bars could be grouped by type of disease, as in Figure 1.3(b). We can more easily see now that three of the six most common infectious diseases reported in California that year were sexually transmitted diseases (chlamydia, gonorrhea, syphilis) and that two were foodborne diseases (campylobacteriosis and salmonellosis).[i]

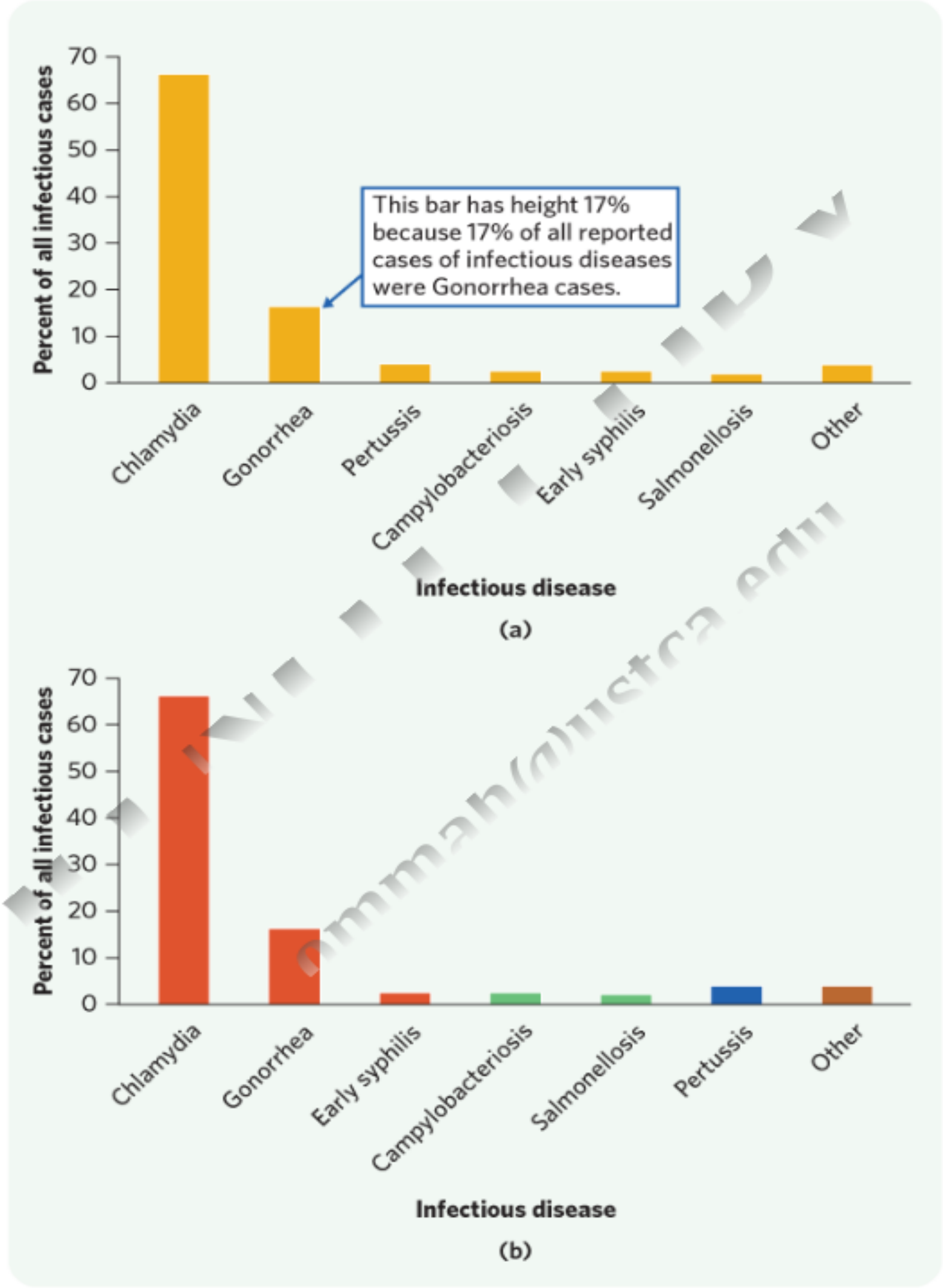

Figure 1. 3

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 1.3 The data on the reported cases of infectious diseases in California for 2014 are displayed in a bar graph format. (a) The bars are sorted according to their height (by order of importance). (b) The bars in the graph are grouped by disease type.

