2.1 Visualizing Distributions: Shape, Center, and Spread

Summaries simplify. In fact, summaries sometimes can oversimplify, which means it is important to know when to use summaries and which summaries to use. Often the right choice depends on the shape of your distribution. To help you build your visual intuition about how shape and summaries are related, this first section of the chapter introduces various shapes and asks you to estimate some summary values visually. Later sections will tell you how to compute summary values numerically.

Distributions come in a variety of shapes. Four of the most common shapes are illustrated in the rest of this section.

The uniform distribution is rectangular.

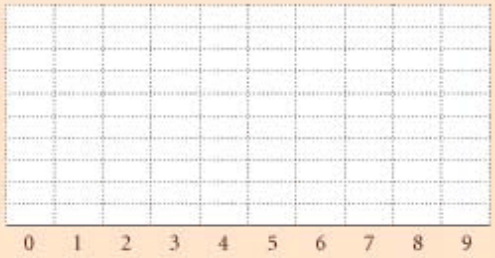
Uniform (Rectangular) Distributions

The plot to the left shows the shape of a *uniform or rectangular distribution*, in which all values occur equally often. How uniform is a sample of values taken from a uniform distribution? In the next activity, you will find out.

ACTIVITY 2.1a Distributing Digits

What you'll need: one page from a phone book for each member of the class, and a box of slips of paper, with one slip for each member of your class, half labeled "phone book" and half "fake it"

1. Suppose your class made a dot plot of the last digits of every phone number in the phone book. (This would take a very long time!) Sketch what you think this plot would look like.
2. Draw a slip of paper from the box.
   - If the slip you drew says "phone book," use the page from the phone book, start at a random spot, and write down the last digit of each of the next 30 phone numbers. Using a full sheet of paper, plot your 30 digits on a dot plot, using a scale like the one here. Use big dots so that they can be seen from across the room.



   - If the slip you drew says "fake it," don't use the page from the phone book but instead make up and plot 30 digits on a dot plot using a scale

(Continued)

like the one on the previous page. Try to make the distribution look like the digits might have come from the phone book.

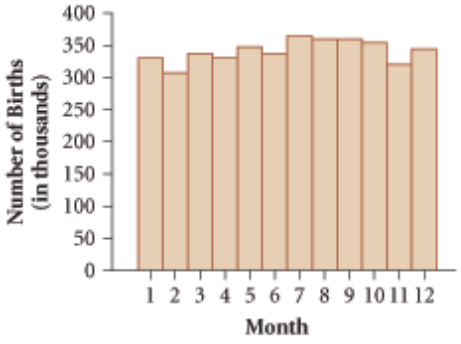
Write your name on the front of the plot, but not which method you used. Don't consult with other students while you are doing this step.

3. Post the dot plots around the room and compare. Which plots are you confident came from the phone book? Which are you confident came from made-up digits? (Don't say anything about your own plot.)
4. Find your plot and write a large P (for "phone book") or F (for "faked it") on the front. Check your predictions from step 3. What differences do you see in the two groups of plots?
5. In this activity, it was important that you sampled from the phone book in such a way that all digits were equally likely to occur. Why did step 1 specify that you use the last digit of the phone number and not, say, the first?

The number of births per month in a year is another set of data you might expect to be fairly uniform. Or, is there a reason to believe that more babies are born in one month than in another? Display 2.1 shows a table and plot of U.S. births (in thousands) for 2003.

Data Set

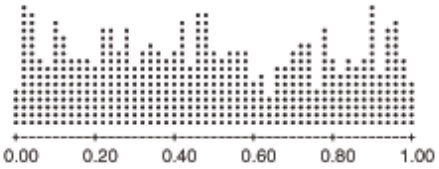
Month	Births (in thousands)
1	330
2	307
3	337
4	330
5	347
6	337
7	364
8	360
9	360
10	354
11	320
12	344



Display 2.1 An example of a (roughly) uniform distribution: births per month in the United States, 2003. [*Source:* Centers for Disease Control and Prevention.]

The plot shows that there is actually little change from month to month; that is, we see a roughly uniform distribution of births across the months. To summarize this distribution, you might write "The distribution of births is roughly uniform over the months January through December, with about 340,000 births per month."

Computers and many calculators generate random numbers between 0 and 1 with a uniform distribution. Display 2.2 shows a dot plot of 1000 random numbers generated by statistical software. There is some variability in the frequencies, but, as expected, about 20% of the random numbers fall between, for example, 0.2 and 0.4. [See **Calculator Note 2A** to learn how to create a distribution of random numbers using your calculator.]



Display 2.2 Dot plot of 1000 random numbers from a uniform distribution. Each dot represents two points.

DISCUSSION

Uniform Distributions

D1. Think of other situations that you would expect to be uniform distributions
   a. over the days of the week
   b. over the digits 0, 1, 2, . . . , 9

D2. Think of situations that you would expect to be very nonuniform distributions
   a. over the months of the year
   b. over the days of the month
   c. over the digits 0, 1, 2, . . . , 9
   d. over the days of the week

Normal Distributions

Activity 2.1b introduces one of the most important common shapes of distributions and one of the common ways this shape is produced. What happens when different people measure the same distance or the same feature of very similar objects? In the activity, you'll measure a tennis ball with a ruler, but the results you get will reflect what happens even if you use very precise instruments under carefully controlled conditions. For example, a 10-gram platinum weight is used for calibration of scales all across the United States. When scientists at the National Institute of Standards and Technology use an analytical balance for the weight's weekly weighing, they face a similar challenge due to variability.

ACTIVITY 2.1b

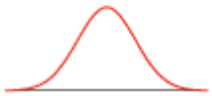
Measuring Diameters

What you'll need: a tennis ball, a ruler with a centimeter scale

1. With your partner, plan a method for measuring the diameter of the tennis ball with the centimeter ruler.

(continued)

2. Using your method, make two measurements of the diameter of your tennis ball to the nearest millimeter.
3. Combine your data with those of the rest of the class and make a dot plot. Speculate first, about the shape you expect for the distribution.
4. *Shape.* What is the approximate shape of the plot? Are there clusters and gaps or unusual values (outliers) in the data?
5. *Center and spread.* Choose two numbers that seem reasonable for completing this sentence: "Our typical diameter measurement is about —?—, give or take about —?—." (More than one reasonable set of choices is possible.)
6. Discuss some possible reasons for the variability in the measurements. How could the variability be reduced? Can the variability be eliminated entirely?

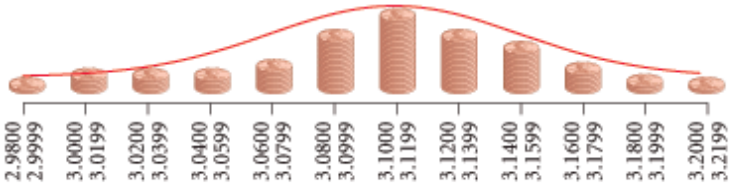


The normal distribution is bell-shaped..

The measurements of the diameter of a tennis ball taken by your class in Activity 2.1b probably were not uniform. More likely, they piled up around some central value, with a few measurements far away on the low side and a few far away on the high side. This common bell shape has an idealized version —the **normal distribution**—that is especially important in statistics.

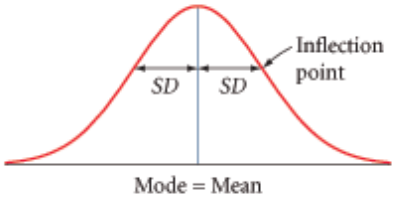
Pennies minted in the United States are supposed to weigh 3.110 g, but a tolerance of 0.130 g is allowed in either direction. Display 2.3 shows a plot of the weights of 100 pennies.

Data Set



Display 2.3 Weights of pennies. [*Source:* W. J. Youden, *Experimentation and Measurement* (National Science Teachers Association, 1985), p. 108.]

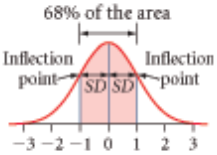
The smooth curve superimposed on the graph of the pennies is an example of a normal curve. No real-world example matches the curve perfectly, but many plots of data are approximately normal. The idealized normal shape is perfectly symmetric—the right side is a mirror image of the left side, as in Display 2.4. There is a single peak, or **mode**, at the line of symmetry, and the curve drops off smoothly on both sides, flattening toward the x-axis but never quite reaching it and stretching infinitely far in both directions. On either side of the mode are inflection points, where the curve changes from concave down to concave up.



Display 2.4 A normal curve, showing the line of symmetry, mode, mean, inflection points, and standard deviation (*SD*).

Use the mean and standard deviation to describe the center and spread of a normal distribution.

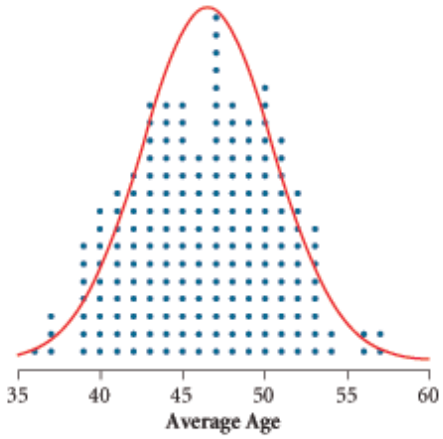
You should use the **mean** (or average) to describe the center of a normal distribution. The mean is the value at the point where the line of symmetry intersects the *x*-axis. You should use the **standard deviation,** *SD* for short, to describe the spread. The *SD* is the horizontal distance from the mean to an inflection point.



It is difficult to locate inflection points, especially when curves are drawn by hand. A more reliable way to estimate the standard deviation is to use areas. For a normal curve, 68% (roughly) of the total area under the curve is between the vertical lines through the two inflection points. In other words, the interval between one standard deviation on either side of the mean accounts for roughly 68% of the area under the normal curve.

Example: Averages of Random Samples

Display 2.5 shows the distribution of average ages computed from 200 sets of five workers chosen at random from the ten hourly workers in Round 2 of the Westvaco case discussed in Chapter 1. Notice that, apart from the bumpiness, the shape is roughly normal. Estimate the mean and standard deviation.



Display 2.5 Distribution of average age for groups of five workers drawn at random.

Solution

The curve in the display has center at 47, and the middle 68% of dots fall roughly between 43 and 51. Thus, the estimated mean is 47, and the estimated standard deviation is 4. A typical random sample of five workers has average age 47 years, give or take about 4 years.

[You can graph a normal curve on your calculator by specifying the mean and standard deviation. See **Calculator Note 2B**.]

In this section, you've seen the three most common ways normal distributions arise in practice:

- through variation in measurements (diameters of tennis balls)
- through natural variation in populations (weights of pennies)
- through variation in averages computed from random samples (average ages)

All three scenarios are common, which makes the normal distribution especially important in statistics.

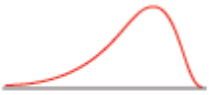
DISCUSSION

Normal Distributions

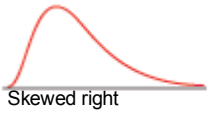
D3. Determine these summary statistics visually.
   a. Estimate the mean and standard deviation of the penny weight data in Display 2.3, and use your estimates to write a summary sentence.
   b. Estimate the mean and standard deviation of your class data from Activity 2.1b.

Skewed Distributions

Both the uniform (rectangular) and normal distributions are symmetric. That is, if you smooth out minor bumps, the right side of the plot is a mirror image of the left side. Not all distributions are symmetric, however. Many common distributions show bunching at one end and a long tail stretching out in the other direction. These distributions are called **skewed**. The direction of the tail tells whether the distribution is **skewed right** (tail stretches right, toward the high values) or **skewed left** (tail stretches left , toward the low values).



Skewed left



Skewed right

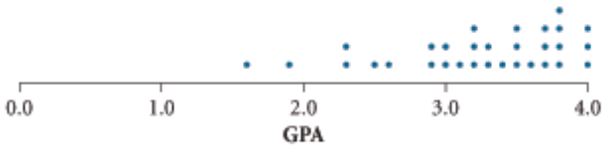
Data Set

Display 2.6 Weights of bears in pounds. [*Source:* MINITAB data set from *MINITAB Handbook,* 3rd ed.]

The dot plot in Display 2.6 shows the weights, in pounds, of 143 wild bears. It is skewed right (toward the higher values) because the tail of the distribution stretches out in that direction. In everyday conversation, you might describe the two parts of

the distribution as "normal" and "abnormal." Usually, bears weigh between about 50 and 250 lb (this part of the distribution even looks approximately normal), but if someone shouts "Abnormal bear loose!" you should run for cover—that unusual bear is likely to be big! The "unusualness" of the distribution is all in one direction.

Often the bunching in a skewed distribution happens because values "bump up against a wall"—either a minimum that values can't go below, such as 0 for measurements and counts, or a maximum that values can't go above, such as 100 for percentages. For example, the distribution in Display 2.7 shows the grade-point averages of college students (mostly first-year students and sophomores) taking an introductory statistics course at the University of Florida. It is skewed left (toward the smaller values). The maximum grade-point average is 4.0, for all A's, so the distribution is bunched at the high end because of this wall. A GPA of 0.0 wouldn't be called a wall, even though GPAs can't go below 0.0, because the values aren't bunched up against it. The skew is to the left : An unusual GPA would be one that is low compared to most GPAs of students in the class.



Display 2.7 Grade-point averages of 62 statistics students. Each dot represents two points.

Use the median along with the lower and upper quartiles to describe the center and spread of a skewed distribution.

Because there is no line of symmetry in a skewed distribution, the ideas of center and spread are not as clear-cut as they are for a normal distribution. To get around this problem, typically you should use the **median** to describe the center of a skewed distribution. To estimate the median from a dot plot, locate the value that divides the dots into two halves, with equal numbers of dots on either side.

You should use the lower and upper quartiles to indicate spread. The **lower quartile** is the value that divides the lower half of the distribution into two halves, with equal numbers of dots on either side. The **upper quartile** is the value that divides the upper half of the distribution into two halves, with equal numbers of dots on either side. The three values—lower quartile, median, and upper quartile—divide the distribution into quarters. This allows you to describe a distribution as in the introduction to this chapter: "The middle 50% of the SAT math scores were between 630 and 720, with half above 680 and half below."

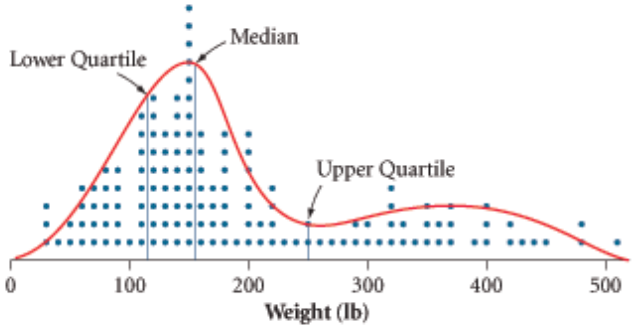
Example: Median and Quartiles for Bear Weights

Divide the bears' weights in Display 2.8 into four groups of equal size, and estimate the median and quartiles. Write a short summary of this distribution.

Solution

There are 143 dots in Display 2.8, so there are about 71 or 72 dots in each half and 35 or 36 in each quarter. The value that divides the dots in half is about 155 lb. The values that divide the two halves in half are roughly 115 and 250. Thus, the middle 50% of the bear weights are between about 115 and 250, with half above about 155 and half below.

Data Set



Display 2.8 Estimating center and spread for the weights of bears.

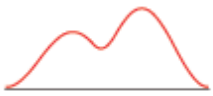
DISCUSSION

Skewed Distributions

D4. Decide whether each distribution described will be skewed. Is there a wall that leads to bunching near it and a long tail stretching out away from it? If so, describe the wall.
   a. the sizes of islands in the Caribbean
   b. the average per capita incomes for the nations of the United Nations
   c. the lengths of pant legs cut and sewn to be 32 in. long
   d. the times for 300 university students of introductory psychology to complete a 1-hour timed exam
   e. the lengths of reigns of Japanese emperors

D5. Which would you expect to be the more common direction of skew, right or left ? Why?

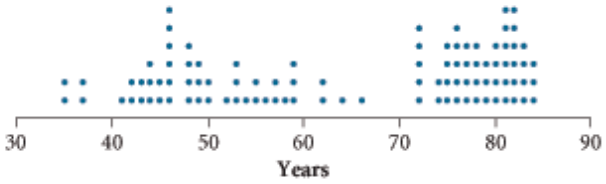
Bimodal Distributions



A bimodal distribution has two peaks.

Many distributions, including the normal distribution and many skewed distributions, have only one peak (**unimodal**), but some have two peaks (**bimodal**) or even more. When your distribution has two or more obvious peaks, or modes, it is worth asking whether your cases represent two or more groups. For example, Display 2.9 shows the life expectancies of females from countries on two continents, Europe and Africa.

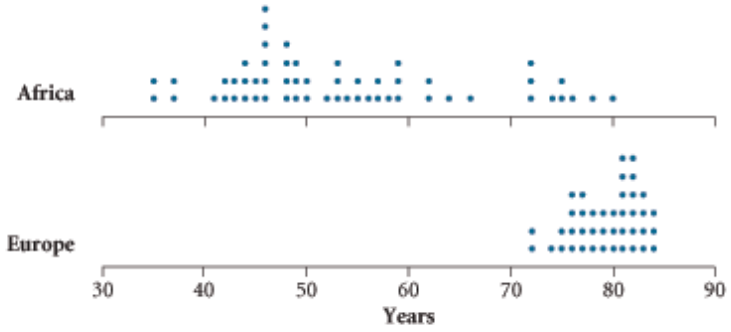
Data Set



Display 2.9 Life expectancy of females by country on two continents. [*Source:* Population Reference Bureau, *World Population Data Sheet*, 2005.]

Europe and Africa differ greatly in their socioeconomic conditions, and the life expectancies reflect those conditions. If you make a separate plot for each of the two continents, the two peaks become essentially one peak in each plot, as in Display 2.10.

Data Set



Display 2.10 Life expectancy of females in Africa and Europe.

Although it makes sense to talk about the center of the distribution of life expectancies for Europe or for Africa, notice that it doesn't really make sense to talk about "the" center of the distribution for both continents together. You could possibly tell the locations of the two peaks, but finding the reason for the two modes and separating the cases into two distributions communicates even more.

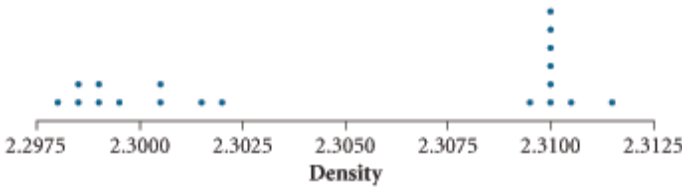
Other Features: Outliers, Gaps, and Clusters

An unusual value, or **outlier**, is a value that stands apart from the bulk of the data. Outliers always deserve special attention. Sometimes they are mistakes (a typing mistake, a measuring mistake), sometimes they are atypical for other reasons (a really big bear, a faulty lab procedure), and sometimes unusual features of the distribution are the key to an important discovery.


Lord Rayleigh

In the late 1800s, John William Strutt, third Baron Rayleigh (English, 1842–1919), was studying the density of nitrogen using samples from the air outside his laboratory (from which known impurities were removed) and samples produced by a chemical procedure in his lab. He saw a pattern in the results that you can observe in the plot of his data in Display 2.11.

