{2}{43} = 0.047$, or 4.7%
45–49	6	$\frac{6}{43} = 0.140$, or 14.0%	8	$\frac{8}{43} = 0.186$, or 18.6%
50–54	13	$\frac{13}{43} = 0.302$, or 30.2%	21	$\frac{21}{43} = 0.488$, or 48.8%
55–59	12	$\frac{12}{43} = 0.279$, or 27.9%	33	$\frac{33}{43} = 0.767$, or 76.7%
60–64	7	$\frac{7}{43} = 0.163$, or 16.3%	40	$\frac{40}{43} = 0.930$, or 93.0%
65–69	3	$\frac{3}{43} = 0.070$, or 7.0%	43	$\frac{43}{43} = 1.000$, or 100%
TOTAL	43			

Step 2: Label and scale your axes and title your graph. Label the horizontal axis "Age at inauguration" and the vertical axis "Relative cumulative frequency." Scale the horizontal axis according to your choice of class intervals and the vertical axis from 0% to 100%.

Step 3: Plot a point corresponding to the relative cumulative frequency in each class interval at the *left endpoint* of the *next* class interval. For example, for the 40–44 interval, plot a point at a height of 4.7% above the age value of 45. This means that 4.7% of presidents were inaugurated before they were 45 years old. Begin your ogive with a point at a height of 0% at the left endpoint of the lowest class interval. Connect consecutive points with a line segment to form the ogive. The last point you plot should be at a height of 100%. Figure 1.12 shows the completed ogive.

How to locate an individual within the distribution:

What about Bill Clinton? He was age 46 when he took office. To find his relative standing, draw a vertical line up from his age (46) on the horizontal axis until it meets the ogive. Then draw a horizontal line from this point of intersection to the vertical axis. Based on Figure 1.13(a), we would estimate that Bill Clinton's age places him at the 10% *relative cumulative frequency* mark. That tells us that about 10% of all U.S. presidents were the same age as or younger than Bill Clinton when they were inaugurated. Put another way, President Clinton was younger than about 90% of all U.S. presidents based on his inauguration age. His age places him at the *10th percentile* of the distribution.

How to locate a value corresponding to a percentile:

• What inauguration age corresponds to the 60th percentile? To answer this question, draw a horizontal line across from the vertical axis at a height of 60% until it meets the ogive. From the point of intersection, draw a vertical line down to the horizontal axis.

In Figure 1.13(b), the value on the horizontal axis is about 57. So about 60% of all presidents were 57 years old or younger when they took office.

• Find the center of the distribution. Since we use the value that has half of the observations above it and half below it as our estimate of center, we simply need to find the 50th percentile of the distribution. Estimating as for the previous question, confirm that 55 is the center.

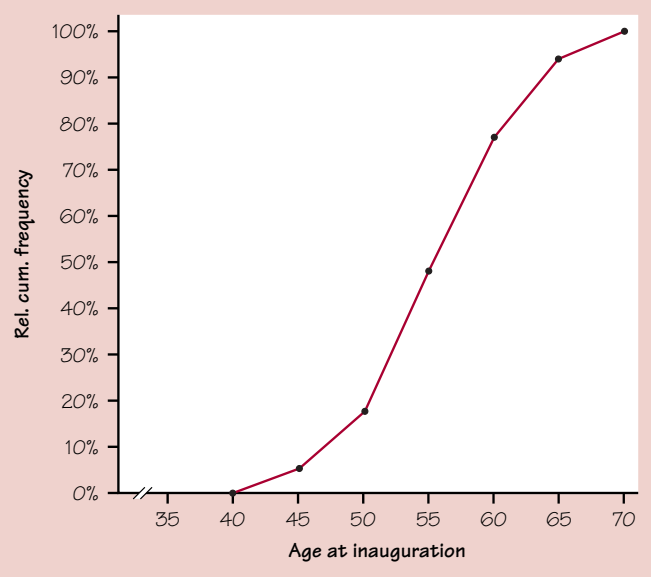
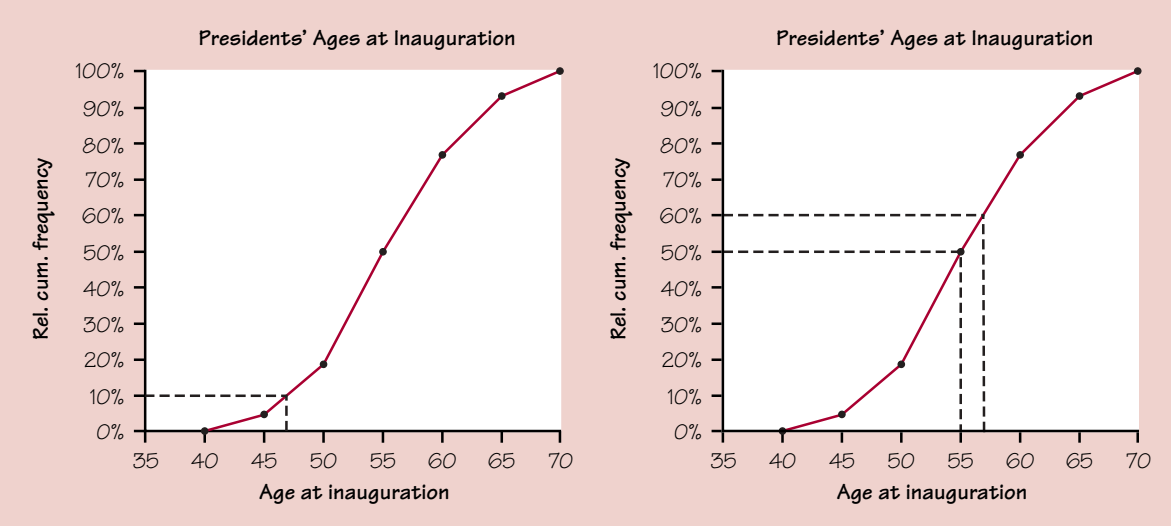


FIGURE 1.12 Relative cumulative frequency plot (ogive) for the ages of U.S. presidents at inauguration.



(a) (b)

FIGURE 1.13 Ogives of presidents' ages at inauguration are used to (a) locate Bill Clinton within the distribution and (b) determine the 60th percentile and center of the distribution.

EXERCISES

1.19 OLDER FOLKS, II In Exercise 1.12 (page 22), you constructed a histogram of the percentage of people aged 65 or older in each state.

(a) Construct a relative cumulative frequency graph (ogive) for these data.

(b) Use your ogive from part (a) to answer the following questions:

- In what percentage of states was the percentage of "65 and older" less than 15%?
- What is the 40th percentile of this distribution, and what does it tell us?
- What percentile is associated with your state?

1.20 SHOPPING SPREE, II Figure 1.14 is an ogive of the amount spent by grocery shoppers in Exercise 1.11 (page 18).

(a) Estimate the center of this distribution. Explain your method.

(b) At what percentile would the shopper who spent $17.00 fall?

(c) Draw the histogram that corresponds to the ogive.

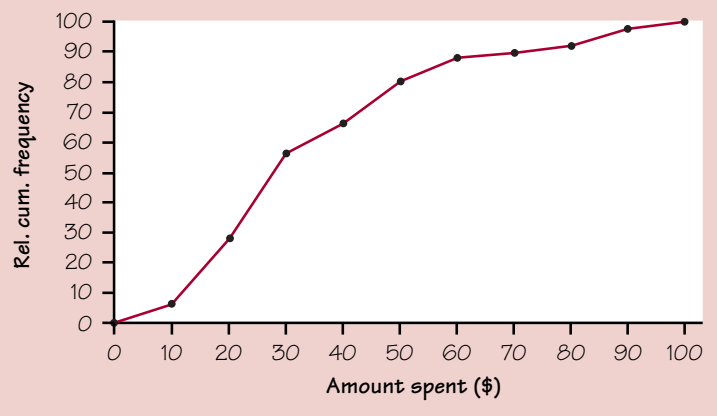


FIGURE 1.14 Amount spent by grocery shoppers in Exercise 1.11.

Time plots

Many variables are measured at intervals over time. We might, for example, measure the height of a growing child or the price of a stock at the end of each month. In these examples, our main interest is change over time. To display change over time, make a time plot.

TIME PLOT

A **time plot** of a variable plots each observation against the time at which it was measured. Always mark the time scale on the horizontal axis and the variable of interest on the vertical axis. If there are not too many points, connecting the points by lines helps show the pattern of changes over time.

trend

seasonal variation

When you examine a time plot, look once again for an overall pattern and for strong deviations from the pattern. One common overall pattern is a **trend,** a long-term upward or downward movement over time. A pattern that repeats itself at regular time intervals is known as **seasonal variation.** The next example illustrates both these patterns.

EXAMPLE 1.9 ORANGE PRICES MAKE ME SOUR!

Figure 1.15 is a time plot of the average price of fresh oranges over the period from January 1990 to January 2000. This information is collected each month as part of the government's reporting of retail prices. The vertical scale on the graph is the orange price index. This represents the price as a percentage of the average price of oranges in the years 1982 to 1984. The first value is 150 for January 1990, so at that time oranges cost about 150% of their 1982 to 1984 average price.

Figure 1.15 shows a clear *trend* of increasing price. In addition to this trend, we can see a strong *seasonal variation*, a regular rise and fall that occurs each year. Orange prices are usually highest in August or September, when the supply is lowest. Prices then fall in anticipation of the harvest and are lowest in January or February, when the harvest is complete and oranges are plentiful. The unusually large jump in orange prices in 1991 resulted from a freeze in Florida. Can you discover what happened in 1999?

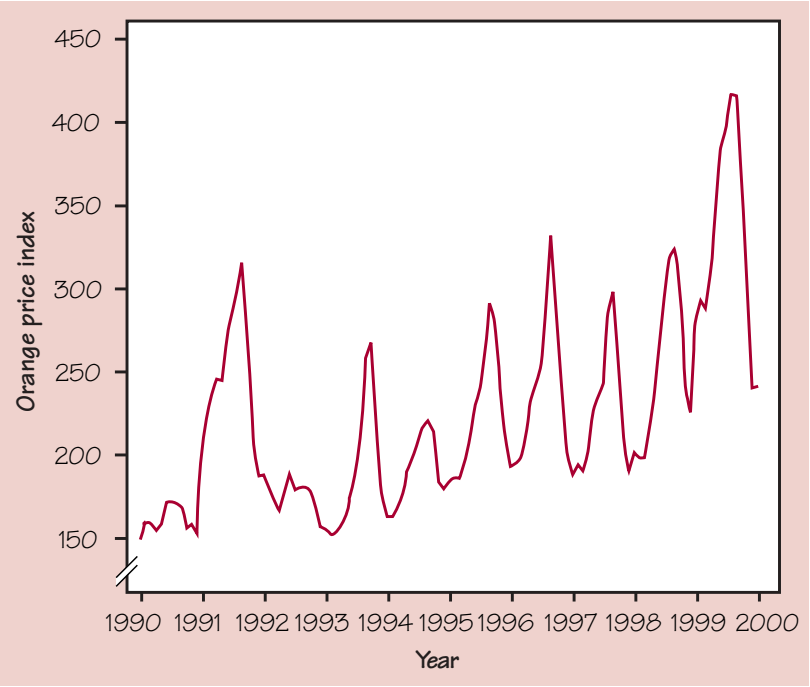


FIGURE 1.15 The price of fresh oranges, January 1990 to January 2000.

EXERCISES

1.21 CANCER DEATHS Here are data on the rate of deaths from cancer (deaths per 100,000 people) in the United States over the 50-year period from 1945 to 1995:

Year:	1945	1950	1955	1960	1965	1970	1975	1980	1985	1990	1995
Deaths:	134.0	139.8	146.5	149.2	153.5	162.8	169.7	183.9	193.3	203.2	204.7

(a) Construct a time plot for these data. Describe what you see in a few sentences.

(b) Do these data suggest that we have made no progress in treating cancer? Explain.

