

Day	Births
Sunday	7,134
Monday	11,430
Tuesday	12,387
Wednesday	12,230
Thursday	12,308
Friday	12,047
Saturday	8,124

Present these data in a well-labeled bar graph. Would it also be correct to make a pie chart? Suggest some possible reasons why there are fewer births on weekends.

1.3 Quantitative variables: histograms

Quantitative variables often take many values. The distribution tells us what values the variable takes and how often it takes these values. A graph of the distribution is clearer if nearby values are grouped together. The most common graph of the distribution of one quantitative variable is a **histogram**.

histogram

EXAMPLE 1.4

Making a Histogram

What percent of your home state's high school students graduate within four years? The No Child Left Behind Act of 2001 uses on-time high school graduation rates as one of its monitoring requirements. However, in 2001 most states were not collecting the necessary data to compute these rates accurately. The Freshman Graduation Rate (FGR) counts the number of high school graduates in a given year for a state and divides this by the number of ninth graders enrolled four years previously. Although the FGR can be computed from readily available data, it neglects high school students moving into and out of a state and may include students who have repeated a grade. Several alternative measures are available that partially correct for these deficiencies, but states have been free to choose their own measure, and the resulting rates can differ by more than 10%. Federal law now requires all states to use a common, more rigorous computation, the *Adjusted Cohort Graduation Rate*, that tracks individual students. Table 1.1 presents the data for 2010–11, the first year in which states used a common formula and for which graduation rates could be compared between states.⁶ Idaho, Kentucky, and Oklahoma received “timeline extensions” and were not required to file in 2010–11.

The *individuals* in this data set are the states. The *variable* is the percent of a state's high school students who graduate within four years. The states vary quite a bit on this variable, from 59% in the District of Columbia to 88% in Iowa. It's much easier to see how your state compares with other states from a graph like a histogram than from the table. To make a histogram of the distribution of this variable, proceed as follows:

Step 1. Choose the classes. Divide the range of the data into classes of equal width. The data in Table 1.1 range from 59 to 88, so we decide to use these classes:

percent on-time graduates between 55 and 60 (55 to <60)

percent on-time graduates between 60 and 65 (60 to <65)

⋮

percent on-time graduates between 85 and 90 (85 to <90)



GRADRATE



WHAT'S THAT NUMBER?

You might think that numbers, unlike words, are universal. Think again. A “billion” in the United States means 1,000,000,000 (nine zeros). In Europe, a “billion” is 1,000,000,000,000 (12 zeros). OK, those are words that describe numbers. But those commas in big numbers are periods in many other languages. This is so confusing that international standards call for spaces instead, so that an American billion is written 1 000 000 000. And the decimal point of the English-speaking world is the decimal comma in many other languages, so that 3.1416 in the United States becomes 3,1416 in Europe. So what is the number 10,642.389? It depends on where you are.

TABLE 1.1 PERCENT OF STATE HIGH SCHOOL STUDENTS GRADUATING ON TIME

STATE	PERCENT	REGION	STATE	PERCENT	REGION	STATE	PERCENT	REGION
Alabama	72	S	Louisiana	71	S	Ohio	80	MW
Alaska	68	W	Maine	84	NE	Oklahoma	—	S
Arizona	78	W	Maryland	83	S	Oregon	68	W
Arkansas	81	S	Massachusetts	83	NE	Pennsylvania	83	NE
California	76	W	Michigan	74	MW	Rhode Island	77	NE
Colorado	74	W	Minnesota	77	MW	South Carolina	74	S
Connecticut	83	NE	Mississippi	75	S	South Dakota	83	MW
Delaware	78	S	Missouri	81	MW	Tennessee	86	S
Florida	71	S	Montana	82	W	Texas	86	S
Georgia	67	S	Nebraska	86	MW	Utah	76	W
Hawaii	80	W	Nevada	62	W	Vermont	87	NE
Idaho	—	W	New Hampshire	86	NE	Virginia	82	S
Illinois	84	MW	New Jersey	83	NE	Washington	76	W
Indiana	86	MW	New Mexico	63	W	West Virginia	76	S
Iowa	88	MW	New York	77	NE	Wisconsin	87	MW
Kansas	83	MW	North Carolina	78	S	Wyoming	80	W
Kentucky	—	S	North Dakota	86	MW	Dist. of Columbia	59	S