Data Set



Display 2.11 Lord Rayleigh's densities of nitrogen. [*Source: Proceedings of the Royal Society* 55 (1894).]

Lord Rayleigh saw two clusters separated by a gap. (There is no formal definition of a **gap** or a **cluster**; you have to use your best judgment about them. For example, some people call a single outlier a cluster of one; others don't. You

also could argue that the value at the extreme right is an outlier, perhaps because of a faulty measurement.)

When Rayleigh checked the clusters, it turned out that the ten values to the left had all come from the chemically produced samples and the nine to the right had all come from the atmospheric samples. What did this great scientist conclude? The air samples on the right might be denser because of something in them besides nitrogen. This hypothesis led him to discover inert gases in the atmosphere.

Summary 2.1: Visualizing Distributions

Distributions have different shapes, and different shapes call for different summaries.

- If your distribution is uniform (rectangular), it's often enough simply to tell the range of the set of values and the approximate frequency with which each value occurs.
- If your distribution is normal (bell-shaped), you can give a good summary with the mean and the standard deviation. The mean lies at the center of the distribution, and the standard deviation (*SD*) is the horizontal distance from the center to the points of inflection, where the curvature changes. To estimate the *SD*, find the distance on either side of the mean that defines the interval enclosing about 68% of the cases.
- If your distribution is skewed, you can give the values (median and quartiles) that divide the distribution into fourths.
- If your distribution is bimodal, reporting a single center isn't useful. One reasonable summary is to locate the two peaks. However, it is even more useful if you can find another variable that divides your set of cases into two groups centered at the two peaks.

Later in this chapter, you will study the various measures of center and spread in more detail and learn how to compute them.

Practice

Practice problems help you master basic concepts and computations. Throughout this textbook, you should work *all* the practice problems for each topic you want to learn. The answers to all practice problems are given in the back of the book.

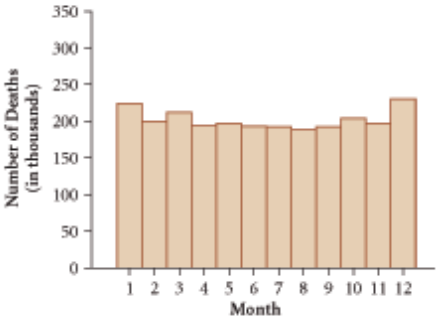
Uniform (Rectangular) Distributions

P1. This diagram shows a uniform distribution on [0, 2], the interval from 0 through 2.

a. What value divides the distribution in half, with half the numbers below that value and half above?
b. What values divide the distribution into quarters?
c. What values enclose the middle 50% of the distribution?
d. What percentage of the values lie between 0.4 and 0.7?
e. What values enclose the middle 95% of the distribution?

P2. The plot in Display 2.12 gives the number of deaths in the United States per month in 2003. Does the number of deaths appear to be uniformly distributed over the months? Give a verbal summary of the way deaths are distributed over the months of the year.

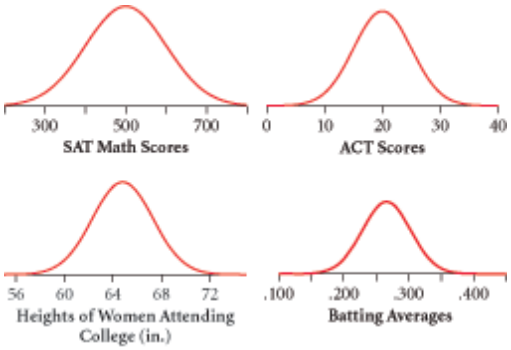
Data Set



Display 2.12 Deaths per month, 2003. [*Source:* Centers for Disease Control and Prevention, *National Vital Statistics Report*, 2004.]

Normal Distributions

P3. For each of the normal distributions in Display 2.13, estimate the mean and standard deviation visually, and use your estimates to write a verbal summary of the form "A typical SAT score is roughly (mean), give or take (*SD*) or so."



Display 2.13 Four distributions that are approximately normal.

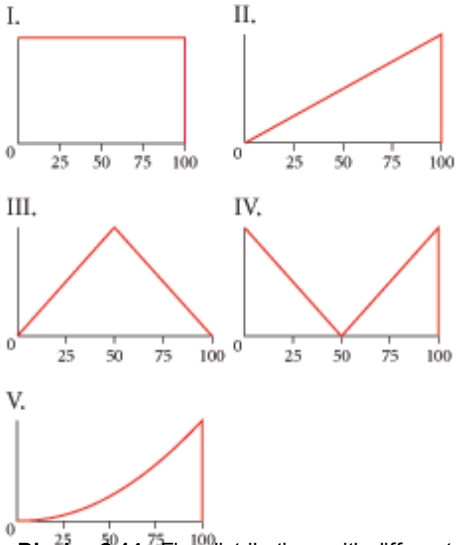
a. SAT math scores
b. ACT scores
c. heights of women attending college
d. single-season batting averages for professional baseball players in the 1910s

Skewed Distributions

P4. Estimate the median and quartiles for the distribution of GPAs in Display 2.7 on page 34. Then write a verbal summary of the same form as in the example.

P5. Match each plot in Display 2.14 with its median and quartiles (the set of values that divide the area under the curve into fourths).

a. 15, 50, 85
b. 50, 71, 87
c. 63, 79, 91
d. 35, 50, 65
e. 25, 50, 75



Display 2.14 Five distributions with different shapes.

Exercises

E1. Describe each distribution as bimodal, skewed right, skewed left , approximately normal, or roughly uniform.
   a. ages of all people who died last year in the United States
   b. ages of all people who got their first driver's license in your state last year
   c. SAT scores for all students in your state taking the test this year
   d. selling prices of all cars sold by General Motors this year

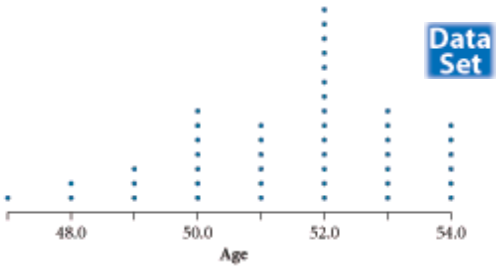
E2. Describe each distribution as bimodal, skewed right, skewed left , approximately normal, or roughly uniform.
   a. the incomes of the world's 100 richest people
   b. the birthrates of Africa and Europe
   c. the heights of soccer players on the last Women's World Cup championship team
   d. the last two digits of telephone numbers in the town where you live
   e. the length of time students used to complete a chapter test, out of a 50-minute class period

The 2003 Women's World Cup Championship team, from Germany

E3. Sketch these distributions.
   a. a uniform distribution that shows the sort of data you would get from rolling a fair die 6000 times
   b. a roughly normal distribution with mean 15 and standard deviation 5
   c. a distribution that is skewed left , with half its values above 20 and half below and with the middle 50% of its values between 10 and 25
   d. a distribution that is skewed right, with the middle 50% of its values between 100 and 1000 and with half the values above 200 and half below

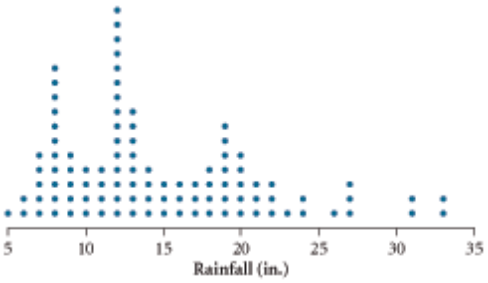
E4. The U.S. Environmental Protection Agency's *National Priorities List Fact Book* tells the number of hazardous waste sites for each U.S. state and territory. For 2006, the numbers ranged from 1 to 138, the middle 50% of the values were between 11 and 32, half the values were above 18, and half were below 18. Sketch what the distribution might look like. [*Source:* U.S. Environmental Protection Agency, www.epa.gov, 2006.]

E5. The dot plot in Display 2.15 gives the ages of the officers who attained the rank of colonel in the Royal Netherlands Air Force.
   a. What are the cases? Describe the variables.
   b. Describe this distribution in terms of shape, center, and spread.
   c. What kind of wall might there be that causes the shape of the distribution? Generate as many possibilities as you can.



Display 2.15 Ages of colonels. Each dot represents two points. [*Source:* Data and Story Library at Carnegie-Mellon University, lib.stat.cmu.edu.]

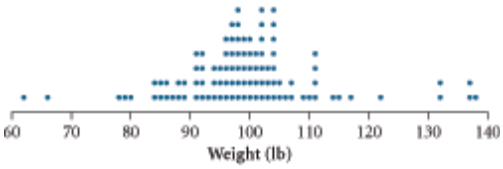
E6. The dot plot in Display 2.16 shows the distribution of the number of inches of rainfall in Los Angeles for the seasons 1899–1900 through 1999–2000.



Display 2.16 Los Angeles rainfall. [*Source:* National Weather Service.]

a. What are the cases? Describe the variables.
b. Describe this distribution in terms of shape, center, and spread.
c. What kind of wall might there be that causes the shape of the distribution? Generate as many possibilities as you can.

E7. The distribution in Display 2.17 shows measurements of the strength in pounds of 22s yarn (22s refers to a standard unit for measuring yarn strength). What is the basic shape of this distribution? What feature makes it uncharacteristic of distributions with that shape?



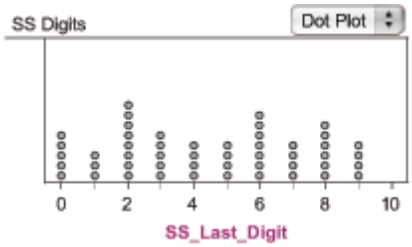
Display 2.17 Strength of yarn. [*Source:* Data and Story Library at Carnegie-Mellon University, lib.stat.cmu.edu.]

E8. Sketch a normal distribution with mean 0 and standard deviation 1. You will study this *standard normal distribution* in Section 2.5.

E9. Make up a scenario (name the cases and variables) whose distribution you would expect to be

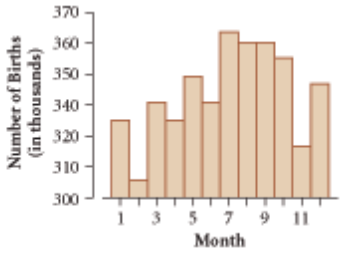
a. skewed right because of a wall. What is responsible for the wall?
b. skewed left because of a wall. What is responsible for the wall?

E10. The plot in Display 2.18 shows the last digit of the Social Security numbers of the students in a statistics class. Describe this distribution.



Display 2.18 Last digit of a sample of Social Security numbers.

E11. Although a uniform distribution gives a reasonable approximation of the actual distribution of births over months (Display 2.1 on page 29), you can "blow up" the graph to see departures from the uniform pattern, as in Display 2.19. Do these deviations from the uniform shape form their own pattern, or do they appear haphazard? If you think there's a pattern, describe it.



Display 2.19 A "blow up" of the distribution of births over months, showing departures from the uniform pattern.

Data Set

E12. Draw a graph similar to that in Display 2.19 for the data on deaths in the United States listed in Display 2.20, and summarize what you find.

Month	Deaths (in thousands)
1	224
2	200
3	212
4	194
5	197
6	193
7	192
8	188
9	192
10	204
11	197
12	230

Display 2.20 Deaths in the United States, 2003. [*Source:* Centers for Disease Control and Prevention.]

Data Set

E13. How do countries compare with respect to the value of the goods they produce? Display 2.21 shows gross domestic product (GDP) per capita, a measure of the total value of all goods and services produced divided by the number of people in a country, and the average number of people per room in housing units, a measure of crowdedness, for a selection of countries in Asia, Europe, and North America. You'll analyze these data in parts a–d.

A family in Albania

Country	Per Capita GDP (in U.S. dollars)	Average Number of People per Room
Austria	31,187	0.7
Azerbaijan	853	2.1
Belgium	29,257	0.6
Bulgaria	2,533	1.0
Canada	27,097	0.5
China	1,100	1.1
Croatia	6,398	1.2
Cyprus	16,038	0.6
Czech Republic	8,834	1.0
Finland	31,069	0.8
France	29,222	0.7
Germany	29,137	0.5
Hungary	8,384	0.8
India	555	2.7
Iraq	594	1.5
Israel	18,101	1.2
Japan	33,819	0.8
Korea, Republic of	11,059	1.1
Kuwait	13,641	1.7
Netherlands	31,759	0.7
Norway	48,881	0.6
Pakistan	498	3.0
Poland	5,355	1.0
Portugal	14,645	0.7
Romania	2,550	1.3
Serbia-Montenegro	1,843	1.2
Slovakia	6,019	1.2
Sri Lanka	913	2.2
Sweden	33,925	0.5
Switzerland	43,486	0.6
Syria	1,497	2.0
Turkey	3,418	1.3
United Kingdom	30,355	0.5
United States	36,924	0.5

Display 2.21 Per capita GDP and crowdedness for a selection of countries. [*Source:* United Nations, unstats.un.org.]

A dot plot of the per capita GDP data is shown in Display 2.22.

Data Set

Display 2.22 Dot plot of per capita GDP.

a. Describe this distribution in terms of shape, center, and spread.

b. Which two countries have the highest per capita GDP? Do they appear to be outliers?

c. A rather large gap appears near the middle of the distribution. Which of the two clusters formed by this gap contains mostly Western European and North American countries? In what part of the world are most of the countries in the other cluster?

d. Is it surprising to find clusters and gaps in data that measure an aspect of the economies of the countries?

E14. The dot plot in Display 2.23 gives a look at how the countries listed in Display 2.21 compare in terms of the crowdedness of their residents.

Data Set

Display 2.23 Dot plot of crowdedness.

a. Describe this distribution in terms of shape, center, and spread.

b. Which countries appear to be outliers? Are they the same as the countries that appeared to be outliers for the per capita GDP data?

c. Where on the dot plot is the cluster that contains mostly Western European and North American countries?

2.2 Graphical Displays of Distributions

Plots should present the essentials quickly and clearly.

As you saw in the previous section, the best way to summarize a distribution often depends on its shape. To see the shape, you need a suitable graph. In this section, you'll learn how to make and interpret three kinds of plots for quantitative variables (dot plot, histogram, and stemplot) and one plot for categorical variables (bar chart).

Cases and Variables, Quantitative and Categorical

Pet cats typically live about 12 years, but some have been known to live 28 years. Is that typical of domesticated predators? What about domesticated nonpredators, such as cows and guinea pigs? What about wild mammals? The rhinoceros, a nonpredator, lives an average of 15 years, with a maximum of about 45 years. The grizzly bear, a wild predator, lives an average of 25 years, with a maximum of about 50 years. Do meat-eaters typically outlive vegetarians in the wild? Often you can find answers to questions like these in a plot of the data.

Many of the examples in this section are based on the data about mammals in Display 2.24. Each row (type of mammal) is a case. As you learned in Chapter 1, the cases in a data set are the people, cities, mammals, or other items being studied.

Measurements and other properties of the cases are organized into columns, with one column for each variable. Thus, *average longevity* and *speed* are variables, and, for example, 30 mi/h is the value of the variable *speed* for the case *grizzly bear*.

Data Set

Mammal	Gestation Period (days)	Average Longevity (years)	Maximum Longevity (years)	Speed (mi/h)	Wild (1 = yes; 0 = no)	Predator (1 = yes; 0 = no)
Baboon	187	20	45	*	1	0
Bear, grizzly	225	25	50	30	1	1
Beaver	105	5	50	*	1	0
Bison	285	15	40	*	1	0
Camel	406	12	50	*	1	0
Cat	63	12	28	30	0	1
Cheetah	*	*	14	70	1	1
Chimpanzee	230	20	53	*	1	0
Chipmunk	31	6	8	*	1	0
Cow	284	15	30	*	0	0
Deer	201	8	20	30	1	0
Dog	61	12	20	39	0	1
Donkey	365	12	47	40	0	0
Elephant	660	35	70	25	1	0
Elk	250	15	27	45	1	0
Fox	52	7	14	42	1	1
Giraffe	425	10	34	32	1	0
Goat	151	8	18	*	0	0
Gorilla	258	20	54	*	1	0
Guinea pig	68	4	8	*	0	0
Hippopotamus	238	41	54	20	1	0
Horse	330	20	50	48	0	0
Kangaroo	36	7	24	40	1	0
Leopard	98	12	23	*	1	1
Lion	100	15	30	50	1	1
Monkey	166	15	37	*	1	0
Moose	240	12	27	*	1	0
Mouse	21	3	4	*	1	0
Opossum	13	1	5	*	1	1
Pig	112	10	27	11	0	0
Puma	90	12	20	*	1	1
Rabbit	31	5	13	35	0	0
Rhinoceros	450	15	45	*	1	0
Sea lion	350	12	30	*	1	1
Sheep	154	12	20	*	0	0
Squirrel	44	10	23	12	1	0
Tiger	105	16	26	*	1	1
Wolf	63	5	13	*	1	1
Zebra	365	15	50	40	1	0

Display 2.24 Facts on mammals. (Asterisks (*) mark missing values.) [*Source: World Almanac and Book of Facts,* 2001, p. 237.]

Counts of *how many* and measurements of *how much* are called **quantitative variables**. *Speed* is a quantitative variable because speed is measured on a numerical scale. A variable that groups cases into categories is called a **categorical variable**. *Predator* is a categorical variable because it groups the mammals into those who eat other animals and those who don't. Although the categories are coded 1 if a mammal preys on other animals and 0 if it does not, these numbers just indicate the appropriate category and are not meant to be quantitative. In Display 2.24, the asterisks (*) mark missing values.

DISCUSSION

Cases and Variables, Quantitative and Categorical

D6. Classify each variable in Display 2.24 as quantitative or categorical.

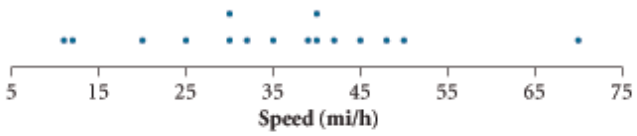
More About Dot Plots

Dot plots show individual cases as dots.

You've already seen dot plots, beginning in Chapter 1. As the name suggests, dot plots show individual cases as dots (or other plotting symbols, such as x). When you read a dot plot, keep in mind that different statistical software packages make dot plots in different ways. Sometimes one dot represents two or more cases, and sometimes values have been rounded. With a small data set, different rounding rules can give different shapes.

Display 2.25 shows a dot plot of the speeds of the mammals from Display 2.24. The gap between the cheetah's speed and that of the other mammals shows up clearly in the dot plot but not in the list of speeds in Display 2.24. Discoveries like this demonstrate why you should always plot your data.

Data Set



Display 2.25 Dot plot of the speeds of mammals.

When are dot plots most useful?

As you saw in Section 2.1, a dot plot shows shape, center, and spread. Dot plots tend to work best when

- you have a relatively small number of values to plot
- you want to see individual values, at least approximately
- you want to see the shape of the distribution
- you have one group or a small number of groups you want to compare

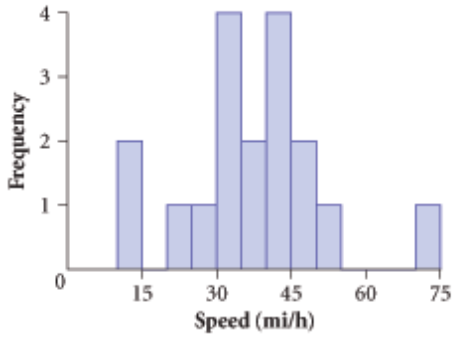
Histograms

Histograms show groups of cases as bars.