Pie charts must include all the categories that make up a whole (one variable in one group), but bar graphs are more flexible: They can be used to compare groups and do not necessarily display all possible outcomes of a variable. Beware that *this flexibility implies that bar graphs often are more challenging to interpret correctly.*

CAUTION

EXAMPLE 1.4 Who is more likely to have suicidal thoughts?

The National Survey on Drug Use and Health reports the percent of adults in the United States who had serious suicidal thoughts in 2014, for each of four age groups:[6]

Age group (years)	Percent who had serious suicidal thoughts
18–25	7.5
26–44	4.2
45–64	3.5
65+	1.6

It's clear that young adults between 18 and 25 years of age are the most susceptible to serious suicidal thoughts.

Figure 1.4(a) is a bar graph of the data in Example 1.4. We can see at a glance that, among adults, the proportion who had serious suicidal thoughts declines with age. We can't make a pie chart to display the data in Example 1.4, because each percent in the table refers to a different age group, not to parts of a single whole.

The **stacked bar graph** in Figure 1.4(b) makes this point more clearly: In each age group, individuals who had serious suicidal thoughts and individuals who did not together make up 100% of that age group.

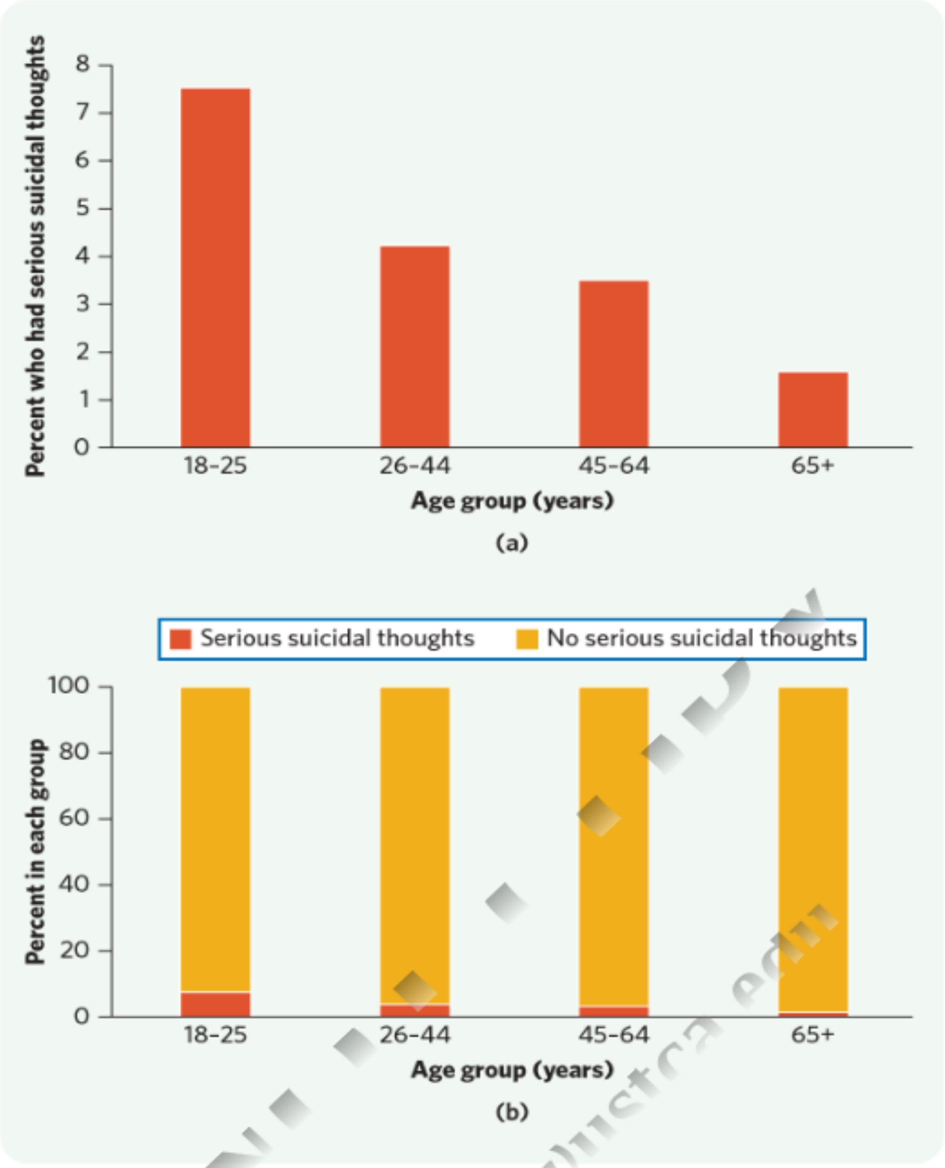

Figure 1.4

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 1.4 Bar graph showing the percent of American adults in each of four age groups who had serious suicidal thoughts in 2014. (a) The graph displays only those individuals in each group who had serious suicidal thoughts. (b) The graph displays both individuals who did and did not have serious suicidal thoughts in each age group, so that the stacked bars make up 100% of each age group.

It's important to understand this point to interpret the data correctly. Notice that the 18 to 25 age group represents a much smaller share of the American adult population than the other age groups listed, and 7.5% of this smaller age group could constitute a relatively small number of individuals who had serious suicidal thoughts. So we can conclude from the data that young adults aged 18 to 25 have the highest rate of serious suicidal thoughts, but not necessarily that they represent a large fraction of all individuals who have serious suicidal thoughts.