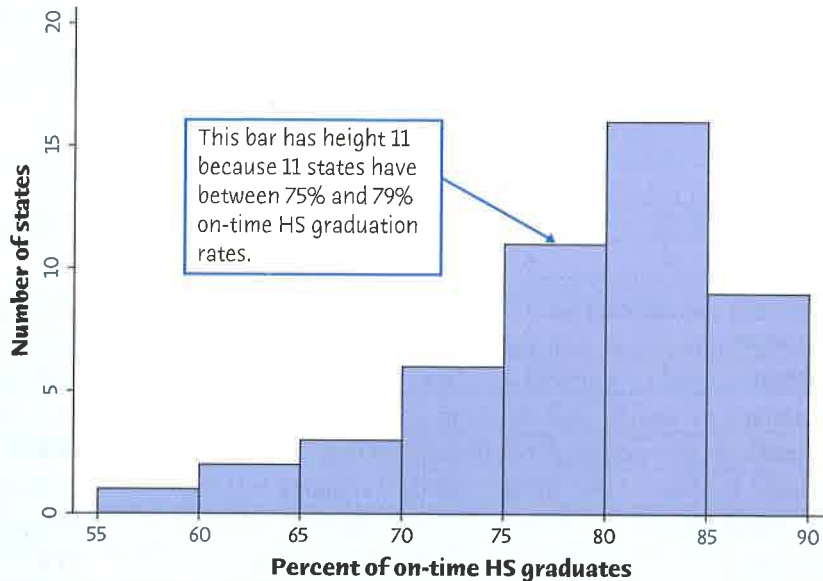
It is important to specify the classes carefully so that each individual falls into exactly one class. Our notation 55 to <60 indicates that the first class includes states with graduation rates starting at 55% and up to, but not including, graduation rates of 60%. Thus, a state with an on-time graduation rate of 60% falls into the second class, whereas a state with an on-time graduation rate of 59% falls into the first class. It is equally correct to use classes 56 to <62, 62 to <68, and so forth. Just be sure to specify the classes precisely so that each individual falls into exactly one class.

Step 2. Count the individuals in each class. Here are the counts:

Class	Count	Class	Count
55 to <60	1	75 to <80	11
60 to <65	2	80 to <85	16
65 to <70	3	85 to <90	9
70 to <75	6		

Check that the counts add to 48, the number of individuals in the data set (the 47 states reporting and the District of Columbia).

Step 3. Draw the histogram. Mark the scale for the variable whose distribution you are displaying on the horizontal axis. That's the percent of a state's high school students who graduate within four years. The scale runs from 55 to 90 because that is the span of the classes we chose. The vertical axis contains the scale of counts. Each bar represents a class. The base of the bar covers the class, and the bar height is the class count. Draw the bars with no horizontal space between them unless a class is empty, so that its bar has height zero. Figure 1.5 is our histogram. Remember, an observation on the boundary of the bars—say, 65—is counted in the bar to its right. ■

**FIGURE 1.5**

Histogram of the distribution of the percent of on-time high school graduates in 47 states and the District of Columbia, for Example 1.4.



Although histograms resemble bar graphs, their details and uses are different. A histogram displays the distribution of a quantitative variable. The horizontal axis of a histogram is marked in the units of measurement for the variable. A bar graph compares the sizes of different quantities. The horizontal axis of a bar graph simply identifies the quantities being compared and need not have any measurement scale. These quantities may be the values of a categorical variable, but they may also be unrelated, like the sources used to learn about music in Example 1.3. Draw bar graphs with blank space between the bars to separate the quantities being compared. Draw histograms with no space, to indicate that all values of the variable are covered. A gap between bars in a histogram indicates that there are no values for that class.