A dot plot shows individual cases as dots above a number line. To make a **histogram**, you divide the number line into intervals, called **bins**, and over each bin construct a bar that has a height equal to the number of cases in that bin. In fact, you can think of a histogram as a dot plot with bars drawn around

the dots and the dots erased. The height of the bar becomes a visual substitute for the number of dots. The plot in Display 2.26 is a histogram of the mammal speeds. Like the dot plot of a distribution, a histogram shows shape, center, and spread. The vertical axis gives the number of cases (the **frequency** or count) represented by each bar. For example, four mammals have speeds of from 30 mi/h up to 35 mi/h.

Data Set



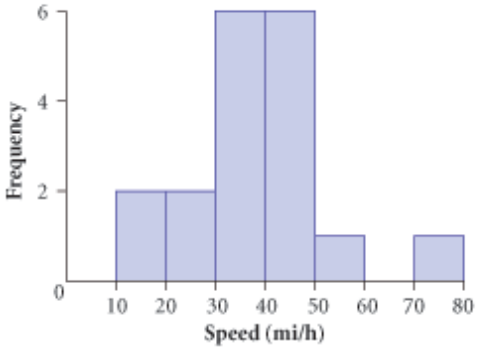
Display 2.26 Histogram of mammal speeds.

Borderline values go in the bar to the right.

Most calculators and statistical soft ware packages place a value that falls at the dividing line between two bars into the bar to the right. For example, in Display 2.26, the bar going from 30 to 35 contains cases for which $30 \leq speed < 35$.

Changing the width of the bars in your histogram can sometimes change your impression of the shape of the distribution. For example, the histogram of the speeds of mammals in Display 2.27 has fewer and wider bars than the histogram in Display 2.26. It shows a more symmetric, bell-shaped distribution, and there appears to be one peak rather than two. There is no "right answer" to the question of which bar width is best, just as there is no rule that tells a photographer when to use a zoom lens for a close-up. Different versions of a picture bring out different features. The job of a data analyst is to find a plot that shows important features of the distribution.

Data Set



Display 2.27 Speeds of mammals using a histogram with wider bars.

You can use your calculator to quickly display histograms with different bar widths. [See **Calculator Note 2C**.] Shown here are the mammal speed data. The numbers below the calculator screens indicate the window settings (minimum x, maximum x, x-scale, minimum y, maximum y, y-scale).

When are histograms most useful?

Histograms work best when

- you have a large number of values to plot
- you don't need to see individual values exactly
- you want to see the general shape of the distribution
- you have only one distribution or a small number of distributions you want to compare
- you can use a calculator or computer to make the plots for you

Relative frequency histograms show proportions instead of counts.

A histogram shows frequencies on the vertical axis. To change a histogram into a **relative frequency histogram**, divide the frequency for each bar by the total number of values in the data set and show these relative frequencies on the vertical axis.

Example: Converting Frequencies to Relative Frequencies

Four of the 18 mammals have speeds from 30 mi/h up to 35 mi/h. Convert the frequency 4 to a relative frequency.

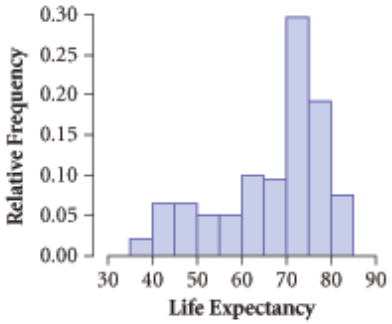
Solution

Four out of 18 is $\frac{4}{18}$ or approximately 0.22. So about 0.22 of the mammals have speeds in this range. ■

Example: Relative Frequency of Life Expectancies

Display 2.28 shows the relative frequency distribution of life expectancies for 203 countries around the world. How many countries have a life expectancy of at least 70 but less than 75 years? What proportion of the countries have a life expectancy of 70 years or more?

Data Set



Display 2.28 Life expectancies of people in countries around the world. [*Source:* Population Reference Bureau, *World Population Data Sheet*, 2005.]

Solution

The bar including 70 years and up to 75 years has a relative frequency of about 0.30, so the number of countries with a life expectancy of at least 70 years but less than 75 years is about 0.30 · 203, or about 61.

The proportion of countries with a life expectancy of 70 years or greater is the sum of the heights of the three bars to the right of 70—about 0.30 + 0.19 + 0.07, or 0.56.

DISCUSSION

Histograms

D7. In what sense does a histogram with narrow bars, as in Display 2.26, give you more information than a histogram with wider bars, as in Display 2.27? In light of your answer, why don't we always make histograms with very narrow bars?

D8. Does using relative frequencies change the shape of a histogram? What information is lost and gained by using a relative frequency histogram rather than a frequency histogram?

Stemplots

The plot in Display 2.29 is a **stem-and-leaf plot**, or **stemplot**, of the mammal speeds. It shows the key features of the distribution and preserves all theoriginal numbers.

Data Set

1	1 2
2	0 5
3	0 0 0 2 5 **9**
4	0 0 0 2 5 8
5	0
6	
7	0

3 | 9 represents 39 mi/h

Display 2.29 Stemplot of mammal speeds.

A stemplot shows cases as digits.

In Display 2.29, the numbers on the left , called the **stems**, are the tens digits of the speeds. The numbers on the right, called the **leaves**, are the ones digits of the speeds. The leaf for the speed 39 mi/h is printed in bold. If you turn your book 90° counterclockwise, you will see that a stemplot looks something like a dot plot or histogram; you can see the shape, center, and spread of the distribution.

The stemplot in Display 2.30 displays the same information, but with **split stems**: Each stem from the original plot has become two stems. If the ones digit is 0, 1, 2, 3, or 4, it is placed on the first line for that stem. If the ones digit is 5, 6, 7, 8, or 9, it is placed on the second line for that stem.

Display 2.30 Stemplot of mammal speeds, using split stems.

Spreading out the stems in this way is similar to changing the width of the bars in a histogram. The goal here, as always, is to find a plot that conveys the essential pattern of the distribution as clearly as possible.

You have compared two data distributions by constructing dot plots on the same scale (see, for example, Display 2.10). Another way to compare two distributions is to construct a back-to-back stemplot. Such a plot for the speeds of predators and nonpredators is shown in Display 2.31. The predators tend to have the faster speeds—or, at least, there are no slow predators!

Display 2.31 Back-to-back stemplot of mammal speeds for predators and nonpredators.

Usually, only two digits are plotted on a stemplot, one digit for the stem and one digit for the leaf. If the values contain more than two digits, the values may be either truncated (the extra digits simply cut off) or rounded. For example, if the speeds had been given to the nearest tenth of a mile, 32.6 mi/h could be either truncated to 32 mi/h or rounded to 33 mi/h.

As with the other types of plots, the rules for making stemplots are flexible. Do what seems to work best to reveal the important features of the data.

The stemplot of mammal speeds in Display 2.32 was made by statistical software. Although it looks a bit different from the handmade plot in Display 2.30, it is essentially the same. In the first two lines, N = 18 means that 18 cases were plotted; N* = 21 means that there were 21 cases in the original data set for which speeds were missing; and Leaf Unit = 1.0 means that the ones digits were graphed as the leaves. The numbers in the left column keep track of the number of cases, counting in from the extremes. The 2 on the left in the first line means that there are two cases on that stem. If you skip down three lines, the 4 on the left means that there are a total of four cases on the first four stems (11, 12, 20, and 25).

Display 2.32 Stem-and-leaf plot of mammal speeds made by statistical software.

When are stem-and-leaf plots most useful?

Stemplots are useful when

- you are plotting a single quantitative variable
- you have a relatively small number of values to plot
- you would like to see individual values exactly, or, when the values contain more than two digits, you would like to see approximate individual values
- you want to see the shape of the distribution clearly
- you have two (or sometimes more) groups you want to compare

Stemplots

D9. What information is given by the numbers in the bottom half of the far left column of the plot in Display 2.32? What does the 2 in parentheses indicate?

D10. How might you construct a stemplot of the data on gestation periods for the mammals listed in Display 2.24? Construct the stemplot and describe the shape of the distribution.

ACTIVITY 2.2a

Do Units of Measurement Affect Your Estimates?

In this activity, you will see whether you and your class estimate lengths better in feet or in meters.

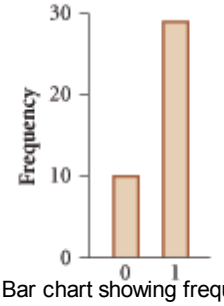
1. Your instructor will split the class randomly into two groups.
2. If you are in the fist group, you will estimate the length of your classroom in feet. If you are in the second group, you will estimate the length of the classroom in meters (without estimating first in feet and then converting to meters). Do this by looking at the length of the room; no pacing the length of the room is allowed.
3. Find an appropriate way to plot the two data sets so that you can compare their shapes, centers, and spreads.
4. Do the students in your class tend to estimate more accurately in feet or in meters? What is the basis for your decision?
5. Why split the class randomly into two groups instead of simply letting the left half of the room estimate in feet and the right half estimate in meters?

Bar Charts for Categorical Data

Bar charts show the frequencies of categorical data as heights of bars.

You now have three different types of plots to use with quantitative variables. What about categorical variables? You could make a dot plot, or you could make what looks like a histogram but is called a **bar chart** or **bar graph**. There is one bar for each category, and the height of the bar tells the frequency. A bar chart has categories on the horizontal axis, whereas a histogram has measurements—values from a quantitative variable.

The bar chart in Display 2.33 shows the frequency of mammals that fall into the categories "wild" and "domesticated," coded 1 and 0, respectively. (Note that the bars are separated so that there is no suggestion that the variable can take on a value of, say, 0.5.)



Display 2.33 Bar chart showing frequency of domesticated (0) and wild (1) mammals.

The relative frequency bar chart in Display 2.34 shows the proportion of the female labor force age 25 and older in the United States who fall into various educational categories. The coding used in the display is as follows:

1: none–8th grade
2: 9th grade–11th grade
3: high school graduate
4: some college, no degree
5: associate degree
6: bachelor's degree
7: master's degree
8: professional degree
9: doctorate degree

Display 2.34 The female labor force age 25 years and older by educational attainment [*Source*: U.S. Census Bureau, March 2005 Current Population Survey, www.census.gov.]

The educational categories in Display 2.34 have a natural order from least education to most education and are coded with the numbers 1 through 9. Note that if you compute the mean of this distribution, there is no reasonable way to interpret it. However, it does make sense to summarize this distribution using the mode: More women fall into the category "high school graduate" than into any other category. Thus, the numbers 1 through 9 are best thought of as representing an ordered categorical variable, not a quantitative variable.

You will learn more about the analysis of categorical data in Chapter 10.

Bar Charts

D11. In the bar chart in Display 2.33, would it matter if the order of the bars were reversed? In the bar chart in Display 2.34, would it matter if the order of the first two bars were reversed? Comment on how we might define two different types of categorical variables.

Summary 2.2: Graphical Displays of Distributions

When a variable is quantitative, you can use dot plots, stemplots (stem-and-leaf plots), and histograms to display the distribution of values. From each plot, you can see the shape, center, and spread of the distribution. However, the amount of

detail varies, and you should choose a plot that fits both your data set and your reason for analyzing it.

- Stemplots can retain the actual data values.
- Dot plots are best used with a small number of values and show roughly where the values lie on a number line.
- Histograms show only intervals of values, losing the actual data values, and are most appropriate for large data sets.

A bar chart shows the distribution of a categorical variable.

When you look at a plot, you should attempt to answer these questions:

- Where did this data set come from?
- What are the cases and the variables?
- What are the shape, center, and spread of this distribution? Does the distribution have any unusual characteristics, such as clusters, gaps, or outliers?
- What are possible interpretations or explanations of the patterns you see in the distribution?

Practice

More About Dot Plots

P6. In the listing of the Westvaco data in Display 1.1 on page 5, which variables are quantitative? Which are categorical?

P7. Select a reasonable scale, and make a dot plot of the gestation periods of the mammals listed in Display 2.24 on page 43. Write a sentence using shape, center, and spread to summarize the distribution of gestation periods for the mammals. What kinds of mammals have longer gestation periods?

Histograms

P8. Make histograms of the average longevities and the maximum longevities from Display 2.24. Describe how the distributions differ in terms of shape, center, and spread. Why do these differences occur?

P9. Convert your histograms from P8 of the average longevities and maximum longevities of the mammals to relative frequency histograms. Do the shapes of the histograms change?

P10. Using the relative frequency histogram of life expectancy in countries around the world (Display 2.28 on page 47), estimate the proportion of countries with a life expectancy of less than 50 years. Then estimate the number of countries with a life expectancy of less than 50 years. Describe the shape, center, and spread of this distribution.

Stemplots

P11. Make a back-to-back stemplot of the average longevities and maximum longevities from Display 2.24 on page 43. Compare the two distributions.

Bar Charts for Categorical Data

P12. Display 2.35, educational attainment of the male labor force, is the counterpart of Display 2.34. What are the cases, and what is the variable? Describe the distribution you see here. How does the distribution of female education compare to the distribution of male education? Why is it better to look at relative frequency bar charts rather than frequency bar charts to make this comparison?

Display 2.35 The male labor force age 25 years and older by educational attainment. [*Source:* U.S. Census Bureau, March 2005 Current Population Survey, www.census.gov.]

P13. Using the Westvaco data in Display 1.1 on page 5, make a bar chart showing the number of workers laid off in each round. In addition to a bar showing layoffs, for each of the five rounds, include a bar showing the number of workers not laid off. Then make a relative frequency bar chart. Describe any patterns you see.

Exercises

E15. The dot plot in Display 2.36 shows the distribution of the ages of pennies in a sample collected by a statistics class.

Display 2.36 Age of pennies. Each dot represents four points.

a. Where did this data set come from? What are the cases and the variables?

b. What are the shape, center, and spread of this distribution?

c. Does the distribution have any unusual characteristics? What are possible interpretations or explanations of the patterns you see in the distribution? That is, why does the distribution have the shape it does?

E16. Suppose you collect this information for each student in your class: age, hair color, number of siblings, gender, and miles he or she lives from school. What are the cases? What are the variables? Classify each variable as quantitative or categorical.

E17. Using your knowledge of the variables and what you think the shape of the distribution might be, match each variable in this list with the appropriate histogram in Display 2.37.

I. scores on a fairly easy examination in statistics
II. heights of a group of mothers and their 12-year-old daughters
III. numbers of medals won by medal–winning countries in the 2004 Summer Olympics
IV. weights of grown hens in a barnyard

Display 2.37 Four histograms with different shapes.

E18. Using the technology available to you, make histograms of the average longevity and maximum longevity data in Display 2.24 on page 43, using bar widths of 4, 8, and 16 years. Comment on the main features of the shapes of these distributions and determine which bar width appears to display these features best.

E19. Rewrite each sentence so that it states a relative frequency rather than a count.

a. Six students in a class of 30 got an A.
b. Out of the 50,732 people at a concert, 24,021 bought a T-shirt.

E20. Convert the histogram in Display 2.38 into a relative frequency histogram.

Display 2.38 Ages of 1000 people.

E21. Display 2.39 shows the distribution of the heights of U.S. males between the ages of 18 and 24. The heights are rounded to the nearest inch.

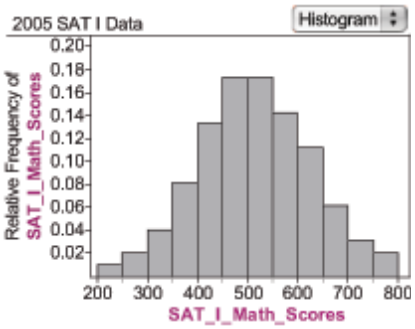
Data Set

Display 2.39 Heights of males, age 18 to 24. [*Source:* U.S. Census Bureau, *Statistical Abstract of the United States,* 1991.]

a. Draw a smooth curve to approximate the histogram.
b. Without doing any computing, estimate the mean and standard deviation.
c. Estimate the proportion of men age 18 to 24 who are 74 in. tall or less.
d. Estimate the proportion of heights that fall below 68 in.
e. Why should you say that the distribution of heights is "approximately" normal rather than simply saying that it is normally distributed?

E22. The histogram in Display 2.40 shows the distribution of SAT I math scores for 2004–2005.
a. Without doing any computing, estimate the mean and standard deviation.
b. Roughly what percentage of the SAT I math scores would you estimate are within one standard deviation of the mean?
c. For SAT I critical reading scores, the shape was similar, but the mean was 10 points lower and the standard deviation was 2 points smaller. Draw a smooth curve to show the distribution of SAT I critical reading scores.

Data Set

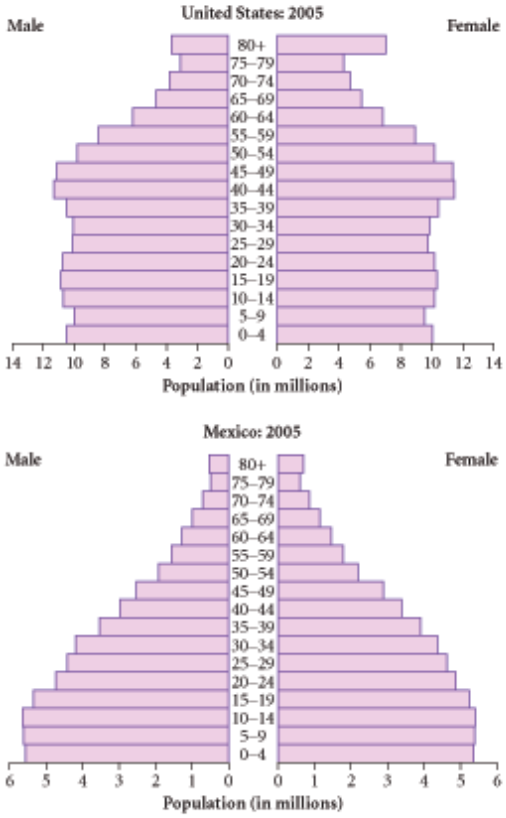


Display 2.40 Relative frequency histogram of SAT I math scores, 2004–2005. [*Source:* College Board Online, www.collegeboard.org.]

E23. In this section, you looked at various characteristics of mammals.
a. Would you predict that wild mammals or domesticated mammals generally have greater longevity?
b. Using the data in Display 2.24 on page 43, make a back-to-back stemplot to compare the average longevities.
c. Write a short summary comparing the two distributions.

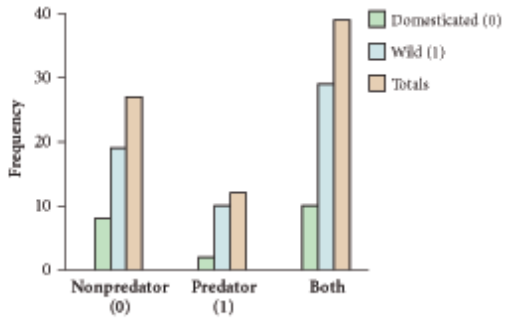
E24. The plots in Display 2.41 show a form of back-to-back histogram called a *population pyramid*. Describe how the population distribution of the United States differs from the population distribution of Mexico.