The ability to juxtapose data for different groups is what makes bar graphs so informative. The data in Example 1.4 were published in the Behavioral Health Barometer, United States, 2015, using the graph in Figure 1.5. This graph shows the percent of individuals who had serious suicidal thoughts, broken down side by side by age group and also by race/ethnicity. The graph makes it clear that age group is an influential factor but that, by comparison, race/ethnicity has very little effect on the disposition to have serious suicidal thoughts.

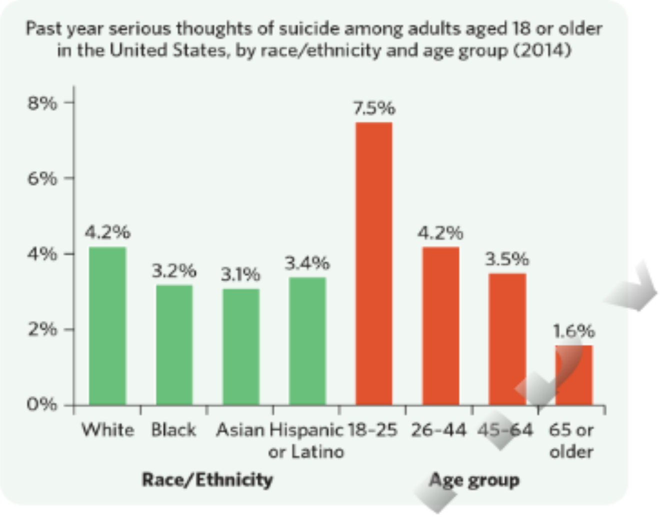

Figure 1.5
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 1.5 Bar graph showing the percent of American adults who had serious suicidal thoughts in 2014, for each of four ethnic groups and for each of four age groups.

STATS IN YOUR WORLD

CONVENIENT GRAPHS
The movie *An Inconvenient Truth* won the Oscar for Best Documentary in 2007. It was an unusual accomplishment for a movie featuring scientific evidence of global warming, supported by a detailed presentation of approximately 30 graphs and data tables.

APPLY YOUR KNOWLEDGE

1.3 Children's food choices. Does the presence of popular cartoon characters on food packages influence children's food choices? A study asked 40 young children (ages four to six) to taste two small pieces of Graham Crackers coming from a package with and a package without a popular cartoon character, and to indicate whether the two foods tasted the same or one tasted better. Unknown to the children, the crackers were the same both times. Here are the findings:[7]

Taste preference	Number of children	Percent
Without character	3	7.5
Taste the same	15	37.5
With character	22	55.0

a. Identify the individuals and the variable or variables in the study.

b. Present these data in a well-labeled bar graph.

c. Would it also be correct to present these data in a single pie chart? Explain your reasoning.

d. What do the data suggest about the influence of cartoon characters on Graham Cracker preference in young children?

young children?