1.22 CIVIL UNREST The years around 1970 brought unrest to many U.S. cities. Here are data on the number of civil disturbances in each three month period during the years 1968 to 1972:

Period		Count	Period		Count
1968	Jan.–Mar.	6	1970	July–Sept.	20
	Apr.–June	46		Oct.–Dec.	6
	July–Sept.	25	1971	Jan.–Mar.	12
	Oct.–Dec.	3		Apr.–June	21
1969	Jan.–Mar.	5		July–Sept.	5
	Apr.–June	27		Oct.–Dec.	1
	July–Sept.	19	1972	Jan.–Mar.	3
	Oct.–Dec.	6		Apr.–June	8
1970	Jan.–Mar.	26		July–Sept.	5
	Apr.–June	24		Oct.–Dec.	5

(a) Make a time plot of these counts. Connect the points in your plot by straight-line segments to make the pattern clearer.

(b) Describe the trend and the seasonal variation in this time series. Can you suggest an explanation for the seasonal variation in civil disorders?

SUMMARY

A data set contains information on a number of **individuals.** Individuals may be people, animals, or things. 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.

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

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

The **distribution** of a variable describes what values the variable takes and how often it takes these values.

To describe a distribution, begin with a graph. Use **bar graphs** and **pie charts** to display categorical variables. **Dotplots, stemplots,** and **histograms** graph the distributions of quantitative variables. An **ogive** can help you determine relative standing within a quantitative distribution.

When examining any graph, look for an **overall pattern** and for notable **deviations** from the pattern.

The **center, spread,** and **shape** describe the overall pattern of a distribution. Some distributions have simple shapes, such as **symmetric** and **skewed.** 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, seasonal variations,** or other changes over time.

SECTION 1.1 EXERCISES

1.23 GENDER EFFECTS IN VOTING Political party preference in the United States depends in part on the age, income, and gender of the voter. A political scientist selects a large sample of registered voters. For each voter, she records gender, age, household income, and whether they voted for the Democratic or for the Republican candidate in the last congressional election. Which of these variables are categorical and which are quantitative?

1.24 What type of graph or graphs would you plan to make in a study of each of the following issues?

(a) What makes of cars do students drive? How old are their cars?

(b) How many hours per week do students study? How does the number of study hours change during a semester?

(c) Which radio stations are most popular with students?

1.25 MURDER WEAPONS The 1999 *Statistical Abstract of the United States* reports FBI data on murders for 1997. In that year, 53.3% of all murders were committed with handguns, 14.5% with other firearms, 13.0% with knives, 6.3% with a part of the body (usually the hands or feet), and 4.6% with blunt objects. Make a graph to display these data. Do you need an "other methods" category?

1.26 WHAT'S A DOLLAR WORTH THESE DAYS? The buying power of a dollar changes over time. The Bureau of Labor Statistics measures the cost of a "market basket" of goods and services to compile its Consumer Price Index (CPI). If the CPI is 120, goods and services that cost $100 in the base period now cost $120. Here are the yearly average values of the CPI for the years between 1970 and 1999. The base period is the years 1982 to 1984.

Year	CPI	Year	CPI	Year	CPI	Year	CPI
1970	38.8	1978	65.2	1986	109.6	1994	148.2
1972	41.8	1980	82.4	1988	118.3	1996	156.9
1974	49.3	1982	96.5	1990	130.7	1998	163.0
1976	56.9	1984	103.9	1992	140.3	1999	166.6

(a) Construct a graph that shows how the CPI has changed over time.

(b) Check your graph by doing the plot on your calculator.

- Enter the years (the last two digits will suffice) into L_1/list1 and enter the CPI into L_2/list2.
- Then set up a statistics plot, choosing the plot type "xyline" (the second type on the TI-83). Use L_1/list1 as X and L_2/list2 as Y. In this graph, the data points are plotted and connected in order of appearance in L_1/list1 and L_2/list2.
- Use the zoom command to see the graph.

(c) What was the overall trend in prices during this period? Were there any years in which this trend was reversed?

(d) In what period during these decades were prices rising fastest? In what period were they rising slowest?

1.27 THE STATISTICS OF WRITING STYLE Numerical data can distinguish different types of writing, and sometimes even individual authors. Here are data on the percent of words of 1 to 15 letters used in articles in Popular Science magazine:[12]

Length:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Percent:	3.6	14.8	18.7	16.0	12.5	8.2	8.1	5.9	4.4	3.6	2.1	0.9	0.6	0.4	0.2

(a) Make a histogram of this distribution. Describe its shape, center, and spread.

(b) How does the distribution of lengths of words used in *Popular Science* compare with the similar distribution in Figure 1.9 (page 26) for Shakespeare's plays? Look in particular at short words (2, 3, and 4 letters) and very long words (more than 10 letters).

1.28 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:[13]

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 in a stemplot. Discuss the shape, center, and spread of the distribution. Are there any outliers? What is your estimate of the density of the earth based on these measurements?

1.29 DRIVE TIME Professor Moore, who lives a few miles outside a college town, records the time he takes to drive to the college each morning. Here are the times (in minutes) for 42 consecutive weekdays, with the dates in order along the rows:

8.25	7.83	8.30	8.42	8.50	8.67	8.17	9.00	9.00	8.17	7.92
9.00	8.50	9.00	7.75	7.92	8.00	8.08	8.42	8.75	8.08	9.75
8.33	7.83	7.92	8.58	7.83	8.42	7.75	7.42	6.75	7.42	8.50
8.67	10.17	8.75	8.58	8.67	9.17	9.08	8.83	8.67		

(a) Make a histogram of these drive times. Is the distribution roughly symmetric, clearly skewed, or neither? Are there any clear outliers?

(b) Construct an ogive for Professor Moore's drive times.

(c) Use your ogive from **(b)** to estimate the center and 90th percentile for the distribution.

(d) Use your ogive to estimate the percentile corresponding to a drive time of 8.00 minutes.

1.30 THE SPEED OF LIGHT Light travels fast, but it is not transmitted instantaneously. Light takes over a second to reach us from the moon and over 10 billion years to reach us from the most distant objects observed so far in the expanding universe. Because radio and radar also travel at the speed of light, an accurate value for that speed is important in communicating with astronauts and orbiting satellites. An accurate value for the speed of light is also important to computer designers because electrical signals travel at light speed. The first reasonably accurate measurements of the speed of light were made over 100 years ago by A. A. Michelson and Simon Newcomb. Table 1.7 contains 66 measurements made by Newcomb between July and September 1882.

Newcomb measured the time in seconds that a light signal took to pass from his laboratory on the Potomac River to a mirror at the base of the Washington Monument and back, a total distance of about 7400 meters. Just as you can compute the speed of a car from the time required to drive a mile, Newcomb could compute the speed of light from the passage time. Newcomb's first measurement of the passage time of light was 0.000024828 second, or 24,828 nanoseconds. (There are 10^9 nanoseconds in a second.) The entries in Table 1.7 record only the deviation from 24,800 nanoseconds.

TABLE 1.7 Newcomb's measurements of the passage time of light

28	26	33	24	34	–44	27	16	40	–2	29	22	24	21
25	30	23	29	31	19	24	20	36	32	36	28	25	21
28	29	37	25	28	26	30	32	36	26	30	22	36	23
27	27	28	27	31	27	26	33	26	32	32	24	39	28
24	25	32	25	29	27	28	29	16	23				

Source: S. M. Stigler, "Do robust estimators work with real data?" *Annals of Statistics*, 5 (1977), pp. 1055–1078.

(a) Construct an appropriate graphical display for these data. Justify your choice of graph.

(b) Describe the distribution of Newcomb's speed of light measurements.

(c) Make a time plot of Newcomb's values. They are listed in order from left to right, starting with the top row.

(d) What does the time plot tell you that the display you made in part **(a)** does not?

Lesson: Sometimes you need to make more than one graphical display to uncover all of the important features of a distribution.

1.2 DESCRIBING DISTRIBUTIONS WITH NUMBERS

Who is baseball's greatest home run hitter? In the summer of 1998, Mark McGwire and Sammy Sosa captured the public's imagination with their pursuit of baseball's single-season home run record (held by Roger Maris). McGwire eventually set a new standard with 70 home runs. Barry Bonds broke Mark McGwire's record when he hit 73 home runs in the 2001 season. How does this accomplishment fit Bonds's career? Here are Bonds's home run counts for the years 1986 (his rookie year) to 2001 (the year he broke McGwire's record):

1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
16	25	24	19	33	25	34	46	37	33	42	40	37	34	49	73

The stemplot in Figure 1.16 shows us the *shape*, *center*, and *spread* of these data. The distribution is roughly symmetric with a single peak and a possible high outlier. The center is about 34 home runs, and the spread runs from 16 to the record 73. Shape, center, and spread provide a good description of the overall pattern of any distribution for a quantitative variable. Now we will learn specific ways to use numbers to measure the center and spread of a distribution.

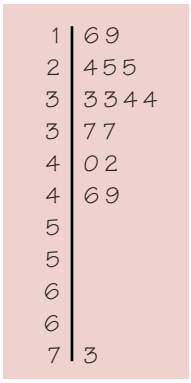


FIGURE 1.16 Number of home runs hit by Barry Bonds in each of his 16 major league seasons.

Measuring center: the mean

A description of a distribution almost always includes a measure of its center or average. The most common measure of center is the ordinary arithmetic average, or *mean*.

THE MEAN $\bar{x}$

To find the **mean** 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 subscripts on the observations x_i are just a way of keeping the n observations distinct. They do not necessarily indicate order or any other special facts about the data. The bar over the x indicates the mean of all the x-values. Pronounce the mean $\bar{x}$ as "x-bar." This notation is very common. When writers who are discussing data use $\bar{x}$ or $\bar{y}$, they are talking about a mean.

EXAMPLE 1.10 BARRY BONDS VERSUS HANK AARON

The mean number of home runs Barry Bonds hit in his first 16 major league seasons is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{16 + 25 + \cdots + 73}{16} = \frac{567}{16} = 35.4375$$

We might compare Bonds to Hank Aaron, the all-time home run leader. Here are the numbers of home runs hit by Hank Aaron in each of his major league seasons:

13	27	26	44	30	39	40	34	45	44	24
32	44	39	29	44	38	47	34	40	20	

Aaron's mean number of home runs hit in a year is

$$\bar{x} = \frac{1}{21}(13 + 27 + \cdots + 20) = \frac{733}{21} = 34.9$$

Barry Bonds's exceptional performance in 2001 stands out from his home run production in the previous 15 seasons. Use your calculator to check that his mean home run production in his first 15 seasons is $\bar{x} = 32.93$. One outstanding season increased Bonds's mean home run count by 2.5 home runs per year.

Example 1.10 illustrates an important fact about the mean as a measure of center: it 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.

resistant measure

Measuring center: the median

In Section 1.1, we used the midpoint of a distribution as an informal measure of center. 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, the number such that half the observations are smaller and the other half are larger. To find the median of a distribution:

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.

3. If the number of observations n is even, the median M is the mean of the two center observations 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 very tedious, however, so that finding the median by hand for larger sets of data is unpleasant. You will need computer software or a graphing calculator to automate finding the median.

EXAMPLE 1.11 FINDING MEDIANS

To find the median number of home runs Barry Bonds hit in his first 16 seasons, first arrange the data in increasing order:

16	19	24	25	25	33	33	**34**	**34**	37	37	40	42	46	49	73

The count of observations $n = 16$ is even. There is no center observation, but there is a center pair. These are the two bold 34s in the list, which have 7 observations to their left in the list and 7 to their right. The median is midway between these two observations. Because both of the middle pair are 34, $M = 34$.

How much does the apparent outlier affect the median? Drop the 73 from the list and find the median for the remaining $n = 15$ years. It is the 8th observation in the edited list, $M = 34$.

How does Bonds's median compare with Hank Aaron's? Here, arranged in increasing order, are Aaron's home run counts:

13	20	24	26	27	29	30
32	34	34	**38**	39	39	40
40	44	44	44	44	45	47

The number of observations is odd, so there is one center observation. This is the median. It is the bold 38, which has 10 observations to its left in the list and 10 observations to its right. Bonds now holds the single-season record, but he has hit fewer home runs in a typical season than Aaron. Barry Bonds also has a long way to go to catch Aaron's career total of 733 home runs.