Data Set



Display 2.41 Population pyramids for the United States and Mexico, 2005. [*Source:* U.S. Census Bureau, International Data Base, www.census.gov.]

E25. Examine the grouped bar chart in Display 2.42, which summarizes some of the information from Display 2.24 on page 43.



Display 2.42 Bar chart for nonpredators and predators, showing frequency of wild and domesticated mammals.

a. Describe what the height of the first three bars represents.
b. How can you tell from this bar chart whether a predator from the list in Display 2.24 is more likely to be wild or domesticated?
c. How can you tell from this bar chart whether a nonpredator or a predator is more likely to be wild?

E26. Make a grouped bar chart similar to that in E25 for the hourly and salaried Westvaco workers (see Display 1.1 on page 5), with bars showing the frequencies of *laid off* and *not laid off* for the two categories of workers.

E27. Consider the mammals' speeds in Display 2.24.
a. Count the number of mammals that have speeds ending in 0 or 5.
b. How many speeds would you expect to end in 0 or 5 just by chance?
c. What are some possible explanations for the fact that your answers in parts a and b are so different?

E28. Look through newspapers and magazines to find an example of a graph that is either misleading or difficult to interpret. Redraw the graph to make it clear.

2.3 Measures of Center and Spread

So far you have relied on visual methods for estimating summary statistics to measure center and spread. In this section, you will learn how to compute exact values of those same summary statistics directly from the data.

Measures of Center

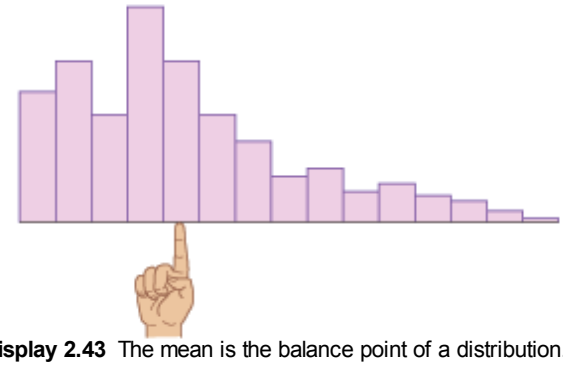
The two most commonly used **measures of center** are the mean and the median.

The **mean**, $\bar{X}$, is the same number that many people call the "average." To compute the mean, sum all the values of x and divide by the number of values, n:

(The symbol $\sum$, for sum, means to add up all the values of x.)

The mean is the balance point.

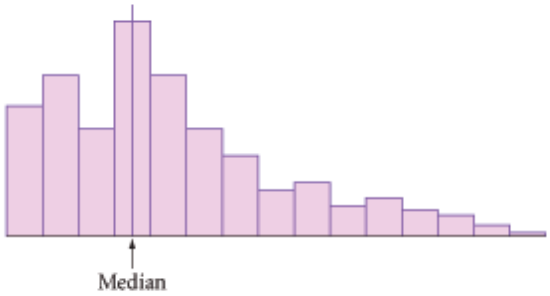
The mean is the balance point of a distribution. To estimate the mean visually on a dot plot or histogram, find where you would have to place a finger below the horizontal axis in order to balance the distribution, as if it were a tray of blocks (see Display 2.43).



Display 2.43 The mean is the balance point of a distribution.

The mean is the balance point.

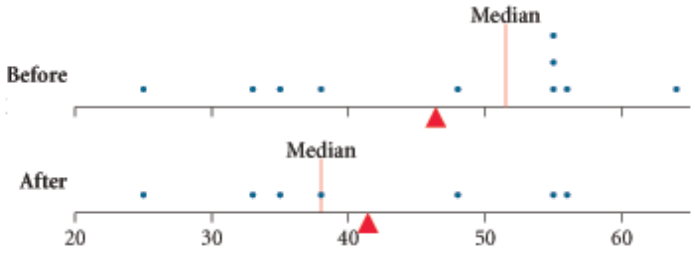
The **median** is the value that divides the data into halves, as shown in Display 2.44. To find it for an odd number of values, list all the values in order and select the middle one. If there are n values and n is odd, you will find the median at position . If n is even, the median is the average of the values on either side of position .



Display 2.44 The median divides the distribution into two equal areas.

Example: Effect of Round 2 Layoffs on Measures of Center

The ages of the hourly workers involved in Round 2 of the layoffs at Westvaco were 25, 33, 35, 38, 48, 55, 55*, 55*, 56, and 64* (* indicates laid off in Round 2). The two dot plots in Display 2.45 show the distributions of hourly workers before and after the second round of layoffs. What was the effect of Round 2 on the mean age? On the median age?



Display 2.45 Ages of Westvaco hourly workers before and after Round 2, showing the means and medians.

Solution

Means

Before: The sum of the ten ages is 464, so the mean age is , or 46.4 years.

After: There are seven ages and their sum is 290, so the mean age is , or 41.4 years.

The layoffs reduced the mean age by 5 years.

Medians

Before: Because there are ten ages, $n = 10$, so = or 5.5, and the median is halfway between the fifth ordered value, 48, and the sixth ordered value, 55. The median is , or 51.5 years.

After: There are seven ages, so = or 4. The median is the fourth ordered value, or 38 years.

The layoffs reduced the median age by 13.5 years.

DISCUSSION

Measures of Center

D12. Find the mean and median of each ordered list, and contrast their behavior.

a. 1, 2, 3
b. 1, 2, 6
c. 1, 2, 9
d. 1, 2, 297

D13. As you saw in D12, typically the mean is affected more than the median by an outlier.

a. Use the fact that the median is the halfway point and the mean is the balance point to explain why this is true.
b. For the distributions of mammal speeds in Display 2.31 on page 48, the means are 43.5 mi/h for predators and 31.5 mi/h for nonpredators. The medians are 40.5 mi/h and 33.5 mi/h, respectively. What about the distributions causes the means to be farther apart than the medians?
c. What about the shapes of the plots in Display 2.45 explains why the means change so much less than the medians?

Measuring Spread Around the Median: Quartiles and *IQR*

Pair a measure of center with a measure of spread.

Use *IQR* as a measure of spread with the median.

You can locate the median of a distribution by dividing your data into a lower and upper half. You can use the same idea to measure spread: Find the values that divide each half in half again. These two values, the lower quartile, Q_1, and the upper quartile, Q_3, together with the median, divide your data into four quarters. The distance between the upper and lower quartiles, called the **interquartile range**, or *IQR*, is a measure of spread.

$$IQR = Q_3 - Q_1$$

San Francisco, California, and Springfield, Missouri, have about the same median temperature over the year. In San Francisco, half the months of the year have a normal temperature above 56.5°F, half below. In Springfield, half the months have a normal temperature above 57°F, half below. If you judge by these medians, the difference hardly matters. But if you visit San Francisco, you better take a jacket, no matter what month you go. If you visit Springfield, take your shorts and a T-shirt in the summer and a heavy coat in the winter. The difference in temperatures between the two cities is not in their centers but in their variability. In San Francisco, the middle 50% of normal monthly temperatures lie in a narrow 9° interval between 52.5°F and 61.5°F, whereas in Springfield the middle 50% of normal monthly temperatures range over a 31° interval, varying from 40.5°F to 71.5°F. In other words, the *IQR* is 9°F for San Francisco and 31°F for Springfield.

Finding the Quartiles and *IQR*

If you have an even number of cases, finding the quartiles is straightforward: Order your observations, divide them into a lower and upper half, and then divide each half in half. If you have an odd number of cases, the idea is the same, but there's a question of what to do with the middle value when you form the upper and lower halves.

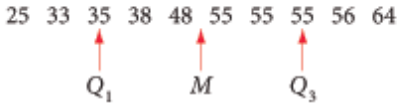
There is no one standard answer. Different statistical software packages use different procedures that can give slightly different values for the quartiles. In this book, the procedure is to omit the middle value when you form the two halves.

Example: Finding the Quartiles and *IQR* for Workers' Ages

Find the quartiles and *IQR* for the ages of the hourly workers at Westvaco before and after Round 2 of the layoffs.

Solution

Before: There are ten ages: 25, 33, 35, 38, 48, 55, 55, 55, 56, 64. Because n is even, the median is halfway between the two middle values, 48 and 55, so it is 51.5. The lower half of the data is made up of the first five ordered values, and the median of the lower half is the third value, so Q_1 is 35. The upper half of the data is the set of the five largest values, and the median of these is again the third value, so Q_3 is 55. The *IQR* is 55 − 35, or 20.



After: After the three workers are laid off in Round 2, there are seven ages: 25, 33, 35, 38, 48, 55, 56. Because n is odd, the median is the middle value, 38. Omit this one number. The lower half of the data is made up of the three ordered values to the left of position 4. The median of these is the second value, so Q_1 is 33. The upper half of the data is the set of the three ordered values to the right of position 4, and the median of these is again the second value, so Q_3 is 55. The *IQR* is 55 − 33, or 22.

DISCUSSION

Finding the Quartiles and IQR

D14. Here are the medians and quartiles for the speeds of the domesticated and wild mammals:

	Q_1	Median	Q_3
Domesticated	30	37	40
Wild	27.5	36	43.5

a. Use the information in Display 2.24 on page 43 to verify these values, and then use them to summarize and compare the two distributions.
b. Why might the speeds of domesticated mammals be less spread out than the speeds of wild mammals?

D15. The following quote is from the mystery *The List of Adrian Messenger*, by Philip MacDonald (Garden City, NY: Doubleday, 1959, p. 188). Detective Firth asks Detective Seymour if eyewitness accounts have provided a description of the murderer:
"Descriptions?" he said. "You must've collected quite a few. How did they boil down?"
"To a no-good norm, sir." Seymour shrugged wearily. "They varied so much, the average was useless."
Explain what Detective Seymour means.

Five-Number Summaries, Outliers, and Boxplots

The visual, verbal, and numerical summaries you've seen so far tell you about the middle of a distribution but not about the extremes. If you include the minimum and maximum values along with the median and quartiles, you get the five-number summary.

The five-number summary for a set of values:

Minimum: the smallest value
Lower or **first quartile, Q_1:** the median of the lower half of the ordered set of values
Median or **second quartile:** the value that divides the ordered set of values into halves
Upper or **third quartile, Q_3:** the median of the upper half of the ordered set of values
Maximum: the largest value

Data Set

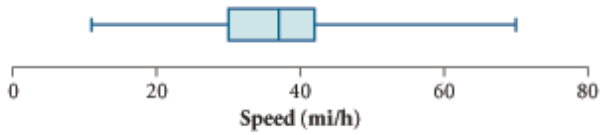
The difference of the maximum and the minimum is called the **range.**
Display 2.46 shows the five-number summary for the speeds of the mammals listed in Display 2.24.

1	1 2	min	11
2	0 5	Q1	30
3	0 0 0 2 5 9	median	37
4	0 0 0 2 5 8	Q3	42
5	0	max	70
6			
7	0		

Display 2.46 Five-number summary for the mammal speeds.

Display 2.47 shows a boxplot of the mammal speeds. A **boxplot** (or **box-and-whiskers plot**) is a graphical display of the five-number summary. The "box" extends from Q_1 to Q_3, with a line at the median. The "whiskers" run from the quartiles to the extreme values.

Data Set



Display 2.47 Boxplot of mammal speeds.

The maximum speed of 70 mi/h for the cheetah is 20 mi/h from the next fastest mammal (the lion) and 28 mi/h from the nearest quartile. It is handy to have a version of the boxplot that shows isolated cases—outliers—such as the cheetah. Informally, outliers are any values that stand apart from the rest. This rule often is used to identify outliers.

A value is an **outlier** if it is more than 1.5 times the *IQR* from the nearest quartile.

1.5 · *IQR* rule for outliers

Note that "more than 1.5 times the *IQR* from the nearest quartile" is another way of saying "either greater than Q_3 plus 1.5 times *IQR* or less than Q_1 minus 1.5 times *IQR*."

Example: Outliers in the Mammal Speeds

Use the 1.5 · *IQR* rule to identify outliers and the largest and smallest non-outliers among the mammal speeds.

Solution

From Display 2.46, $Q_1 = 30$ and $Q_3 = 42$, so the *IQR* is 42 − 30 or 12, and 1.5 · *IQR* equals 18.

At the low end:

$$Q_1 - 1.5 \cdot IQR = 30 - 18 = 12$$

The pig, at 11 mi/h, is an outlier.

The squirrel, at 12 mi/h, is the smallest non-outlier.

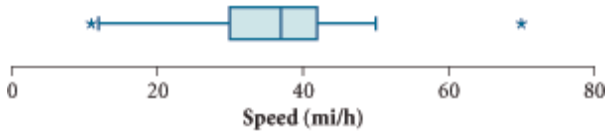
At the high end:

$$Q_3 + 1.5 \cdot IQR = 42 + 18 = 60$$

The cheetah, at 70 mi/h, is an outlier.
The lion, at 50 mi/h, is the largest non-outlier.

Data Set

A **modified boxplot**, shown in Display 2.48, is like the basic boxplot except that the whiskers extend only as far as the largest and smallest non-outliers (sometimes called **adjacent values**) and any outliers appear as individual dots or other symbols.



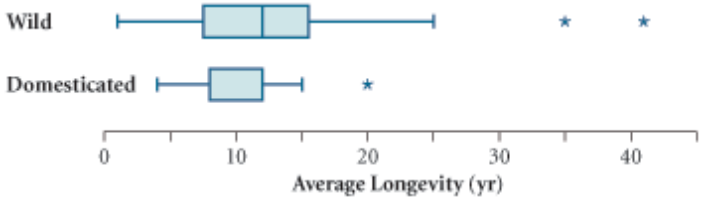
Display 2.48 Modified boxplot of mammal speeds.

Boxplots are particularly useful for comparing several distributions.

Data Set

Example: Boxplots That Show Outliers

Display 2.49 shows side-by-side modified boxplots of average longevity for wild and domesticated mammals. Compare the two distributions.

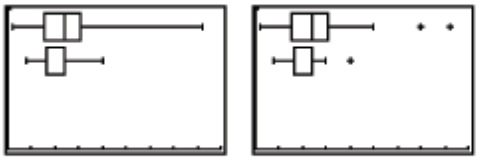


Display 2.49 Comparison of average longevity.

Solution

The boxplot for domesticated animals has no median line. So many domesticated animals had an average longevity of 12 years that it is both the median and the upper quartile. These plots show that species of domestic mammals typically have median average longevities of about 12 years, with about the middle half of these average longevities falling between 8 and 12 years. The average longevity of wild mammals centers at about the same place, but the wild mammal average longevities have more variability, with the middle half between about 7.5 and 15.5 years. Both shapes are roughly symmetric except for some unusually large average longevities—two wild mammals have average longevities of more than 30 years.

[See **Calculator Note 2D** to learn how to display regular and modified boxplots and five-number summaries on your calculator.]



When are boxplots most useful?

Boxplots are useful when you are plotting a single quantitative variable and

- you want to compare the shapes, centers, and spreads of two or more distributions
- you don't need to see individual values, even approximately
- you don't need to see more than the five-number summary but would like outliers to be clearly indicated

DISCUSSION

Five-Number Summaries, Outliers, and Boxplots

D16. Test your ability to interpret boxplots by answering these questions.

a. Approximately what percentage of the values in a data set lie within the box? Within the lower whisker, if there are no outliers? Within the upper whisker, if there are no outliers?

b. How would a boxplot look for a data set that is skewed right? Skewed left? Symmetric?

c. How can you estimate the *IQR* directly from a boxplot? How can you estimate the range?

d. Is it possible for a boxplot to be missing a whisker? If so, give an example. If not, explain why not.

e. Contrast the information you can learn from a boxplot with what you can learn from a histogram. List the advantages and disadvantages of each type of plot.

Measuring Spread Around the Mean: The Standard Deviation

There are various ways you can measure the spread of a distribution around its mean. Activity 2.3a gives you a chance to create a measure of your own.

ACTIVITY 2.3a

Comparing Hand Spans: How Far Are You from the Mean?

What you'll need: a ruler

1. Spread your hand on a ruler and measure your hand span (the distance from the tip of your thumb to the tip of your little finger when you spread your fingers) to the nearest half centimeter.
2. Find the mean hand span for your group.
3. Make a dot plot of the results for your group. Write names or initials above the dots to identify the cases. Mark the mean with a wedge (▲) below the number line.
4. Give two sources of variability in the measurements. That is, give two reasons why all the measurements aren't the same.
5. How far is your hand span from the mean hand span of your group? How far from the mean are the hand spans of the others in your group?
6. Make a plot of differences from the mean. Again label the dots with names or initials. What is the mean of these differences? Tell how to get the second plot from the first without computing any diff erences.
7. Using the idea of differences from the mean, invent at least two measures that give a "typical" distance from the mean.
8. Compare your measures with those of the other groups in your class. Discuss the advantages and disadvantages of each group's method.

The differences from the mean, $x - \bar{x}$, are called **deviations.** The mean is the balance point of the distribution, so the set of deviations from the mean will always sum to zero.

Deviations from the mean sum to zero:

Example: Deviations from the Mean

Find the deviations from the mean for the predators' speeds and verify that the sum of these deviations is 0. Which predator's speed is farthest from the mean?

Solution

The speeds 30, 30, 39, 42, 50, and 70 mi/h have mean 43.5 mi/h. The deviations from the mean are

$$30 - 43.5 = -13.5$$
$$30 - 43.5 = -13.5$$
$$39 - 43.5 = -4.5$$
$$42 - 43.5 = -1.5$$
$$50 - 43.5 = 6.5$$
$$70 - 43.5 = 26.5$$

The sum of the deviations is $-13.5 + (-13.5) + (-4.5) + (-1.5) + 6.5 + 26.5$, which equals 0. The cheetah's speed, 70 mi/h, is farthest from the mean.

How can you use the deviations from the mean to get a measure of spread? You can't simply find the average of the deviations, because you will get 0 every time. As you might have suggested in the activity, you could find the average of the absolute values of the deviations. That gives a perfectly reasonable measure of spread, but it does not turn out to be very easy to use or very useful. Think of how hard it is to deal with an equation that has sums of absolute values in it, for example, $y = |x - 1| + |x - 2| + |x - 3|$. On the other hand, if you square the deviations, which also gets rid of the negative signs, you get a sum of squares. Such a sum is always quadratic no matter how many terms there are, for example, $y = (x - 1)^2 + (x - 2)^2 + (x - 3)^2 = 3x^2 - 12x + 14$.

The measure of spread that incorporates the square of the deviations is the standard deviation, abbreviated *SD* or *s*, that you met in Section 2.1. Because sums of squares really are easy to work with mathematically, the *SD* offers important advantages that other measures of spread don't give you. You will learn more about these advantages in Chapter 7. The formula for the standard deviation, *s*, is given in the box.

Formula for the Standard Deviation, *s*

The square of the standard deviation, s^2, is called the **variance**.

Calculators might label the two versions σ_n and σ_{n-1}, or σ_n and *s*.

It might seem more natural to divide by *n* to get the average of the squared deviations. In fact, two versions of the standard deviation formula are used: One divides by the sample size, *n*; the other divides by $n - 1$. Dividing by $n - 1$ gives a slightly larger value. This is useful because otherwise the standard deviation computed from a sample would tend to be smaller than the standard deviation of the population from which the sample came. (You will learn more about this in Chapter 7.) In practice, dividing by $n - 1$ is almost always used for real data even if they aren't a sample from a larger population.