1.4 More on children's food choices. The study in Exercise 1.3 also asked the 40 children to taste small pieces of gummy fruit snacks and baby carrots presented in packages with and in packages without a popular cartoon character. For each food type, the children indicated which of the two options they would prefer to eat for a snack. (Note that this is a different question from the one asked in Exercise 1.3.) The number and percent of children choosing the version with a cartoon on the package are displayed in the following table:

Food item	Number of children choosing the cartoon version	Percent choosing the cartoon version
Graham Crackers	35	87.5
Gummy fruit snacks	34	85.0
Baby carrots	29	72.5

a. Identify the individuals and the variable or variables in the study.

b. Make a well-labeled bar graph of the data.

c. Would it be correct to present these data in a single pie chart? Explain your reasoning.

d. What can you conclude from these findings?

Identify the individuals and the variable or variables in the study.

- Present these data in a well-labeled bar graph.
- Would it also be correct to present these data in a single pie chart? Explain your reasoning.
- What do the data suggest about the influence of cartoon characters on Graham Cracker preference in young children?
- **1.4 More on children's food choices.** The study in Exercise 1.3 also asked the 40 children to taste small pieces of gummy fruit snacks and baby carrots presented in packages with and in packages without a popular cartoon character. For each food type, the children indicated which of the two options they would prefer to eat for a snack. (Note that this is a different question from the one asked in Exercise 1.3.) The number and percent of children choosing the version with a cartoon on the package are displayed in the following table:

Food item	Number of children choosing the cartoon version	Percent choosing the cartoon version
Graham Crackers	35	87.5
Gummy fruit snacks	34	85.0
Baby carrots	29	72.5

a. Identify the individuals and the variable or variables in the study.

b. Make a well-labeled bar graph of the data.

c. Would it be correct to present these data in a single pie chart? Explain your reasoning.

d. What can you conclude from these findings?

QUANTITATIVE VARIABLES:

Histograms

Quantitative variables can take many values. The distribution of a variable tells us what values the variable takes and how often it takes these values. When there are more than just a few data points, a graph of the distribution is often easier to interpret if nearby values are grouped together. The most common graph of the distribution of one quantitative variable is a **histogram.**

EXAMPLE 1.5 Making a histogram: sharks

Stephen Frink Collection/Alamy

The great white shark, *Carcharodon carcharias*, is a large ocean predator at the top of the food chain. Here are the lengths in feet of 44 great whites:[8]

18.7	12.3	18.6	16.4	15.7	18.3	14.6	15.8	14.9	17.6	12.1
16.4	16.7	17.8	16.2	12.6	17.8	13.8	12.2	15.2	14.7	12.4
13.2	15.8	14.3	16.6	9.4	18.2	13.2	13.6	15.3	16.1	13.5
19.1	16.2	22.8	16.8	13.6	13.2	15.7	19.7	18.7	13.2	16.8

The *individuals* in this data set are individual sharks, and the *variable* is body length in feet. To make a histogram of the distribution of this variable, follow these steps:

STEP 1. Choose the classes. Divide the range of the data into classes of equal width. The data range from 9.4 to 22.8 feet, so we decide to use these classes:

$$\begin{array}{rcl} 9.0 & < & \text{individuals with body length} \leq 11.0 \\ 11.0 & < & \text{individuals with body length} \leq 13.0 \\ & \vdots & \\ 21.0 & < & \text{individuals with body length} \leq 23.0 \end{array}$$

In this example, we chose to exclude the lower bound and include the upper bound in making the histogram classes. Choosing, instead, to include the lower bound and exclude the upper bound would also have been a valid option. What matters is specifying the classes precisely so that each individual falls into exactly one class. You can explain the nature of the class boundaries in the legend accompanying your histogram.

Based on the chosen class definition, the smallest shark, measuring 9.4 feet, falls into the first class; a shark measuring exactly 11.0 feet would still fall in that first class, but a shark measuring 11.1 feet would fall into the second class.

STEP 2. Count the individuals in each class. Here are the counts and corresponding percents:

Class	Count	Percent
9.1 to 11.0	1	2
11.1 to 13.0	5	11
13.1 to 15.0	12	27
15.1 to 17.0	15	34
17.1 to 19.0	8	18
19.1 to 21.0	2	5
21.1 to 23.0	1	2

Check that the counts add to 44, the number of individuals in the data (the 44 great white sharks) and that the percents add to 100 up to *roundoff error*. Here the rounded percents add to 99.