Our eyes respond to the *area* of the bars in a histogram.⁷ Because the classes are all the same width, area is determined by height, and all classes are fairly represented. There is no one right choice of the classes in a histogram. Too few classes will give a “skyscraper” graph, with all values in a few classes with tall bars. Too many will produce a “pancake” graph, with most classes having one or no observations. Neither choice will give a good picture of the shape of the distribution. You must use your judgment in choosing classes to display the shape. Statistics software will choose the classes for you. The software’s choice is usually a good one, but you can change it if you want. The histogram function in the *One-Variable Statistical Calculator* applet on the text 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

- 1.6 Foreign Born.** How are foreign-born residents distributed in the United States? The country as a whole has 13.0% foreign-born residents, but the states vary from 1.3% in West Virginia to 27.1% in California. Table 1.2 presents the data for all 50 states and the District of Columbia.⁸ Make a histogram of

TABLE 1.2 PERCENT OF STATE POPULATION BORN OUTSIDE THE UNITED STATES, 2011

STATE	PERCENT	STATE	PERCENT	STATE	PERCENT
Alabama	3.4	Louisiana	3.8	Ohio	3.9
Alaska	6.2	Maine	3.3	Oklahoma	5.5
Arizona	13.4	Maryland	13.7	Oregon	9.5
Arkansas	4.3	Massachusetts	14.9	Pennsylvania	5.9
California	27.1	Michigan	6.1	Rhode Island	13.5
Colorado	9.7	Minnesota	7.4	South Carolina	4.7
Connecticut	13.3	Mississippi	2.3	South Dakota	2.9
Delaware	8.6	Missouri	4.1	Tennessee	4.7
Florida	19.4	Montana	2.0	Texas	16.5
Georgia	9.6	Nebraska	6.2	Utah	8.4
Hawaii	18.2	Nevada	19.2	Vermont	3.9
Idaho	5.9	New Hampshire	5.3	Virginia	11.1
Illinois	13.9	New Jersey	21.3	Washington	13.4
Indiana	4.6	New Mexico	10.2	West Virginia	1.3
Iowa	4.3	New York	22.2	Wisconsin	4.8
Kansas	6.7	North Carolina	7.3	Wyoming	2.9
Kentucky	3.3	North Dakota	2.4	Dist. of Columbia	13.6

the percents using classes of width 5% starting at 0.0%. That is, the first bar covers 0.0% to <5.0%, the second covers 5.0% to <10.0%, and so on. (Make this histogram by hand, even if you have software, to be sure you understand the process. You may then want to compare your histogram with your software's choice.)  FOREIGN



1.7 Choosing Classes in a Histogram. The data set menu that accompanies the *One-Variable Statistical Calculator* applet includes the data on foreign-born residents in the states from Table 1.2. Choose these data, then click on the “Histogram” tab to see a histogram.

- How many classes does the applet choose to use? (You can click on the graph outside the bars to get a count of classes.)
- 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 the bar to find out.) Make a rough sketch of this histogram.
- Click and drag to the right. What is the greatest number of classes you can get? How many observations does the largest class have?
- You see that the choice of classes changes the appearance of a histogram. Drag back and forth until you get the histogram that you think best displays the distribution. How many classes did you use? Why do you think this is best?

1.4 Interpreting histograms

Making a statistical graph is not an end in itself. *The purpose of graphs 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 **variability**. You will sometimes see variability referred to as **spread**.

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

One way to describe the center of a distribution is by its *midpoint*, the value with roughly half the observations taking smaller values and half taking larger values. To find the midpoint, order the observations from smallest to largest, making sure to include repeated observations as many times as they appear in the data. First cross off the largest and smallest observations, then the largest and smallest of those remaining, and continue this process. If there were an odd number of observations initially, you will be left with a single observation, which is the midpoint. If there were an even number of observations initially, you will be left with two observations, and their average is the midpoint.

For now, we will describe the variability of a distribution by giving the *smallest and largest values*. We will learn better ways to describe center and variability in Chapter 2. The overall shape of a distribution can often be described in terms of symmetry or skewness, defined as follows.