Example: The Standard Deviation of Mammal Longevity

Compute the standard deviation of the average longevity of domesticated mammals from Display 2.24 on page 43.

Solution

The table in Display 2.50 is a good way to organize the steps. First find the mean longevity, $\bar{x}$, and then subtract it from each observed value x to get the deviations, $x - \bar{x}$. Square each deviation to get $(x - \bar{x})^2$.

Data Set

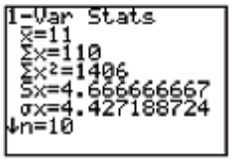
Case	Longevity, x	Mean, $\bar{x}$	Deviation, $x - \bar{x}$	Squared Deviation, $(x - \bar{x})^2$
Cat	12	11	1	1
Cow	15	11	4	16
Dog	12	11	1	1
Donkey	12	11	1	1
Goat	8	11	−3	9
Guinea pig	4	11	−7	49
Horse	20	11	9	81
Pig	10	11	−1	1
Rabbit	5	11	−6	36
Sheep	12	11	1	1
Total	**110**	**110**	**0**	**196**

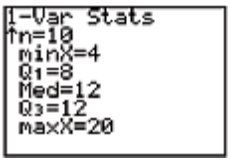
Display 2.50 Steps in computing the standard deviation.

To compute the standard deviation, sum the squared deviations, divide the sum by $n - 1$, and finally take the square root:

[You can organize the steps of calculating the standard deviation on your calculator. See **Calculator Note 2E**.]

Your calculator will compute the summary statistics for a set of data. [See **Calculator Note 2F**.] Here are the summary statistics for the domesticated mammal longevity data. Note that the standard deviation calculated in the previous example is denoted as Sx. Note also that the five-number summary is shown.





DISCUSSION

The Standard Deviation

D17. Refer to the previous example for mammal longevities.

a. Does 4.67 years seem like a typical distance from the mean of 11 years for the average longevities in the example?

b. The average longevities are measured in years. What is the unit of measurement for the mean? For the standard deviation? For the variance? For the interquartile range? For the median?

D18. The standard deviation, if you look at it the right way, is a generalization of the usual formula for the distance between two points. How does the formula for the standard deviation remind you of the formula for the distance between two points?

D19. What effect does dividing by $n - 1$ rather than by n have on the standard deviation? Does which one you divide by matter more with a large number of values or with a small number of values?

Summaries from a Frequency Table

To find the mean of the numbers 5, 5, 5, 5, 5, 5, 8, 8, and 8, you could sum them and divide their sum by how many numbers there are. However, you could get the same answer faster by taking advantage of the repetitions:

You can use formulas to find the mean and standard deviation of values in a frequency table, like the one in Display 2.51 in the example on the next page.

Formulas for the Mean and Standard Deviation of Values in a Frequency Table

If each value x occurs with frequency f, the mean of a frequency table is given by

The standard deviation is given by

where n is the sum of the frequencies, or

Example: Mean and Standard Deviation of Coin Values

Suppose you have five pennies, three nickels, and two dimes. Find the mean value of the coins and the standard deviation.

Solution

The table in Display 2.51 shows a way to organize the steps in computing the mean using the formula for the mean of values in a frequency table.

	Value x	Frequency f	$x \cdot f$
Penny	1	5	5
Nickel	5	3	15
Dime	10	2	20
Sum		10	40

$$\bar{x} = \frac{\Sigma x \cdot f}{n} = \frac{40}{10} = 4$$

Display 2.51 Steps in computing the mean of a frequency table.

Display 2.52 gives an extended version of the table, designed to organize the steps in computing both the mean and the standard deviation.

	Value x	Frequency f	$x \cdot f$	$x - \bar{x}$	$(x - \bar{x})^2$	$(x - \bar{x})^2 \cdot f$
Penny	1	5	5	−3	9	45
Nickel	5	3	15	1	1	3
Dime	10	2	20	6	36	72
Sum		10	40			120

$$s = \sqrt{\frac{\Sigma (x - \bar{x})^2 \cdot f}{n - 1}} = \sqrt{\frac{120}{9}} \approx 3.65$$

Display 2.52 Steps in computing the standard deviation of values in a frequency table.

[See **Calculator Note 2F** to learn how to compute numerical summaries from a frequency table.]

DISCUSSION

Summaries from a Frequency Table

D20. Explain why the formula for the standard deviation in the box on page 67 gives the same result as the formula on page 65.

Summary 2.3: Measures of Center and Spread

Your first step in any data analysis should always be to look at a plot of your data, because the shape of the distribution will help you determine what summary measures to use for center and spread.

- To describe the center of a distribution, the two most common summaries are the median and the mean. The median, or halfway point, of a set of ordered values is either the middle value (if n is odd) or halfway between the two middle values (if n is even). The mean, or balance point, is the sum of the values divided by the number of values.
- To measure spread around the median, use the interquartile range, or *IQR*, which is the width of the middle 50% of the data values and equals the distance from the lower quartile to the upper quartile. The quartiles are the medians of the lower half and upper half of the ordered list of values.
- To measure spread around the mean, use the standard deviation. To compute the standard deviation for a data set of size n, first find the deviations from the mean, then square them, sum the squared deviations, divide by $n - 1$, and take the square root.

A boxplot is a useful way to compare the general shape, center, and spread of two or more distributions with a large number of values. A modified boxplot also shows outliers. An outlier is any value more than 1.5 times the *IQR* from the nearest quartile.

Practice

Measures of Center

P14. Find the mean and median of these ordered lists.

a. 1 2 3 4
b. 1 2 3 4 5
c. 1 2 3 4 5 6
d. 1 2 3 4 5 . . . 97 98
e. 1 2 3 4 5 . . . 97 98 99

P15. Five 3rd graders, all about 4 ft tall, are standing together when their teacher, who is 6 ft tall, joins the group. What is the new mean height? The new median height?

P16. The stemplots in Display 2.53 show the life expectancies (in years) for females in the countries of Africa and Europe. The means are 53.6 years for Africa and 79.3 years for Europe.

a. Find the median life expectancy for each set of countries.
b. Is the mean or the median smaller for each distribution? Why is this so?

Stem-and-leaf of Life
Exp Africa
N = 56 Leaf Unit = 1.0

2	3	55
4	3	77
4	4	3
5	4	1
9	4	2233
14	4	44455
20	4	666666
27	4	8888999
(2)	5	00
27	5	2333
23	5	455
20	5	677
17	5	8999
13	6	
13	6	22
11	6	4
10	6	6
9	6	
9	7	
9	7	222
6	7	455
3	7	6
2	7	8
1	8	0

Stem-and-leaf of Life
Exp Europe
N = 41 Leaf Unit = 1.0

2	7	22
5	7	455
13	7	66667777
19	7	888999
(9)	8	000111111
13	8	2222223333
3	8	444

6 | 8 represents 68 years

Data Set

Display 2.53 Female life expectancies in Africa and Europe. [*Source:* Population Reference Bureau, *World Population Data Sheet*, 2005.]

Measuring Spread Around the Median: Quartiles and *IQR*

P17. Find the quartiles and *IQR* for these ordered lists.

a. 1 2 3 4 5 6
b. 1 2 3 4 5 6 7
c. 1 2 3 4 5 6 7 8
d. 1 2 3 4 5 6 7 8 9

P18. Display 2.54 shows a back-to-back stemplot of the average longevity of predators and nonpredators.

Data Set

Predators		Nonpredators
1	0	3 4
7 5	·	5 5 6 7 8 8
2 2 2 2 2	1	0 0 0 2 2 2 2
6 5	·	5 5 5 5 5 5
	2	0 0 0 0
5	·	
	3	
	·	5
	4	1

1 | 5 stands for 15 years

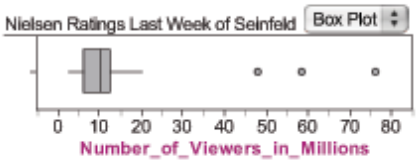
Display 2.54 Average longevities of predators and nonpredators.

a. By counting on the plot, find the median and quartiles for each group of mammals.
b. Write a pair of sentences summarizing and comparing the shape, center, and spread of the two distributions.

Five-Number Summaries, Outliers, and Boxplots

P19. The boxplot in Display 2.55 shows the number of viewers who watched the 101 prime-time network television shows in the week that *Seinfeld* aired its last new episode.

Data Set



Display 2.55 Modified boxplot of number of viewers of prime-time television shows.

a. *Seinfeld* had more viewers than any other show. About how many viewers did it have?
b. Estimate the median number of viewers, and use this median in a sentence.

P20. Use the medians and quartiles from D14 on page 60 and the data in Display 2.24 on page 43 to construct side-by-side boxplots of the speeds of wild and domesticated mammals. (Don't show outliers in these plots.)

Data Set

P21. The stemplot of average mammal longevities is shown in Display 2.56.

Data Set

0	1 3 4	1 | 5 stands for 15 years
·	5 5 5 6 7 7 8 8	
1	0 0 0 2 2 2 2 2 2 2 2 2	
·	5 5 5 5 5 5 5 6	
2	0 0 0 0	
·	5	
3		
·	5	
4	1	

Display 2.56 Average longevity (in years) of 38 mammals.

a. Use the stemplot to find the five-number summary.
b. Find the *IQR*.
c. Compute $Q_1 - 1.5 \cdot IQR$. Identify any outliers (give the animal name and longevity) at the low end.
d. Identify an outlier at the high end and the largest non-outlier.
e. Draw a modified boxplot.

The Standard Deviation

P22. Verify that the sum of the deviations from the mean is 0 for the numbers 1, 2, 4, 6, and 9. Find the standard deviation.

P23. Without computing, match each list of numbers in the left column with its standard deviation in the right column. Check any answers you aren't sure of by computing.

a. 1 1 1 1	i. 0
b. 1 2 2	ii. 0.058
c. 1 2 3 4 5	iii. 0.577
d. 10 20 20	iv. 1.581
e. 0.1 0.2 0.2	v. 3.162
f. 0 2 4 6 8	vi. 3.606
g. 0 0 0 0 5 6 6 8 8	vii. 5.774

Summaries from a Frequency Table

P24. Display 2.57 shows the data on family size for two representative sets of 100 families, one set from 1967 and the other from 1997.

1967		1997	
Number of Children	Number of Families	Number of Children	Number of Families
0	5	0	15
1	10	1	22
2	21	2	25
3	28	3	18
4	17	4	10
5	7	5	2
6	4	6	4
7	3	7	2
8	5	8	2

Display 2.57 Number of children in a sample of families, 1967 and 1997. [*Source:* U.S. Census Bureau, www.census.gov.]

a. Try to visualize the shapes of the two distributions. Are they symmetric, skewed left , or skewed right?

b. Find the median number of children per family for 1967.

c. Use the formulas to compute the mean and standard deviation of the 1967 distribution.

P25. Refer to Display 2.57.

a. Use the formulas for the mean and standard deviation of values in a frequency table to compute the mean number of children per family and the standard deviation for the 1997 distribution.

b. Find the median number of children for 1997.

c. What are the positions of the quartiles in an ordered list of 100 numbers? Find the quartiles for the 1967 distribution and compute the *IQR*. Do the same for the 1997 distribution.

d. Write a comparison of the shape, center, and spread of the distributions for the two years.

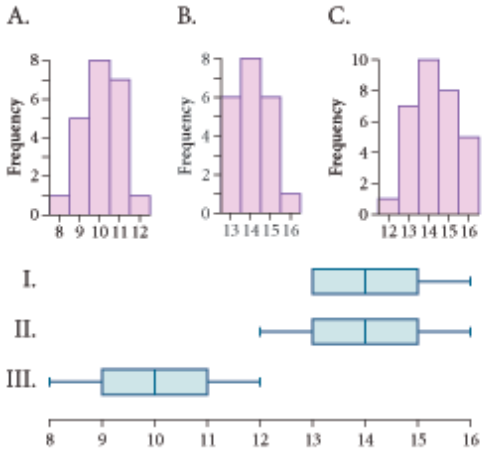
Exercises

E29. The mean of a set of seven values is 25. Six of the values are 24, 47, 34, 10, 22, and 28. What is the 7th value?

E30. The sum of a set of values is 84, and the mean is 6. How many values are there?

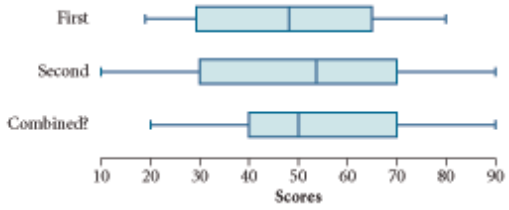
E31. Three histograms and three boxplots appear in Display 2.58. Which boxplot displays the same information as

a. histogram A?
b. histogram B?
c. histogram C?



Display 2.58 Match the histograms with their boxplots.

E32. The test scores of 40 students in a first-period class were used to construct the first boxplot in Display 2.59, and the test scores of 40 students in a second-period class were used to construct the second. Can the third plot be a boxplot of the combined scores of the 80 students in the two classes? Why or why not?



Display 2.59 Boxplots of three sets of test scores.

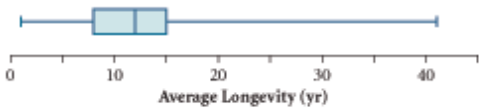
E33. Make side-by-side boxplots of the speeds of predators and nonpredators. (The stemplot in Display 2.31 on page 48 shows the values ordered.) Are the boxplots or the back-to-back stemplot in Display 2.31 better for comparing these speeds? Explain.

E34. The U.S. Supreme Court instituted a temporary ban on capital punishment between 1967 and 1976. Between 1977 and 2000, 31 states carried out 683 executions. (The other 19 states either did not have a death penalty or executed no one.) The five states that executed the most prisoners were Texas (239), Virginia (81), Florida (50), Missouri (46), and Oklahoma (30). The remaining 26 states carried out these numbers of executions: 26, 25, 23, 23, 23, 22, 16, 12, 11, 8, 8, 7, 6, 4, 3, 3, 3, 3, 2, 2, 2, 1, 1, 1, 1, 1. For all 50 states, what was the mean number of executions per state? The median number? What were the quartiles? Draw a boxplot, showing any outliers, of the number of executions for all 50 states. [*Source:* U.S. Department of Justice, *Bulletin: Capital Punishment 2000*.]

E35. Make a back-to-back stemplot comparing the ages of those retained and those laid off among the salaried workers in the engineering department at Westvaco (see Display 1.1 on page 5). Find the medians and quartiles, and use them to write a verbal comparison of the two distributions.

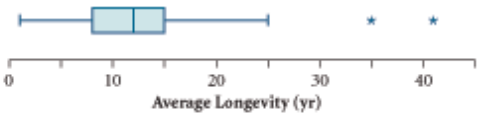
E36. The boxplots below show the average longevity of mammals, from Display 2.24.

a. Using only the basic boxplot in Display 2.60, show that there must be at least one outlier in the set of average longevity.



Display 2.60 Boxplot of average longevity of mammals.

b. How many outliers are there in the modified boxplot of average longevity in Display 2.61?



Display 2.61 Modified boxplot of average longevity of mammals, showing outliers.

c. How many outliers are shown in Display 2.49 on page 62? How can that be, considering the boxplot in Display 2.61?

E37. No computing should be necessary to answer these questions.

a. The mean of each of these sets of values is 20, and the range is 40. Which set has the largest standard deviation? Which has the smallest?

I. 0 10 20 30 40
II. 0 0 20 40 40
III. 0 19 20 21 40

b. Two of these sets of values have a standard deviation of about 5. Which two?

I. 5 5 5 5 5 5
II. 10 10 10 20 20 20
III. 6 8 10 12 14 16 18 20 22
IV. 5 10 15 20 25 30 35 40 45

E38. The standard deviation of the first set of values listed here is about 32. What is the standard deviation of the second

set? Explain. (No computing should be necessary.)
16 23 34 56 78 92 93
20 27 38 60 82 96 97

E39. Consider the set of the heights of all female National Collegiate Athletic Association (NCAA) athletes and the set of the heights of all female NCAA basketball players. Which distribution will have the larger mean? Which will have the larger standard deviation? Explain.

E40. Consider the data set 15, 8, 25, 32, 14, 8, 25, and 2. You can replace any one value with a number from 1 to 10. How would you make this replacement

a. to make the standard deviation as large as possible?
b. to make the standard deviation as small as possible?
c. to create an outlier, if possible?

E41. Another measure of center that sometimes is used is the **midrange**. To find the midrange, compute the mean of the largest value and the smallest value.
The statistics in this computer output summarize the number of viewers of prime-time television shows (in millions) for the week of the last new *Seinfeld* episode.

Variable	N	Mean	Median	TrMean	StDev	SEMean
Viewers	101	11.187	10.150	9.831	9.896	0.985
Variable	Min	Max	Q1	Q3		
Viewers	2.320	76.260	6.160	12.855		

a. Using these summary statistics alone, compute the midrange both with and without the value representing the Seinfeld episode. (*Seinfeld* had the largest number of viewers and *Seinfeld Clips*, with 58.53 million viewers, the second largest.) Is the midrange affected much by outliers? Explain.
b. Compute the mean of the ratings without the *Seinfeld* episode, using only the summary statistics in the computer output.

E42. In computer output like that in E41, TrMean is the **trimmed mean**. It typically is computed by removing the largest 5% of values and the smallest 5% of values from the data set and then computing the mean of the remaining middle 90% of values. (The percentage that is cut off at each end can vary depending on the software.)

a. Find the trimmed mean of the maximum longevities in Display 2.24 on page 43.
b. Is the trimmed mean affected much by outliers?

E43. This table shows the weights of the pennies in Display 2.3 on page 31. For example, the four pennies in the second interval, 3.0000 g to 3.0199 g, are grouped at the midpoint of this interval, 3.01.

Data Set

Weight	Frequency
2.99	1
3.01	4
3.03	4
3.05	4
3.07	7
3.09	17
3.11	24
3.13	17
3.15	13
3.17	6
3.19	2
3.21	1

a. Find the mean weight of the pennies.
b. Find the standard deviation.
c. Does the standard deviation appear to represent a typical deviation from the mean?

E44. Suppose you have five pennies, six nickels, four dimes, and five quarters.
a. Sketch a dot plot of the values of the 20 coins, and use it to estimate the mean.
b. Compute the mean using the formula for the mean of values in a frequency table.
c. Estimate the SD from your plot: Is it closest to 0, 5, 10, 15, or 20?
d. Compute the standard deviation using the formula for the standard deviation of values in a frequency table.

E45. On the first test of the semester, the scores of the first-period class of 30 students had a mean of 75 and a median of 70. The scores of the second-period class of 22 students had a mean of 70 and a median of 68.

a. To the nearest tenth, what is the mean test score of all 52 students? If you cannot calculate the mean of the two classes combined, explain why.

b. What is the median test score of all 52 students? If you cannot find the median of the two classes combined, explain why.

E46. The National Council on Public Polls rebuked the press for its coverage of a Gallup poll of Islamic countries. According to the Council:

News stories based on the Gallup poll reported results in the aggregate without regard to the population of the countries they represent. Kuwait, with less than 2 million Muslims, was treated the same as Indonesia, which has over 200 million Muslims. The "aggregate" quoted in the media was actually the average for the countries surveyed regardless of the size of their populations.