STEP 3. Draw the histogram. Graphs are typically created with statistical software rather than hand-drawn. The principle, however, is exactly the same for both approaches. Mark the scale for the variable whose distribution you are displaying on the horizontal axis—the shark body length in feet. The scale runs from 9 to 23 feet because that is the span of the classes we chose. The vertical axis contains either the scale of *counts* or the scale of *percents*. A histogram of percents rather than counts is convenient when we want to compare several distributions. It would, for instance, allow us to more easily compare the lengths of several shark species regardless of how many sharks were examined from each species. Each bar in the histogram represents a class. The base of the bar covers the class, and the bar height is the class percent. The horizontal axis represents a continuum of values (here body lengths), and therefore the histogram does not leave any horizontal space between the bars unless a class is empty (in which case the bar has height zero). Figure 1.6(a) is our histogram.

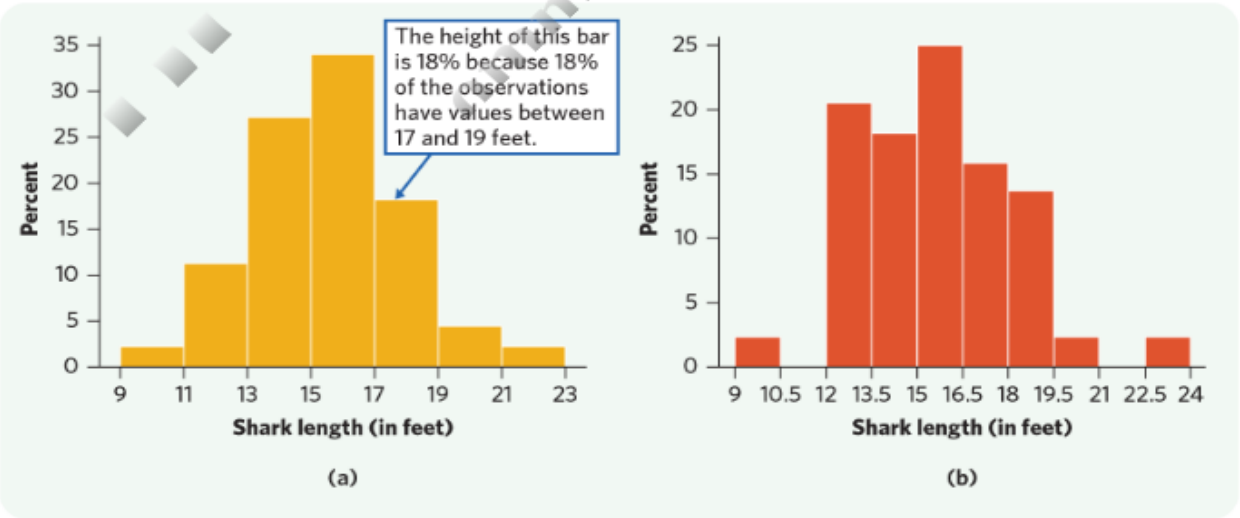

Figure 1.6

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 1.6 Histograms of the body length in feet for 44 great white sharks. (a) This histogram has 7 classes. (b) This histogram has 10 classes and shows more detail.

Although histograms resemble bar graphs in some aspects, their details and uses are very different. A histogram displays the distribution of one quantitative variable. The horizontal axis of a histogram is marked in the units of measurement for the variable. A histogram for a continuous quantitative variable should be drawn with no extra space between consecutive classes, to indicate that all values of the variable are covered. In contrast, a bar graph compares different items that may be ordered any way we choose, as in Figure 1.3 (a, b). To emphasize this, a bar graph should separate the items being compared by leaving some blank space between the bars.

Our eyes respond to the *area* of the bars in a histogram.[9] Because the classes are all the same width, area is determined by height and all classes are fairly represented. There is no one "right" choice of the classes in a histogram. Too few classes will give a "skyscraper" graph, with all values in a few classes with tall bars. Too many classes will produce a "pancake" graph, with most classes having one or no observations. Neither choice will give a good picture of the shape of the distribution. You must use your judgment in choosing classes to display the shape. Statistics software will choose the classes for you (but *be aware that different software programs use different conventions for displaying the class boundaries*). The software's choice is usually a good one, but you can change it if you want. Figure 1.6(b) is a histogram of the shark data from Example 1.5 using class sizes of 1.5 feet rather than 2 feet. The histogram function in the *One-Variable Statistical Calculator* applet on the companion website allows you to change the number of classes by dragging with the mouse, so that it is easy to see how the choice of classes affects the histogram.