Symmetric and Skewed Distributions

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

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

EXAMPLE 1.5 Describing a Distribution

Look again at the histogram in Figure 1.5. To describe the distribution, we want to look at its overall pattern and any deviations.

Shape: The distribution has a *single peak*, which represents states in which between 80% and 85% of students graduate high school on time. The distribution is *skewed to the left*. A majority of states have more than 75% of students graduating high school on time, but several states have much lower percents, so the graph extends quite far to the left of its peak.

Center: Arranging the observations from Table 1.1 in order of size shows that 80% is the midpoint of the distribution. There are a total of 48 observations, and if we cross off the 23 highest graduation rates and the 23 lowest graduation rates, we are left with two graduation rates, both of which are 80%. The center is their average, which is 80%.

Variability: The graduation rates range from 59% to 88%, which shows considerable variability in graduation rates among the states.

Outliers: Figure 1.5 shows no observations outside the overall single-peaked, left-skewed pattern of the distribution. Figure 1.6 is another histogram of the same distribution, with classes of width 3% rather than 5%. Now there are three states that



GRADRATE

CHAPTER 1 EXERCISES

- 1.23 Medical students.** Students who have finished medical school are assigned to residencies in hospitals to receive further training in a medical specialty. Here is part of a hypothetical database of students seeking residency positions. USMLE is the student's score on Step 1 of the national medical licensing examination.

NAME	MEDICAL SCHOOL	SEX	AGE	USMLE	SPECIALTY SOUGHT
Abrams, Laurie	Florida	F	28	238	Family medicine
Brown, Gordon	Meharry	M	25	205	Radiology
Cabrera, Maria	Tufts	F	26	191	Pediatrics
Ismael, Miranda	Indiana	F	32	245	Internal medicine

- (a) What individuals does this data set describe?
 (b) In addition to the student's name, how many variables does the data set contain? Which of these variables are categorical, and which are quantitative? If a variable is quantitative, what units is it measured in?

- 1.24 Buying a refrigerator.** *Consumer Reports* will have an article comparing refrigerators in the next issue. Some of the characteristics to be included in the report are the brand name and model; whether it has a top, bottom, or side-by-side freezer; the estimated energy consumption per year (kilowatts); whether or not it is Energy Star compliant; the width, depth, and height in inches; and both the freezer and refrigerator net capacity in cubic feet. Which of these variables are categorical, and which are quantitative? Give the units for the quantitative variables and the categories for the categorical variables. What are the individuals in the report?

- 1.25 What color is your car?** The most popular colors for cars and light trucks vary with region and over time. In North America white remains the top color choice, with silver the top choice in South America and white the top choice worldwide for the third consecutive year. Here is the distribution of the top colors for vehicles sold globally in 2013:¹⁶



COLOR	POPULARITY
White	25%
Black	18%
Silver	18%
Gray	12%
Red	9%
Beige, brown	8%
Blue	7%
Other colors	

Fill in the percent of vehicles that are in other colors. Make a graph to display the distribution of color popularity.

- 1.26 Facebook, Twitter, and LinkedIn users.** After years of explosive growth in number of users of social networking sites in all age ranges and demographics, it is hard to argue that social media haven't changed forever how we interact and connect online. Although Facebook is still the dominant player in social networking, both Twitter and LinkedIn have continued to increase their usage. Here is the age distribution of the users for the three sites in 2013:¹⁷



SOCIALNT

AGE GROUP	FACEBOOK USERS	TWITTER USERS	LINKEDIN USERS
13 to 17 years	10%	10%	4%
18 to 24 years	14%	18%	10%
25 to 34 years	19%	22%	20%
35 to 44 years	17%	17%	18%
45 to 54 years	17%	15%	20%
55 to 64 years	13%	11%	17%
Over 65 years	10%	7%	11%

- (a) Draw a bar graph for the age distribution of Facebook users. The leftmost bar should correspond to "13 to 17," the next bar to "18 to 24," and so on. Do the same for Twitter and LinkedIn, using the same scale for the percent axis.
 (b) Describe the most important difference in the age distribution of the audience for these three social networking sites. How does this difference show up in the bar graphs? Do you think it was important to order the bars by age to make the comparison easier? Why or why not?
 (c) Explain why it is appropriate to use a pie chart to display any of these distributions. Draw a pie chart for each distribution. Do you think it is easier to compare the three distributions with bar graphs or pie charts? Explain your reasoning.