Comparing the mean and the median

Examples 1.10 and 1.11 illustrate an important difference between the mean and the median. The one high value pulls Bonds's mean home run count up from 32.93 to 35.4375. The median is not affected at all. The median, unlike the mean, is *resistant*. If Bonds's record 73 had been 703, his median would not change at all. The 703 just counts as one observation above the center, no matter how far above the center it lies. The mean uses the actual value of each observation and so will chase a single large observation upward.

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. For example, the distribution of house prices is strongly skewed to the right. There are many moderately priced houses and a few very expensive mansions. The few expensive houses pull the mean up but do not affect the median. The mean price of new houses sold in 1997 was $176,000, but the median price for these same houses was only $146,000. Reports about house prices, incomes, and other strongly skewed distributions usually give the median ("midpoint") rather than the mean ("arithmetic average"). However, if you are a tax assessor interested in the total value of houses in your area, use the mean. The total value is the mean times the number of houses; it has no connection with the median. The mean and median measure center in different ways, and both are useful.

EXERCISES

1.31 Joey's first 14 quiz grades in a marking period were

86	84	91	75	78	80	74	87	76	96	82	90	98	93

(a) Use the formula to calculate the mean. Check using "one-variable statistics" on your calculator.

(b) Suppose Joey has an unexcused absence for the fifteenth quiz and he receives a score of zero. Determine his final quiz average. What property of the mean does this situation illustrate? Write a sentence about the effect of the zero on Joey's quiz average that mentions this property.

(c) What kind of plot would best show Joey's distribution of grades? Assume an 8-point grading scale (A: 93 to 100, B: 85 to 92, etc.). Make an appropriate plot, and be prepared to justify your choice.

1.32 SSHA SCORES The Survey of Study Habits and Attitudes (SSHA) is a psychological test that evaluates college students' motivation, study habits, and attitudes toward school. A private college gives the SSHA to a sample of 18 of its incoming first-year women students. Their scores are

154	109	137	115	152	140	154	178	101
103	126	126	137	165	165	129	200	148

(a) Make a stemplot of these data. The overall shape of the distribution is irregular, as often happens when only a few observations are available. Are there any potential outliers? About where is the center of the distribution (the score with half the scores above it and half below)? What is the spread of the scores (ignoring any outliers)?

(b) Find the mean score from the formula for the mean. Then enter the data into your calculator. You can find the mean from the home screen as follows:

TI-83	TI-89
• Press 2nd STAT (LIST) ▶ ▶ (MATH).	• Press CATALOG then 5 (M).
• Choose 3:mean(, enter list name, press ENTER.	• Choose mean(, type list name, press ENTER.

(c) Find the median of these scores. Which is larger: the median or the mean? Explain why.

1.33 Suppose a major league baseball team's mean yearly salary for a player is $1.2 million, and that the team has 25 players on its active roster. What is the team's annual payroll for players? If you knew only the median salary, would you be able to answer the question? Why or why not?

1.34 Last year a small accounting firm paid each of its five clerks $22,000, two junior accountants $50,000 each, and the firm's owner $270,000. What is the mean salary paid at this firm? How many of the employees earn less than the mean? What is the median salary? Write a sentence to describe how an unethical recruiter could use statistics to mislead prospective employees.

1.35 U.S. INCOMES The distribution of individual incomes in the United States is strongly skewed to the right. In 1997, the mean and median incomes of the top 1% of Americans were $330,000 and $675,000. Which of these numbers is the mean and which is the median? Explain your reasoning.

Measuring spread: the quartiles

The mean and median provide two different measures of the center of a distribution. But a measure of center alone can be misleading. The Census Bureau reports that in 2000 the median income of American households was $41,345. Half of all households had incomes below $41,345, and half had higher incomes. But these figures do not tell the whole story. Two nations with the same median household income are very different if one has extremes of wealth and poverty and the other has little variation among households. 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.

range

One way to measure spread is to calculate the ***range***, which is the difference between the largest and smallest observations. For example, the number of home runs Barry Bonds has hit in a season has a *range* of 73 – 16 = 57. The range shows the full spread of the data. But it depends on only the smallest observation and the largest observation, which may be outliers. We can improve our description of spread by also looking at the spread of the middle half of the data. The *quartiles* mark out the middle half. Count up the ordered list of observations, starting from the smallest. The *first quartile* lies one-quarter of the way up the list. The *third quartile* lies three-quarters of the way up the list. In other words, the first quartile is larger than 25% of the observations, and the third quartile is larger than 75% of the observations. The second quartile is the median, which is larger than 50% of the observations. That is the idea of quartiles. We need a rule to make the idea exact. The rule for calculating the quartiles uses the rule for the median.

THE QUARTILES Q_1 and Q_3

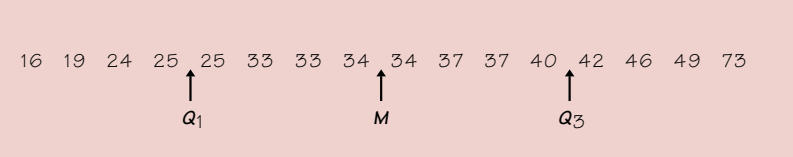
To calculate the *quartiles*

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.

Here is an example that shows how the rules for the quartiles work for both odd and even numbers of observations.

EXAMPLE 1.12 FINDING QUARTILES

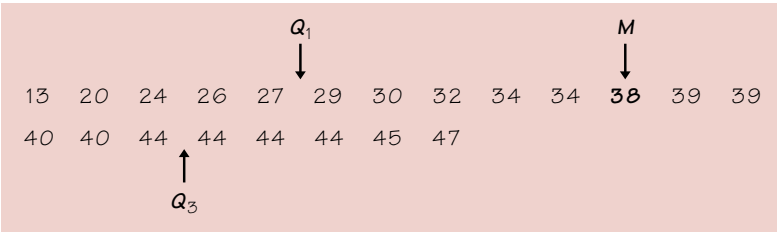
Barry Bonds's home run counts (arranged in order) are



There is an even number of observations, so the median lies midway between the middle pair, the 8th and 9th in the list. The first quartile is the median of the 8 observations to the left of $M = 34$. So $Q_1 = 25$. The third quartile is the median of the 8 observations to the right of M. $Q_3 = 41$. Note that we don't include M when we're computing the quartiles.

The quartiles are *resistant*. For example, Q_3 would have the same value if Bonds's record 73 were 703.

Hank Aaron's data, again arranged in increasing order, are



In Example 1.11, we determined that the median is the bold 38 in the list. The first quartile is the median of the 10 observations to the left of $M = 38$. This is the mean of the 5th and 6th of these 10 observations, so $Q_1 = 28$. $Q_3 = 44$. The overall median is left out of the calculation of the quartiles.

Be careful when, as in these examples, several observations take the same numerical value. Write down all of the observations and apply the rules just as if they all had distinct values. Some software packages use a slightly different rule to find the quartiles, so computer results may be a bit different from your own work. Don't worry about this. The differences will always be too small to be important.

The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the *interquartile range*.

THE INTERQUARTILE RANGE (IQR)

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

$$IQR = Q_3 - Q_1$$

If an observation falls between Q_1 and Q_3, then you know it's neither unusually high (upper 25%) or unusually low (lower 25%). The IQR is the basis of a rule of thumb for identifying suspected outliers.

OUTLIERS: THE 1.5 × *IQR* CRITERION

Call an observation an outlier if it falls more than $1.5 \times IQR$ above the third quartile or below the first quartile.

EXAMPLE 1.13 DETERMINING OUTLIERS

We suspect that Barry Bonds's 73 home run season is an outlier. Let's test.

$$IQR = Q_3 - Q_1 = 41 - 25 = 16$$
$$Q_3 + 1.5 \times IQR = 41 + (1.5 \times 16) = 65 \text{ (upper cutoff)}$$
$$Q_1 - 1.5 \times IQR = 25 - (1.5 \times 16) = 1 \text{ (lower cutoff)}$$

Since 73 is above the upper cutoff, Bonds's record-setting year was an outlier.

The five-number summary and boxplots

The smallest and largest 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 data set 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 from Example 1.12 are

16 25 34 41 73

for Bonds and

13 28 38 44 47

for Aaron. The five-number summary of a distribution leads to a new graph, the ***boxplot.*** Figure 1.17 shows boxplots for the home run comparison.

boxplot

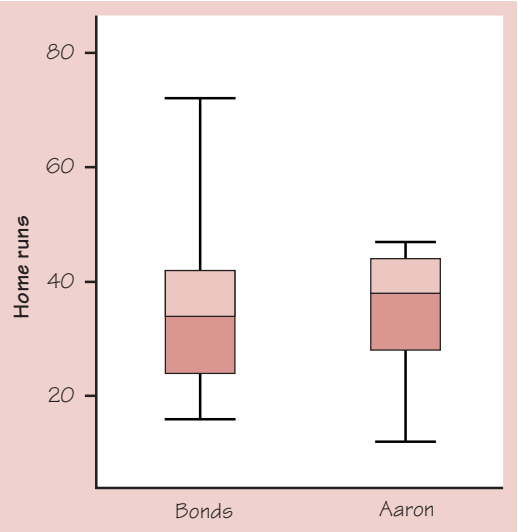


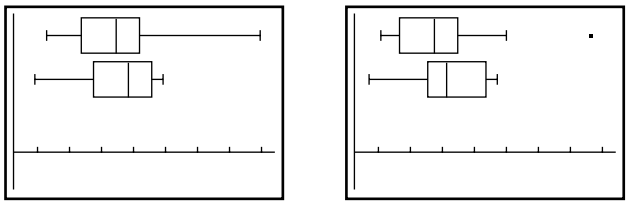
FIGURE 1.17 Side-by-side boxplots comparing the numbers of home runs per year by Barry Bonds and Hank Aaron.

Because boxplots show less detail than histograms or stemplots, they are best used for side-by-side comparison of more than one distribution, as in Figure 1.17. You can draw boxplots either horizontally or vertically. Be sure to include a numerical scale in the graph. 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. We see from Figure 1.17 that Aaron and Bonds are about equally consistent when we look at the middle 50% of their home run distributions.

A boxplot also gives an indication of the symmetry or skewness of a distribution. In a symmetric distribution, the first and third quartiles are equally distant from the median. In most distributions that are skewed to the right, however, the third quartile will be farther above the median than the first quartile is below it. The extremes behave the same way, but remember that they are just single observations and may say little about the distribution as a whole. In Figure 1.17, we can see that Aaron's home run distribution is skewed to the left. Barry Bonds's distribution is more difficult to describe.

Outliers usually deserve special attention. Because the regular boxplot conceals outliers, we will adopt the ***modified boxplot***, which plots outliers as isolated points. Figures 1.18(a) and (b) show regular and modified boxplots for the home runs hit by Bonds and Aaron. The regular boxplot suggests a very large spread in the upper 25% of Bonds's distribution. The modified boxplot shows that if not for the outlier, the distribution would show much less variability. Because the modified boxplot shows more detail, when we say "boxplot" from now on, we will mean "modified boxplot." Both the TI-83 and the TI-89 give you a choice of regular or modified boxplot. When you construct a (modified) boxplot by hand, extend the "whiskers"

modified boxplot



(a) (b)

FIGURE 1.18 Regular (a) and modified (b) boxplots comparing the home run production of Barry Bonds and Hank Aaron.

out to the largest and the smallest data points that are not outliers. Then plot outliers as isolated points.

BOXPLOT (MODIFIED)

A **modified boxplot** is a graph of the five-number summary, with outliers plotted individually.

- A central box spans the quartiles.
- A line in the box marks the median.
- Observations more than $1.5 \times IQR$ outside the central box are plotted individually.
- Lines extend from the box out to the smallest and largest observations that are not outliers.