The percentage of people in Kuwait who thought the September 11 terrorist attacks were morally justified was 36%, while the percentage in Indonesia was 4%. [*Source:* www.ncpp.org.]

a. Suppose that the poll covered only these two countries and that the people surveyed were representative of the entire country. What percentage of all the people in these two countries thought that the terrorist attacks were morally justified?

b. What percentage would have been reported by the press?

2.4 Working with Summary Statistics

Summary statistics are very useful, but only when they are used with good judgment. This section will teach you how to tell which summary statistic to use, how changing units of measurement and the presence of outliers affect your summary statistics, and how to interpret percentiles.

Which Summary Statistic?

Plot first, then look for summaries.

Which summary statistics should you use to describe a distribution? Should you use mean and standard deviation? Median and quartiles? Something else? The right choice can depend on the shape of your distribution, so you should always start with a plot. For normal distributions, the mean and standard deviation are nearly always the most suitable. For skewed distributions, the median and quartiles are often the most useful, in part because they have a simple interpretation based on dividing a data set into fourths.

Sometimes, however, the mean and standard deviation will be the right choice even if you have a skewed distribution. For example, if you have a representative sample of house prices for a town and you want to use your sample to estimate the total value of all the town's houses, the mean is what you want, not the median. In Chapter 7, you'll see why the mean and standard deviation are the most useful choices when doing statistical inference.

Choosing the right summary statistics is something you will get better at as you build your intuition about the properties of these statistics and how they behave in various situations.

DISCUSSION

Which Summary Statistic?

D21. Explain how to determine the total amount of property taxes if you know the number of houses, the mean value, and the tax rate. In what sense is knowing the mean equivalent to knowing the total?

D22. When a measure of center for the income of a community's residents is given, that number is usually the median. Why do you think that is the case?

The Effects of Changing Units

This discussion illustrates some important properties of summary statistics. It will also help you develop your intuition about how the geometry and the arithmetic of working with data are related.

The lowest temperature on record for Washington, D.C., is −15°F. How does that temperature compare with the lowest recorded temperatures for capitals of other countries? Display 2.62 gives data for the few capitals whose record low temperatures turn out to be whole numbers on both the Fahrenheit and Celsius scales.

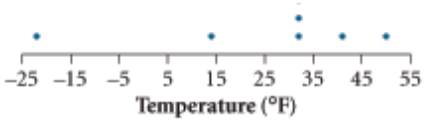
Data Set

City	Country	Temperature (°F)
Addis Ababa	Ethiopia	32
Algiers	Algeria	32
Bangkok	Thailand	50
Madrid	Spain	14
Nairobi	Kenya	41
Brazilia	Brazil	32
Warsaw	Poland	−22

Display 2.62 Record low temperatures for seven capitals.
[*Source:* National Climatic Data Center, 2002.]

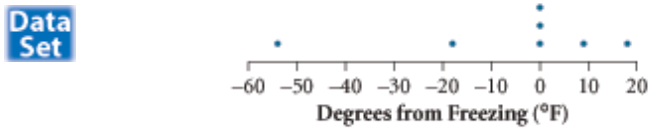
The dot plot in Display 2.63 shows that the temperatures are centered at about 32°F, with an outlier at −22°F. The spread and shape are hard to determine with only seven values.

Data Set



Display 2.63 Dot plot for record low temperatures in degrees Fahrenheit for seven capitals.

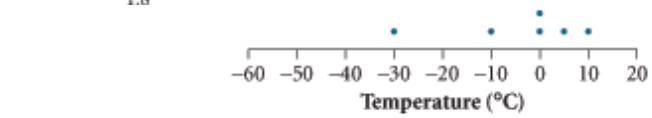
What happens to the shape, center, and spread of this distribution if you convert each temperature to the number of degrees above or below freezing, 32°F? To find out, subtract 32 from each value, and plot the new values. Display 2.64 shows that the center of the dot plot is now at 0 rather than 32 but the spread and shape are unchanged.



Display 2.64 Dot plot of the number of degrees Fahrenheit above or below freezing for record low temperatures for the seven capitals.

Adding (or subtracting) a constant to each value in a set of data doesn't change the spread or the shape of a distribution but slides the entire distribution a distance equivalent to the constant. Thus, the transformation amounts to a recentering of the distribution.

What happens to the shape and spread of this distribution if you convert each temperature to degrees Celsius? The Celsius scale measures temperature based on the number of degrees above or below freezing, but it takes 1.8°F to make 1°C. To convert, divide each value in Display 2.64 by 1.8 (or, equivalently, multiply by $\frac{1}{1.8}$), and plot the new values. Display 2.65 shows that the center of the new dot plot is still at 0 and the shape is the same. However, the spread has shrunk by a factor of $\frac{1}{1.8}$.



Display 2.65 Dot plot for record low temperatures in degrees Celsius for the seven capitals.

Multiplying each value in a set of data by a positive constant doesn't change the basic shape of the distribution. Both the mean and the spread are multiplied by that number. This transformation amounts to a rescaling of the distribution.

[See **Calculator Note 2G** to explore on your calculator the effects of changing units.]

Recentering and Rescaling a Data Set

Recentering a data set—adding the same number c to all the values in the set—doesn't change the shape or spread but slides the entire distribution by the amount *c*, adding *c* to the median and the mean.

Rescaling a data set—multiplying all the values in the set by the same positive number *d*—doesn't change the basic shape but stretches or shrinks the distribution, multiplying the spread (*IQR* or standard deviation) by *d* and multiplying the center (median or mean) by *d*.

DISCUSSION

The Effects of Changing Units

D23. Suppose a U.S. dollar is worth 9.4 Mexican pesos.

a. A set of prices, in U.S. dollars, has mean \$20 and standard deviation \$5. Find the mean and standard deviation of the prices expressed in pesos.

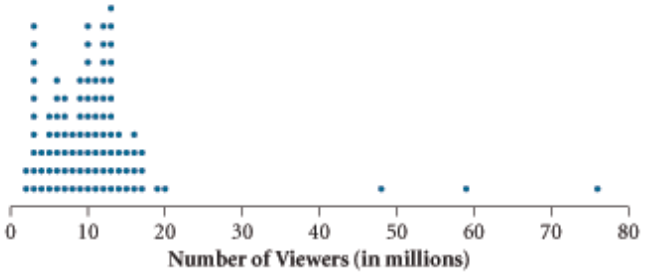
b. Another set of prices, in Mexican pesos, has a median of 94 pesos and quartiles of 47 pesos and 188 pesos. Find the median and quartiles of the same prices expressed in U.S. dollars.

The Influence of Outliers

A summary statistic is **resistant to outliers** if the summary statistic is not changed very much when an outlier is removed from the set of data. If the summary statistic tends to be affected by the removal of outliers, it is **sensitive to outliers.**

Display 2.66 shows a dot plot of the number of viewers of prime-time television shows (in millions) in a particular week. (A boxplot of these data is shown in Display 2.55 on page 70.) The three highest values—the three shows with the largest numbers of viewers—are outliers.

Data Set



Display 2.66 Number of viewers of prime-time television shows in a particular week.

The printout in Display 2.67 gives the summary statistics for all 101 shows.

Variable	N	Mean	Median	StDev
Ratings	101	11.187	10.150	9.896

Variable	Min	Max	Q1	Q3
Ratings	2.320	76.260	6.160	12.855

Display 2.67 Printout of summary statistics for number of viewers.

The printout in Display 2.68 gives the summary statistics for the number of viewers when the three outliers are removed from the set of data. Compare this printout with the one in Display 2.67 and notice which summary statistics are most sensitive to the outliers.

Display 2.68 Summary statistics for number of viewers without outliers.

DISCUSSION

The Influence of Outliers

D24. Are these measures of center for the number of television viewers affected much by the three outliers? (Refer to Displays 2.66–2.68.) Explain.
a. mean b. median

D25. Are these measures of spread for the number of television viewers affected much by the three outliers? Explain why or why not.
a. range b. standard deviation c. interquartile range

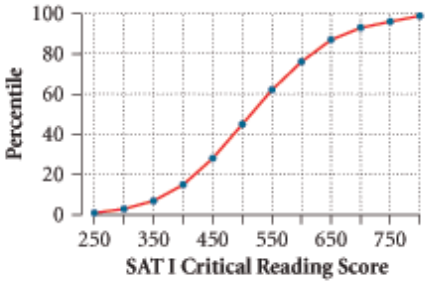
Percentiles and Cumulative Relative Frequency Plots

Percentiles measure position within a data set. The first quartile, Q_1, of a distribution is the 25th percentile—the value that separates the lowest 25% of the ordered values from the rest. The median is the 50th percentile, and Q_3 is the 75th percentile. You can define other percentiles in the same way. The 10th percentile, for example, is the value that separates the lowest 10% of ordered values in a distribution from the rest. In general, a value is at the kth **percentile** if k% of all values are less than or equal to it.

For large data sets, you might see data listed in a table or plotted in a graph, like those for the SAT I critical reading scores in Display 2.69. Such a plot is sometimes called a **cumulative percentage plot** or a **cumulative relative frequency plot.** The table shows that, for example, 28% of the students received a score of 450 or lower and about 13% received a score between 400 and 450.

Data Set

Score	Percentile	Score	Percentile
800	99+	450	28
750	96	400	15
700	93	350	7
650	87	300	3
600	76	250	1
550	62	200	—
500	45		



Display 2.69 Cumulative relative frequency plot of SAT I critical reading scores and percentiles, 2004–2005. [*Source:* The College Board, www.collegeboard.org.]

[See **Calculator Note 2H** to learn how to construct cumulative relative frequency plots on your calculator.]

DISCUSSION

Percentiles and Cumulative Relative Frequency Plots

D26. Refer to Display 2.69.
   a. Use the plot to estimate the percentile for an SAT I critical reading score of 425.
   b. What two values enclose the middle 90% of SAT scores? The middle 95%?
   c. Use the table to estimate the score that falls at the 40th percentile.

D27. What proportion of cases lie between the 5th and 95th percentiles of a distribution? What percentiles enclose the middle 95% of the cases in a distribution?

Summary 2.4: Working with Summary Statistics

Knowing which summary statistic to use depends on what use you have for that summary statistic.

If a summary statistic doesn't change much whether you include or exclude outliers from your data set, it is said to be resistant to outliers.

- The median and quartiles are resistant to outliers.
- The mean and standard deviation are sensitive to outliers.

Recentering a data set—adding the same number c to all the values—slides the entire distribution. It doesn't change the shape or spread but adds c to the median and the mean. Rescaling a data set—multiplying all the values by the same nonzero number d—is like stretching or squeezing the distribution. It doesn't change the basic shape but multiplies the spread (*IQR* or standard deviation) by $|d|$ and multiplies the measure of center (median or mean) by d.

The percentile of a value tells you what percentage of all values lie at or below the given value. The 30th percentile, for example, is the value that separates the distribution into the lowest 30% of values and the highest 70% of values.

Practice

Which Summary Statistic?

P26. A community in Nevada has 9751 households, with a median house price of $320,000 and a mean price of $392,059.
   a. Why is the mean larger than the median?
   b. The property tax rate is about 1.15%. What total amount of taxes will be assessed on these houses?
   c. What is the average amount of taxes per house?

P27. A news release at www.polk.com stated that the median age of cars being driven in 2004 was 8.9 years, the oldest to date. The median was 8.3 years in 2000 and 7.7 years in 1995
   a. Why were medians used in this news story?
   b. What reasons might there be for the increase in the median age of cars? (The median age in 1970 was only 4.9 years!)

The Effects of Changing Units

P28. The mean height of a class of 15 children is 48 in., the median is 45 in., the standard deviation is 2.4 in., and the interquartile range is 3 in. Find the mean, standard deviation, median, and interquartile range if

a. you convert each height to feet
b. each child grows 2 in.
c. each child grows 4 in. and you convert the heights to feet

P29. Compute the means and standard deviations (use the formula for s) of these sets of numbers. Use recentering and rescaling wherever you can to avoid or simplify the arithmetic.

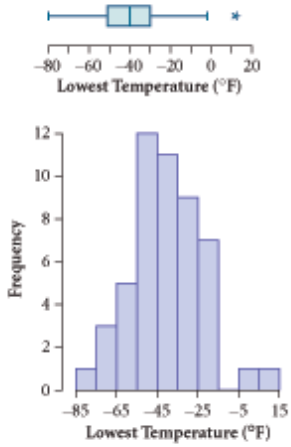
a. 1 2 3
b. 11 12 13
c. 10 20 30
d. 105 110 115
e. −800 −900 −1000

The Influence of Outliers

P30. The histogram and boxplot in Display 2.70 and the summary statistics in Display 2.71 show the record low temperatures for the 50 states.

Data Set

a. Hawaii has a lowest recorded temperature of 12°F. The boxplot shows Hawaii as an outlier. Verify that this is justified.
b. Suppose you exclude Hawaii from the data set. Copy the table in Display 2.71, substituting the value (or your best estimate if you don't have enough information to compute the value) of each summary statistic with Hawaii excluded.



Data Set

Display 2.70 Record low temperatures for the 50 states. [*Source:* National Climatic Data Center, 2002, www.ncdc.noaa.gov.]

Summary of Lowest Temperature	
Count	50
Mean	-40.3800
Median	-40
StdDev	17.6946
Min	-80
Max	12
Range	92
Lower 25 %tile	-51
Upper 75 %tile	-30

Data Set

Display 2.71 Summary statistics for lowest temperatures for the 50 states.

Percentiles and Cumulative Relative Frequency Plots

P31. Estimate the quartiles and the median of the SAT I critical reading scores in Display 2.69 on page 78, and then use these values to draw a boxplot of the distribution. What is the *IQR*?

Exercises

E47. Discuss whether you would use the mean or the median to measure the center of each set of data and why you prefer the one you chose.

a. the prices of single-family homes in your neighborhood
b. the yield of corn (bushels per acre) for a sample of farms in Iowa
c. the survival time, following diagnosis, of a sample of cancer patients

E48. *Mean versus median.*

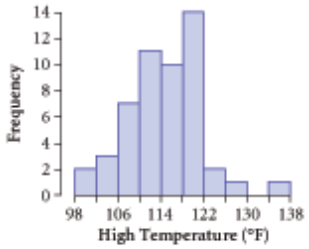
a. You are tracing your family tree and would like to go back to the year 1700. To estimate how many generations back you will have to trace, would you need to know the median length of a generation or the mean length of a generation?

b. If a car trip takes 3 h, do you need to know the mean speed or the median speed in order to find the total distance?

c. Suppose all trees in a forest are right circular cylinders with radius 3 ft. The heights vary, but the mean height is 45 ft, the median is 43 ft, the *IQR* is 3 ft, and the standard deviation is 3.5 ft. From this information, can you compute the total volume of wood in all the trees?

E49. The histogram in Display 2.72 shows record high temperatures for the 50 states.

Data Set



Display 2.72 Record high temperatures for the 50 U.S. states. [*Source:* National Climatic Data Center, 2002, www.ncdc.noaa.gov.]

a. Suppose each temperature is converted from degrees Fahrenheit, F, to degrees Celsius, C, using the formula

If you make a histogram of the temperatures in degrees Celsius, how will it differ from the one in Display 2.72?

b. The summary statistics in Display 2.73 are for record high temperatures in degrees Fahrenheit. Make a similar table for the temperatures in degrees Celsius.

Variable	N	Mean	Median	StDev
HighTemp	50	114.10	114.00	6.69
Variable	Min	Max	Q1	Q3
HighTemp	100.00	134.00	110.00	118.00

Display 2.73 Summary statistics for record high temperatures for the 50 U.S. states.

c. Are there any outliers in the data in °C?

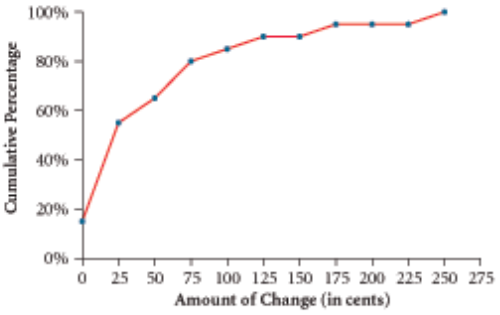
E50. Tell how you could use recentering and rescaling to simplify the computation of the mean and standard deviation for this list of numbers:

5478.1 5478.3 5478.3 5478.9 5478.4 547

E51. Suppose a constant c is added to each value in a set of data, x_1, x_2, x_3, x_4, and x_5. Prove that the mean increases by c by comparing the formula for the mean of the original data to the formula for the mean of the recentered data.

E52. Suppose a constant c is added to each value in a set of data, x_1, x_2, x_3, x_4, and x_5. Prove that the standard deviation is unchanged by comparing the formula for the standard deviation of the original data to the formula for the standard deviation of the recentered data.

E53. The cumulative relative frequency plot in Display 2.74 shows the amount of change carried by a group of 200 students. For example, about 80% of the students had \$0.75 or less.



Display 2.74 Cumulative percentage plot of amount of change.

a. From this plot, estimate the median amount of change.

b. Estimate the quartiles and the interquartile range.

c. Is the original set of amounts of change skewed right, skewed left, or symmetric?

d. Does the data set look as if it should be modeled by a normal distribution? Explain your reasoning.

E54. Use Display 2.74 to make a boxplot of the amounts of change carried by the students.

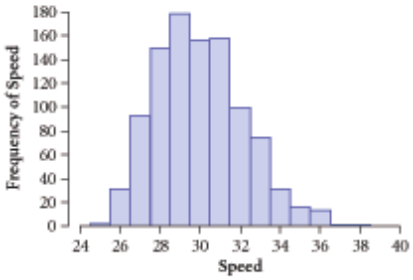
E55. Did you ever wonder how speed limits on roadways are determined? Most government jurisdictions set speed limits by this standard practice, described on the website of the Michigan State Police.

Data Set

> Speed studies are taken during times that represent normal free-flow traffic. Since modified speed limits are the maximum allowable speeds, roadway conditions must be close to ideal. The primary basis for establishing a proper, realistic speed limit is the nationally recognized method of using the 85th percentile speed. This is the speed at or below which 85% of the traffic moves. [*Source:* www.michigan.gov.]

The 85th percentile speed typically is rounded down to the nearest 5 miles per hour. The table and histogram in Display 2.75 give the measurements of the speeds of 1000 cars on a stretch of road in Mellowville with no curviness or other additional factors. At what speed would the speed limit be set if the guidelines described were followed?

Speed (mi/h)	Count
25	2
26	31
27	92
28	149
29	178
30	156
31	157
32	99
33	74
34	31
35	16
36	13
37	1
38	1
Total	1000



Display 2.75 Speed of 1000 cars in Mellowville.

E56. Refer to the distribution of speeds in E55. Make a cumulative relative frequency plot of these speeds.

Data Set

E57. The cumulative relative frequency plot in Display 2.76 gives the ages of the CEOs (Chief Executive Officers) of the 500 largest U.S. companies. Does A, B, or C give its median and quartiles? Using the diagram, explain why your choice is correct.