- 1.27 Deaths among young people.** Among persons aged 15 to 24 years in the United States, the leading causes of death and number of deaths in 2011 were: accidents, 12,032; suicide, 4688; homicide, 4508; cancer, 1609; heart disease, 948; congenital defects, 429.¹⁸

(a) Make a bar graph to display these data.
 (b) To make a pie chart, you need one additional piece of information. What is it?

- 1.28 Hispanic origins.** According to the 2010 U.S. Census, 308.7 million people resided in the United States on April 1, 2010, of which 50.5 million (or 16%) were of Hispanic origin. What countries do they come from? Figure 1.13 is a pie chart to show the country of origin of Hispanics in the United States in 2010.¹⁹ About what percent of Hispanics are Mexican? Puerto Rican? You see that it is hard to determine numbers from a pie chart. Bar graphs are much easier to use. (The U.S. Census Bureau includes the percents in many of its pie charts.)

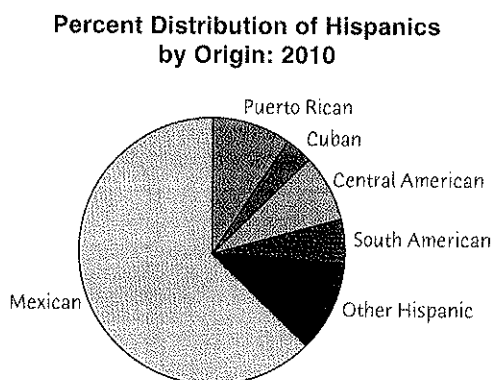


FIGURE 1.13

Pie chart of the national origins of Hispanic residents of the United States, for Exercise 1.28.

- 1.29 Canadian students rate their universities.** The National Survey of Student Engagement asked students

UNIVERSITY	EXCELLENT RATING
Toronto	21%
York	18%
Alberta	23%
Ottawa	11%
Western Ontario	38%
British Columbia	18%
Calgary	14%
McGill	26%
Waterloo	36%
Concordia	21%

at many universities, "How would you evaluate your entire educational experience at this university?" Here are the percents of senior-year students at Canada's 10 largest primarily English-speaking universities who responded "Excellent".²⁰



(a) The list is arranged in order of undergraduate enrollment. Make a bar graph with the bars in order of student rating.
 (b) Explain carefully why it is not correct to make a pie chart of these data.

- 1.30 Do adolescent girls eat fruit?** We all know that fruit is good for us. Many of us don't eat enough. Figure 1.14 is a histogram of the number of servings of fruit per day claimed by 74 seventeen-year-old girls in a study in Pennsylvania.²¹ Describe the shape, center, and variability of this distribution. Are there any outliers? What percent of these girls ate six or more servings per day? How many of these girls ate fewer than two servings per day?

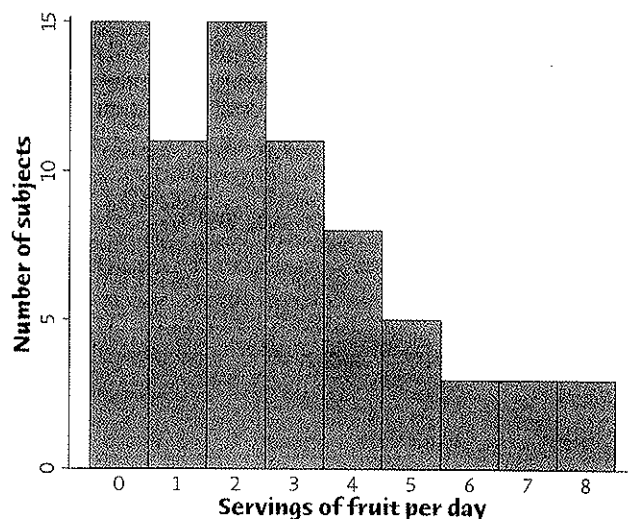


FIGURE 1.14

The distribution of fruit consumption in a sample of 74 seventeen-year-old girls, for Exercise 1.30.