TECHNOLOGY TOOLBOX *Calculator boxplots and numerical summaries*

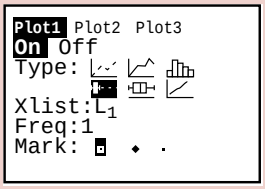
The TI-83 and TI-89 can plot up to three boxplots in the same viewing window. Both calculators can also calculate the mean, median, quartiles, and other one-variable statistics for data stored in lists. In this example, we compare Barry Bonds to Babe Ruth, the "Sultan of Swat." Here are the numbers of home runs hit by Ruth in each of his seasons as a New York Yankee (1920 to 1934):

54	59	35	41	46	25	47	60	54	46	49	46	41	34	22

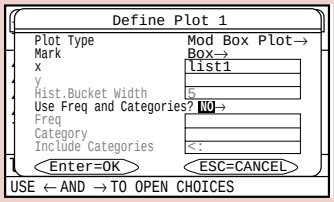
1. Enter Bonds's home run data in L_1/list1 and Ruth's in L_2/list2.
2. Set up two statistics plots: Plot 1 to show a modified boxplot of Bonds's data and Plot 2 to show a modified boxplot of Ruth's data.

TECHNOLOGY TOOLBOX *Calculator boxplots and numerical summaries (continued)*

TI-83

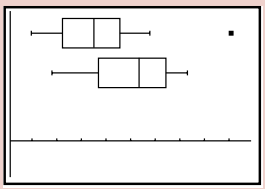


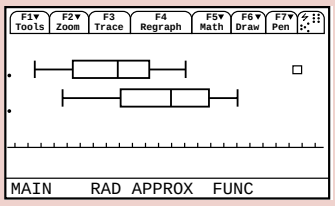
TI-89



3. Use the calculator's zoom feature to display the side-by-side boxplots.

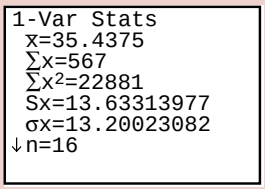
- Press ZOOM and select 9:ZoomStat.
- Press F5 (ZoomData).

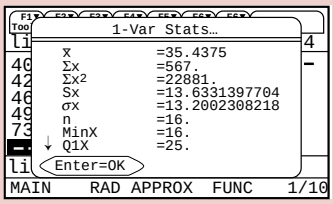




4. Calculate numerical summaries for each set of data.

- Press STAT ▶ (CALC) and select 1:1-Var Stats
- Press ENTER. Now press 2nd 1 (L1) and ENTER.
- Press F4 (Calc) and choose 1:1-Var Stats.
- Type list1 in the list box. Press ENTER.





5. Notice the down arrow on the left side of the display. Press ▼ to see Bonds's other statistics. Repeat the process to find the Babe's numerical summaries.

EXERCISES

1.36 SSHA SCORES Here are the scores on the Survey of Study Habits and Attitudes (SSHA) for 18 first-year college women:

154 109 137 115 152 140 154 178 101 103 126 126 137 165 165 129 200 148

and for 20 first-year college men:

108 140 114 91 180 115 126 92 169 146 109 132 75 88 113 151 70 115 187 104

(a) Make side-by-side boxplots to compare the distributions.

(b) Compute numerical summaries for these two distributions.

(c) Write a paragraph comparing the SSHA scores for men and women.

1.37 HOW OLD ARE PRESIDENTS? Return to the data on presidential ages in Table 1.4 (page 19). In Example 1.6, we constructed a histogram of the age data.

(a) From the shape of the histogram (Figure 1.7, page 20), do you expect the mean to be much less than the median, about the same as the median, or much greater than the median? Explain.

(b) Find the five-number summary and verify your expectation from (a).

(c) What is the range of the middle half of the ages of new presidents?

(d) Construct by hand a (modified) boxplot of the ages of new presidents.

(e) On your calculator, define Plot 1 to be a histogram using the list named PREZ that you created in the Technology Toolbox on page 22. Define Plot 2 to be a (modified) boxplot also using the list PREZ. Use the calculator's zoom command to generate a graph. To remove the overlap, adjust your viewing window so that Ymin = –6 and Ymax = 22. Then graph. Use TRACE to inspect values. Press the up and down cursor keys to toggle between plots. Is there an outlier? If so, who was it?

1.38 Is the interquartile range a resistant measure of spread? Give an example of a small data set that supports your answer.

1.39 SHOPPING SPREE, III Figure 1.19 displays computer output for the data on amount spent by grocery shoppers in Exercise 1.11 (page 18).

(a) Find the total amount spent by the shoppers.

(b) Make a boxplot from the computer output. Did you check for outliers?

DataDesk

```
Summaryof    spending
No Selector

Percentile   25

          Count    50
           Mean    34.7022
         Median    27.8550
         StdDev    21.6974
            Min     3.11000
            Max    93.3400
Lower ith %tile    19.2700
Upper ith %tile    45.4000
```

Minitab

```
Descriptive Statistics

Variable      N     Mean   Median   TrMean    StDev   SEMean
spending     50    34.70    27.85    32.92    21.70     3.07

Variable    Min      Max       Q1       Q3
spending   3.11    93.34    19.06    45.72
```

FIGURE 1.19 Numerical descriptions of the unrounded shopping data from the Data Desk and Minitab software.

Measuring spread: the standard deviation

The five-number summary is not the most common numerical description of a distribution. That distinction belongs to the combination of the mean to measure center and the *standard deviation* to measure spread. The standard deviation measures spread by looking at how far the observations are from their mean.

THE STANDARD DEVIATION s

The **variance** s^2 of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of n observations $x_1, x_2, \ldots, x_n$ is

$$s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1}$$

or, more compactly,

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

The **standard deviation** s is the square root of the variance s^2:

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

In practice, use software or your calculator to obtain the standard deviation from keyed-in data. Doing a few examples step-by-step will help you understand how the variance and standard deviation work, however. Here is such an example.

EXAMPLE 1.14 METABOLIC RATE

A person's metabolic rate is the rate at which the body consumes energy. Metabolic rate is important in studies of weight gain, dieting, and exercise. Here are the metabolic rates of 7 men who took part in a study of dieting. (The units are calories per 24 hours. These are the same calories used to describe the energy content of foods.)

1792	1666	1362	1614	1460	1867	1439

The researchers reported $\bar{x}$ and s for these men.

First find the mean:

$$\bar{x} = \frac{1792+1666+1362+1614+1460+1867+1439}{7} = \frac{11{,}200}{7} = 1600 \text{ calories}$$

To see clearly the nature of the variance, start with a table of the deviations of the observations from this mean.

Observations x_i	Deviations $x_i - \bar{x}$	Squared deviations $(x_i - \bar{x})^2$
1792	$1792 - 1600 = 192$	$192^2 = 36{,}864$
1666	$1666 - 1600 = 66$	$66^2 = 4{,}356$
1362	$1362 - 1600 = -238$	$(-238)^2 = 56{,}644$
1614	$1614 - 1600 = 14$	$14^2 = 196$
1460	$1460 - 1600 = -140$	$(-140)^2 = 36{,}864$
1867	$1867 - 1600 = 267$	$267^2 = 71{,}289$
1439	$1439 - 1600 = -161$	$(-161)^2 = 25{,}921$
	sum $= 0$	sum $= 214{,}870$

The variance is the sum of the squared deviations divided by one less than the number of observations:

$$s^2 = \frac{214{,}870}{6} = 35{,}811.67$$

The standard deviation is the square root of the variance:

$$s = \sqrt{35{,}811.67} = 189.24 \text{ calories}$$

Compare these results for s^2 and s with those generated by your calculator or computer.

Figure 1.20 displays the data of Example 1.14 as points above the number line, with their mean marked by an asterisk (*). The arrows show two of the deviations from the mean. These deviations show how spread out the data are about their mean. Some of the deviations will be positive and some negative because observations fall on each side of the mean. In fact, *the sum of the deviations of the observations from their mean will always be zero*. Check that this is true in Example 1.14. So we cannot simply add the deviations to get an overall measure of spread. Squaring the deviations makes them all nonnegative, so that observations far from the mean in either direction will have large positive squared deviations. The variance s^2 is the average squared deviation. The variance is large if the observations are widely spread about their mean; it is small if the observations are all close to the mean.

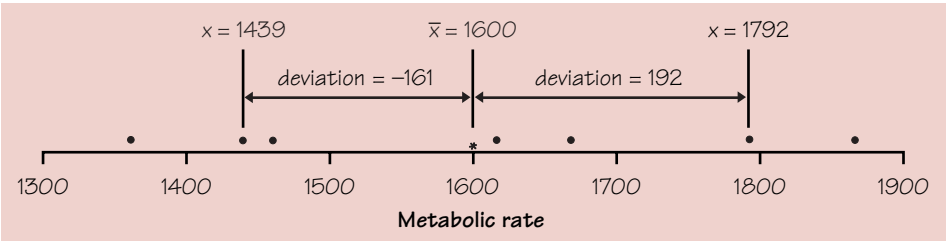


FIGURE 1.20 Metabolic rates for seven men, with their mean (*) and the deviations of two observations from the mean.

Because the variance involves squaring the deviations, it does not have the same unit of measurement as the original observations. Lengths measured in centimeters, for example, have a variance measured in squared centimeters. Taking the square root remedies this. The standard deviation s measures spread about the mean in the original scale.

If the variance is the average of the squares of the deviations of the observations from their mean, why do we average by dividing by $n - 1$ rather than n? Because the sum of the deviations is always zero, the last deviation can be found once we know the other $n - 1$ deviations. So we are not averaging n unrelated numbers. Only $n - 1$ of the squared deviations can vary freely, and we average by dividing the total by $n - 1$. The number $n - 1$ is called the ***degrees of freedom*** of the variance or of the standard deviation. Many calculators offer a choice between dividing by n and dividing by $n - 1$, so be sure to use $n - 1$.

degrees of freedom

Leaving the arithmetic to a calculator allows us to concentrate on what we are doing and why. What we are doing is measuring spread. Here are the basic properties of the standard deviation s as a measure of spread.

PROPERTIES OF THE STANDARD DEVIATION

- s measures spread about the mean and should be used only when the mean is chosen as the measure of center.
- $s = 0$ only when there is *no spread*. This happens only when all observations have the same value. Otherwise, $s > 0$. As the observations become more spread out about their mean, s gets larger.
- s, like the mean $\bar{x}$, is not resistant. Strong skewness or a few outliers can make s very large. For example, the standard deviation of Barry Bonds's home run counts is 13.633. (Use your calculator to verify this.) If we omit the outlier, the standard deviation drops to 9.573.

You may rightly feel that the importance of the standard deviation is not yet clear. We will see in the next chapter that the standard deviation is the natural measure of spread for an important class of symmetric distributions, the normal distributions. The usefulness of many statistical procedures is tied to distributions of particular shapes. This is certainly true of the standard deviation.

Choosing measures of center and spread

How do we choose between the five-number summary and $\bar{x}$ and s to describe the center and spread of a distribution? Because the two sides of a strongly skewed distribution have different spreads, no single number such as s describes the spread well. The five-number summary, with its two quartiles and two extremes, does a better job.

CHOOSING A SUMMARY

The five-number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with strong outliers. Use $\bar{x}$ and s only for reasonably symmetric distributions that are free of outliers.

Do remember that a graph gives the best overall picture of a distribution. Numerical measures of center and spread report specific facts about a distribution, but they do not describe its entire shape. Numerical summaries do not disclose the presence of multiple peaks or gaps, for example. **Always plot your data.**

EXERCISES

1.40 PHOSPHATE LEVELS The level of various substances in the blood influences our health. Here are measurements of the level of phosphate in the blood of a patient, in milligrams of phosphate per deciliter of blood, made on 6 consecutive visits to a clinic:

5.6	5.2	4.6	4.9	5.7	6.4

A graph of only 6 observations gives little information, so we proceed to compute the mean and standard deviation.

(a) Find the mean from its definition. That is, find the sum of the 6 observations and divide by 6.

(b) Find the standard deviation from its definition. That is, find the deviations of each observation from the mean, square the deviations, then obtain the variance and the standard deviation. Example 1.14 shows the method.

(c) Now enter the data into your calculator to obtain $\bar{x}$ and s. Do the results agree with your hand calculations? Can you find a way to compute the standard deviation without using one-variable statistics?