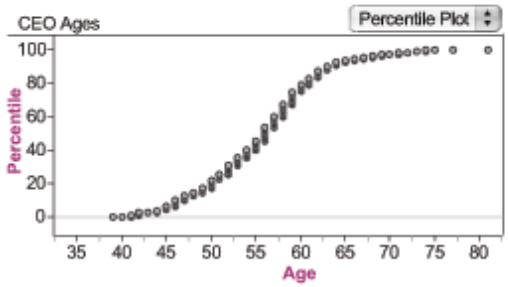
Data Set

A. Q_1 51; median 56; Q_3 60

B. Q_1 50; median 60; Q_3 70

C. Q_1 25; median 50; Q_3 75

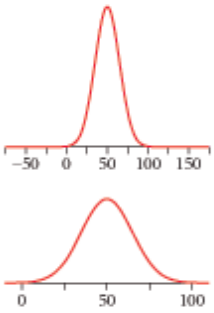
Data Set



Display 2.76 Cumulative relative frequency plot of CEO ages. [*Source:* www.forbes.com]

E58. Refer to the distribution of ages in E57. Can you give the median and quartiles of the distribution of ages in months? If so, do it. If not, explain why not.

2.5 The Normal Distribution



These are both the same normal curve.

You have seen several reasons why the normal distribution is so important:

- It tells you how variability in repeated measurements often behaves (diameters of tennis balls).
- It tells you how variability in populations often behaves (weights of pennies, SAT scores).
- It tells you how means (and some other summary statistics) computed from random samples behave (the Westvaco case, Activity 1.2a).

In this section, you will learn that if you know that a distribution is normal (shape), then the mean (center) and standard deviation (spread) tell you everything else about the distribution. The reason is that, whereas skewed distributions come in many different shapes, there is only one normal shape. It's true that one normal distribution might appear tall and thin while another looks short and fat. However, the x-axis of the tall, thin distribution can be stretched out so that it looks exactly the same as the short, fat one.

The Standard Normal Distribution

Because all normal distributions have the same basic shape, you can use recentering and rescaling to change any normal distribution to the one with mean 0 and standard deviation 1. Solving problems involving normal distributions depends on this important property.

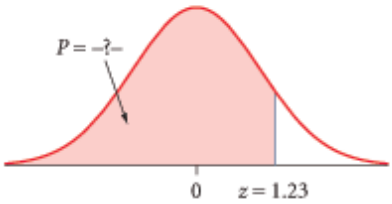
The normal distribution with mean 0 and standard deviation 1 is called the **standard normal distribution.** In this distribution, the variable along the horizontal axis is called a *z*-score.

The standard normal distribution is symmetric, with total area under the curve equal to 1, or 100%. To find the percentage, P, that describes the area to the left of the corresponding z-score, you can use the z-table or your calculator.

The next two examples show you how to use the z-table, Table A on pages 824–825.

Example: Finding the Percentage When You Know the z-Score

Find the percentage, P, of values less than $z = 1.23$, the shaded area in Display 2.77. Find the percentage greater than $z = 1.23$.



Display 2.77 The percentage of values less than $z = 1.23$.

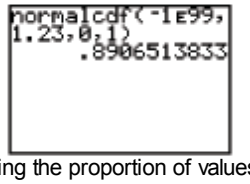
Solution

	Tail probability p		
z	.02	.03	.04
1.20	.8888	**.8907**	.8925

Think of 1.23 as 1.2 +0.03. In Table A on pages 824–825, find the row labeled 1.2 and the column headed .03. Where this row and column intersect, you find the number .8907. That means that 89.07% of standard normal scores are less than 1.23.

The total area under the curve is 1, so the proportion of values greater than $z = 1.23$ is $1 - 0.8907$, or 0.1093, which is 10.93%.

A graphing calculator will give you greater accuracy in finding the proportion of values that lie between two specified values in a standard normal distribution. For example, you can find the proportion of values that are less than 1.23 in a standard normal distribution like this:



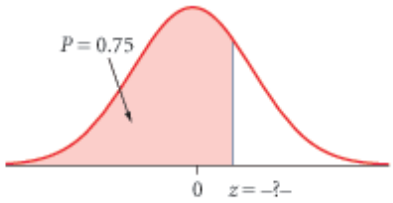
[To learn more about calculating the proportion of values between two z-scores, see **Calculator Note 2I**.]

Example: Finding the z-Score When You Know the Percentage

Find the z-score that falls at the 75th percentile of the standard normal distribution, that is, the z-score that divides the bottom 75% of values from the rest.

Solution

First make a sketch of the situation, as in Display 2.78.



Display 2.78 The *z*-score that corresponds to the 75th percentile.

	Tail probability *p*		
z	.06	.07	.08
.60	.7454	**.7486**	.7517

Look for .7500 in the body of Table A. No value in the table is exactly equal to .7500. The closest value is .7486. The value .7486 sits at the intersection of the row labeled .60 and the column headed .07, so the corresponding *z*-score is roughly 0.60 + 0.07, or 0.67.

You can use a graphing calculator to find the 75th percentile of a standard normal distribution like this:

[To learn more about finding the z-score that has a specified proportion of values below it, see **Calculator Note 2J**.]

DISCUSSION

The Standard Normal Distribution

D28. For the standard normal distribution,

a. what is the median?
b. what is the lower quartile?
c. what z-score falls at the 95th percentile?
d. what is the *IQR*?

Standard Units: How Many Standard Deviations Is It from Here to the Mean?

Converting to standard units, or **standardizing**, is the two-step process of recentering and rescaling that turns any normal distribution into the standard normal distribution.

First you recenter all the values of the normal distribution by subtracting the mean from each. This gives you a distribution with mean 0. Then you rescale by

dividing all the values by the standard deviation. This gives you a distribution with standard deviation 1. You now have a standard normal distribution. You can also think of the two-step process of standardizing as answering two questions: How far above or below the mean is my score? How many standard deviations is that?

The **standard units** or ***z*-score** is the number of standard deviations that a given x-value lies above or below the mean.

How far and which way to the mean?

$$x - \text{mean}$$

How many standard deviations is that?

Example: Computing a *z*-Score

In a recent year, the distribution of SAT I math scores for the incoming class at the University of Georgia was roughly normal, with mean 610 and standard deviation 69. What is the z-score for a University of Georgia student who got 560 on the math SAT?

Solution

A score of 560 is 50 points below the mean of 610. This is or 0.725 standard deviation below the mean. Alternatively, using the formula,

So the student's z-score is −0.725.

To "unstandardize," think in reverse. Alternatively, you can solve the z-score formula for x and get

$$x = \text{mean} + z \cdot SD$$

Example: Finding the Value When You Know the *z*-Score

What was a University of Georgia student's SAT I math score if his or her score was 1.6 standard deviations above the mean?

Solution

The score that is 1.6 standard deviations above the mean is

$$x = \text{mean} + z \cdot SD = 610 + 1.6(69) \approx 720$$

Example: Using *z*-Scores to Make a Comparison

In the United States, heart disease kills roughly one-and-a-quarter times as many people as cancer. If you look at the death rate per 100,000 residents by state, the distributions for the two diseases are roughly normal, provided you leave out Alaska and Utah, which are outliers because of their unusually young populations. The means and standard deviations for all 50 states are given here.

	Mean	*SD*
Heart disease	238	52
Cancer	196	31

Alaska had 88 deaths per 100,000 residents from heart disease, and 111 from cancer. Explain which death rate is more extreme compared to other states. [*Source*: Centers for Disease Control, *National Vital Statistics Report*, vol. 53, no. 5, October 12, 2004.]

Solution

$$z_{\text{heart}} = \frac{88 - 238}{52} \approx -2.88$$

$$z_{\text{cancer}} = \frac{111 - 196}{31} \approx -2.74$$

Alaska's death rate for heart disease is 2.88 standard deviations below the mean. The death rate for cancer is 2.74 standard deviations below the mean. These rates are about equally extreme, but the death rate for heart disease is slightly more extreme.

DISCUSSION

Standard Units

D29. Standardizing is a process that is similar to other processes you have seen already.

a. You're driving at 60 mi/h on the interstate and are now passing the marker for mile 200, and your exit is at mile 80. How many hours from your exit are you?

b. What two arithmetic operations did you do to get the answer in part a? Which operation corresponds to recentering? Which corresponds to rescaling?

Solving the Unknown Percentage Problem and the Unknown Value Problem

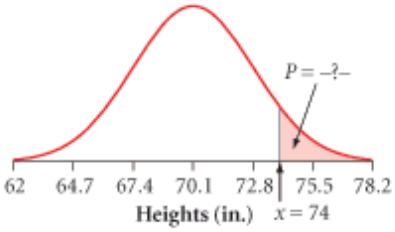
Now you know all you need to know to analyze situations involving two related problems concerning a normal distribution: finding a percentage when you know the value, and finding the value when you know the percentage.

Example: Percentage of Males Taller Than 74 Inches

For groups of similar individuals, heights often are approximately normal in their distribution. For example, the heights of 18- to 24-year-old males in the United States are approximately normal, with mean 70.1 in. and standard deviation 2.7 in. What percentage of these males are more than 74 in. tall? [*Source:* U.S. Census Bureau, *Statistical Abstract of the United States*, 1991.]

Solution

First make a sketch of the situation, as in Display 2.79. Draw a normal shape above a horizontal axis. Place the mean in the middle on the axis. Then mark and label the points that are two standard deviations either side of the mean, 64.7 and 75.5, so that about 95% of the values lie between them. Next, mark and label the points that are one and three standard deviations either side of the mean (67.4 and 72.8, and 62 and 78.2). Finally, estimate the location of the given value of x and mark it on the axis.



Display 2.79 The percentage of heights greater than 74 in.

Standardize:

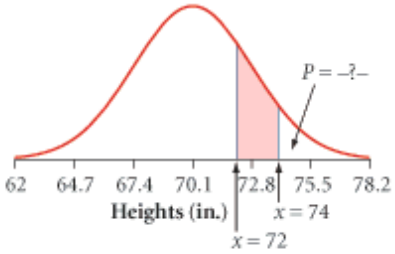
Look up the proportion: The area to the left of the z-score 1.44 is 0.9251, so the proportion of males taller than 74 in. is 1 − 0.9251, or 0.0749 or 7.49%. ■

Example: Percentage of Males Between 72 and 74 Inches Tall

The heights of 18- to 24-year-old males in the United States are approximately normal, with mean 70.1 in. and standard deviation 2.7 in. What percentage of these males are between 72 and 74 in. tall?

Solution

First make a sketch, as in Display 2.80.

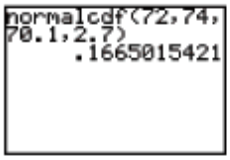


Display 2.80 The percentage of male heights between 72 and 74 in.

Standardize: From the previous example, a height of 74 in. has a z-score of 1.44. For a height of 72 in.,

Look up the proportion: The area to the left of the z-score 1.44 is 0.9251. The area to the left of the z-score 0.70 is 0.7580. The area you want is the area between these two z-scores, which is 0.9251 − 0.7580, or 0.1671. So the percentage of 18- to 24-year-old males between 72 and 74 in. tall is about 16.71%.

You can also use a graphing calculator to find this value. [See **Calculator Note 2I** for more details.]



Example: 75th Percentile of Female Heights

The heights of females in the United States who are between the ages of 18 and 24 are approximately normally distributed, with mean 64.8 in. and standard deviation 2.5 in. What height separates the shortest 75% from the tallest 25%?

Solution

First make a sketch, as in Display 2.81.

Display 2.81 The 75th percentile in height for women age 18 to 24.

Look up the z-score: If the proportion P is 0.75, then from Table A, you find that z is approximately 0.67.
Unstandardize:

$$x = \text{mean} + z \cdot SD = 64.8 + 0.67(2.5) \approx 66.475 \text{ in.}$$

For an unknown percentage problem:
First standardize by converting the given value to a z-score:

$$z = \frac{x - \text{mean}}{SD}$$

Then look up the percentage.
For an unknown value problem, reverse the process:
First look up the z-score corresponding to the given percentage. Then unstandardize:

$$x = \text{mean} + z \cdot SD$$

Solving the Unknown Percentage Problem and the Unknown Value Problem

D30. *Age of cars.* The cars in Clunkerville have a mean age of 12 years and a standard deviation of 8 years. What percentage of cars are more than 4 years old? (Warning: This is a trick question.)

Central Intervals for Normal Distributions

You learned in Section 2.1 that if a distribution is roughly normal, about 68% of the values lie within one standard deviation of the mean. It is helpful to memorize this fact as well as the others in the box on the next page.

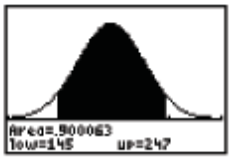
Example: Middle 90% of Death Rates from Cancer

According to the table on page 87, the death rates per 100,000 residents from cancer are approximately normal, with mean 196 and *SD* 31. The middle 90% of death rates are between what two numbers?

Solution

The middle 90% of values in this distribution lie within 1.645 standard deviations of the mean, 196. That is, about 90% of the values lie in the interval $196 \pm 1.645(31)$, or between about 145 and 247.

You can confirm this result using your calculator: Shade the area under this normal curve between 145 and 247 and calculate the area. [See **Calculator Note 2K.**]



[100, 300, 50, −0.004, 0.015, 1]

DISCUSSION

Central Intervals for Normal Distributions

D31. Use Table A on pages 824–825 to verify that 99.7% of the values in a normal distribution lie within three standard deviations of the mean.

Summary 2.5: The Normal Distribution

The standard normal distribution has mean 0 and standard deviation 1. All normal distributions can be converted to the standard normal distribution by converting to standard units:

- First, recenter by subtracting the mean.
- Then rescale by dividing by the standard deviation:

$$z = \frac{x - \text{mean}}{SD}$$

Standard units z tell how far a value x is from the mean, measured in standard deviations. If you know z, you can find x by using the formula x = mean + z • *SD*.

If your population is approximately normal, you can compute z and then use Table A or your calculator to find the corresponding proportion. Be sure to make a sketch so that you know whether to use the proportion in the table or to subtract that proportion from 1.

For any normal distribution,

- 68% of the values lie within 1 standard deviation of the mean
- 90% of the values lie within 1.645 standard deviations of the mean
- 95% of the values lie within 1.960 (or about 2) standard deviations of the mean
- 99.7% (or almost all) of the values lie within 3 standard deviations of the mean

Practice

The Standard Normal Distribution

P32. Find the percentage of values below each given *z*-score in a standard normal distribution.

a. –2.23 b. –1.67 c. – 0.40 d. 0.80

P33. Find the z-score that has the given percentage of values below it in a standard normal distribution.

a. 32% b. 41% c. 87% d. 94%

P34. What percentage of values in a standard

a. −1.46 and 1.46?
b. −3 and 3?

P35. For a standard normal distribution, what interval contains

a. the middle 90% of *z*-scores?
b. the middle 95% of *z*-scores?

Standard Units

P36. Refer to the table in the example on page 87.

a. California had 196 deaths from heart disease and 154 deaths from cancer per 100,000 residents. Which rate is more extreme compared to other states? Why?
b. Florida had 295 deaths from heart disease and 234 deaths from cancer per 100,000 residents. Which rate is more extreme?
c. Colorado had an unusually low rate of heart disease, 143 deaths per 100,000 residents. Hawaii had an unusually low rate of cancer, 156 deaths per 100,000 residents. Which is more extreme?

Solving the Unknown Percentage Problem and the Unknown Value Problem

P37. The heights of 18- to 24-year-old males in the United States are approximately normal, with mean 70.1 in. and standard deviation 2.7 in. The heights of 18- to 24-year-old females are also approximately normally distributed and have mean 64.8 in. and standard deviation 2.5 in.

a. Estimate the percentage of U.S. males between 18 and 24 who are 6 ft tall or taller.
b. How tall does a U.S. woman between 18 and 24 have to be in order to be at the 35th percentile of heights?

Central Intervals for Normal Distributions

P38. Refer to the table in the example on page 87.

a. The middle 90% of the states' death rates from heart disease fall between what two numbers?
b. The middle 68% of death rates from heart disease fall between what two numbers?

P39. Refer to the information in P37. Which of the following heights are outside the middle 95% of the distribution? Which are outside the middle 99%?

A. a male who is 79 in. tall
B. a female who is 68 in. tall
C. a male who is 65 in. tall
D. a female who is 65 in. tall

Exercises

E59. What percentage of values in a standard normal distribution fall

a. below a *z*-score of 1.00? 2.53?
b. below a *z*-score of −1.00? −2.53?
c. above a *z*-score of −1.5?
d. between *z*-scores of −1 and 1?

E60. On the same set of axes, draw two normal curves with mean 50, one having standard deviation 5 and the other having standard deviation 10.

E61. *Standardizing*. Convert each of these values to standard units, *z*. (Do not use a calculator. These are meant to be done in your head.)

a. $x = 12$, mean 10, *SD* 1
b. $x = 12$, mean 10, *SD* 2
c. $x = 12$, mean 9, *SD* 2
d. $x = 12$, mean 9, *SD* 1
e. $x = 7$, mean 10, *SD* 3
f. $x = 5$, mean 10, *SD* 2

E62. *Unstandardizing*. Find the value of *x* that was converted to the given *z*-score.

a. $z = 2$, mean 20, *SD* 5
b. $z = -1$, mean 25, *SD* 3
c. $z = -1.5$, mean 100, *SD* 10
d. $z = 2.5$, mean −10, *SD* 0.2

E63. SAT I critical reading scores are scaled so that they are approximately normal, with mean about 505 and standard deviation about 111.

a. Find the probability that a randomly selected student has an SAT I critical reading score
   i. between 400 and 600
   ii. over 700
   iii. below 450
b. What SAT I critical reading scores fall in the middle 95% of the distribution?

E64. SAT I math scores are scaled so that they are approximately normal, with mean about 511 and standard deviation about 112. A college wants to send letters to students scoring in the top 20% on the exam. What SAT I math score should the college use as the dividing line between those who get letters and those who do not?

E65. *Height limitations for flight attendants.* To work as a flight attendant for United Airlines, you must be between 5 ft 2 in. and 6 ft tall. [*Source:* www.ual.com.] The mean height of 18- to 24-year-old males in the United States is about 70.1 in., with a standard deviation of 2.7 in. The mean height of 18- to 24-year-old females is about 64.8 in., with a standard deviation of 2.5 in. Both distributions are approximately normal. What percentage of men this age meet the height limitation? What percentage of women this age meet the height limitation?

E66. *Where is the next generation of male professional basketball players coming from?*

a. The mean height of 18- to 24-year-old males in the United States is approximately normally distributed, with mean 70.1 in. and standard deviation 2.7 in. Use this information to approximate the percentage of men in the United States between the ages of 18 and 24 who are as tall as or taller than each basketball player listed here. Then, using the fact that there are about 13 million men between the ages of 18 and 24 in the United States, estimate how many are as tall as or taller than each player.
   i. Shawn Marion, 6 ft 7 in.
   ii. Allen Iverson, 6 ft 0 in.
   iii. Shaquille O'Neal, 7 ft 1 in.
b. Distributions of real data that are approximately normal tend to have heavier "tails" than the ideal normal curve. Does this mean your estimates in part a are too small, too big, or just right?

E67. *Puzzle problems.* Problems that involve computations with the normal distribution have four quantities: mean, standard deviation, value x, and proportion P below value x. Any three of these values are enough to determine the fourth. Think of each row in this table as little puzzles, and find the missing value in each case. This isn't the sort of thing you are likely to run into in practice, but solving the puzzles can help you become more skilled at working with the normal distribution.