- 1.31 IQ test scores.** Figure 1.15 (see page 40) is a stemplot of the IQ test scores of 78 seventh-grade students in a rural midwestern school.²²




(a) Four students had low scores that might be considered outliers. Ignoring these, describe the shape, center, and variability of the remainder of the distribution.
 (b) We often read that IQ scores for large populations are centered at 100. What percent of these 78 students have scores above 100?

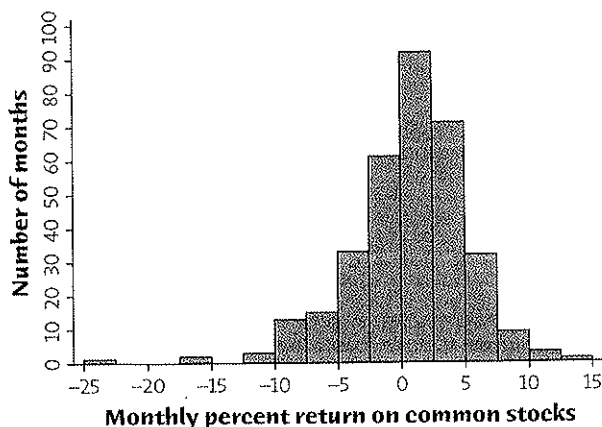
7	2 4
7	7 9
8	
8	6 9
9	0 1 3 3
9	6 7 7 8
10	0 0 2 2 3 3 3 3 4 4
10	5 5 5 6 6 6 7 7 7 8 9
11	0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 4 4 4 4
11	5 5 6 8 8 9 9 9
12	0 0 3 3 4 4
12	6 7 7 8 8 8
13	0 2
13	6

FIGURE 1.15

The distribution of IQ scores for 78 seventh-grade students, for Exercise 1.31.

1.32 Returns on common stocks. The return on a stock is the change in its market price plus any dividend payments made. Total return is usually expressed as a percent of the beginning price. Figure 1.16 is a histogram of the distribution of the monthly returns for all stocks listed on U.S. markets from January 1985 to December 2013 (348 months).²³ The extreme low outlier is the market crash of October 1987, when stocks lost 23% of their value in one month. The other two low outliers are 16% during August 1998, a month when the Dow Jones Industrial Average experienced its second-largest drop in history to that time, and the financial crisis in October 2008, when stocks lost 17% of their value.  **STOCKRET**

(a) Ignoring the outliers, describe the overall shape of the distribution of monthly returns.

**FIGURE 1.16**

The distribution of monthly percent returns on U.S. common stocks from January 1985 to December 2013, for Exercise 1.32.

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

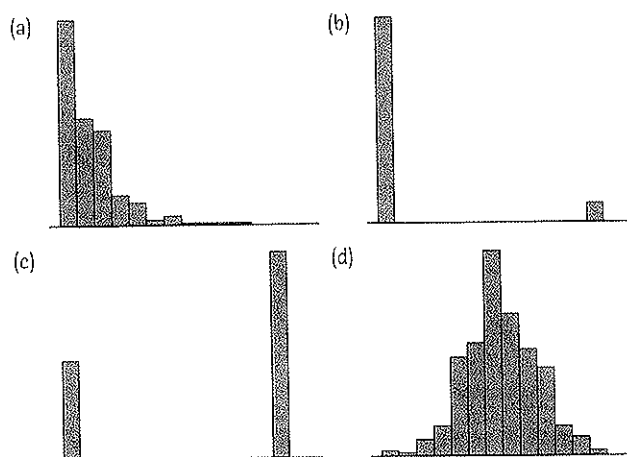
(c) Approximately what were the smallest and largest monthly returns, leaving out the outliers? (This is one way to describe the variability of the distribution.)