1.41 ROGER MARIS New York Yankee Roger Maris held the single-season home run record from 1961 until 1998. Here are Maris's home run counts for his 10 years in the American League:

14	28	16	39	61	33	23	26	8	13

(a) Maris's mean number of home runs is $\bar{x} = 26.1$. Find the standard deviation s from its definition. Follow the model of Example 1.14.

(b) Use your calculator to verify your results. Then use your calculator to find $\bar{x}$ and s for the 9 observations that remain when you leave out the outlier. How does the outlier affect the values of $\bar{x}$ and s? Is s a resistant measure of spread?

1.42 OLDER FOLKS, III In Exercise 1.12 (page 22), you made a histogram displaying the percentage of residents aged 65 or older in each of the 50 U.S. states. Do you prefer the five-number summary or $\bar{x}$ and s as a brief numerical description? Why? Calculate your preferred description.

1.43 This is a standard deviation contest. You must choose four numbers from the whole numbers 0 to 10, with repeats allowed.

(a) Choose four numbers that have the smallest possible standard deviation.

(b) Choose four numbers that have the largest possible standard deviation.

(c) Is more than one choice possible in either **(a)** or **(b)**? Explain.

Changing the unit of measurement

The same variable can be recorded in different units of measurement. Americans commonly record distances in miles and temperatures in degrees Fahrenheit. Most of the rest of the world measures distances in kilometers and temperatures in degrees Celsius. Fortunately, it is easy to convert from one unit of measurement to another. In doing so, we perform a *linear transformation*.

LINEAR TRANSFORMATION

A linear transformation changes the original variable x into the new variable x_{new} given by an equation of the form

$$x_{new} = a + bx$$

Adding the constant a shifts all values of x upward or downward by the same amount.

Multiplying by the positive constant b changes the size of the unit of measurement.

EXAMPLE 1.15 LOS ANGELES LAKERS' SALARIES

Table 1.8 gives the approximate base salaries of the 14 members of the Los Angeles Lakers basketball team for the year 2000. You can calculate that the mean is $\bar{x}$ = \$4.14 million and that the median is M = \$2.6 million. No wonder professional basketball players have big houses!

TABLE 1.8 Year 2000 salaries for the Los Angeles Lakers

Player	Salary	Player	Salary
Shaquille O'Neal	\$17.1 million	Ron Harper	\$2.1 million
Kobe Bryant	\$11.8 million	A. C. Green	\$2.0 million
Robert Horry	\$5.0 million	Devean George	\$1.0 million
Glen Rice	\$4.5 million	Brian Shaw	\$1.0 million
Derek Fisher	\$4.3 million	John Salley	\$0.8 million
Rick Fox	\$4.2 million	Tyronne Lue	\$0.7 million
Travis Knight	\$3.1 million	John Celestand	\$0.3 million

Figure 1.21(a) is a stemplot of the salaries, with millions as stems. The distribution is skewed to the right and there are two high outliers. The very high salaries of Kobe Bryant and Shaquille O'Neal pull up the mean. Use your calculator to check that s = \$4.76 million, and that the five-number summary is

\$0.3 million	\$1.0 million	\$2.6 million	\$4.5 million	\$17.1 million

(a) Suppose that each member of the team receives a \$100,000 bonus for winning the NBA Championship (which the Lakers did in 2000). How will this affect the shape, center, and spread of the distribution?

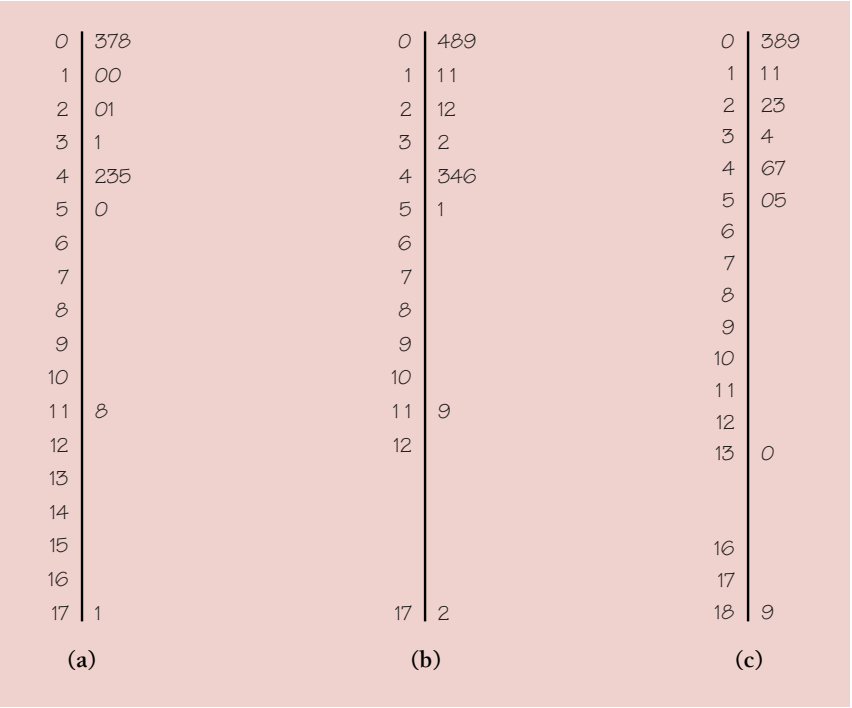


FIGURE 1.21 Stemplots of the salaries of Los Angeles Lakers players, from Table 1.8.

Since \$100,000 = \$0.1 million, each player's salary will increase by \$0.1 million. This linear transformation can be represented by $x_{new} = 0.1 + 1x$, where x_{new} is the salary after the bonus and x is the player's base salary. Increasing each value in Table 1.8 by 0.1 will also increase the mean by 0.1. That is, $\bar{x}_{new}$ = \$4.24 million. Likewise, the median salary will increase by 0.1 and become M = \$2.7 million.

What will happen to the spread of the distribution? The standard deviation of the Lakers' salaries after the bonus is still s = \$4.76 million. With the bonus, the five-number summary becomes

$0.4 million	$1.1 million	$2.7 million	$4.6 million	$17.2 million

Both before and after the salary bonus, the *IQR* for this distribution is $3.5 million. *Adding a constant amount to each observation does not change the spread.* The shape of the distribution remains unchanged, as shown in Figure 1.21(b).

(b) Suppose that, instead of receiving a $100,000 bonus, each player is offered a 10% increase in his base salary. John Celestand, who is making a base salary of $0.3 million, would receive an additional (0.10)($0.3 million) = $0.03 million. To obtain his new salary, we could have used the linear transformation $x_{new} = 0 + 1.10x$, since multiplying the current salary (x) by 1.10 increases it by 10%. Increasing all 14 players' salaries in the same way results in the following list of values (in millions):

$0.33	$0.77	$0.88	$1.10	$1.10	$2.20	$2.31
$3.41	$4.62	$4.73	$4.95	$5.50	$12.98	$18.81

Use your calculator to check that $\bar{x}_{new}$ = $4.55 million, s_{new} = $5.24 million, M_{new} = $2.86 million, and the five-number summary for x_{new} is

$0.33	$1.10	$2.86	$4.95	$18.81

Since $4.14(1.10) = $4.55 and $2.6(1.10) = $2.86, you can see that both measures of center (the mean and median) have increased by 10%. This time, the spread of the distribution has increased, too. Check for yourself that the standard deviation and the *IQR* have also increased by 10%. The stemplot in Figure 1.21(c) shows that the distribution of salaries is still right-skewed.

Linear transformations do not change the shape of a distribution. As you saw in the previous example, changing the units of measurement can affect the center and spread of the distribution. Fortunately, the effects of such changes follow a simple pattern.

EFFECT OF A LINEAR TRANSFORMATION

To see the effect of a linear transformation on measures of center and spread, apply these rules:

- Multiplying each observation by a positive number b multiplies both measures of center (mean and median) and measures of spread (standard deviation and *IQR*) by b.
- Adding the same number a (either positive or negative) to each observation adds a to measures of center and to quartiles but does not change measures of spread.

EXERCISES

1.44 COCKROACHES! Maria measures the lengths of 5 cockroaches that she finds at school. Here are her results (in inches):

1.4	2.2	1.1	1.6	1.2

(a) Find the mean and standard deviation of Maria's measurements.

(b) Maria's science teacher is furious to discover that she has measured the cockroach lengths in inches rather than centimeters. (There are 2.54 cm in 1 inch.) She gives Maria two minutes to report the mean and standard deviation of the 5 cockroaches in centimeters. Maria succeeded. Will you?

(c) Considering the 5 cockroaches that Maria found as a small sample from the population of all cockroaches at her school, what would you estimate as the average length of the population of cockroaches? How sure of your estimate are you?

1.45 RAISING TEACHERS' PAY A school system employs teachers at salaries between $30,000 and $60,000. The teachers' union and the school board are negotiating the form of next year's increase in the salary schedule. Suppose that every teacher is given a flat $1000 raise.

(a) How much will the mean salary increase? The median salary?

(b) Will a flat $1000 raise increase the spread as measured by the distance between the quartiles?

(c) Will a flat $1000 raise increase the spread as measured by the standard deviation of the salaries?

1.46 RAISING TEACHERS' PAY, II Suppose that the teachers in the previous exercise each receive a 5% raise. The amount of the raise will vary from $1500 to $3000, depending on present salary. Will a 5% across-the-board raise increase the spread of the distribution as measured by the distance between the quartiles? Do you think it will increase the standard deviation?

Comparing distributions

An experiment is carried out to compare the effectiveness of a new cholesterol-reducing drug with the one that is currently prescribed by most doctors. A survey is conducted to determine whether the proportion of males who are likely to vote for a political candidate is higher than the proportion of females who are likely to vote for the candidate. Students taking AP Calculus AB and AP Statistics are curious about which exam is harder. They have information on the distribution of scores earned on each exam from the year 2000. In each of these situations, we are interested in comparing distributions. This section presents some of the more common methods for making statistical comparisons.

EXAMPLE 1.16 COOL CAR COLORS

Table 1.9 gives information about the color preferences of vehicle purchasers in 1998.

TABLE 1.9 Colors of cars and trucks purchased in 1998

Color	Full-sized or intermediate-sized car	Light truck or van
Medium or dark green	16.4%	15.5%
White	15.6%	22.5%
Light brown	14.1%	6.1%
Silver	11.0%	6.2%
Black	8.9%	11.5%

Source: *The World Almanac and Book of Facts*, 2000.

Figure 1.22 is a graph that can be used to compare the color distributions for cars and trucks. By placing the bars side-by-side, we can easily observe the similarities and differences within each of the color categories. White seems to be the favorite color of most truck buyers, while car purchasers favor medium or dark green. What other similarities and differences do you see?

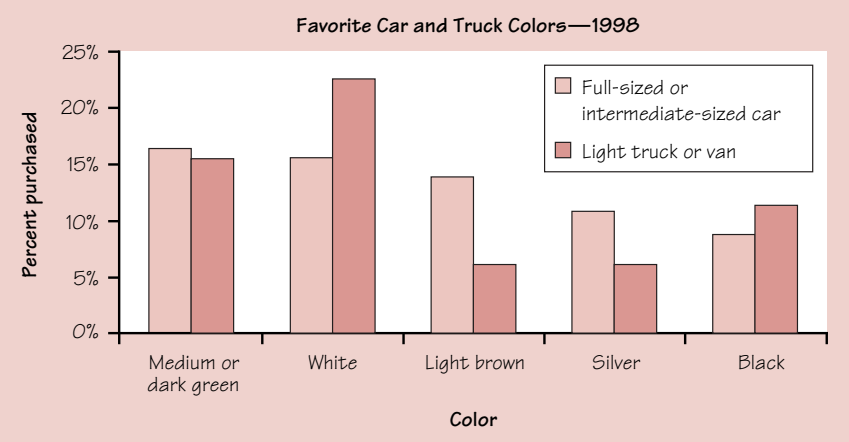


FIGURE 1.22 Side-by-side bar graph of most-popular car and truck colors from 1998.