APPLY YOUR KNOWLEDGE

1.5 Prescriptions of opioid pain relievers. Opioid pain relievers are prescribed at a higher rate in the United States than in any other nation, even though abuse of these medications can result in addiction and fatal overdoses. The CDC examined opioid pain reliever prescriptions in each state to find out how variable prescription rates are across the nation. Here are the 2012 state prescription rates, in number of prescriptions per 100 people, listed in increasing order:[10]

52.0	57.0	59.5	61.6	62.9	65.1	66.5	67.4	67.9	69.6
70.8	71.2	71.7	72.4	72.7	72.8	73.8	74.3	74.3	74.7
76.1	77.3	77.5	79.4	82.0	82.4	85.1	85.6	85.8	88.2
89.2	89.6	90.7	90.8	93.8	94.1	94.8	96.6	100.1	101.8
107.0	109.1	115.8	118.0	120.3	127.8	128.4	137.6	142.8	142.9

Make a histogram of the state opioid pain reliever prescription rates using classes of width 10 starting at 50.0 prescriptions per 100 people. (Make this histogram by hand even if you have software, to be sure you understand the process. You may want to compare your histogram with your software's choice.)

1.6 Choosing classes in a histogram. The data set menu that accompanies the *One-Variable Statistical Calculator* applet includes a data set called "healing of skin wounds." Choose these data, then click on the "Histogram" tab to see a histogram.

a. How many classes does the applet choose to use? (You can click on the graph outside the bars to get a count of classes.)

b. Click on the graph and drag to the left. What is the smallest number of classes you can get? What are the lower and upper bounds of each class? (Click on each bar to find out.) Make a rough sketch of this histogram.

c. Click and drag to the right. What is the largest number of classes you can get? How many observations does the class with the highest count have?

d. The choice of classes changes the appearance of a histogram. Drag back and forth until you get the histogram you think best displays the distribution. How many classes did you use?

Here are the 2012 state prescription rates, in number of prescriptions per 100 people, listed in increasing order:[10]

52.0	57.0	59.5	61.6	62.9	65.1	66.5	67.4	67.9	69.6
70.8	71.2	71.7	72.4	72.7	72.8	73.8	74.3	74.3	74.7
76.1	77.3	77.5	79.4	82.0	82.4	85.1	85.6	85.8	88.2
89.2	89.6	90.7	90.8	93.8	94.1	94.8	96.6	100.1	101.8
107.0	109.1	115.8	118.0	120.3	127.8	128.4	137.6	142.8	142.9

Make a histogram of the state opioid pain reliever prescription rates using classes of width 10 starting at 50.0 prescriptions per 100 people. (Make this histogram by hand even if you have software, to be sure you understand the process. You may want to compare your histogram with your software's choice.)

- **1.6 Choosing classes in a histogram.** The data set menu that accompanies the *One-Variable Statistical Calculator* applet includes a data set called "healing of skin wounds." Choose these data, then click on the "Histogram" tab to see a histogram.

APPLET

a. How many classes does the applet choose to use? (You can click on the graph outside the bars to get a count of classes.)

b. Click on the graph and drag to the left. What is the smallest number of classes you can get? What are the lower and upper bounds of each class? (Click on each bar to find out.) Make a rough sketch of this histogram.

c. Click and drag to the right. What is the largest number of classes you can get? How many observations does the class with the highest count have?

d. The choice of classes changes the appearance of a histogram. Drag back and forth until you get the histogram you think best displays the distribution. How many classes did you use?

INTERPRETING HISTOGRAMS

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

EXAMINING A HISTOGRAM

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

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

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

We will learn how to describe center and spread numerically in Chapter 2. For now, we can describe the center of a distribution by its *midpoint,* the value at which roughly half the observations take smaller values and half take larger values. We can describe the spread of a distribution by giving the *smallest and largest values*.

EXAMPLE 1.6 Describing a distribution: sharks

Look again at the histogram in Figure 1.6(a).