Mean	SD	x	Proportion
3	1	2	—a—
10	2	—b—	0.18
—c—	3	6	0.09
10	—d—	12	0.60

E68. *More puzzle problems.* In each row of this table, assume the distribution is normal. Knowing any two of the mean, standard deviation, Q_1, and Q_3 is enough to determine the other two. Complete the table.

Mean	SD	Q_1	Q_3
10	5	—a—	—b—
—c—	—d—	120	180
—e—	10	100	—f—
10	—g—	—h—	11

E69. ACT scores are approximately normally distributed, with mean 18 and standard deviation 6. Without using your calculator, roughly what percentage of scores are between 12 and 24? Between 6 and 30? Above 24? Below 24? Above 6? Below 6?

E70. A group of subjects tested a certain brand of foam earplug. The number of decibels (dB) that noise was reduced for these subjects was

approximately normally distributed, with mean 30 dB and standard deviation 3.6 dB. The middle 95% of noise reductions were between what two values?

E71. The heights of 18- to 24-year-old males in the United States are approximately normally distributed with mean 70.1 in. and standard deviation 2.7 in.

a. If you select a U.S. male between ages 18 and 24 at random, what is the approximate probability that he is less than 68 in. tall?

b. There are roughly 13 million 18- to 24-year-old males in the United States. About how many are between 67 and 68 in. tall?

c. Find the height of 18- to 24-year-old males that falls at the 90th percentile.

E72. If the measurements of height are transformed from inches into feet, will that change the shape of the distribution in E71? Describe the distribution of male heights in terms of feet rather than inches.

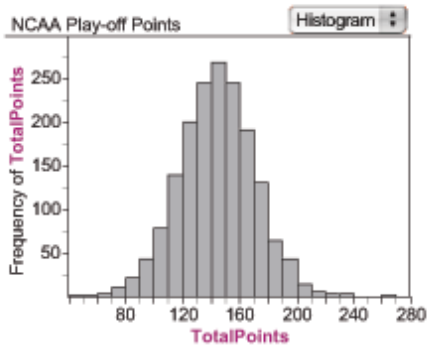
E73. The *British monarchy*. Over the 1200 years of the British monarchy, the average reign of kings and queens has lasted 18.5 years, with a standard deviation of 15.4 years.

a. What can you say about the shape of the distribution based on the information given?

b. Suppose you made the mistake of assuming a normal distribution. What fraction of the reigns would you estimate lasted a negative number of years?

c. Use your work in part b to suggest a rough rule for using the mean and standard deviation of a set of positive values to check whether it is possible that a distribution might be approximately normal.

E74. *NCAA scores*. The histogram in Display 2.82 was constructed from the total of the scores of both teams in all NCAA basketball play-off games over a 57-year period. Data Set



Display 2.82 Total points scored in NCAA play-off games. [*Source:* www.ncaa.com.]

a. Approximate the mean of this distribution.

b. Approximate the standard deviation of this distribution.

c. Between what two values do the middle 95% of total points scored lie?

d. Suppose you choose a game at random from next year's NCAA play-offs. What is the approximate probability that the total points scored in this game will exceed 150? 190? Do you see any potential weaknesses in your approximations?

Chapter Summary

Distributions come in various shapes, and the appropriate summary statistics (for center and spread) usually depend on the shape, so you should always start with a plot of your data.

Common symmetric shapes include the uniform (rectangular) distribution and the normal distribution. There are also various skewed distributions. Bimodal distributions often result from mixing cases of two kinds.

Dot plots, stemplots, and histograms show distributions graphically and let you estimate center and spread visually from the plot.

For approximately normal distributions, you ordinarily use the mean (balance point) and standard deviation as the measure of center and spread. If you know the mean and standard deviation of a normal distribution, you can use z-scores and Table A or your calculator to find the percentage of values in any interval.

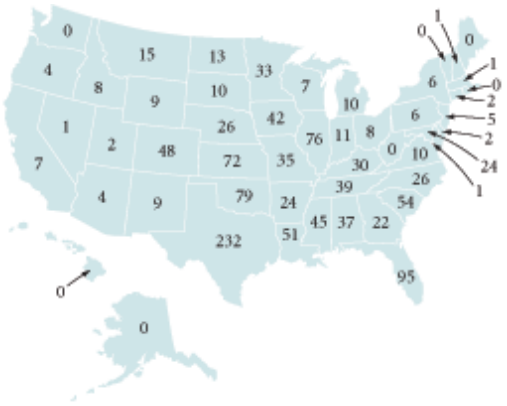
The mean and standard deviation are not resistant—their values are sensitive to outliers. For a description of a skewed distribution, you should consider using the median (halfway point) and quartiles (medians of the lower and upper halves of the data) as summary statistics.

Later on, when you make inferences about the entire population from a sample taken from that population, the sample mean and standard deviation will be the most useful summary statistics, even if the population is skewed.

Review Exercises

E75. The map in Display 2.83, from the U.S. National Weather Service, gives the number of tornadoes by state, including the District of Columbia.

Data Set



Display 2.83 The number of tornadoes per state in a recent year. [*Source:* www.ncdc.noaa.gov.]

a. Make a stemplot of the number of tornadoes.

b. Write the five-number summary.

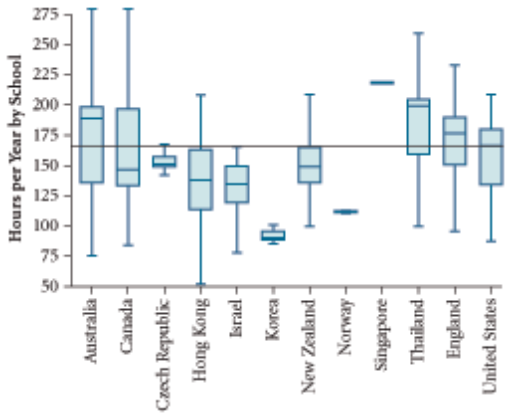
c. Identify any outliers.

d. Draw a boxplot.

e. Compare the information in your stemplot with the information in your boxplot. Which plot is more informative?

f. Describe the shape, center, and spread of the distribution of the number of tornadoes.

E76. Display 2.84 shows some results of the Third International Mathematics and Science study for various countries. Each case is a school.



Display 2.84 Boxplots of mathematics instruction time by country for 9-year-olds. [*Source:* Report #8, April 1998, of the Third International Mathematics and Science Study (TIMSS), p. 6.]

a. Estimate the median for the United States. Use this median value in a sentence that makes it clear what the median represents in this context.

b. Why are there only lines and no boxes for Norway and Singapore?

c. Describe how the distribution of values for the United States compares to the distributions of values for the other countries.

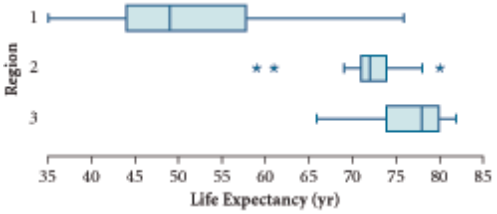
E77. A university reports that the middle 50% of the SAT I math scores of its students were between 585 and 670, with half the scores above 605 and half below.

a. What SAT I math scores would be considered outliers for that university?

b. What can you say about the shape of this distribution?

E78. These statistics summarize a set of television ratings from a week without any special programming. Are there any outliers among the 113 ratings?

E79. The boxplots in Display 2.85 show the life expectancies for the countries of Africa, Europe, and the Middle East. The table shows a few of the summary statistics for each of the three data sets.



Group	Mean	Median	StdDev
A	76.44	78	4.15
B	72	72	5.24
C	52.20	49	11.04

Display 2.85 Life expectancies for the countries of Africa, Europe, and the Middle East. [*Source:* Population Reference Bureau, *World Population Data Sheet, 2005*.]

a. From your knowledge of the world, match the boxplots to the correct region.

b. Match the summary statistics (for Groups A–C) to the correct boxplot (for Regions 1–3).

E80. The National Climatic Data Center records high and low temperatures by state since 1890. Stem-and-leaf plots of the years each state had its lowest temperature and the years each state had its highest temperature are shown in Display 2.86. What do the stems represent? What do the leaves represent? Compare the two distributions with respect to shape, center, spread, and any interesting features.

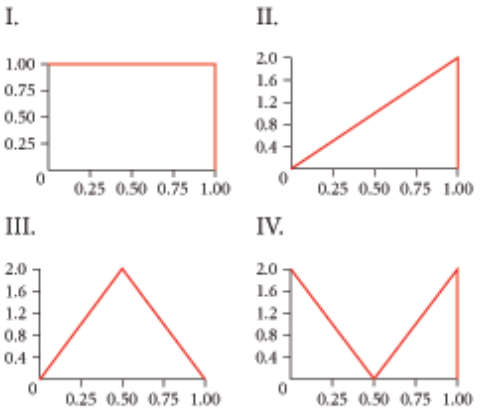
Data Set

Display 2.86 Stem-and-leaf plots of record low and high temperatures of states. [*Source:* National Climatic Data Center, 2002.]

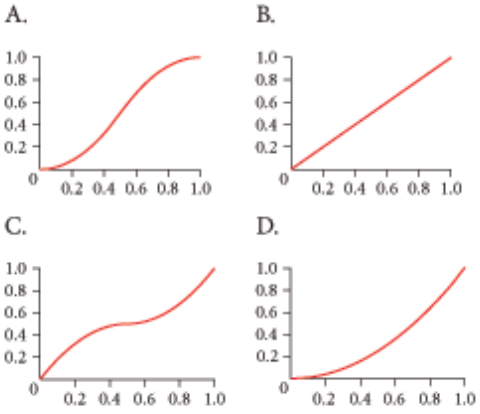
E81. A distribution is symmetric with approximately equal mean and median. Is it necessarily the case that about 68% of the values are within one standard deviation of the mean? If yes, explain why. If not, give an example.

E82. Display 2.87 shows two sets of graphs. The first set shows smoothed histograms I–IV for four distributions. The second set shows the corresponding cumulative relative frequency plots, in scrambled order A–D. Match each plot in the first set with its counterpart in the second set.

Distributions



Cumulative relative frequency plots



Display 2.87 Four distributions with different shapes and their cumulative relative frequency plots.

E83. The average number of pedestrian deaths annually for 41 metropolitan areas is given in Display 2.88.

Data Set

Metro Area	Average Annual Deaths
Atlanta	84
Baltimore	66
Boston	22
Charlotte, NC	29
Chicago	180
Cincinnati	23
Cleveland	36
Columbus, OH	20
Dallas	76
Denver	28
Detroit	107
Fort Lauderdale	58
Houston	101
Indianapolis	24
Kansas City	27
Los Angeles	299
Miami	100
Milwaukee	19
Minneapolis	35
Nassau-Suffolk, NY	80
Newark, NJ	51
New Orleans	47
New York	310
Norfolk, VA	25
Orlando, FL	48
Philadelphia	120
Phoenix	79
Pittsburgh	33
Portland, OR	34
Riverside, CA	92
Rochester, NY	17
Sacramento, CA	37
Salt Lake City	28
San Antonio	37
San Diego	96
San Francisco	43
San Jose, CA	33
Seattle	37
St. Louis	51
Tampa	85
Washington, DC	98

Display 2.88 Average annual pedestrian deaths. [*Source*: Environmental Working Group and the Surface Transportation Policy Project. Compiled from National Highway Traffic Safety Administration and U.S. Census data. *USA Today*, April 9, 1997.]

a. What is the median number of deaths? Write a sentence explaining the meaning of this median.

b. Is any city an outlier in terms of the number of deaths? If so, what is the city, and what are some possible explanations?

c. Make a plot of the data that you think will show the distribution in a useful way. Describe why you chose that plot and what information it gives you about average annual pedestrian deaths.

d. In which situations might giving the death rate be more meaningful than giving the number of deaths?

E84. The side-by-side boxplots in Display 2.89 give the percentage of 4th-grade-age children who are still in school on various continents according to the United Nations. Each case is a country. The four regions marked 1, 2, 3, and 4 are Africa, Asia, Europe, and South/Central America, not necessarily in that order.

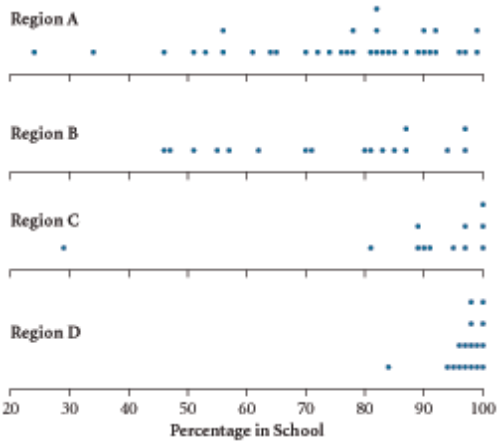
a. Which region do you think corresponds to which number?

b. Is the distribution of values for any region skewed left ? Skewed right? Symmetrical?

Data Set

Display 2.89 Boxplots of the percentage of 4th-grade-age children still in school in countries of the world, by continent. [*Source: 1993 Information Please Almanac.*]

Display 2.90 shows dot plots of the same data.



Display 2.90 Dot plots of the percentage of 4th grade- age children still in school in countries of the world, by continent.

c. Match each dot plot to the corresponding boxplot.

d. In what ways do the boxplots and dot plots give diff erent impressions? Why does this happen? Which type of plot gives a better impression of the distributions?

E85. The first AP Statistics Exam was given in 1997. The distribution of scores received by the 7667 students who took the exam is given in Display 2.91. Compute the mean and standard deviation of the scores.

Data Set

Score	Number of Students
5	1205
4	1696
3	1873
2	1513
1	1380

Display 2.89 Scores on the first AP Statistics Exam. [*Source:* The College Board.]

E86. For the countries of Europe, many average life expectancies are approximately the same, as you can see from the stemplot in Display 2.53 on page 69. Use the formulas for the summary statistics of values in a frequency table to compute the mean and standard deviation of the life expectancies for the countries of Europe.

E87. Construct a set of data in which all values are larger than 0, but one standard deviation below the mean is less than 0.

E88. Without computing, what can you say about the standard deviation of this set of values: 4, 4, 4, 4, 4, 4, 4, 4?

E89. In this exercise, you will compare how dividing by n versus $n - 1$ affects the SD for various values of n. So that you don't have to compute the sum of the squared deviations each time, assume that this sum is 400.

a. Compare the standard deviation that would result from
   i. dividing by 10 versus dividing by 9
   ii. dividing by 100 versus dividing by 99
   iii. dividing by 1000 versus dividing by 999

b. Does the decision to use n or $n - 1$ in the formula for the standard deviation matter very much if the sample size is large?

E90. If two sets of test scores aren't normally distributed, it's possible to have a larger z-score on Test II than on Test I yet be in a lower percentile on Test II than on Test I. The computations in this exercise will illustrate this point.

a. On Test I, a class got these scores: 11, 12, 13, 14, 15, 16, 17, 18, 19, 20. Compute the z-score and the percentile for the student who got a score of 19.

b. On Test II, the class got these scores: 1, 1, 1, 1, 1, 1, 1, 18, 19, 20. Compute the z-score and the percentile for the student who got a score of 18.

c. Do you think the student who got a score of 19 on Test I or the student who got a score of 18 on Test II did better relative to the rest of the class?

E91. The average income, in dollars, of people in each of the 50 states was computed for 1980 and for 2000. Summary statistics for these two distributions are given in Display 2.92.

	1980	2000
Mean	9,725	28,336
Standard Deviation	1,503	4,413
Minimum	7,007	21,007
Lower Quartile	8,420	25,109
Median	9,764	28,045
Upper Quartile	10,746	30,871
Maximum	14,866	41,495

Display 2.84 Summary statistics of the average income, in dollars, for the 50 states for 1980 and 2000. [*Source*: U.S. Census Bureau, *Statistical Abstract of the United States*, 2004–2005.]

a. Explain the meaning of $7,007 for the minimum in 1980.

b. Are any states outliers for either year?

c. In 2000 the average personal income in Alabama was $23,768, and in 1980 it was $7,836. Did the income in Alabama change much in relation to the other states? Explain your reasoning.

E92. For these comparisons, you will either use the SAT I critical reading scores in Display 2.69 on page 78 or assume that the scores have a normal distribution with mean 505 and standard deviation 111.

a. Estimate the percentile for an SAT I critical reading score of 425 using the cumulative relative frequency plot. Then find the percentile for a score of 425 using a z-score. Are the two values close?

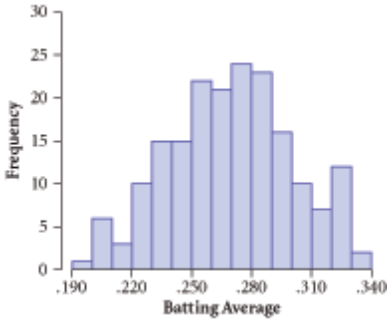
b. Estimate the SAT I critical reading score that falls at the 40th percentile, using the table in Display 2.69. Then find the 40th percentile using a z-score. Are the two values close?

c. Estimate the median from the cumulative relative frequency plot. Is this value close to the median you would get by assuming a normal distribution of scores?

d. Estimate the quartiles and the interquartile range using the plot. Find the quartiles and interquartile range assuming a normal distribution of scores.

E93. For 17-year-olds in the United States, blood cholesterol levels in milligrams per deciliter have an approximately normal distribution with mean 176 mg/dL and standard deviation 30 mg/dL. The middle 90% of the cholesterol levels are between what two values?

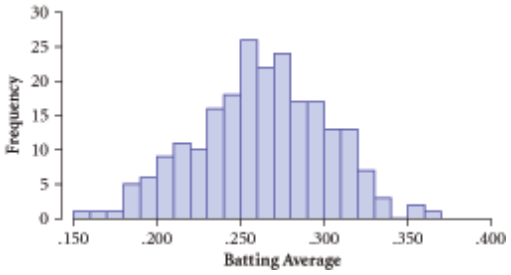
E94. Display 2.93 shows the distribution of batting averages for all 187 American League baseball players who batted 100 times or more in a recent season. (A batting "average" is the fraction of times that a player hits safely—that is, the hit results in a player advancing to a base—usually reported to three decimal places.)



Display 2.93 American League batting averages. [*Source:* CBS SportsLine.com, www.sportsline.com.]

a. Do the batting averages appear to be approximately normally distributed?
b. Approximate the mean and standard deviation of the batting averages from the histogram.
c. Use your mean and SD from part b to compute an estimate of the percentage of players who batted over .300 (or 300).
d. Now use the histogram to estimate the percentage of players who batted over .300. Compare to your estimate from part c.

E95. How good are batters in the National League? Display 2.94 shows the distribution of batting averages for all 223 National Leaguers who batted 100 times or more in a recent season.



Display 2.94 National League batting averages. [*Source:* CBS SportsLine.com, www.sportsline.com.]

a. Approximate the mean and standard deviation of the batting averages from the histogram.
b. Compare the distributions of batting averages for the two leagues. (See E94 for the American League.) What are the main differences between the two distributions?
c. A batter hitting .300 in the National League is traded to a team in the American League. What batting average could be expected of him in his new league if he maintains about the same position in the distribution relative to his peers?