An effective graphical display for comparing two fairly small quantitative data sets is a *back-to-back stemplot*. Example 1.17 shows you how.

EXAMPLE 1.17 SWISS DOCTORS

A study in Switzerland examined the number of cesarean sections (surgical deliveries of babies) performed in a year by doctors. Here are the data for 15 male doctors:

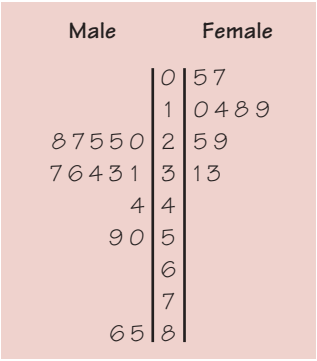
27	50	33	25	86	25	85	31	37	44	20	36	59	34	28

The study also looked at 10 female doctors. The number of cesareans performed by these doctors (arranged in order) were

5	7	10	14	18	19	25	29	31	33

We can compare the number of cesarean sections performed by male and female doctors using a back-to-back stemplot. Figure 1.23 shows the completed graph. As you can see, the stems are listed in the middle and leaves are placed on the left for male doctors and on the right for female doctors. It is usual to have the leaves increase in value as they move away from the stem.

NUMBER OF CESAREAN SECTIONS PERFORMED BY MALE AND FEMALE DOCTORS



Key:
|2| 5 means that a female doctor performed 25 cesarean sections that year
0 |5| means that a male doctor performed 50 cesarean sections that year

FIGURE 1.23 Back-to-back stemplot of the number of cesarean sections performed by male and female Swiss doctors.

The distribution of the number of cesareans performed by female doctors is roughly symmetric. For the male doctors, the distribution is skewed to the right. More than half of the female doctors in the study performed fewer than 20 cesarean sections in a year. The minimum number of cesareans performed by any of the male doctors was 20. Two male physicians performed an unusually high number of cesareans, 85 and 86.

Here are numerical summaries for the two distributions:

	$\bar{x}$	s	Min.	Q_1	M	Q_3	Max.	IQR
Male doctors	41.333	20.607	20	27	34	50	86	23
Female doctors	19.1	10.126	5	10	18.5	29	33	19

The mean and median numbers of cesarean sections performed are higher for the male doctors. Both the standard deviation and the *IQR* for the male doctors are much larger than the corresponding statistics for the female doctors. So there is much greater variability in the number of cesarean sections performed by male physicians. Due to the apparent outliers in the male doctor data and the lack of symmetry of their distribution of cesareans, we should use the medians and *IQR*s in our numerical comparisons.

We have already seen that boxplots can be useful for comparing distributions of quantitative variables. Side-by-side boxplots, like those in the Technology Toolbox on page 47, help us quickly compare shape, center, and spread.

EXERCISES

1.47 GET YOUR HOT DOGS HERE! "Face it. A hot dog isn't a carrot stick." So said *Consumer Reports*, commenting on the low nutritional quality of the all-American frank. Table 1.10 shows the magazine's laboratory test results for calories and milligrams of sodium (mostly due to salt) in a number of major brands of hot dogs. There are three types: beef, "meat" (mainly pork and beef, but government regulations allow up to 15% poultry meat), and poultry. Because people concerned about their health may prefer low-calorie, low-sodium hot dogs, we ask: "Are there any systematic differences among the three types of hot dogs in these two variables?" Use side-by-side boxplots and numerical summaries to help you answer this question. Write a paragraph explaining your findings.

TABLE 1.10 Calories and sodium in three types of hot dogs

Beef hot dogs		Meat hot dogs		Poultry hot dogs	
Calories	Sodium	Calories	Sodium	Calories	Sodium
186	495	173	458	129	430
181	477	191	506	132	375
176	425	182	473	102	396
149	322	190	545	106	383
184	482	172	496	94	387
190	587	147	360	102	542
158	370	146	387	87	359
139	322	139	386	99	357
175	479	175	507	170	528
148	375	136	393	113	513
152	330	179	405	135	426
111	300	153	372	142	513
141	386	107	144	86	358
153	401	195	511	143	581
190	645	135	405	152	588
157	440	140	428	146	522
131	317	138	339	144	545
149	319				
135	298				
132	253				

Source: *Consumer Reports*, June 1986, pp.366–367

1.48 WHICH AP EXAM IS EASIER: CALCULUS AB OR STATISTICS? The table below gives the distribution of grades earned by students taking the Calculus AB and Statistics exams in 2000.[14]

	5	4	3	2	1
Calculus AB	16.8%	23.2%	23.5%	19.6%	16.8%
Statistics	9.8%	21.5%	22.4%	20.5%	25.8%

(a) Make a graphical display to compare the AP exam grades for Calculus AB and Statistics.

(b) Write a few sentences comparing the two distributions of exam grades. Do you now know which exam is easier? Why or why not?

1.49 WHO MAKES MORE? A manufacturing company is reviewing the salaries of its full-time employees below the executive level at a large plant. The clerical staff is almost entirely female, while a majority of the production workers and technical staff are male. As a result, the distributions of salaries for male and female employees may be quite different. Table 1.11 gives the frequencies and relative frequencies for women and men.

(a) Make histograms for these data, choosing a vertical scale that is most appropriate for comparing the two distributions.

(b) Describe the shape of the overall salary distributions and the chief differences between them.

(c) Explain why the total for women is greater than 100%.

TABLE 1.11 Salary distributions of female and male workers in a large factory

Salary ($1000)	Women		Men	
	Number	%	Number	%
10–15	89	11.8	26	1.1
15–20	192	25.4	221	9.0
20–25	236	31.2	677	27.9
25–30	111	14.7	823	33.6
30–35	86	11.4	365	14.9
35–40	25	3.3	182	7.4
40–45	11	1.5	91	3.7
45–50	3	0.4	33	1.4
50–55	2	0.3	19	0.8
55–60	0	0.0	11	0.4
60–65	0	0.0	0	0.0
65–70	1	0.1	3	0.1
Total	756	100.1	2451	100.0

1.50 BASKETBALL PLAYOFF SCORES Here are the scores of games played in the California Division I-AAA high school basketball playoffs:[15]

71–38 52–47 55–53 76–65 77–63 65–63 68–54 64–62
87–47 64–56 78–64 58–51 91–74 71–41 67–62 106–46

On the same day, the final scores of games in Division V-AA were

98–45 67–44 74–60 96–54 92–72 93–46
98–67 62–37 37–36 69–44 86–66 66–58

(a) Construct a back-to-back stemplot to compare the number of points scored by Division I-AAA and Division V-AA basketball teams.

(b) Compare the shape, center, and spread of the two distributions. Which numerical summaries are most appropriate in this case? Why?

(c) Is there a difference in "margin of victory" in Division I-AAA and Division V-AA playoff games? Provide appropriate graphical and numerical support for your answer.

SUMMARY

A numerical summary of a distribution should report its **center** and its **spread,** or **variability.**

The **mean** $\bar{x}$ and the **median** M describe the center of a distribution in different ways. The mean is the arithmetic average of the observations, and the median is the midpoint of the values.

When you use the median to indicate the center of a distribution, describe its spread by giving the **quartiles.** The **first quartile** Q_1 has one-fourth of the observations below it, and the **third quartile** Q_3 has three-fourths of the observations below it. An extreme observation is an **outlier** if it is smaller than $Q_1 - (1.5 \times IQR)$ or larger than $Q_3 + (1.5 \times IQR)$.

The **five-number summary** consists of the median, the quartiles, and the high and low extremes and provides a quick overall description of a distribution. The median describes the center, and the quartiles and extremes show the spread.

Boxplots based on the five-number summary are useful for comparing two or more distributions. The box spans the quartiles and shows the spread of the central half of the distribution. The median is marked within the box. Lines extend from the box to the smallest and the largest observations that are not outliers. Outliers are plotted as isolated points.

The **variance** s^2 and especially its square root, **the standard deviation** s, are common measures of spread about the mean as center. The standard deviation s is zero when there is no spread and gets larger as the spread increases.

The mean and standard deviation are strongly influenced by outliers or skewness in a distribution. They are good descriptions for symmetric distributions and are most useful for the normal distributions, which will be introduced in the next chapter.

The median and quartiles are not affected by outliers, and the two quartiles and two extremes describe the two sides of a distribution separately. The five-number summary is the preferred numerical summary for skewed distributions.

When you add a constant a to all the values in a data set, the mean and median increase by a. Measures of spread do not change. When you multiply all the values in a data set by a constant b, the mean, median, IQR, and standard deviation are multiplied by b. These **linear transformations** are quite useful for changing units of measurement.

Back-to-back stemplots and **side-by-side boxplots** are useful for comparing quantitative distributions.

SECTION 1.2 EXERCISES

1.51 MEAT HOT DOGS Make a stemplot of the calories in meat hot dogs from Exercise 1.47 (page 59). What does this graph reveal that the boxplot of these data did not? *Lesson:* Be aware of the limitations of each graphical display.

1.52 EDUCATIONAL ATTAINMENT Table 1.12 shows the educational level achieved by U.S. adults aged 25 to 34 and by those aged 65 to 74. Compare the distributions of educational attainment graphically. Write a few sentences explaining what your display shows.

TABLE 1.12 Educational attainment by U.S. adults aged 25 to 34 and 65 to 74

	Number of people (thousands)	
	Ages 25–34	Ages 65–74
Less than high school	4474	4695
High school graduate	11,546	6649
Some college	7376	2528
Bachelor's degree	8563	1849
Advanced degree	3374	1266
Total	35,333	16,987

Source: Census Bureau, *Educational Attainment in the United States*, March 2000.

1.53 CASSETTE VERSUS CD SALES Has the increasing popularity of the compact disc (CD) affected sales of cassette tapes? Table 1.13 shows the number of cassettes and CDs sold from 1990 to 1999.

TABLE 1.13 Sales (in millions) of full-length cassettes and CDs, 1990–1999

	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Full-length cassettes	54.7	49.8	43.6	38.0	32.1	25.1	19.3	18.2	14.8	8.0
Full-length CDs	31.1	38.9	46.5	51.1	58.4	65.0	68.4	70.2	74.8	83.2

Source: The Recording Industry Association of America, *1999 Consumer Profile*.

Make a graphical display to compare cassette and CD sales. Write a few sentences describing what your graph tells you.

1.54 $\bar{x}$ AND s ARE NOT ENOUGH The mean $\bar{x}$ and standard deviation s measure center and spread but are not a complete description of a distribution. Data sets with different shapes can have the same mean and standard deviation. To demonstrate this fact, use your calculator to find $\bar{x}$ and s for the following two small data sets. Then make a stemplot of each and comment on the shape of each distribution.

Data A:	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
Data B:	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

1.55 In each of the following settings, give the values of a and b for the linear transformation $x_{new} = a + bx$ that expresses the change in measurement units. Then explain how the transformation will affect the mean, the *IQR*, the median, and the standard deviation of the original distribution.

(a) You collect data on the power of car engines, measured in horsepower. Your teacher requires you to convert the power to watts. One horsepower is 746 watts.

(b) You measure the temperature (in degrees Fahrenheit) of your school's swimming pool at 20 different locations within the pool. Your swim team coach wants the summary statistics in degrees Celsius ($^\circ$ F = (9/5)$^\circ$ C + 32).

(c) Dr. Data has given a very difficult statistics test and is thinking about "curving" the grades. She decides to add 10 points to each student's score.

1.56 A change of units that multiplies each unit by b, such as the change $x_{new} = 0 + 2.54x$ from inches x to centimeters x_{new}, multiplies our usual measures of spread by b. This is true of the *IQR* and standard deviation. What happens to the variance when we change units in this way?

1.57 BETTER CORN Corn is an important animal food. Normal corn lacks certain amino acids, which are building blocks for protein. Plant scientists have developed new corn varieties that have more of these amino acids. To test a new corn as an animal food, a group of 20 one-day-old male chicks was fed a ration containing the new corn. A control group of another 20 chicks was fed a ration that was identical except that it contained normal corn. Here are the weight gains (in grams) after 21 days:[16]

Normal corn				New corn			
380	321	366	356	361	447	401	375
283	349	402	462	434	403	393	426
356	410	329	399	406	318	467	407
350	384	316	272	427	420	477	392
345	455	360	431	430	339	410	326

(a) Compute five-number summaries for the weight gains of the two groups of chicks. Then make boxplots to compare the two distributions. What do the data show about the effect of the new corn?