SHAPE: The distribution is **unimodal.** That is, it has a **single peak,** which represents sharks with a body length between 15 and 17 feet. The distribution is also *symmetric*. In mathematics, the two sides of symmetric patterns are exact mirror images. Real data are almost never exactly symmetric. We are content to describe the histogram in Figure 1.6(a) as roughly symmetric.

CENTER: The counts in Example 1.5 show that 18 of the 44 sharks (41%) have a body length of 15 feet or less, but the count increases to 33 out of 44 (75%) for lengths of 17 feet or less. So the midpoint of the distribution is about 15 to 17 feet.

SPREAD: The spread is from 9.4 to 22.8 feet, but only 1 shark is shorter than 11 feet and only 1 is more than 21 feet long.

OUTLIERS: In Figure 1.6(a), the observations greater than 21 feet or less than 11 feet are part of the continuous range of body lengths and do not stand apart from the overall distribution. This histogram, with only 7 classes, hides some of the detail in the distribution. Figure 1.6(b) is a histogram of the same data with 10 classes instead. It reveals that the shortest and longest sharks, at 9.4 and 22.8 feet, respectively, do stand apart from the other observations.

Once you have spotted possible outliers, look for an explanation. Some outliers are due to mistakes, such as typing 19.4 as 9.4. Other outliers point to the special nature of some observations. The smallest shark might be a juvenile exhibiting some adult features, for instance. An outlier could also simply be an unusual but perfectly legitimate observation. For instance, the largest shark could be just that: an unusually large shark. Human height, for instance, has a somewhat homogeneous distribution within a gender and ethnicity, but some individuals, such as the famous basketball player Shaquille O'Neal, clearly stand out from the norm.

Comparing Figures 1.6(a) and (b) reminds us that *the choice of classes in a histogram can influence the appearance of a distribution.* Both histograms portray a symmetric distribution with one peak, but only Figure 1.6(b) shows the mild outliers.

CAUTION

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

SYMMETRIC AND SKEWED DISTRIBUTIONS

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

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

Here are more examples of describing the overall pattern of a histogram.

EXAMPLE 1.7 A pertussis epidemic

Pertussis, also known as whooping cough, is a highly contagious respiratory disease characterized by uncontrollable, violent coughing. Pertussis can be fatal, especially among infants too young to be vaccinated. As mentioned in Example 1.3, California experienced an unusually large epidemic of pertussis in 2014, with 11,203 cases reported in that year. Here are the rates of pertussis cases per 100,000 persons reported in each of the 59 California counties for all of 2014:[11]

0.0	0.0	0.0	0.0	0.0	0.0	2.8	3.4	3.5	4.7
5.2	5.6	7.2	8.1	8.2	9.5	10.0	11.3	13.6	14.5
14.6	14.9	15.3	15.5	16.6	17.4	17.4	17.4	18.5	18.8
19.2	19.6	20.3	20.8	25.1	27.8	29.5	29.9	30.0	30.7
30.7	30.9	33.4	34.0	37.2	37.3	40.9	41.8	44.1	48.6
59.9	60.4	63.4	64.7	71.3	99.2	107.1	109.8	143.0	

The histogram in Figure 1.7 indicates that the distribution of pertussis rates is *single-peaked* and obviously strongly *skewed to the right.* Most counties had a relatively small rate, between 0 and 40 cases per 100,000 persons, and they make up the peak of the distribution (a single peak can span several classes, since the choice of classes is somewhat arbitrary). However, some counties has higher rates, so that the graph extends substantially to the right of its peak. This is not surprising considering that epidemics tend to cluster geographically. Here, the center (half above, half below) lies in the first class (0 to 20 cases per 100,000 persons). The pertussis rates range from 0 to 143 cases per 100,000 persons, though the county with the largest rate (Sonoma County) is a clear *outlier.*

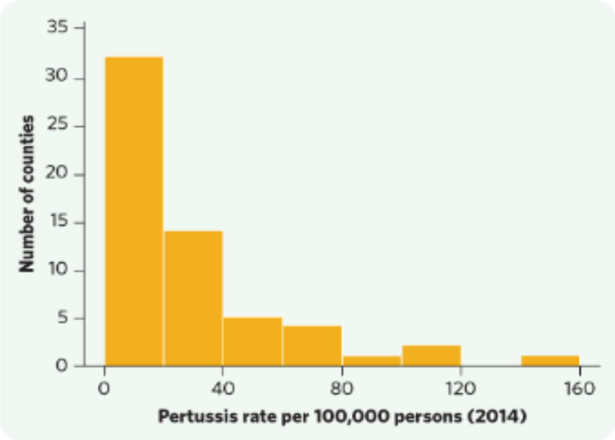

Figure 1.7
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 1.7 Histogram of 2014 rates of pertussis (in number of cases per 100,000 persons) in the 59 California counties. Classes include the lower-bound value but exclude the upper-bound value.