(d) A return less than zero means that stocks lost value in that month. About what percentage of all months had returns less than zero?

1.33 Name that variable. A survey of a large college class asked the following questions:

1. Are you female or male? (In the data, male = 0, female = 1.)
2. Are you right-handed or left-handed? (In the data, right = 0, left = 1.)
3. What is your height in inches?
4. How many minutes do you study on a typical weeknight?

Figure 1.17 shows histograms of the student response: in scrambled order and without scale marking. Which graph goes with each variable? Explain your reasoning.

**FIGURE 1.17**

Histograms of four distributions, for Exercise 1.33.



1.34 Food oils and health. Fatty acids, despite their unpleasant name, are necessary for human health. Two types of essential fatty acids, called omega-3 and omega-6, are not produced by our bodies and so must be obtained from our food. Food oils, widely used in food processing and cooking, are major sources of these compounds. There is some evidence that a healthy diet should have more omega-3 than omega-6. Table 1.1 gives the ratio of omega-3 to omega-6 in some common food oils.²⁴ Values greater than 1 show that an oil has more omega-3 than omega-6.  **FOODOILS**

TABLE 1.4 OMEGA-3 FATTY ACIDS AS A FRACTION OF OMEGA-6 FATTY ACIDS IN FOOD OILS

OIL	RATIO	OIL	RATIO
Perilla	5.33	Flaxseed	3.56
Walnut	0.20	Canola	0.46
Wheat germ	0.13	Soybean	0.13
Mustard	0.38	Grape seed	0.00
Sardine	2.16	Menhaden	1.96
Salmon	2.50	Herring	2.67
Mayonnaise	0.06	Soybean, hydrogenated	0.07
Cod liver	2.00	Rice bran	0.05
Shortening (household)	0.11	Butter	0.64
Shortening (industrial)	0.06	Sunflower	0.03
Margarine	0.05	Corn	0.01
Olive	0.08	Sesame	0.01
Shea nut	0.06	Cottonseed	0.00
Sunflower (oleic)	0.05	Palm	0.02
Sunflower (linoleic)	0.00	Cocoa butter	0.04

- (a) Make a histogram of these data, using classes bounded by the whole numbers from 0 to 6.
- (b) What is the shape of the distribution? How many of the 30 food oils have more omega-3 than omega-6? What does this distribution suggest about the possible health effects of modern food oils?
- (c) Table 1.4 contains entries for several fish oils (cod, herring, menhaden, salmon, sardine). How do these values support the idea that eating fish is healthy?

1.35 Where are the nurses? Table 1.5 gives the number of active nurses per 100,000 people in each state.²⁵  **NURSES**

- (a) Why is the number of nurses per 100,000 people a better measure of the availability of nurses than a simple count of the number of nurses in a state?
- (b) Make a histogram that displays the distribution of nurses per 100,000 people. Write a brief description of the distribution. Are there any outliers? If so, can you explain them?

TABLE 1.5 NURSES PER 100,000 PEOPLE, BY STATE

STATE	NURSES	STATE	NURSES	STATE	NURSES
Alabama	912	Louisiana	894	Ohio	1001
Alaska	756	Maine	1053	Oklahoma	712
Arizona	544	Maryland	869	Oregon	795
Arkansas	774	Massachusetts	1210	Pennsylvania	1017
California	641	Michigan	841	Rhode Island	1007
Colorado	761	Minnesota	1017	South Carolina	795
Connecticut	994	Mississippi	868	South Dakota	1215
Delaware	977	Missouri	958	Tennessee	894
Florida	814	Montana	748	Texas	662
Georgia	653	Nebraska	1010	Utah	625
Hawaii	753	Nevada	574	Vermont	912
Idaho	642	New Hampshire	970	Virginia	750
Illinois	812	New Jersey	907	Washington	774
Indiana	864	New Mexico	580	West Virginia	938
Iowa	990	New York	859	Wisconsin	905
Kansas	867	North Carolina	886	Wyoming	812
Kentucky	923	North Dakota	1097	Dist. of Columbia	1380