(b) The researchers actually reported means and standard deviations for the two groups of chicks. What are they? How much larger is the mean weight gain of chicks fed the new corn?

(c) The weights are given in grams. There are 28.35 grams in an ounce. Use the results of part (b) to compute the means and standard deviations of the weight gains measured in ounces.

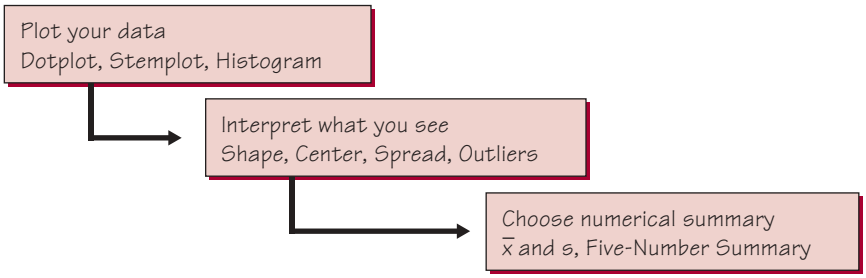
1.58 Which measure of center, the mean or the median, should you use in each of the following situations?

(a) Middletown is considering imposing an income tax on citizens. The city government wants to know the average income of citizens so that it can estimate the total tax base.

(b) In a study of the standard of living of typical families in Middletown, a sociologist estimates the average family income in that city.

CHAPTER REVIEW

Data analysis is the art of describing data using graphs and numerical summaries. The purpose of data analysis is to describe the most important features of a set of data. This chapter introduces data analysis by presenting statistical ideas and tools for describing the distribution of a single variable. The figure below will help you organize the big ideas.



Here is a review list of the most important skills you should have acquired from your study of this chapter.

A. DATA

1. Identify the individuals and variables in a set of data.

2. Identify each variable as categorical or quantitative. Identify the units in which each quantitative variable is measured.

B. DISPLAYING DISTRIBUTIONS

1. Make a bar graph and a pie chart of the distribution of a categorical variable. Interpret bar graphs and pie charts.

2. Make a dotplot of the distribution of a small set of observations.

3. Make a stemplot of the distribution of a quantitative variable. Round leaves or split stems as needed to make an effective stemplot.

4. Make a histogram of the distribution of a quantitative variable.

5. Construct and interpret an ogive of a set of quantitative data.

C. INSPECTING DISTRIBUTIONS (QUANTITATIVE VARIABLES)

1. Look for the overall pattern and for major deviations from the pattern.

2. Assess from a dotplot, stemplot, or histogram whether the shape of a distribution is roughly symmetric, distinctly skewed, or neither. Assess whether the distribution has one or more major peaks.

3. Describe the overall pattern by giving numerical measures of center and spread in addition to a verbal description of shape.

4. Decide which measures of center and spread are more appropriate: the mean and standard deviation (especially for symmetric distributions) or the five-number summary (especially for skewed distributions).

5. Recognize outliers.

D. TIME PLOTS

1. Make a time plot of data, with the time of each observation on the horizontal axis and the value of the observed variable on the vertical axis.

2. Recognize strong trends or other patterns in a time plot.

E. MEASURING CENTER

1. Find the mean $\bar{x}$ of a set of observations.

2. Find the median M of a set of observations.

3. Understand that the median is more resistant (less affected by extreme observations) than the mean. Recognize that skewness in a distribution moves the mean away from the median toward the long tail.

F. MEASURING SPREAD

1. Find the quartiles Q_1 and Q_3 for a set of observations.

2. Give the five-number summary and draw a boxplot; assess center, spread, symmetry, and skewness from a boxplot. Determine outliers.

3. Using a calculator, find the standard deviation s for a set of observations.

4. Know the basic properties of s: $s \geq 0$ always; $s = 0$ only when all observations are identical; s increases as the spread increases; s has the same units as the original measurements; s is increased by outliers or skewness.

G. CHANGING UNITS OF MEASUREMENT (LINEAR TRANSFORMATIONS)

1. Determine the effect of a linear transformation on measures of center and spread.

2. Describe a change in units of measurement in terms of a linear transformation of the form $x_{\text{new}} = a + bx$.

H. COMPARING DISTRIBUTIONS

1. Use side-by-side bar graphs to compare distributions of categorical data.

2. Make back-to-back stemplots and side-by-side boxplots to compare distributions of quantitative variables.

3. Write narrative comparisons of the shape, center, spread, and outliers for two or more quantitative distributions.

CHAPTER 1 REVIEW EXERCISES

1.59 Each year *Fortune* magazine lists the top 500 companies in the United States, ranked according to their total annual sales in dollars. Describe three other variables that could reasonably be used to measure the "size" of a company.

1.60 ATHLETES' SALARIES Here is a small part of a data set that describes major league baseball players as of opening day of the 1998 season:

Player	Team	Position	Age	Salary
⋮				
Perez, Eduardo	Reds	First base	28	300
Perez, Neifi	Rockies	Shortstop	23	210
Pettitte, Andy	Yankees	Pitcher	25	3750
Piazza, Mike	Dodgers	Catcher	29	8000
⋮				

(a) What individuals does this data set describe?

(b) In addition to the player's name, how many variables does the data set contain? Which of these variables are categorical and which are quantitative?

(c) Based on the data in the table, what do you think are the units of measurement for each of the quantitative variables?

1.61 HOW YOUNG PEOPLE DIE The number of deaths among persons aged 15 to 24 years in the United States in 1997 due to the seven leading causes of death for this age group were accidents, 12,958; homicide, 5793; suicide, 4146; cancer, 1583; heart disease, 1013; congenital defects, 383; AIDS, 276.[17]

(a) Make a bar graph to display these data.

(b) What additional information do you need to make a pie chart?

1.62 NEVER ON SUNDAY? The Canadian Province of Ontario carries out statistical studies of the working of Canada's national health care system in the province. The bar graphs in Figure 1.24 come from a study of admissions and discharges from community hospitals in Ontario.[18] They show the number of heart attack patients admitted and discharged on each day of the week during a 2-year period.

(a) Explain why you expect the number of patients admitted with heart attacks to be roughly the same for all days of the week. Do the data show that this is true?

(b) Describe how the distribution of the day on which patients are discharged from the hospital differs from that of the day on which they are admitted. What do you think explains the difference?

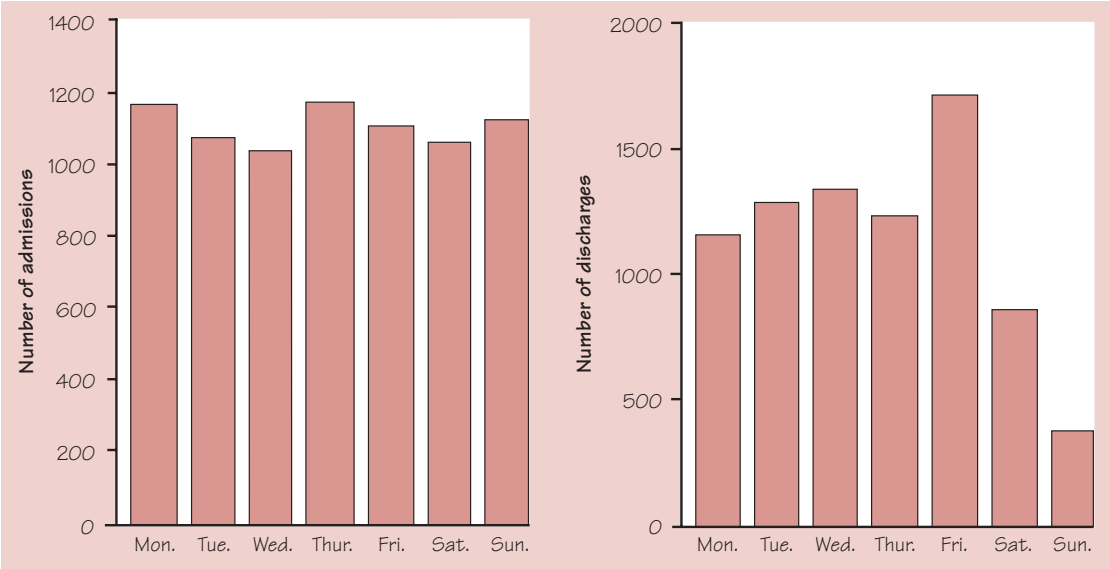


FIGURE 1.24 Bar graphs of the number of heart attack victims admitted and discharged on each day of the week by hospitals in Ontario, Canada.

1.63 PRESIDENTIAL ELECTIONS Here are the percents of the popular vote won by the successful candidate in each of the presidential elections from 1948 to 2000:

Year:	1948	1952	1956	1960	1964	1968	1972	1976	1980	1984	1988	1992	1996	2000
Percent:	49.6	55.1	57.4	49.7	61.1	43.4	60.7	50.1	50.7	58.8	53.9	43.2	49.2	47.9

(a) Make a stemplot of the winners' percents. (Round to whole numbers and use split stems.)

(b) What is the median percent of the vote won by the successful candidate in presidential elections? (Work with the unrounded data.)

(c) Call an election a landslide if the winner's percent falls at or above the third quartile. Find the third quartile. Which elections were landslides?

1.64 HURRICANES The histogram in Figure 1.25 (next page) shows the number of hurricanes reaching the east coast of the United States each year over a 70-year period.[19] Give a brief description of the overall shape of this distribution. About where does the center of the distribution lie?

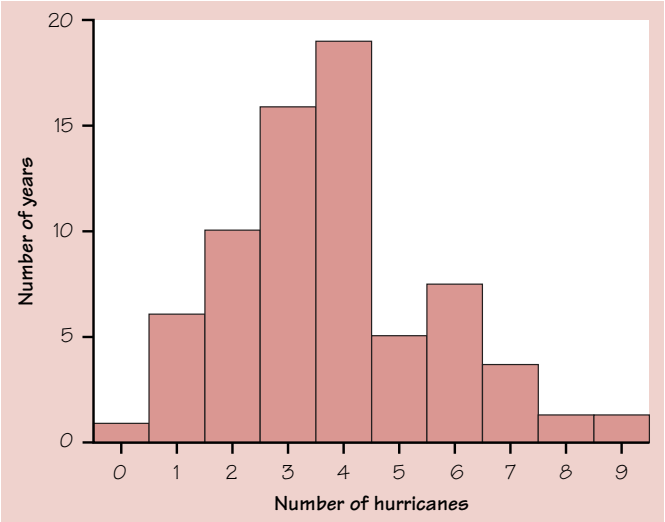


FIGURE 1.25 The distribution of the annual number of hurricanes on the U.S. east coast over a 70-year period, for Exercise 1.64.

1.65 DO SUVS WASTE GAS? Table 1.3 (page 17) gives the highway fuel consumption (in miles per gallon) for 32 model year 2000 midsize cars. We constructed a dotplot for these data in Exercise 1.8. Table 1.14 shows the highway mileages for 26 four-wheel-drive model year 2000 sport utility vehicles.

(a) Give a graphical and numerical description of highway fuel consumption for SUVs. What are the main features of the distribution?

(b) Make boxplots to compare the highway fuel consumption of midsize cars and SUVs. What are the most important differences between the two distributions?

TABLE 1.14 Highway gas mileages for model year 2000 four-wheel-drive SUVs

Model	MPG	Model	MPG
BMW X5	17	Kia Sportage	22
Chevrolet Blazer	20	Land Rover	17
Chevrolet Tahoe	18	Lexus LX470	16
Dodge Durango	18	Lincoln Navigator	17
Ford Expedition	18	Mazda MPV	19
Ford Explorer	20	Mercedes-Benz ML320	20
Honda Passport	20	Mitsubishi Montero	20
Infinity QX4	18	Nissan Pathfinder	19
Isuzu Amigo	19	Nissan Xterra	19
Isuzu Trooper	19	Subaru Forester	27
Jeep Cherokee	20	Suzuki Grand Vitara	20
Jeep Grand Cherokee	18	Toyota RAV4	26
Jeep Wrangler	19	Toyota 4Runner	21

1.66 DR. DATA RETURNS! Dr. Data asked her students how much time they spent using a computer during the previous week. Figure 1.26 is an ogive of her students' responses.