STATS IN YOUR WORLD

THE VITAL FEW

Skewed distributions can show us where to concentrate our efforts. Ten percent of the cars on the road account for half of all carbon dioxide emissions. A histogram of CO_2 emissions would show many cars with small or moderate values and a few with very high values. Cleaning up or replacing these high-emission cars would reduce pollution at a cost much lower than that of programs aimed at all cars. Statisticians who work at improving quality in industry often rely on this principle: Distinguish "the vital few" from "the trivial many."

EXAMPLE 1.8 Lyme disease

Lyme disease is a bacterial infection spread through the bite of an infected blacklegged tick. Left untreated, it can cause lifelong complications. Figure 1.8 displays data about the age in years of 241,931 individuals diagnosed with Lyme disease in the United States between 1992 and 2006.[12] We can't call this irregular distribution either symmetric or skewed. The major feature of the overall pattern is the presence of two main peaks; that is, the data have a **bimodal** distribution corresponding to two **clusters** of individuals—children and adults.

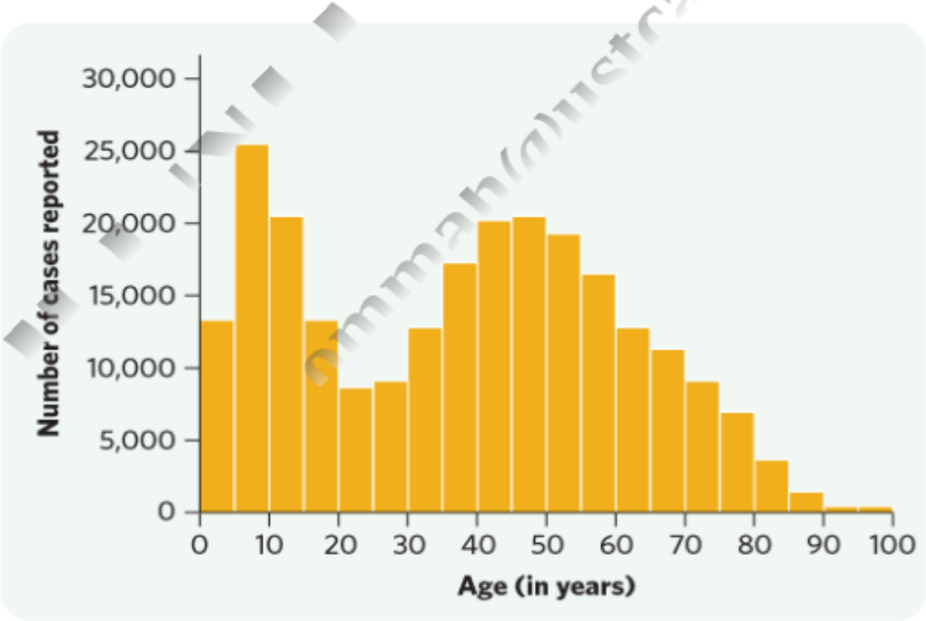

Figure 1.8
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 1.8 Histogram of patient age in years for 241,931 cases of Lyme disease reported in the United States between 1992 and 2006. Notice the two separate peaks around 10 years and 45 years.

Clusters suggest that several types of individuals are mixed in the data set. The first cluster has a clear peak around 5 to 15 years, whereas the peak of the second cluster is at 45 to 50 years. Because ticks are peak around 5 to 15 years, whereas the peak of the second cluster is at 45 to 50 years. Because ticks are found outdoors in woody or grassy areas, these two clusters might reflect the outdoor activities of children and the adults accompanying them.

Giving a single center and spread for this distribution would be misleading, because the data suggest two age groups. It would be better to describe the two groups separately.

■

CAUTION