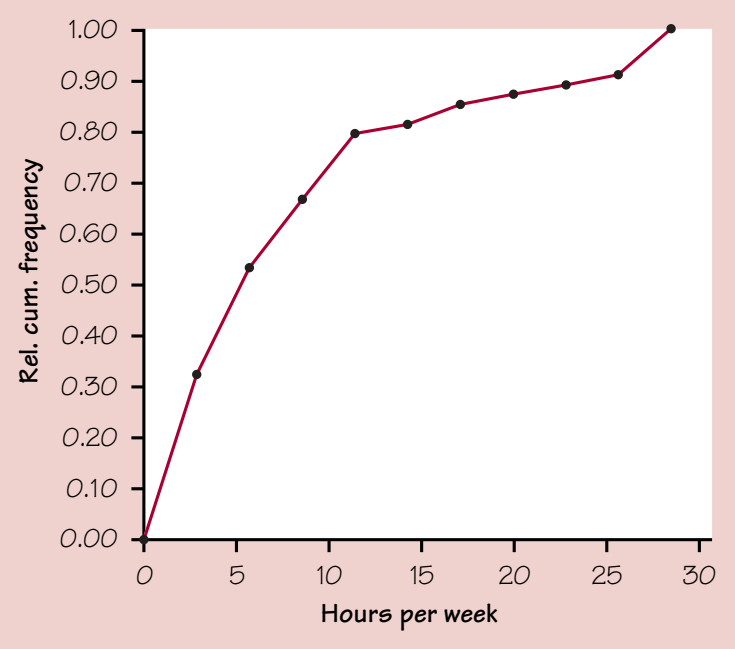


FIGURE 1.26 Ogive of weekly computer use by Dr. Data's statistics students.

(a) Construct a relative frequency table based on the ogive. Then make a histogram.

(b) Estimate the median, Q_1, and Q_3 from the ogive. Then make a boxplot. Are there any outliers?

(c) At what percentile does a student who used her computer for 10 hours last week fall?

1.67 WAL-MART STOCK The rate of return on a stock is its change in price plus any dividends paid. Rate of return is usually measured in percent of the starting value. We have data on the monthly rates of return for the stock of Wal-Mart stores for the years 1973 to 1991, the first 19 years Wal-Mart was listed on the New York Stock Exchange. There are 228 observations.

Figure 1.27 (next page) displays output from statistical software that describes the distribution of these data. The stems in the stemplot are the tens digits of the percent returns. The leaves are the ones digits. The stemplot uses split stems to give a better display. The software gives high and low outliers separately from the stemplot rather than spreading out the stemplot to include them.

(a) Give the five-number summary for monthly returns on Wal-Mart stock.

(b) Describe in words the main features of the distribution.

(c) If you had $1000 worth of Wal-Mart stock at the beginning of the best month during these 19 years, how much would your stock be worth at the end of the month? If you had $1000 worth of stock at the beginning of the worst month, how much would your stock be worth at the end of the month?

(d) Find the interquartile range (IQR) for the Wal-Mart data. Are there any outliers according to the 1.5 × IQR criterion? Does it appear to you that the software uses this criterion in choosing which observations to report separately as outliers?

```
Mean  =  3.064
Standard deviation  =  11.49

N  = 228   Median  =  3.4691
Quartiles  =  -2.950258,  8.4511

Decimal point is 1 place to the right of the colon

Low:  -34.04255  -31.25000  -27.06271  -26.61290

 -1 : 985
 -1 : 444443322222110000
 -0 : 999988777666666665555
 -0 : 44444444333333322222222222211111111100
  0 : 000001111111111112222223333333344444444
  0 : 55555555555555555555566666666666777777788888888888999999
  1 : 000000000111111111222333344444
  1 : 55566667889
  2 : 011334

High:  32.01923  41.80531  42.05607  57.89474  58.67769
```

FIGURE 1.27 Output from software describing the distribution of monthly returns from Wal-Mart stock.

1.68 A study of the size of jury awards in civil cases (such as injury, product liability, and medical malpractice) in Chicago showed that the median award was about $8000. But the mean award was about $69,000. Explain how this great difference between the two measures of center can occur.

1.69 You want to measure the average speed of vehicles on the interstate highway on which you are driving. You adjust your speed until the number of vehicles passing you equals the number you are passing. Have you found the mean speed or the median speed of vehicles on the highway?

TABLE 1.15 Data on education in the United States for Exercises 1.70 to 1.73

State	Region	Population (1000)	SAT Verbal	SAT Math	Percent taking	Percent no HS diploma	Teachers' pay ($1000)
AL	ESC	4,447	561	555	9	33.1	32.8
AK	PAC	627	516	514	50	13.4	51.7
AZ	MTN	5,131	524	525	34	21.3	34.4
AR	WSC	2,673	563	556	6	33.7	30.6
CA	PAC	33,871	497	514	49	23.8	43.7

TABLE 1.15 Data on education in the United States, for Exercises 1.70 to 1.73 (*continued*)

State	Region	Population (1000)	SAT Verbal	SAT Math	Percent taking	Percent no HS diploma	Teachers' pay ($1000)
CO	MTN	4,301	536	540	32	15.6	37.1
CT	NE	3,406	510	509	80	20.8	50.7
DE	SA	784	503	497	67	22.5	42.4
DC	SA	572	494	478	77	26.9	46.4
FL	SA	15,982	499	498	53	25.6	34.5
GA	SA	8,186	487	482	63	29.1	37.4
HI	PAC	1,212	482	513	52	19.9	38.4
ID	MTN	1,294	542	540	16	20.3	32.8
IL	ENC	12,419	569	585	12	23.8	43.9
IN	ENC	6,080	496	498	60	24.4	39.7
IA	WNC	2,926	594	598	5	19.9	34.0
KS	WNC	2,688	578	576	9	18.7	36.8
KY	ESC	4,042	547	547	12	35.4	34.5
LA	WSC	4,469	561	558	8	31.7	29.7
ME	NE	1,275	507	503	68	21.2	34.3
MD	SA	5,296	507	507	65	21.6	41.7
MA	NE	6,349	511	511	78	20.0	43.9
MI	ENC	9,938	557	565	11	23.2	49.3
MN	WNC	4,919	586	598	9	17.6	39.1
MS	ESC	2,845	563	548	4	35.7	29.5
MO	WNC	5,595	572	572	8	26.1	34.0
MT	MTN	902	545	546	21	19.0	30.6
NE	WNC	1,711	568	571	8	18.2	32.7
NV	MTN	1,998	512	517	34	21.2	37.1
NH	NE	1,236	520	518	72	17.8	36.6
NJ	MA	8,414	498	510	80	23.3	50.4
NM	MTN	1,819	549	542	12	24.9	30.2
NY	MA	18,976	495	502	76	25.2	49.0
NC	SA	8,049	493	493	61	30.0	33.3
ND	WNC	642	594	605	5	23.3	28.2
OH	ENC	11,353	534	568	25	24.3	39.0
OK	WSC	3,451	567	560	8	25.4	30.6
OR	PAC	3,421	525	525	53	18.5	42.2
PA	MA	12,281	498	495	70	25.3	47.7
RI	NE	1,048	504	499	70	28.0	44.3
SC	SA	4,012	479	475	61	31.7	33.6
SD	WNC	755	585	588	4	22.9	27.3
TN	ESC	5,689	559	553	13	32.9	35.3
TX	WSC	20,852	494	499	50	27.9	33.6
UT	MTN	2,233	570	568	5	14.9	33.0
VT	NE	609	514	506	70	19.2	36.3
VA	SA	7,079	508	499	65	24.8	36.7
WA	PAC	5,894	525	526	52	16.2	38.8
WV	SA	1,808	527	512	18	34.0	33.4
WI	ENC	5,364	584	595	7	21.4	39.9
WY	MTN	494	546	551	10	17.0	32.0

Source: U.S. Census Bureau Web site, http://www.census.gov, 2001.

Table 1.15 presents data about the individual states that relate to education. Study of a data set with many variables begins by examining each variable by itself. Exercises 1.70 to 1.73 concern the data in Table 1.15.

1.70 POPULATION OF THE STATES Make a graphical display of the population of the states. Briefly describe the shape, center, and spread of the distribution of population. Explain why the shape of the distribution is not surprising. Are there any states that you consider outliers?

1.71 HOW MANY STUDENTS TAKE THE SAT? Make a stemplot of the distribution of the percent of high school seniors who take the SAT in the various states. Briefly describe the overall shape of the distribution. Find the midpoint of the data and mark this value on your stemplot. Explain why describing the center is not very useful for a distribution with this shape.

1.72 HOW MUCH ARE TEACHERS PAID? Make a graph to display the distribution of average teachers' salaries for the states. Is there a clear overall pattern? Are there any outliers or other notable deviations from the pattern?

1.73 PEOPLE WITHOUT HIGH SCHOOL EDUCATIONS The "Percent no HS" column gives the percent of the adult population in each state who did not graduate from high school. We want to compare the percents of people without a high school education in the northeastern and the southern states. Take the northeastern states to be those in the MA (Mid-Atlantic) and NE (New England) regions. The southern states are those in the SA (South Atlantic) and ESC (East South Central) regions. Leave out the District of Columbia, which is a city rather than a state.

(a) List the percents without high school for the northeastern and for the southern states from Table 1.15. These are the two data sets we want to compare.

(b) Make numerical summaries and graphs to compare the two distributions. Write a brief statement of what you find.

NOTES AND DATA SOURCES

1. Data from *Beverage Digest*, February 18, 2000.
2. Seat-belt data from the National Highway and Traffic Safety Administration, *NOPUS Survey*, 1998.
3. Data from the 1997 *Statistical Abstract of the United States.*
4. Data on accidental deaths from the Centers for Disease Control Web site, www.cdc.gov.
5. Data from the *Los Angeles Times*, February 16, 2001.
6. Based on experiments performed by G. T. Lloyd and E. H. Ramshaw of the CSIRO Division of Food Research, Victoria, Australia, 1982–83.
7. Maribeth Cassidy Schmitt, from her Ph.D. dissertation, "The effects of an elaborated directed reading activity on the metacomprehension skills of third graders," Purdue University, 1987.
8. Data from "America's best small companies," *Forbes*, November 8, 1993.
9. The Shakespeare data appear in C. B. Williams, *Style and Vocabulary: Numerological Studies*, Griffin, London, 1970.

10. Data from John K. Ford, "Diversification: how many stocks will suffice?" *American Association of Individual Investors Journal*, January 1990, pp. 14–16.
11. Data on frosts from C. E. Brooks and N. Carruthers, *Handbook of Statistical Methods in Meteorology*, Her Majesty's Stationery Office, London, 1953.
12. These data were collected by students as a class project.
13. Data from S. M. Stigler, "Do robust estimators work with real data?" *Annals of Statistics*, 5 (1977), pp. 1055–1078.
14. Data obtained from The College Board.
15. Basketball scores from the *Los Angeles Times*, February 16, 2001.
16. Based on summaries in G. L. Cromwell et al., "A comparison of the nutritive value of *opaque*-2, *floury*-2, and normal corn for the chick," *Poultry Science*, 57 (1968), pp. 840–847.
17. Centers for Disease Control and Prevention, *Births and Deaths: Preliminary Data for 1997*, Monthly Vital Statistics Reports, 47, No. 4, 1998.
18. Based on Antoni Basinski, "Almost never on Sunday: implications of the patterns of admission and discharge for common conditions," Institute for Clinical Evaluative Sciences in Ontario, October 18, 1993.
19. Hurricane data from H. C. S. Thom, *Some Methods of Climatological Analysis*, World Meteorological Organization, Geneva, Switzerland, 1966.